remote sensing

MDPI

Article

A Scene Classification Model Based on Global-Local Features and Attention in Lie Group Space

Chengjun Xu [1,2,*], Jingqian Shu [1], Zhenghan Wang [1] and Jialin Wang [1]

[1] School of Software, Jiangxi Normal University, Nanchang 330022, China; 005627@jxnu.edu.cn (J.S.); 202226703007@jxnu.edu.cn (Z.W.); 202226703038@jxnu.edu.cn (J.W.)
[2] School of Remote Sensing and Information Engineering, Wuhan University, Wuhan 430072, China
* Correspondence: 2018102160001@whu.edu.cn

Abstract: The efficient fusion of global and local multi-scale features is quite important for remote sensing scene classification (RSSC). The scenes in high-resolution remote sensing images (HRRSI) contain many complex backgrounds, intra-class diversity, and inter-class similarities. Many studies have shown that global features and local features are helpful for RSSC. The receptive field of a traditional convolution kernel is small and fixed, and it is difficult to capture global features in the scene. The self-attention mechanism proposed in transformer effectively alleviates the above shortcomings. However, such models lack local inductive bias, and the calculation is complicated due to the large number of parameters. To address these problems, in this study, we propose a classification model of global-local features and attention based on Lie Group space. The model is mainly composed of three independent branches, which can effectively extract multi-scale features of the scene and fuse the above features through a fusion module. Channel attention and spatial attention are designed in the fusion module, which can effectively enhance the crucial features in the crucial regions, to improve the accuracy of scene classification. The advantage of our model is that it extracts richer features, and the global-local features of the scene can be effectively extracted at different scales. Our proposed model has been verified on publicly available and challenging datasets, taking the AID as an example, the classification accuracy reached 97.31%, and the number of parameters is 12.216 M. Compared with other state-of-the-art models, it has certain advantages in terms of classification accuracy and number of parameters.

Keywords: attention mechanism; feature fusion; global feature; Lie Group; local feature; remote sensing scene classification

Citation: Xu, C.; Shu, J.; Wang, Z.; Wang, J. A Scene Classification Model Based on Global-Local Features and Attention in Lie Group Space. *Remote Sens.* **2024**, *16*, 2323. https://doi.org/10.3390/rs16132323

Academic Editors: Yang-Lang Chang, Toshifumi Moriyama, Ying-Nong Chen and Kuo-Chin Fan

Received: 6 May 2024
Revised: 20 June 2024
Accepted: 24 June 2024
Published: 25 June 2024

1. Introduction

Remote sensing scene classification (RSSC) is a fundamental task for the interpretation of high-resolution remote sensing images (HRRSIs) [1–3]. In recent years, with the rapid progress of satellite remote sensing technology, RSSC has also made significant progress [4–6]. It has been widely used in various scenarios, such as urban planning, natural disaster prediction, and environmental detection [7–11]. An HRRSI contains a variety of complex information about the structure of the Earth's surface. The challenge of RSSC is mainly to extract effective features from HRRSI, pay attention to crucial feature information, and realize the differentiation of different scenes.

Deep learning has received widespread attention from scholars due to its autonomous feature learning ability in the field of computer vision (CV), such as image classification, segmentation, and object detection [12–14]. Convolutional neural network (CNN) is considered to be the most widely used deep learning model technology, and RSSC based on deep learning has become one of the current mainstream types. Therefore, to improve the accuracy of scene classification, scholars have proposed many improved methods based on CNN models, and these methods have achieved better classification performance [15,16].

In addition to the CNN mentioned above, scholars have also proposed a deep learning model called transformer, which has been widely applied in many fields [17]. The transformer was originally applied in natural language processing (NLP), and its proposed self-attention mechanism can effectively capture the dependency relationships between long input sequences. After the success of NLP, inspired by it, scholars proposed the vision transformer (ViT) model, which can effectively learn the contextual features of images and the correlation relationships between different positions. Furthermore, scholars have also proposed a series of ViT-based RSSC models [18–20].

However, the transformer-based models also have limitations. Specifically, firstly, transformer-based models have more computational parameters compared to CNN-based models, which limits the use of transformer-based models in some cases where computational resources are insufficient [1]. Secondly, transformer-based models can effectively extract global relationships in images, but they may ignore smaller local-target-object feature information in HRRSIs [21]. Finally, transformer-based models ignore local induction bias in images, which typically require larger datasets. In the field of remote sensing, the number of HRRSIs is much less than that of natural images, so there may be overfitting issues [1]. Therefore, in this study, we still used a CNN-based model to explore RSSC.

In fields such as HRRSI classification and segmentation, a multi-scale fusion of global and local features is required. Some recent studies, such as ViTAE [22], StoHisNet [23], Transfuse [24], CMT [25], and Comformer [26], have improved the accuracy of classification and segmentation to a certain extent, mainly through the extracted features and attention mechanisms. By studying the successful work of other scholars and the models we proposed earlier, we believe that a better scene classification model should have the following characteristics: (1) integrating multi-scale global features and local features, (2) effective spatial and channel attention mechanisms, and (3) fewer parameters and better computational performance. Therefore, in this study, we pay more attention to the global and local features of the scene, focus on the more crucial regions in the scene through an efficient attention mechanism, eliminate the interference of irrelevant regions, and reduce the number of parameters of the model to improve the computational performance.

Inspired by previous studies such as Swin-Transformer [27], in this study, we propose a novel multi-scale branching model based on Lie Group space. In this model, we design two branches for extracting multi-scale global features and local features, respectively, design efficient spatial and channel attention mechanisms, and perform fusion operations through global and local fusion modules. In summary, the main contributions of this study are as follows:

1. We propose a multi-scale branching model for RSSC. In this model, it aims to extract the multi-scale and more discriminative features of the scene in a more fine-grained manner.
2. We propose global and local fusion modules to achieve efficient fusion between features at different scales. The module contains spatial and channel attention mechanisms and shortcut connections, which can effectively improve the model to focus on the crucial regions and ignore the irrelevant regions.
3. Compared with some existing models, our proposed model is more lightweight and achieves a better balance between classification accuracy and computational performance.

2. Related Works

2.1. RSSC Based on Features of Different Levels

According to the different levels of features, the RSSC model can be divided into four types [2,3,28–32]: (1) RSSC model based on low-level features, (2) RSSC model based on middle-level features, (3) RSSC model based on high-level features, and (4) RSSC model based on object-level features. Early RSSC models were mainly designed from low-level features such as local binary pattern (LBP) [33], color, gradient, and shape features [34]. However, RSSC based on the low-level features cannot effectively handle and represent the features in complex scenes. Then, scholars proposed an RSSC model based on middle-level features. This model mainly creates and encodes local feature dictionaries through

local descriptors to achieve a feature representation of scenes. The bag of visual words (BoVW) [35] is one of the representatives of the middle-level model because of its simplicity and ease of implementation [36]. The object-level features model is mainly analyzed from the object-oriented classification results, focusing on the relationship of objects rather than the shallower features of the scene (low-level and middle-level features) [37,38]. Due to the complex geometric structure and spatial layout of HRRSIs, RSSC models based on shallower features do not have advantages in terms of classification accuracy and efficiency.

Different from the mentioned RSSC model, the high-level feature-based RSSC model achieves the autonomous learning of features through the deep neural network, which can effectively improve the accuracy of classification. Xu et al. [29] proposed a scene classification model based on Lie Group manifold space, which effectively extracted shallower and high-level features in the scene. Zhang et al. [39] proposed a gradient-boosting-based random convolution model, which incorporates convolution operations at different depths. Lu et al. [40] proposed a novel aggregated CNN model that incorporates supervised convolutional feature encoding and progressive fusion strategies. Liu et al. [41] proposed a novel multi-scale CNN (MCNN) framework to improve the accuracy of scene classification.

2.2. RSSC Based on Attention Mechanism

The above models have shown high classification accuracy in RSSC. However, the above model does not fully consider the crucial features in the scene, and it is easy to be disturbed by irrelevant features. To address such deficiencies, scholars have proposed the attention mechanism. This mechanism makes the model pay more attention to crucial feature information in crucial regions of the scene and ignore irrelevant regions, and has been successfully applied to many fields, such as scene classification [42–46] and few-shot learning [47,48]. Hu et al. [44] proposed the SENet model, which includes squeeze-and-excite (SE) modules for extracting global contextual feature information. Then, BAM [49] and CBAM [46] attention mechanisms are proposed, which are weighted mainly from channel and spatial dimensions. To further reduce the computational complexity of the model, SA-Net [50] divides the features into several sub-features along the channel dimension and calculates the attention values in channel and spatial dimensions.

Scholars have also proposed a large number of RSSC models based on attention mechanisms. Li et al. [51] proposed a local-global context-aware generative dual-region adversarial model, which introduced globally aware self-attention and locally aware self-attention. Wang et al. [52] proposed the ARCNet model, which is a recurrent neural network based on attention intelligence and can adaptively select crucial regions. Yu et al. [53] proposed an improved model based on SE, mainly by enhancing features of different layers and fusing features. Chen et al. [43] proposed the MBLNet model, which is based on the ResNet50 model and can extract both channel and spatial attention.

In summary, based on our previous research, we decided to thoroughly explore the global and local features of the scene and integrate the above features. The model we propose is based on a hierarchical architecture, which includes both global and local branches that do not interfere with each other, and an efficient feature fusion module is designed. Our model retains the advantages of CNN and transformer models while effectively extracting global and local features.

3. Method

3.1. Overall Framework

In this study, we propose a novel RSSC model, which is a multi-scale global-local feature and attention scene classification model based on Lie Group space, as shown in Figure 1. Different from the existing methods, the model adopts a branch parallel structure, which mainly consists of a global feature extraction (GFE) module, local feature extraction (LFE) module, and global-local fusion (GLF) module. The global features mainly reflect the overall situation of an HRRSI and are statistical features of the entire HRRSI, such as contrast. Local features reflect the characteristics of local regions in an HRRSI, such as

texture structure. Local features contain more detailed feature information, which can help us process the details in HRRSIs. Local features can serve as supplements to global features. Some existing methods (such as CNN), when dealing with high-resolution images, increase the number of hidden layers and the number of parameters, slow calculation, and may also encounter overfitting problems. Therefore, in the feature extraction stage, we utilize two independent branches for global and local feature extraction to prevent mutual interference in the above feature extraction process. In the GFE, we use adaptive pooling, parallel dilated convolution, and Lie Group kernel functions to simulate the global space. In the LFE module, we utilize parallel dilated convolution to replace the traditional convolution operation to achieve multi-scale feature extraction. In the GLF module, we adopt efficient global and local fusion methods to achieve feature fusion at different stages, capture the dependency between features in channel dimension and spatial dimension, and combine grouping features with an attention mechanism to enhance semantic features in spatial and channel dimension. More module details are described below.

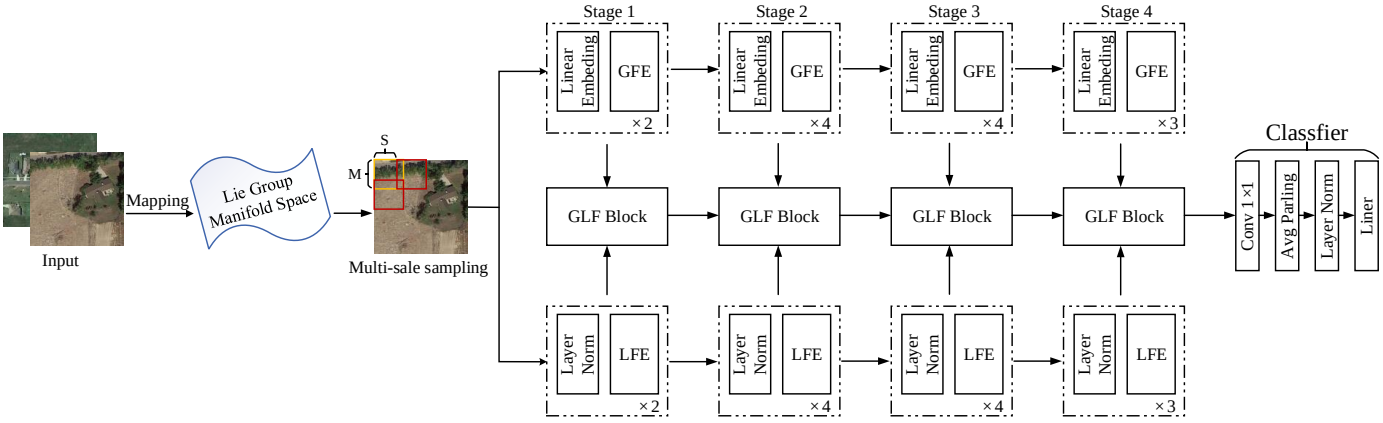

Figure 1. The overall architecture of our model. The model adopts an independent three-branch structure, which mainly includes a global feature extraction (GFE) module, local feature extraction (LFE) module, and global-local fusion (GLF) module.

3.2. Multi-Scale Sampling of HRRIs

The image resolution of HRRSI datasets varies from $0.2m$ to $30m$, which causes a large difference in scale. Considering that the resolution of the actual HRRSI is fixed, it may lead to a significant loss of accuracy in HRRSI transferring. To reduce the scale difference, in this study, we adopt the multi-scale sampling method to obtain different scale data samples of the same scene. Previous studies have verified that uniform grid sampling is effective for the characterization of HRRSIs [28,29]. Therefore, based on previous studies, we first conducted a uniform grid sampling of HRRSIs with block size M and spacing S. Secondly, multi-scale sampling was performed for each region according to the matching of the generated patches with the most overlapping regions. For details, please refer to our previous research results [28].

3.3. Hierarchical Parallel Model

To improve the classification accuracy of RSSC, we integrate global and local features at different levels and scales. Based on the previous research [28,29], we designed a hierarchical parallel framework. In Figure 1, the first branch is used for global feature information extraction, the second branch is used for local feature information extraction, and the third branch is used for the fusion of the above feature information.

The above design is mainly to effectively extract global and local feature information, and preserve the extracted feature information to the greatest extent so that their operations will not affect each other. The previous experimental results show that the branch structure is conducive to feature extraction [28], the parallel branch structure can effectively enhance

the characterization of features, and the multi-scale features extracted by the model are more robust and can provide better features for downstream tasks [3,28,29].

3.4. Global Feature Extraction

Due to the shooting angle, shooting distance, and imaging methods, there are various remote sensing scenes with strong inter-class similarity and significant intra-class differences. Therefore, the extraction of global feature information is quite important. The Swin-Transformer [27] model proposed the windows multi-head self-attention (W-MAS) mechanism. Compared with the traditional multi-head attention mechanism (MSA) in transformer, W-MAS divides the feature map into $M \times M$ sizes and then performs self-attention on the divided feature maps, improving calculation efficiency. Therefore, the W-MAS mechanism was introduced in this study.

Traditional models typically utilize global average pooling to simulate the global context environment, but these operations are not comprehensive enough for HRRSIs with complex geometric structures and spatial layouts. To better simulate the global context environment of scenes in HRRSIs, we adopted an adaptive pooling operation. More specifically, in Figure 2, adaptive pooling operation is first used to extract global contextual feature information. To accelerate the convergence of the model, batch normalization (BN) operation was adopted. The W-MAS mechanism is adopted, and its window size is 5×5, $H = 3$. Then, a larger kernel parallel dilated convolution is adopted; the kernel size is 5×5. Previous research results have verified that larger kernel parallel dilated convolution has a larger receptive field, and the number of parameters is much smaller than traditional convolution [2,28,29,31]. After a 1×1 convolution operation, in the following operations, nonlinear activation is usually introduced to enhance the representativeness of the model. In our model, the Lie Group Sigmoid activation function is applied instead of the traditional Sigmoid activation function, mainly to improve the robustness and computational performance of the model [54,55]. In addition, we also adopted a residual connection manner to further enhance the representation of features, as shown below.

$$GF_i' = LGS(Conv(PDConv(W - MSA(BN(AP(GF_{i-1}))))))) + GF_{i-1} \quad (1)$$

where GF represents the input feature map, GF_{i-1} represents the previous input feature map, GF_i represents the i^{th} feature map, AP represents the adaptive pooling, BN represents the batch normalization, $W - MAS$ represents the windows multi-head self-attention, and $PDConv$ represents the parallel dilated convolution with 7×7 convolution kernel. $Conv$ represents a convolution with a convolution kernel of 1×1, and LGS represents a Lie Group Sigmoid activation function. The global feature information extracted from the above will be input into the global and local feature fusion module.

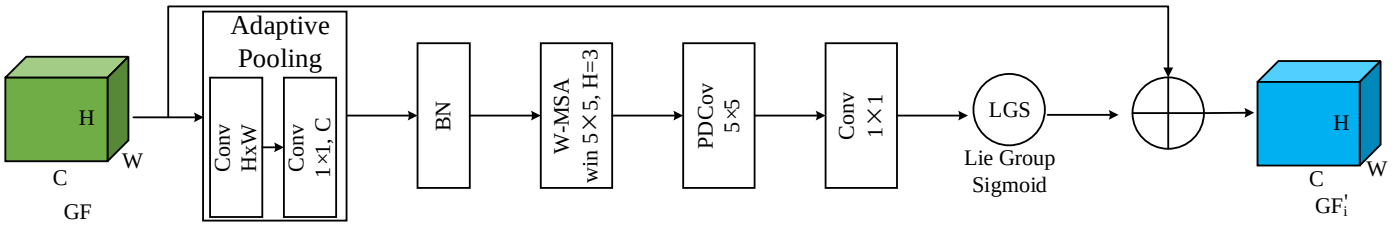

Figure 2. Global feature extraction framework. The framework consists of adaptive pooling, BN, W-MASA, various convolution, and Lie Group Sigmoid activation functions.

3.5. Local Feature Extraction

Local features in HRRSIs are also quite important. As a supplement to the global features, local features can effectively enhance the features of the scene. As shown in Figure 3, the feature map is firstly divided into four partitions along the channel dimension, as shown below.

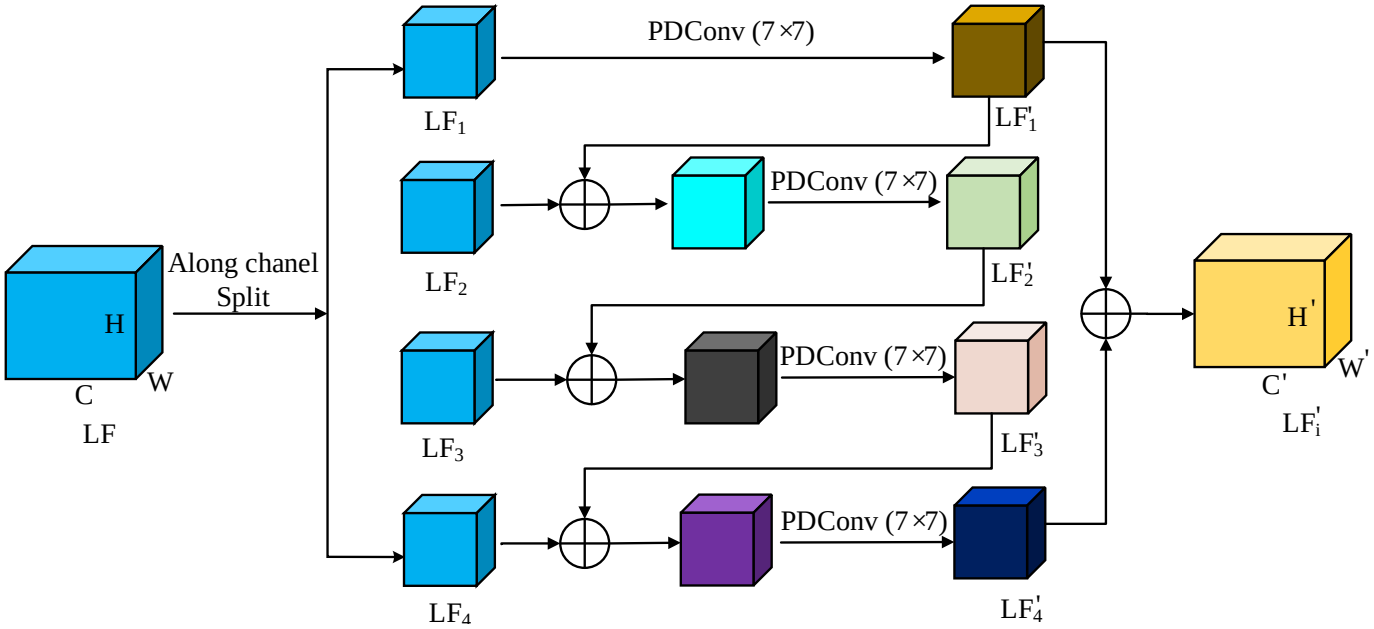

Figure 3. Local feature extraction framework. The framework firstly divides the feature map into four partitions along the channel dimension and then performs parallel dilated convolution operations, respectively. The obtained feature map is then reused with the feature map of another partition to improve the interaction of feature information and achieve multi-scale feature extraction.

$$LF_1, LF_2, LF_3, LF_4 = S(LF_i) \tag{2}$$

where LF_i represents the i^{th} input feature map, and LF_1, LF_2, LF_3, and LF_4, respectively, represent the four partitions obtained using the split function $S(\cdot)$.

In the first partition, LF_1, a larger kernel-parallel-dilated convolution is utilized to extract features, as follows:

$$LF_1' = PDConv(LF_1) \tag{3}$$

Then, the extracted feature map LF_1' is added to the second partition $LF2$, and the new feature map LF_2' is obtained by using the same convolution operation as above for the added feature. By repeating the above operation, the new feature maps of the four partitions can be obtained, and they are connected as follows:

$$LF_i' = LF_{i-1}' + LF_i \tag{4}$$

$$LF_i' = C_{ch}(LF_1' + LF_2' + LF_3' + LF_4') \tag{5}$$

where $C_{ch}(\cdot)$ indicates the channelwise concatenation operation.

Compared with the traditional convolution operation, the above operation has the following advantages:

1. The above operation can effectively increase the feature information interaction between different partition modules.
2. We have achieved multi-channel and multi-dimensional feature extraction in a finer granularity manner, effectively expanding the receptive field while suppressing the increase in model parameters, improving the computational efficiency of the model, and reducing the number of model parameters.
3. More specifically, traditional multi-scale feature extraction mainly adopts multiple parallel branch structures, each branch contains a fixed kernel (such as 3×3, 5×5), without considering the relationship between the feature maps of different branches. Compared with our model, the feature maps obtained from each branch are fed to the next branch, which achieves the reuse of features and promotes the interaction of feature information between different channels. In addition, the receptive field can be

effectively expanded in this way, such as in two consecutive 7×7 convolution operations, the receptive field of the first convolution is 7×7 (each output value addresses 49 values in the input feature map), the receptive field of the second convolution is effectively 9×9 (each output value addresses 81 values in the input feature map), with the same size kernel itself, but the receptive field is enlarged.

3.6. Global-Local Fusion

3.6.1. Channel Attention

Each channel in the feature map usually contains the response of the corresponding feature. Channel attention enables the model to pay attention to different regions in the scene, and adaptively assigns different weights to each channel. This process can be regarded as feature screening [56–58].

As shown in Figure 4, the feature map is firstly divided into four partitions along the channel dimension, and the four partitions are flattened, which is mainly to mine the correlation between features. GF_1' and GF_4' are then operated through average pooling and maximum pooling, respectively. After the GF_2' and GF_3' pass through the multi-layer perceptron (MLP), the multiplied operation will be performed with the above feature map, respectively, as shown below.

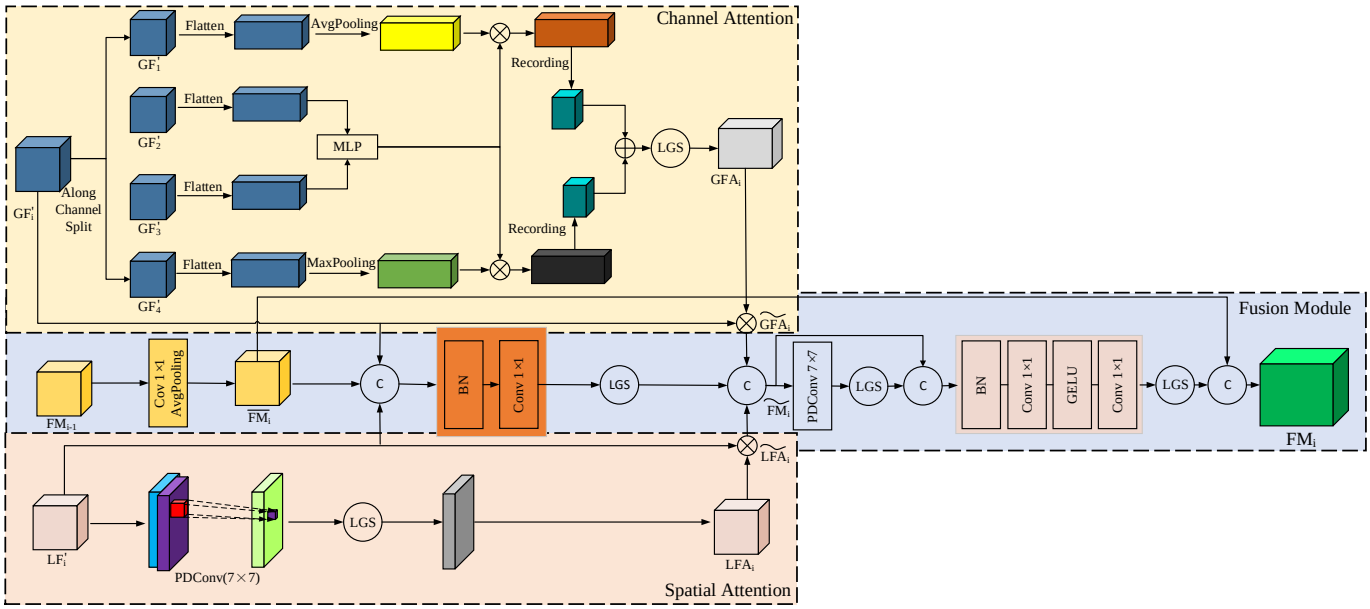

Figure 4. Global-local fusion framework. The framework is composed of channel attention, spatial attention, and fusion modules. The attention mechanism is implemented by spatial attention and channel attention, and the fusion module can effectively integrate the captured attention.

$$GFA_i = LGS(Red(Avg(Fla(GF_1')) \cdot MLP(Fla(GF_2'), Fla(GF_3'))) + Red(Max(Fla(GF_4')) \cdot MLP(Fla(GF_2'), Fla(GF_3')))) \quad (6)$$

where GF_1', GF_2', GF_3', and GF_4' represent the four partitions divided along the channel dimension, respectively, Fla represents the flattening operation, Avg represents the average pooling operation, Max represents the maximum pooling operation, and MLP represents the multi-layer perceptron operation. Red indicates the Reordering operation, while GFA_i represents the feature maps obtained through channel attention mechanism.

3.6.2. Spatial Attention

Spatial attention can be used as a supplement to the above, mainly to enhance the semantic distribution of features in the spatial dimension, that is, enabling the model to selectively focus on crucial regions in the scene while ignoring irrelevant regions. At the

same time, spatial attention is also the process by which the model screens important regions and decides what to focus on. Our model utilizes this mechanism to capture different semantic features for each feature map.

As shown in Figure 3, a 7×7 parallel dilated convolution operation is performed, and a Lie Group Sigmoid kernel function is added to the nonlinear transformation. This operation can effectively preserve the feature information of spatial dimensions, and the use of a 7×7 parallel dilated convolution operation can effectively reduce the number of model parameters and improve computational performance. The specific operation is as follows.

$$LAF_i = LGS(PDConv(LF_i')) \tag{7}$$

where LFA_i represents the feature maps obtained through spatial attention mechanism.

3.6.3. Fusion Module

The main function of this module is to efficiently integrate global features, local features, features obtained through channel attention, features obtained through spatial attention, and features obtained from the previous level, as shown in Figure 3. The fusion operation is as follows:

$$\overline{FM_i} = Avg(Conv(FM_{i-1})) \tag{8}$$

$$\widetilde{GFA_i} = GFA_i \otimes GF_i' \tag{9}$$

$$\widetilde{LFA_i} = LFA_i \otimes LF_i' \tag{10}$$

$$\widetilde{FM_i} = Conc(LGS(Conv(BN(Conc(GF_i', LF_i', \overline{FM_i})))), \widetilde{GFA_i}, \widetilde{LFA_i}) \tag{11}$$

$$FM_i = Conc(LGS(Conv(GELU(Conv(BN(Conc(LGS(PDConv(\widetilde{FM_i}), \widetilde{FM_i}))))))), \overline{FM_i}) \tag{12}$$

where FM_{i-1} represents the feature map generated in the previous stage of fusion; FM_i represents the feature map generated after the fusion; $\widetilde{GFA_i}$ represents the feature map after the fusion of channel dimensions, that is, the element-wise multiple of the feature map before the channel attention operation and the feature map obtained after the channel attention mechanism is used; $\widetilde{LFA_i}$ represents the feature map after the fusion of spatial dimensions, that is, the element-wise multiple of the feature map before the spatial attention operation and the feature map obtained after using the spatial attention mechanism; *Conc* represents the concatenation operation; *BN* represents the batch normalization; and *GELU* represents the GELU activation function.

4. Experiments

4.1. Experimental Environment

4.1.1. Datasets

We chose three publicly available and representative datasets, such as the aerial image dataset (AID) [10], the remote sensing image classification benchmark (RSICB-256) [59], and the Northwestern Polytechnical University remote sensing image scene classification (NWPU-RESISC45) dataset [37]. The AID is a large-scale aerial imagery dataset collected from Google Earth images, containing 30 scene categories. The RSICB-256 dataset integrates sensor and Google Earth images, containing 35 scene categories. The NWPU-RESISC45 dataset was created by Northwestern Polytechnical University and contains 45 scene categories. The above dataset contains a large number of scenes, and the scenes of different categories have high similarity while the scenes within the same category have significant variations. Therefore, the above dataset poses certain challenges. We compared our proposed model with some classic and state-of-the-art (SOTA) in terms of parameter quantity, classification accuracy, and computational performance.

4.1.2. Experimental Parameter Setting and Evaluation Metrics

All experiments were carried out under the same training parameters; the specific parameters are shown in Table 1. Referring to the previous model [28–31,42,43,60,61], we adopted different settings for the above three datasets. Specifically, the training ratio is 20% and 50% for the AID, 50% for the RSICB-256 dataset, and 10% and 20% for the NWPU-RESISC45 dataset. The results obtained in the experiment are set according to the parameters in the references. We report the average results and standard deviations of ten replicates performed independently to reduce the effect of randomness.

Table 1. Setting of experimental environment and other parameters.

Item	Content
CPU	Inter Core i7-4700 CPU with 2.70 GHz ×12
Memory	32 GB
Operating system	CentOS 7.8 64 bit
Hard disk	1TB
GPU	Nvidia Titan-X ×2
Python	3.7.2
PyTorch	1.4.0
CUDA	10.0
Learning rate	10^{-3}
Momentum	0.73
Weight decay	5×10^{-4}
Batch	16
Saturation	1.7
Subdivisions	64

We chose overall accuracy (OA), confusion matrix (CM), giga multiply-accumulation operations per second (GMACs), and the number of model parameters as evaluation metrics. The above metrics mainly reflect the accuracy of the classification model, the easily confused scenarios in each dataset, and the evaluation model size and computational performance.

4.2. Comparison with SOTA Models

4.2.1. Experimental Results of AID

The experimental results on the AID are shown in Table 2. From the experimental results, we obtain the following findings:

1. Our proposed model achieved 95.09% and 97.31% at a training ratio of 20% and 50%, which improved 2.65%, 2.8%, and 1.98% compared to ResNet50 [62], ResNet50+CBAM [1], and ResNet50+HFAM [1], respectively. The experimental results indicate that our proposed model can achieve better classification results.
2. The model with an added attention mechanism has higher classification accuracy compared to traditional models without an added attention mechanism. For example, the ResNet50+CBAM [1] model improved by 0.13% compared to the ResNet50 [62] model, and the VGG16+HFAM [1] model improved by 6.25% compared to the VGG-VD-16 [10] model. The experimental results verified the positive role of the attention mechanism in scene classification.
3. In our proposed model, crucial feature information of key regions in the scene can be selectively focused on. According to the experimental results, our model improved by 0.67%, 0.4%, and 1.98% compared to the Fine-tune MobileNet V2 [63], DS-SURF-LLC+Mean-Std-LLC+MO-CLBP-LLC [64], and ResNet50+HFAM [1] models, respectively. Therefore, we believe that by combining channel attention and spatial attention, we can obtain more discriminative features and achieve better classification performance.

Table 2. Overall accuracies (%) of thirty-two kinds of methods and our method under the training ratios of 20% and 50% in AID.

Models	Training Ratios	
	20%	50%
CaffeNet [10]	86.72 ± 0.45	88.91 ± 0.26
VGG-VD-16 [10]	85.81 ± 0.25	89.36 ± 0.36
GoogLeNet [10]	83.27 ± 0.36	85.67 ± 0.55
Fusion by addition [65]	–	91.79 ± 0.26
LGRIN [30]	94.74 ± 0.23	97.65 ± 0.25
TEX-Net-LF [66]	93.91 ± 0.15	95.66 ± 0.17
DS-SURF-LLC+Mean-Std-LLC+MO-CLBP-LLC [64]	94.69 ± 0.22	96.57 ± 0.27
LiG with RBF kernel [55]	94.32 ± 0.23	96.22 ± 0.25
ADPC-Net [67]	88.61 ± 0.25	92.21 ± 0.26
VGG19 [62]	86.83 ± 0.26	91.83 ± 0.38
ResNet50 [1]	92.16 ± 0.18	95.51 ± 0.15
ResNet50+SE [1]	92.77 ± 0.18	95.84 ± 0.22
ResNet50+CBAM [1]	92.29 ± 0.15	95.38 ± 0.16
ResNet50+HFAM [1]	93.11 ± 0.20	95.86 ± 0.15
InceptionV3 [62]	92.65 ± 0.19	94.97 ± 0.22
DenseNet121 [68]	92.91 ± 0.25	94.65 ± 0.25
DenseNet169 [68]	92.39 ± 0.35	93.46 ± 0.27
MobileNet [69]	87.91 ± 0.16	91.23 ± 0.16
EfficientNet [70]	87.37 ± 0.16	89.41 ± 0.15
Two-stream deep fusion Framework [71]	92.42 ± 0.38	94.62 ± 0.27
Fine-tune MobileNet V2 [63]	94.42 ± 0.25	96.11 ± 0.25
SE-MDPMNet [63]	93.77 ± 0.16	97.23 ± 0.16
Two-stage deep feature Fusion [72]	–	93.87 ± 0.35
Contourlet CNN [73]	–	96.87 ± 0.42
LCPP [74]	91.12 ± 0.35	93.35 ± 0.35
RSNet [75]	94.62 ± 0.27	96.78 ± 0.56
SPG-GAN [76]	92.31 ± 0.17	94.53 ± 0.38
TSAN [77]	89.67 ± 0.23	92.16 ± 0.25
LGDL [29]	93.97 ± 0.16	97.29 ± 0.35
VGG16+CBAM [1]	91.91 ± 0.35	95.53 ± 0.07
VGG16+SE [1]	91.98 ± 0.31	95.45 ± 0.19
VGG16+HFAM [1]	92.06 ± 0.16	95.78 ± 0.21
Proposed	95.09 ± 0.15	97.31 ± 0.23

Figure 5 shows the CM for a training ratio of 50%. Our model can correctly identify most scenes in the AID, and some scenes achieve 100% accuracy, such as “Beach” and “Forest”. However, some scenes have also been confused, such as “School” and “Commercial”. Through further analysis, we find that the features in these two types of scenarios are highly similar, leading to confusion.

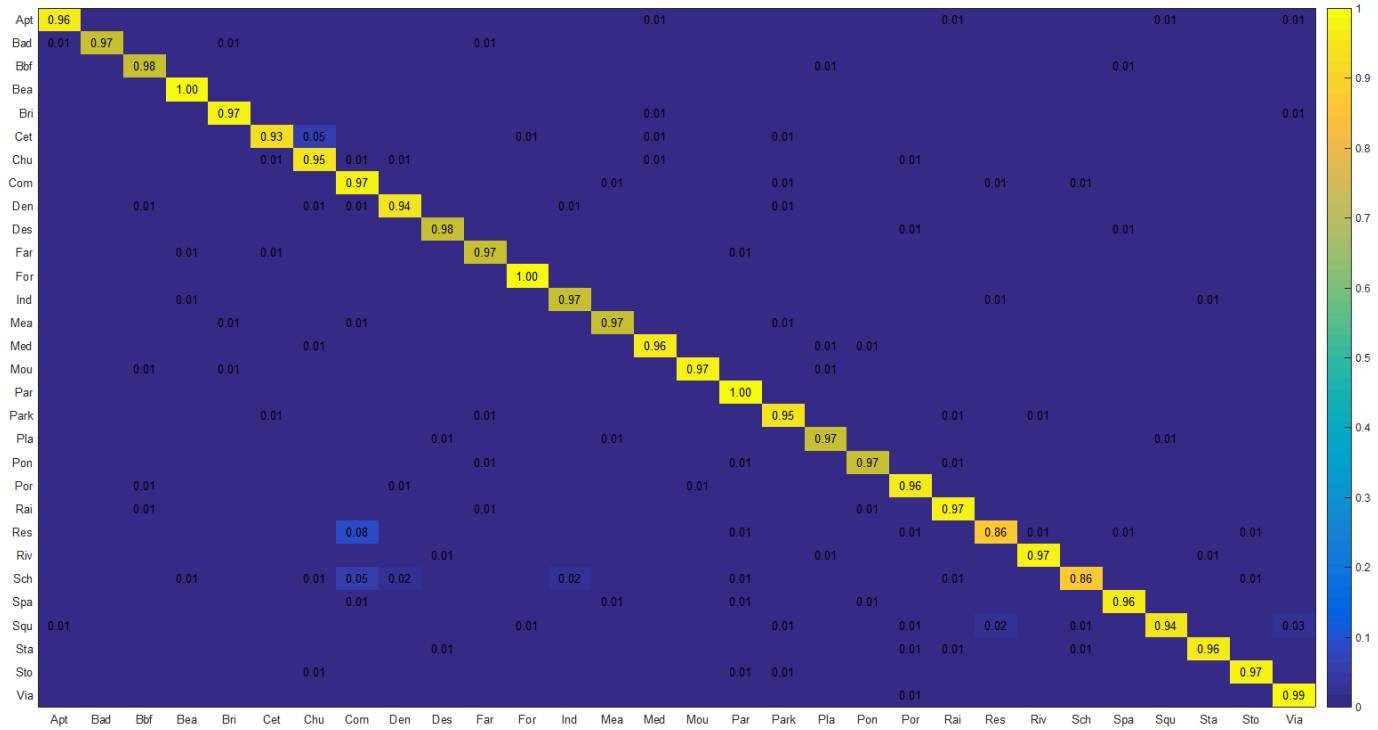

Figure 5. Confusion matrix on AID.

4.2.2. Experimental Results of RSICB-256

To further verify the validity of our model, a large number of experiments were conducted on the RSICB-256 dataset, and the experimental results are shown in Table 3. The following can be found from the table:

1. When the training ratio is 50%, our proposed model reaches 97.72%. And 0.2%, 1.37%, and 0.07% are, respectively, increased compared with VGG16+HFAM [1], SE-MDPMNet [63], and ResNet50+HFAM [1]. The experimental results further verify the validity of our model.
2. In general, the attention mechanism can improve the accuracy of classification. From our experiments, we found that our model improves 1.19%, 2%, and 1.55%, respectively, compared to models that use other attention mechanisms, such as ResNet50+SE [1], ResNet50+CBAM [1], and VGG16+CBAM [1].
3. The ViT-based model has achieved better performance because it uses the global attention mechanism to simulate the global environment. In our model, we utilize both the global attention mechanism and the local attention mechanism, and compared with the ViT-based model, the classification accuracy has been improved to some extent.

Table 3. Overall accuracies (%) of thirty-five kinds of methods and our method under the training ratios of 50% in RSICB-256.

Models	50%
CaffeNet [10]	91.37 ± 0.23
VGG-VD-16 [10]	92.44 ± 0.25
GoogLeNet [10]	89.87 ± 0.36
Fusion by addition [65]	93.36 ± 0.25

Table 3. *Cont.*

Models	50%
LGRIN [30]	97.55 ± 0.23
TEX-Net-LF [66]	95.34 ± 0.15
DS-SURF-LLC+Mean-Std-LLC+ MO-CLBP-LLC [64]	96.43 ± 0.25
LiG with RBF kernel [55]	95.37 ± 0.26
ADPC-Net [67]	92.19 ± 0.25
VGG19 [62]	92.67 ± 0.35
ResNet50 [1]	96.37 ± 0.15
ResNet50+SE [1]	96.53 ± 0.29
ResNet50+CBAM [1]	95.72 ± 0.26
ResNet50+HFAM [1]	97.65 ± 0.22
InceptionV3 [62]	94.53 ± 0.22
DenseNet121 [68]	94.21 ± 0.26
DenseNet169 [68]	93.27 ± 0.28
MobileNet [69]	91.33 ± 0.17
EfficientNet [70]	92.25 ± 0.18
Two-stream deep fusion Framework [71]	94.57 ± 0.25
Fine-tune MobileNet V2 [63]	95.83 ± 0.26
SE-MDPMNet [63]	96.35 ± 0.26
Two-stage deep feature Fusion [72]	94.89 ± 0.39
Contourlet CNN [73]	95.39 ± 0.29
LCPP [74]	93.72 ± 0.37
RSNet [75]	95.89 ± 0.41
SPG-GAN [76]	94.57 ± 0.35
TSAN [77]	93.12 ± 0.26
LGDL [29]	97.36 ± 0.32
ViT-B-16 [78]	97.37 ± 0.25
T2T-ViT-12 [79]	97.50 ± 0.22
PVT-V2-B0 [80]	97.45 ± 0.26
VGG16+CBAM [1]	96.17 ± 0.35
VGG16+SE [1]	95.87 ± 0.15
VGG16+HFAM [1]	97.52 ± 0.20
Proposed	97.72 ± 0.25

Figure 6 shows the CM on the RSICB-256 dataset. From Figure 6, we can find that for most scenarios on the RSICB-256 dataset, the classification accuracy of our proposed model is greater than 90%. This experimental result once again verifies the effectiveness of the global attention mechanism and the local attention mechanism in our model.

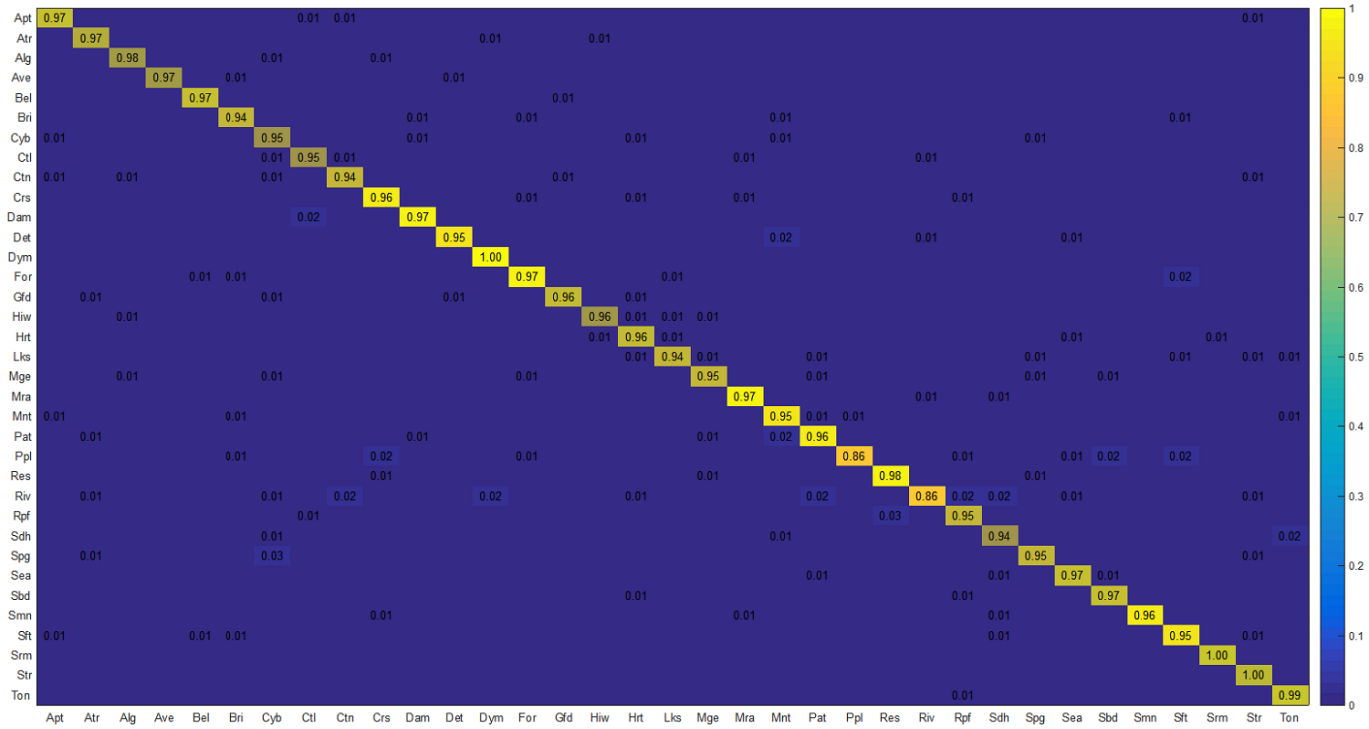

Figure 6. Confusion matrix on RSICB-256 dataset.

4.2.3. Experimental Results of NWPU-RESISC45

Compared with the above two datasets, the NWPU-RESISC45 dataset contains more scene categories and has a high degree of similarity between scenes and a large difference within scenes, which puts forward higher requirements for the performance of the model. Table 4 shows the following specific results:

1. Since the category of scenes has increased compared with the above two datasets, and the training ratio is only 10% and 20%, the classification results of all models have decreased compared with the above two datasets.
2. Compared with other models, our proposed model still has higher classification accuracy. Specifically, when the training ratio is 10%, the ratios of 0.16%, 2.54%, and 1.11% are increased, respectively, compared to ResNet50+EAM [81], ResNet101+HFAM [1], and ViT-B-16 [78]. When the training ratio is 20%, the ratios of 1.44%, 1.14%, and 4.03% are increased, respectively, compared to PVT-V2-B0 [80], LiG with RBF kernel [55], and ResNet101 [1]. Experimental results show that our proposed model is also effective on datasets with multiple scenes.
3. Under the same training ratio, the ViT-based model (such as ViT-B-16 [78], T2T-ViT-12 [79], PVT-V2-B0 [80]) achieves higher classification accuracy than the classical CNN model (such as GoogLeNet [82]), mainly because the ViT-based model makes up for the shortcomings of the classical CNN model in the global context. However, in addition to the global context feature information, the proposed model also utilizes a local spatial attention mechanism to extract local detail feature information, further filling the gap in local feature information. Furthermore, since the transformer-based method lacks convolutional inductive bias, it requires more training data samples. However, with a training ratio of 10% and 20%, for example, the classification accuracy of ViT-B-16 [78] is 90.96% and 93.36%, respectively. In terms of classification accuracy, the classification accuracy of the transformer-based method is lower than that of our model. In terms of computational complexity, however, it is more complex than our proposed model.

Table 4. Overall accuracies (%) of twenty-one kinds of methods and our method under the training ratios of 10% and 20% in NWPU-RESISC45.

Models	Training Ratios	
	10%	20%
GoogLeNet [82]	76.19 ± 0.38	78.48 ± 0.26
SCCov [83]	89.30 ± 0.35	92.10 ± 0.25
ACNet [61]	91.09 ± 0.13	92.42 ± 0.16
ViT-B-16 [78]	90.96 ± 0.08	93.36 ± 0.17
T2T-ViT-12 [79]	90.62 ± 0.18	93.19 ± 0.10
PVT-V2-B0 [80]	89.72 ± 0.16	92.95 ± 0.09
LGRIN [30]	91.91 ± 0.15	94.43 ± 0.16
LiG with RBF kernel [55]	90.23 ± 0.13	93.25 ± 0.12
ResNet50 [1]	87.43 ± 0.29	88.93 ± 0.12
ResNet50+EAM [81]	91.91 ± 0.22	94.29 ± 0.09
ResNet50+SE [1]	89.09 ± 0.14	91.37 ± 0.25
ResNet50+CBAM [1]	88.11 ± 0.39	90.27 ± 0.15
ResNet50+HFAM [1]	89.16 ± 0.06	91.49 ± 0.23
ResNet101 [1]	87.97 ± 0.44	90.36 ± 0.17
ResNet101+SE [1]	89.39 ± 0.14	91.46 ± 0.25
ResNet101+CBAM [1]	88.33 ± 0.26	90.47 ± 0.15
ResNet101+HFAM [1]	89.53 ± 0.29	91.67 ± 0.18
VGG16 [1]	86.44 ± 0.41	88.57 ± 0.16
VGG16+SE [1]	86.65 ± 0.26	88.75 ± 0.22
VGG16+CBAM [1]	86.84 ± 0.24	89.32 ± 0.15
VGG16+HFAM [1]	87.16 ± 0.22	90.21 ± 0.22
Proposed	92.07 ± 0.25	94.39 ± 0.25

Figure 7 shows the CM on the NWPU-RESISC45 dataset when the training ratio is 20%. Compared with the first two datasets, the NWPU-RESISC45 dataset is more challenging, so the classification accuracy of the 45 categories of scenes is reduced to a certain extent. Further analysis of the easily confused scenes shows that they contain a large number of the same target objects. For example, "Dense_residential" and "Medium_residential" scenes both contain trees, so their feature maps are relatively similar, resulting in confusion.

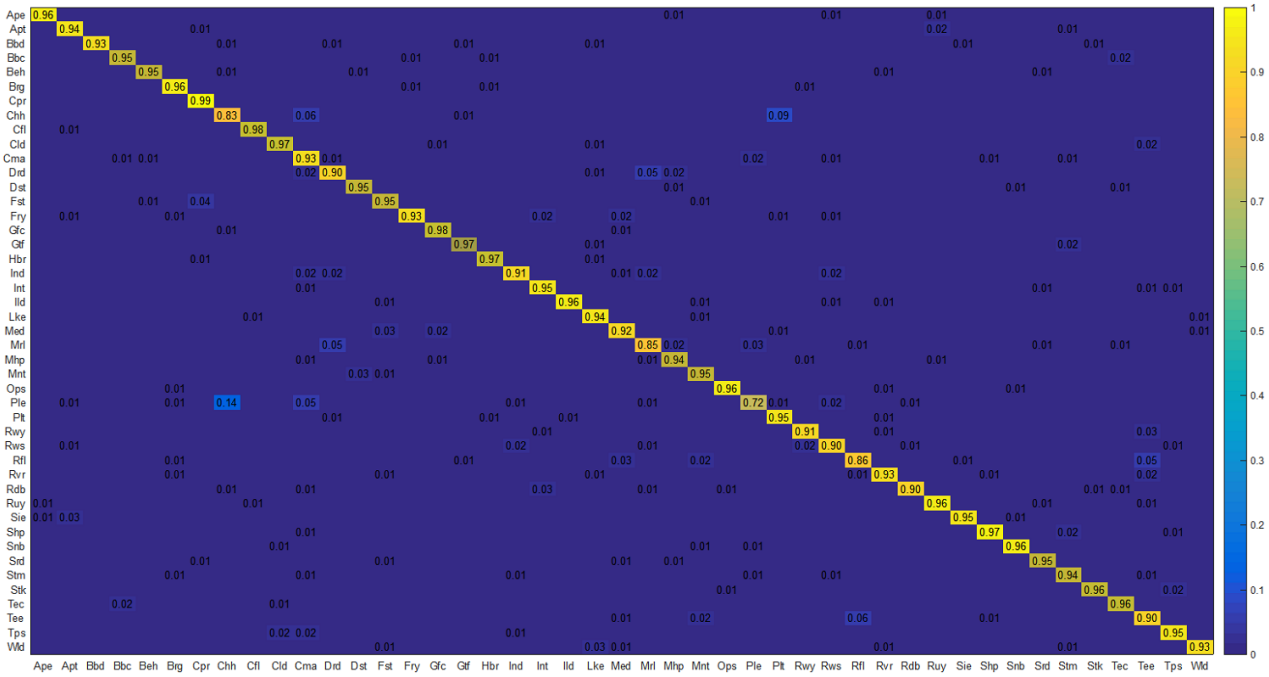

Figure 7. Confusion matrix on NWPU-RESISC45 dataset.

4.3. Comparison of the Number of Model Parameters and Computational Performance

In addition to comparing with the SOTA model, we also selected twelve models (such as VGG-VD-16 [10], ResNet50+CBAM [1], and ResNet50+HFAM [1]) for comparison. Specifically, we compared the classification accuracy, number of parameters, GMACs, and velocity of the model. These metrics mainly reflect the size, computational performance, and classification accuracy of the model.

As shown in Table 5, the details of the different models and our proposed model are listed. From Table 5, we can see the following:

1. When the classification accuracy of most models reaches 90%, they have a large number of parameters. Specifically, the OA of ResNet50+SE [1] is 95.84%, the parameter size is 26.28 M, the OA of Contourlet CNN [73] is 96.87%, the parameter size is 12.6 M, and our classification accuracy reaches 97.31%, but the parameters are smaller than theirs.
2. Compared with ResNet50+CBAM [1], VGG-VD-16 [10], and ResNet50+HFAM [1], our model decreased by 0.6592, 6.4765, and 0.5825, respectively, in the GMAC metric. The experimental results show that our model achieved better results in the above metrics, and the validity of our model is verified once again.

Table 5. Evaluation of size of models.

Models	Acc (%)	Parameters (M)	GMACs (G)	Velocity (Samples/s)
CaffeNet [10]	88.91	60.97	3.6532	32
GoogLeNet [10]	85.67	7	0.7500	37
VGG-VD-16 [10]	89.36	138.36	7.7500	35
LiG with RBF kernel [55]	96.22	2.07	0.2351	43
ResNet50 [1]	95.51	25.58	1.8555	38
ResNet50+SE [1]	95.84	26.28	1.9325	38
ResNet50+CBAM [1]	95.38	26.29	1.9327	38
ResNet50+HFAM [1]	95.86	25.58	1.8556	38
Inception V3 [62]	94.97	45.37	2.4356	21
Contourlet CNN [73]	96.87	12.6	1.0583	35
SPG-GAN [76]	94.53	87.36	2.1322	29
TSAN [77]	92.16	381.67	3.2531	32
Proposed	97.31	12.216	1.2735	45

4.4. Ablation Experiment

Taking the AID with a training ratio of 20% as an example, we evaluated the impact of each module on the accuracy of model classification. Starting from global feature extraction, we added local feature extraction, channel attention, spatial attention, and global-local fusion modules. The experimental results are shown in Table 6. From Table 6, we find that after adding the local feature extraction module, the model improved by 1.6% compared to utilizing only global features. The classification accuracy utilizing dual attention mechanisms (i.e., channel attention and spatial attention) is improved compared to utilizing one attention mechanism. The experimental results also verified the effectiveness of the above modules.

Table 6. Component ablation experiment results on AID.

Component	OA
Global Feature Extraction	88.65%
+Local Feature Extraction	90.25%
+Channel Attention	91.07%
+Spatial Attention	92.26%
+Fusion Module	95.09%

To verify the performance of the global-local fusion module at different stages, we also conducted ablation experiments, and the experimental results are shown in Table 7. If fusion is only used in the fourth stage, the classification accuracy obtained is lower. From the experimental results, we find that the more stages of fusion there are, the better the classification accuracy. The experimental results have verified the effectiveness of our model fusion mechanism, which can more comprehensively integrate global and local features, and consider the channel attention mechanism and spatial attention mechanism, effectively improving the accuracy of scene classification.

Table 7. Stage fusion ablation experiment results on AID.

Model	Stage 1	Stage 2	Stage 3	Stage 4	OA
Ours				✓	90.32%
			✓	✓	91.15%
		✓	✓	✓	93.27%
	✓	✓	✓	✓	95.09%

4.5. Visual Comparison of Different Attention Mechanisms

To further verify that our proposed attention mechanism can effectively capture feature information in the scene, we chose different attention mechanisms for comparison. Taking the AID with a training ratio of 20% as an example, the experimental results are shown in Figure 8. From Figure 8, we find that our proposed attention mechanism is more capable of obtaining features of key regions in the scene compared to other attention mechanisms and more accurately covers the key regions. The experimental results show that our proposed model can better integrate global and local features and help the model focus on more important regions, which is beneficial for the model to extract more discriminative features.

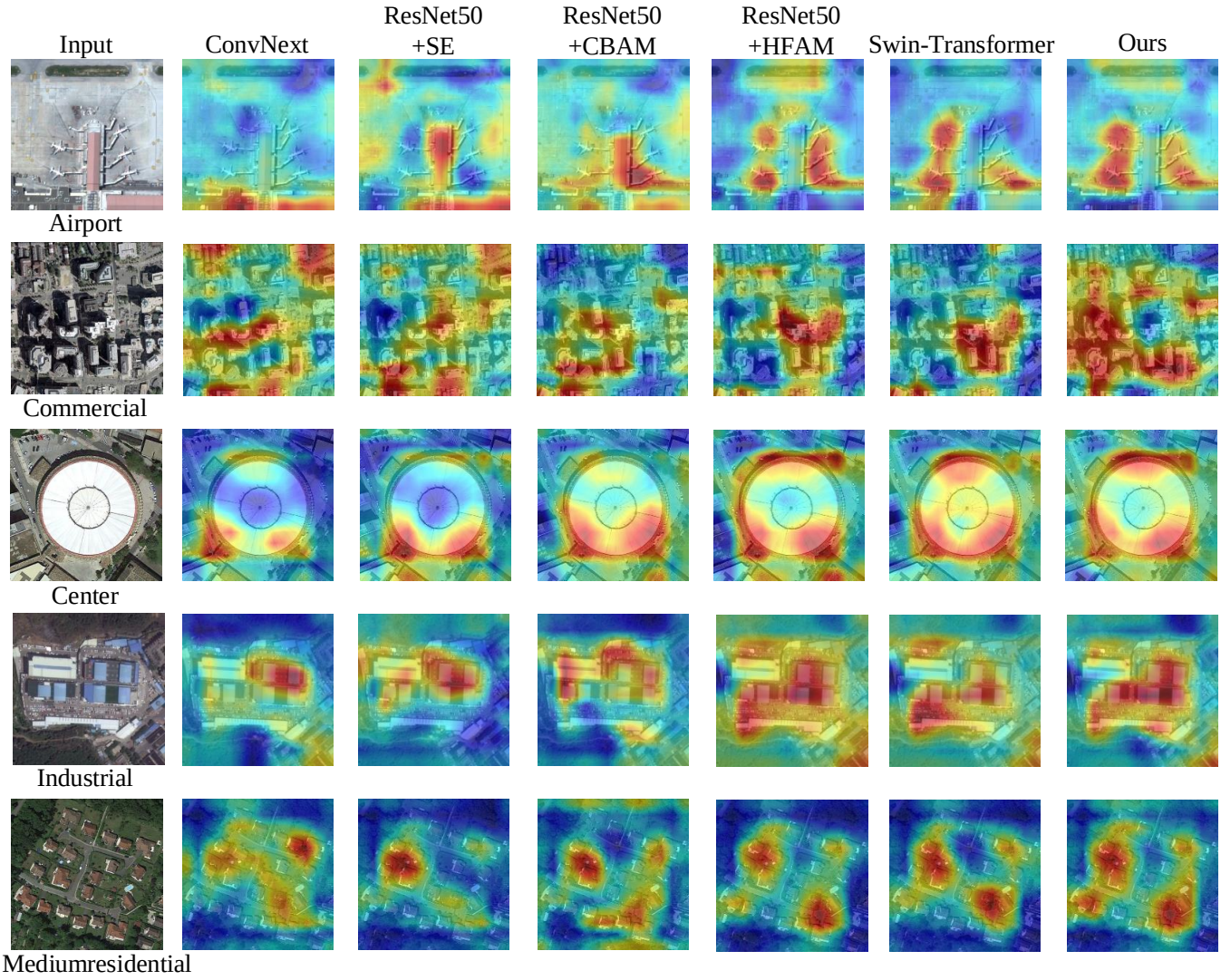

Figure 8. Ablation experiments were performed using Grad-CAM [84] to visualize the effects of other attention mechanisms on AID.

5. Conclusions

In this study, we propose a novel framework model for remote sensing scene classification, which mainly includes three branches: global extraction module, local extraction module, and global-local fusion module. The global and local extraction modules we designed can efficiently extract multi-scale features and combine channel attention and spatial attention to effectively focus on key regions in the scene. In addition, to improve computational performance, the model adopts parallel dilated convolution, effectively increasing the receptive field and reducing the number of feature parameters. A large number of experimental results have verified the effectiveness of our model, which can effectively distinguish different scenarios. Our model has significant improvement compared with other models. Taking AID data as examples, the classification accuracy can be improved by 8.4% compared with the traditional model, and it also improves by 1.93% compared with other attention mechanisms. In addition, our model has also decreased by 48.754 M compared to other models in terms of parameter quantity and other indicators. We believe that this research work can provide basic services for other applications in remote sensing image interpretation. In future research, we will explore the fusion model of Lie Group space learning with feature extraction, attention mechanism, and other related methods to further improve the classification accuracy and reduce the complexity of the model.

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

Funding: This work was supported by the National Natural Science Foundation of China under grant number 42261068 ("Research on Urban Land-use Scene Classification Based on Lie Group Spatial Learning and Heterogeneous Feature Modeling of Multi-source Data").

Data Availability Statement: The data associated with this research are available online. The RSICB-256 dataset is available for download at https://github.com/lehaifeng/RSI-CB (accessed on 12 November 2023). The AID is available for download at https://captain-whu.github.io/AID/ (accessed on 15 December 2021). The NWPU-RESISC45 dataset is available for download at http://www.escience.cn/people/Junwei_\Han/NWPURE-SISC45.html (accessed on 16 October 2020).

Acknowledgments: The authors would like to thank the five anonymous reviewers for carefully reviewing this study and giving valuable comments to improve this study.

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

Abbreviations

AID	Aerial Image Dataset
BAM	Bottleneck Attention Module
BN	Batch Normalization
BoVW	Bag of Visual Words
CBAM	Convolutional Block Attention Module
CM	Confusion Matrix
CNN	Convolutional Neural Network
CV	Computer Vision
GELU	Gaussian Error Linear Unit
GFE	Global Feature Extraction
GLF	Global-Local Fusion
GMACs	Giga Multiply-Accumulation operations per Second
HRRSI	High-Resolution Remote Sensing Image
LBP	Local Binary Pattern
LFE	Local Feature Extraction
LN	Layer Normalization
MCNN	Multi-scale Convolutional Neural Network
MSA	Multi-head Attention
NLP	Natural Language Processing
NWPU-RESISC	Northwestern Polytechnical University Remote Sensing Image Scene Classification
OA	Overall Accuracy
SE	Squeeze and Excite
RSSC	Remote Sensing Scene Classification
RSICB	Remote Sensing Image Classification Benchmark
ViT	Vision Transformer
W-MAS	Windows Multi-head Self-attention

References

1. Wan, Q.; Qiao, Z.; Yu, Y.; Liu, Z.; Wang, K.; Li, D. A Hyperparameter-Free Attention Module Based on Feature Map Mathematical Calculation for Remote-Sensing Image Scene Classification. *IEEE Trans. Geosci. Remote Sens.* **2023**, *62*, 5600318. [CrossRef]
2. Xu, C.; Shu, J.; Zhu, G. Multi-Feature Dynamic Fusion Cross-Domain Scene Classification Model Based on Lie Group Space. *Remote Sens.* **2023**, *15*, 4790. [CrossRef]

3. Xu, C.; Shu, J.; Zhu, G. Adversarial Remote Sensing Scene Classification Based on Lie Group Feature Learning. *Remote Sens.* **2023**, *15*, 914. [CrossRef]
4. 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]
5. Bai, L.; Liu, Q.; Li, C.; Zhu, C.; Ye, Z.; Xi, M. A lightweight and multiscale network for remote sensing image scene classification. *IEEE Geosci. Remote Sens. Lett.* **2022**, *19*, 8012605. [CrossRef]
6. Bai, L.; Liu, Q.; Li, C.; Ye, Z.; Hui, M.; Jia, X. Remote sensing image scene classification using multiscale feature fusion covariance network with octave convolution. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 5620214. [CrossRef]
7. Vetrivel, A.; Gerke, M.; Kerle, N.; Nex, F.; Vosselman, G. Disaster damage detection through synergistic use of deep learning and 3D point cloud features derived from very high resolution oblique aerial images and multiple-kernel-learning. *ISPRS J. Photogramm. Remote Sens.* **2018**, *140*, 45–59. [CrossRef]
8. Zheng, K.; Gao, L.; Hong, D.; Zhang, B.; Chanussot, J. NonRegSRNet: A nonrigid registration hyperspectral super-resolution network. *IEEE Trans. Geosci. Remote Sens.* **2021**, *60*, 5520216. [CrossRef]
9. Li, J.; Hong, D.; Gao, L.; Yao, J.; Zheng, K.; Zhang, B.; Chanussot, J. Deep learning in multimodal remote sensing data fusion: A comprehensive review. *Int. J. Appl. Earth Observ. Geoinf.* **2022**, *112*, 102926. [CrossRef]
10. Xia, G.S.; Hu, J.; Hu, F.; Shi, B.; Bai, X.; Zhong, Y.; Lu, X. AID: A benchmark dataset for performance evaluation of aerial scene classification. *IEEE Trans. Geosci. Remote Sens.* **2017**, *55*, 3965–3981. [CrossRef]
11. Wang, Q.; Liu, S.; Chanussot, J.; Li, X. Hybrid feature aligned network for salient object detection in optical remote sensing imagery. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 5624915. [CrossRef]
12. Wang, Z.; Chen, J.; Hoi, S.C.H. Deep learning for image super-resolution: A survey. *IEEE Trans. Pattern Anal. Mach. Intell.* **2020**, *43*, 3365–3387. [CrossRef] [PubMed]
13. Wang, X.; Wang, S.; Ning, C.; Zhou, H. Enhanced feature pyramid network with deep semantic embedding for remote sensing scene classification. *IEEE Trans. Geosci. Remote Sens.* **2021**, *59*, 7918–7932. [CrossRef]
14. Su, Y.; Gao, L.; Jiang, M.; Plaza, A.; Sun, X.; Zhang, B. NSCKL: Normalized spectral clustering with kernel-based learning for semisupervised hyperspectral image classification. *IEEE Trans. Cybern.* **2022**, *53*, 6649–6662. [CrossRef]
15. Qin, A.; Chen, F.; Li, Q.; Tang, L.; Yang, F.; Zhao, Y.; Gao, C. Deep Updated Subspace Networks for Few-Shot Remote Sensing Scene Classification. *IEEE Trans. Geosci. Remote Sens.* **2024**, *62*, 5606714. [CrossRef]
16. Ma, A.; Wan, Y.; Zhong, Y.; Wang, J.; Zhang, L. SceneNet: Remote sensing scene classification deep learning network using multi-objective neural evolution architecture search. *ISPRS J. Photogramm. Remote Sens.* **2021**, *172*, 171–188. [CrossRef]
17. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. *Proc. Conf. Neural Inf. Process. Syst.* **2017**, *30*, 5998–6008. [CrossRef]
18. Bazi, Y.; Bashmal, L.; Rahhal, M.M.A.; Dayil, R.A.; Ajlan, N.A. Vision transformers for remote sensing image classification. *Remote Sens.* **2021**, *13*, 516. [CrossRef]
19. Lv, P.; Wu, W.; Zhong, Y.; Du, F.; Zhang, L. SCViT: A spatial-channel feature preserving vision transformer for remote sensing image scene classification. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 4409512. [CrossRef]
20. 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*, 4409512. [CrossRef]
21. Huo, X.; Sun, G.; Tian, S.; Wang, Y.; Yu, L.; Long, J.; Zhang, W.; Li, A. HiFuse: Hierarchical multi-scale feature fusion network for medical image classification. *Biomed Signal Process.* **2024**, *87*, 105534. [CrossRef]
22. Xu, Y.; Zhang, Q.; Zhang, J.; Tao, D. Vitae: Vision transformer advanced by exploring intrinsic inductive bias. *Biomed Signal Process.* **2021**, *34*, 28522–28535. [CrossRef]
23. Fu, B.; Zhang, M.; He, J.; Cao, Y.; Guo, Y.; Wang, R. StoHisNet: A hybrid multi-classification model with CNN and transformer for gastric pathology images. *Biomed Signal Process.* **2021**, *34*, 28522–28535. [CrossRef]
24. Zhang, Y.; Liu, H.; Hu, Q. Transfuse: Fusing transformers and cnns for medical image segmentation. *MICCAI 2021* **2021**, 14–24. [CrossRef]
25. Guo, J.; Han, K.; Wu, H.; Tang, Y.; Chen, X.; Wang, Y.; Xu, C. Cmt: Convolutional neural networks meet vision transformers. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 12175–12185. [CrossRef]
26. Peng, Z.; Huang, W.; Gu, S.; Xie, L.; Wang, Y.; Jiao, J.; Ye, Q. Conformer: Local features coupling global representations for visual recognition. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Virtual, 11–17 October 2021; pp. 367–376. [CrossRef]
27. Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; Guo, B. Swin transformer: Hierarchical vision transformer using shifted windows. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, QC, Canada, 10–17 October 2021; pp. 10012–10022. [CrossRef]
28. Xu, C.; Shu, J.; Zhu, G. Scene Classification Based on Heterogeneous Features of Multi-Source Data. *Remote Sens.* **2023**, *15*, 325. [CrossRef]
29. 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]

30. Xu, C.; Zhu, G.; Shu, J. A Lightweight and Robust Lie Group-Convolutional Neural Networks Joint Representation for Remote Sensing Scene Classification. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 5501415. [CrossRef]
31. Xu, C.; Zhu, G.; Shu, J. Lie Group spatial attention mechanism model for remote sensing scene classification. *Int. J. Remote Sens.* **2022**, *43*, 2461–2474. [CrossRef]
32. Huang, B.; Zhao, B.; Song, Y. Urban land-use mapping using a deep convolutional neural network with high spatial resolution multispectral remote sensing imagery. *Remote Sens. Environ.* **2018**, *214*, 73–86. [CrossRef]
33. Ojala, T.; Pietikainen, M.; Maenpaa, T. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *IEEE Trans. Pattern Anal. Mach. Intell.* **2002**, *24*, 971–987. [CrossRef]
34. dos Santos, J.A.; Penatti, O.A.; Torres, R.D.S. Evaluating the potential of texture and color descriptors for remote sensing image retrieval and classification. *ICCV* **2010**, *2*, 203–208. [CrossRef]
35. 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, New York, NY, USA, 2–5 November 2010; pp. 270–279. [CrossRef]
36. 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**, *6*, 747–751. [CrossRef]
37. Cheng, G.; Han, J.; Lu, X. Remote sensing image scene classification: Benchmark and state of the art. *Proc. IEEE* **2017**, *105*, 1865–1883. [CrossRef]
38. Li, M.; Stein, A.; Bijker, W.; Zhan, Q. Urban land use extraction from very high resolution remote sensing imagery using a Bayesian network. *ISPRS J. Photogramm. Remote Sens.* **2016**, *122*, 192–205. [CrossRef]
39. Zhang, F.; Du, B.; Zhang, L. Scene classification via a gradient boosting random convolutional network framework. *IEEE Trans. Geosci. Remote Sens.* **2016**, *54*, 1793–1802. [CrossRef]
40. 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]
41. Liu, Y.; Zhong, Y.; Qin, Q. Scene classification based on multiscale convolutional neural network. *IEEE Trans. Geosci. Remote Sens.* **2018**, *56*, 7109–7121. [CrossRef]
42. Tang, X.; Li, M.; Ma, J.; Zhang, X.; Liu, F.; Jiao, L. EMTCAL: Efficient multiscale transformer and cross-level attention learning for remote sensing scene classification. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 5626915. [CrossRef]
43. Chen, S.-B.; Wei, Q.-S.; Wang, W.-Z.; Tang, J.; Luo, B.; Wang, Z.-Y. Remote sensing scene classification via multi-branch local attention network. *IEEE Trans. Image Process.* **2022**, *31*, 99–109. [CrossRef]
44. Hu, J.; Shen, L.; Sun, G. Squeeze-and-excitation networks. *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.* **2018**, *31*, 7132–7141. [CrossRef]
45. Hu, J.; Shen, L.; Albanie, S.; Sun, G.; Vedaldi, A. Gather-excite: Exploiting feature context in convolutional neural networks. *Proc. Adv. Neural Inf. Process. Syst.* **2018**, *31*, 9423–9433. [CrossRef]
46. Woo, S.; Park, J.; Lee, J.-Y.; Kweon, I.S. CBAM: Convolutional block attention module. In *Computer Vision—ECCV*; Springer: Munich, Germany, 2018; pp. 3–19. [CrossRef]
47. Song, H.; Deng, B.; Pound, M.; Özcan, E.; Triguero, I. A fusion spatial attention approach for few-shot learning. *Inf. Fusion.* **2022**, *81*, 187–202. [CrossRef]
48. Qin, Z.; Wang, H.; Mawuli, C.B.; Han, W.; Zhang, R.; Yang, Q.; Shao, J. Multi-instance attention network for few-shot learning. *Inf. Fusion.* **2022**, *611*, 464–475. [CrossRef]
49. Park, J.; Woo, S.; Lee, J.-Y.; Kweon, I.S. BAM: Bottleneck attention module. *arXiv* **2018**, arXiv:1807.06514. [CrossRef]
50. Zhang, Q.-L.; Yang, Y.-B. SA-Net: Shuffle attention for deep convolutional neural networks. In Proceedings of the ICASSP 2021—2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Toronto, ON, Canada, 6–11 June 2021; pp. 2235–2239. [CrossRef]
51. Li, H.; Deng, W.; Zhu, Q.; Guan, Q.; Luo, J.C. Local-Global Context-Aware Generative Dual-Region Adversarial Networks for Remote Sensing Scene Image Super-Resolution. *IEEE Trans. Geosci. Remote Sens.* **2024**, *62*, 5402114. [CrossRef]
52. Wang, Q.; Liu, S.; Chanussot, J.; Li, X. Scene classification with recurrent attention of VHR remote sensing images. *IEEE Trans. Geosci. Remote Sens.* **2019**, *57*, 1155–1167. [CrossRef]
53. 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 Observ. Remote Sens.* **2020**, *13*, 6372–6383. [CrossRef]
54. Xu, C.; G, Z.; Shu, J. Robust Joint Representation of Intrinsic Mean and Kernel Function of Lie Group for Remote Sensing Scene Classification. *IEEE Geosci. Remote Sens. Lett.* **2020**, *118*, 796–800. [CrossRef]
55. Xu, C.; G, Z.; Shu, J. A Lightweight Intrinsic Mean for Remote Sensing Classification With Lie Group Kernel Function. *IEEE Geosci. Remote Sens. Lett.* **2020**, *18*, 1741–1745. [CrossRef]
56. Chattopadhay, A.; Sarkar, A.; Howlader, P.; Balasubramania, V.N. Grad-CAM++: Generalized gradient-based visual explanations for deep convolutional networks. *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)* **2018**, *14*, 839–847. [CrossRef]
57. der Maaten, L.V.; Hinton, G. Visualizing data using t-SNE. *J. Mach. Learn. Res.* **2008**, *9*, 2579–2605. Available online: http://jmlr.org/papers/volume9/vandermaaten08a/vandermaaten08a.pdf (accessed on 5 May 2024).

58. Zhao, Y.; Chen, Y.; Rong, Y.; Xiong, S.; Lu, X. Global-Group Attention Network With Focal Attention Loss for Aerial Scene Classification. *IEEE Trans. Geosci. Remote Sens.* **2024**, *62*, 1–14. [CrossRef]
59. Li, H.; Dou, X.; Tao, C.; Hou, Z.; Chen, J.; Peng, J.; Deng, M.; Zhao, L. RSI-CB: A large-scale remote sensing image classification benchmark using crowdsourced data. *Sensors* **2020**, *20*, 1594. [CrossRef] [PubMed]
60. Xu, K.; Huang, H.; Deng, P.; Li, Y. Deep feature aggregation framework driven by graph convolutional network for scene classification in remote sensing. *IEEE Trans. Neural Netw. Learn. Syst.* **2022**, *33*, 5751–5765. [CrossRef]
61. Tang, X.; Ma, Q.; Zhang, X.; Liu, F.; Ma, J.; Jiao, L. Attention consistent network for remote sensing scene classification. *IEEE J. Sel. Top. Appl. Earth Observ. Remote Sens.* **2021**, *14*, 2030–2045. [CrossRef]
62. Li, W.; Wang, Z.; Wang, Y.; Wu, J.; Wang, J.; Jia, Y.; Gui, G. Classification of high spatial resolution remote sensing scenes methodusing transfer learning and deep convolutional neural network. *IEEE J. Sel. Top. Appl. Earth Observ. Remote Sens.* **2020**, *13*, 1986–1995. [CrossRef]
63. 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 Observ. Remote Sens.* **2019**, *12*, 2636–2653. [CrossRef]
64. Wang, X.; Xu, M.; Xiong, X.; Ning, C. Remote Sensing Scene Classification Using Heterogeneous Feature Extraction and Multi-Level Fusion. *IEEE Access* **2020**, *8*, 217628–217641. [CrossRef]
65. Chaib, S.; Liu, H.; Gu, Y.; Yao, H. Deep feature fusion for VHR remote sensing scene classification. *IEEE Trans. Geosci. Remote Sens.* **2017**, *55*, 4775–4784. [CrossRef]
66. Anwer, R.M.; Khan, F.S.; van de Weijer, J.; Molinier, M.; Laaksonen, J. Binary patterns encoded convolutional neural networks for texture recognition and remote sensing scene classification. *ISPRS J. Photogramm. Remote Sens.* **2018**, *138*, 74–85. [CrossRef]
67. Bi, Q.; Qin, K.; Zhang, H.; Xie, J.; Li, Z.; Xu, K. APDC-Net: Attention pooling-based convolutional network for aerial scene classification. *Remote Sens. Lett.* **2019**, *9*, 1603–1607. [CrossRef]
68. Aral, R.A.; Keskin, Ş.R.; Kaya, M.; Hacıömeroğlu, M. Classification of trashnet dataset based on deep learning models. In Proceedings of the 2018 IEEE International Conference on Big Data (Big Data), Seattle, WA, USA, 10–13 December 2018; pp. 1986–1995. [CrossRef]
69. 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]
70. Pour, A.M.; Seyedarabi, H.; Jahromi, S.H.A.; Javadzadeh, A. Automatic Detection and Monitoring of Diabetic Retinopathy using Efficient Convolutional Neural Networks and Contrast Limited Adaptive Histogram Equalization. *IEEE Access* **2020**, *8*, 136668–136673. [CrossRef]
71. Yu, Y.; Liu, F. A two-stream deep fusion framework for high-resolution aerial scene classification. *Comput. Intell. Neurosci.* **2018**, *2018*, 1986–1995. [CrossRef] [PubMed]
72. Liu, Y.; Liu, Y.; Ding, L. Scene classification based on two-stage deep feature fusion. *IEEE Geosci. Remote Sens. Lett.* **2018**, *15*, 183–186. [CrossRef]
73. Liu, M.; Jiao, L.; Liu, X.; Li, L.; Liu, F.; Yang, S. C-CNN: Contourlet convolutional neural networks. *IEEE Trans. Neural Netw. Learn. Syst.* **2020**, *32*, 2636–2649. [CrossRef]
74. Sun, X.; Zhu, Q.; Qin, Q. A Multi-Level Convolution Pyramid Semantic Fusion Framework for High-Resolution Remote Sensing Image Scene Classification and Annotation. *IEEE Access* **2021**, *9*, 18195–18208. [CrossRef]
75. Wang, J.; Zhong, Y.; Zheng, Z.; Ma, A.; Zhang, L. RSNet: The search for remote sensing deep neural networks in recognition tasks. *IEEE Trans. Geosci. Remote Sens.* **2020**, *59*, 2520–2534. [CrossRef]
76. 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]
77. Zheng, J.; Wu, W.; Yuan, S.; Zhao, Y.; Li, W.; Zhang, L.; Dong, R.; Fu, H. A Two-Stage Adaptation Network (TSAN) for Remote Sensing Scene Classification in Single-Source-Mixed-Multiple-Target Domain Adaptation (S^2M^2T DA) Scenarios. *IEEE Trans. Geosci. Remote Sens.* **2021**, *60*, 5609213. [CrossRef]
78. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. An image is worth 16×16 words: Transformers for image recognition at scale. In Proceedings of the 9th International Conference on Learning Representations, Virtual Event, 3–7 May 2021; pp. 1–22. [CrossRef]
79. Yuan, L.; Chen, Y.; Wang, T.; Yu, W.; Shi, Y.; Jiang, Z.H.; Tay, F.E.; Feng, J.; Yan, S. Tokens-to-token vit: Training vision transformers from scratch on imagenet. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, QC, Canada, 10–17 October 2021; pp. 538–547. [CrossRef]
80. Wang, W.; Xie, E.; Li, X.; Fan, D.P.; Song, K.; Liang, D.; Lu, T.; Luo, P.; Shao, L. Pvt v2: Improved baselines with pyramid vision transformer. *Comput. Vis. Media* **2022**, *8*, 415–424. [CrossRef]
81. 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]
82. Szegedy, C.; Liu, W.; Jia, Y.; Sermanet, P.; Reed, S.; Anguelov, D.; Erhan, D.; Vanhoucke, V.; Rabinovich, A. Going deeper with convolutions. In Proceedings of the IEEE Conference on Computer Vision and Pattern RECOGNITION (CVPR), Boston, MA, USA, 7–12 June 2015; pp. 1–9. [CrossRef]

83. 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.* **2020**, *31*, 1461–1474. [CrossRef]
84. Selvaraju, R.R.; Cogswell, M.; Das, A.; Vedantam, R.; Parikh, D.; Batra, D. Grad-CAM: Visual explanations from deep networks via gradient-based localization. In Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV 2017), Venice, Italy, 22–29 October 2017; pp. 618–626. [CrossRef]