Remote Sensing Image Scene Classification Based on Multidimensional Attention and Feature Enhancement

Chengrui Liu, Hong Dai*, Shuang Wang, and Junhong Chen

***Abstract*—Remote sensing image scene classification is a challenging task that involves automatically assigning labels to remote sensing images based on predefined categories. The inherent intra-class diversity and inter-class similarity of remote sensing images make it difficult for classification models to capture the discriminative key information necessary for accurate labeling, resulting in classification confusion. This paper proposes a novel method called Multidimensional Attention and Feature Enhancement (MA-FE) to address this issue. The proposed MA-FE method comprehensively captures essential features in different dimensions of channel and position through the Multidimensional Attention (MA) module, which integrates and combines the captured features. The Feature Enhancement (FE) module then amplifies the discriminative features to suppress the interference of useless information, thus improving the representation ability of the model. We conducted detailed experiments on three public remote sensing datasets and performed a comparative evaluation with multiple remote sensing scene classification methods proposed in recent years. The overall accuracies of the proposed MA-FE method on these datasets were 99.66%, 95.68%, and 93.21%, respectively. Our experimental results demonstrate that the proposed MA-FE method is more effective in extracting complex features in remote sensing images than other methods, thereby proving its effectiveness.**

***Index Terms*—remote sensing images, scene classification, multidimensional attention, feature enhancement**

I. Introduction

With the rapid advancement of satellite sensing technology, acquiring high-resolution image data of the Earth's surface through remote sensing technology has become increasingly convenient. These remote sensing images contain rich semantic information, enabling researchers to perform more comprehensive metric analysis of the Earth's surface [1]. One of the key goals of remote sensing image scene classification is to automatically assign images into predefined categories [2-4], such as vegetation, water bodies, buildings, etc., using pixel points or regions in the remote sensing image, which is an essential way to comprehend remote sensing images. In recent years, remote sensing image scene classification has emerged as an important research area in the field of remote sensing image processing, and has found extensive applications in various domains, such as natural resource management, urban planning, environmental monitoring, military surveying, among others [5-8].

Improving accuracy in remote sensing image scene classification is a challenging task, primarily due to two main reasons. Firstly, the aerial overhead angle of such images can capture multiple categories of disturbed feature types within one image, leading to increased complexity and potential confusion in classification targets, such as stadiums and railway stations. Secondly, remote sensing images depict objects amidst complex backgrounds and varying scales, leading to increased similarity in appearance and characteristics among distinct feature categories. For instance, both schools and parks possess vegetation, making it difficult for the classification network to differentiate between them accurately [9]. It is clear that the accuracy of remote sensing image scene classification is closely related to the feature extraction method from the images.

In the early days of processing remote sensing images, researchers typically extracted features using handmade structures such as pixels, textures, and spectra, these methods usually relied on professionals. Then they used these features as input to a classifier to achieve scene classification. Several methods were proposed to extract such features, including the scale-invariant feature transform (SIFT) method by Lowe et al. [10], the local binary patterns (LSPs) method by Ojala et al. [11, 12], the histogram of oriented gradients (HOG) method by Dalal et al. [13], the bag of visual words (BOVW) method by Yang et al. [14], and the spatial pyramid matching (SPM) method by Lazebnik et al. [15]. However, these manual feature extraction methods are mainly designed for shallow local information, and the operations are relatively tedious and lack the effective perception of higher-level semantic information, which is unfavorable for the complex background of remote sensing images. Deep learning has become the mainstream choice for scene classification due to its powerful ability to learn and extract discriminative and abstract features of high-level semantic information [16, 17].

Manuscript received May 12, 2023; revised September 5, 2023. The research work was supported by Graduate student science and technology innovation project of University of Science and Technology Liaoning (No. LKDYC202221).

Chengrui Liu is a postgraduate student of University of Science and Technology Liaoning, Anshan, Liaoning, CO 114051, China. (e-mail:641588835@qq.com).

Hong Dai* is a professor of School of Computer Science and Software Engineering, University of Science and Technology Liaoning, Anshan, Liaoning, China. (corresponding author to provide phone: +086-186-4226-8599; fax:0412-5929818; e-mail: dear_red9@163.com).

Shuang Wang is a postgraduate student of University of Science and Technology Liaoning, Anshan, Liaoning, CO 114051, China. (e-mail: 1657669526@qq.com).

Junhong Chen is a postgraduate student of University of Science and Technology Liaoning, Anshan, Liaoning, CO 114051, China. (e-mail:shopkeeperakashi@hotmail.com).

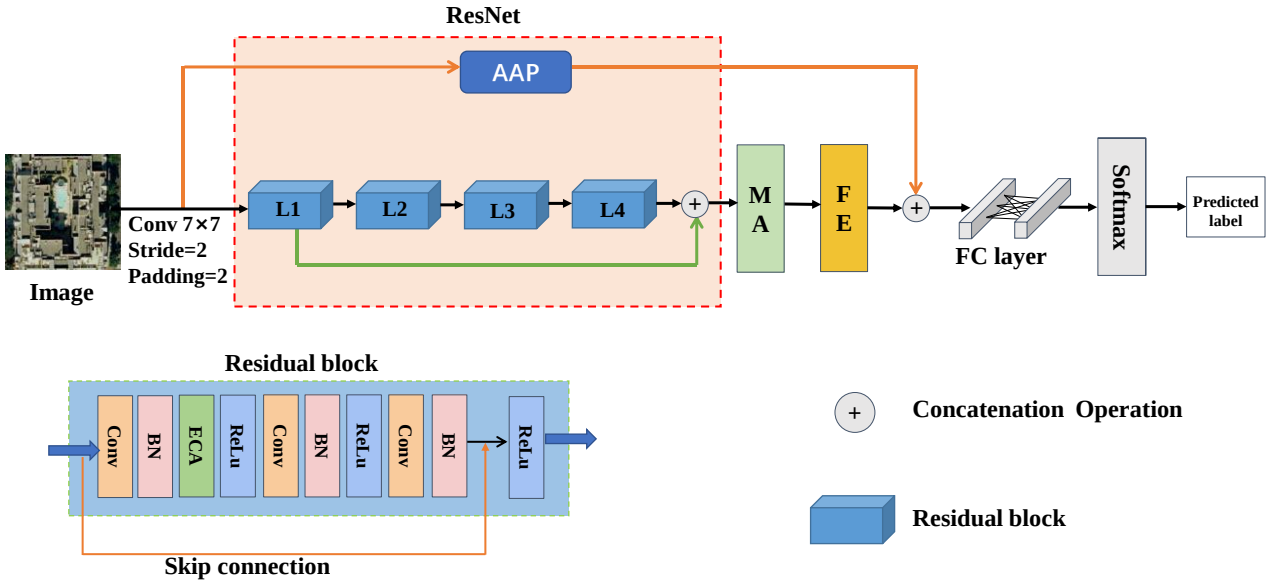

Fig. 1 Schematic diagram of MA-FE model

In conclusion, to enhance the model's emphasis on critical feature map areas, diminish the impact of irrelevant information, enhance the model's capacity for discrimination, and consequently enhance classification accuracy, this paper introduces a method involving multidimensional attention and feature enhancement. This method consists of three parts. The first part is to extract the channel features through a pre-trained convolutional neural network model. The second part is to extract the location features of the image and combine them with the channel features using the multidimensional attention module. The third part is to enhance the features in the focused attention regions using a feature enhancement module. This aims to improve the overall representativeness and adaptability of the model while reducing the influence of interfering features and improving the classification accuracy.

II. Related Work

In recent years, Convolutional Neural Networks (CNNs) have gained great popularity in the field of scene classification[18]. Classical networks like GoogleNet [3], MobileNet [19], and EfficientNet [20] have been developed to improve classification accuracy by learning high-level information. He et al. [21] proposed a residual block to enhance the model's memory capacity, effectively reducing the probability of gradient explosion and gradient disappearance during weight training. However, CNN's ability to generalize and capture key discriminative regions is weak. Therefore, CNNs are often combined with attention mechanisms. Hou et al. [22] proposed a Coordinate Attention (CA) module that effectively extracts location information by integrating feature vectors of two directional coordinates to prevent overfitting. Cao et al. [23] proposed a VGG_VD16 with SAFF, combining a pre-trained VGG network with Self-Attentive Feature Fusion (SAFF) with aggregated weighting capability to extract scene features. Tang et al. [24] proposed an Attention-Consistent module (ACNet) for feature extraction. Wang et al.[25] proposed an effective channel attention that emphasizes the important information of features from the perspective of channels, thereby improving classification accuracy. However, the above approaches face two problems. Firstly, most attention mechanisms extract features in a single dimension of space or location, making the model weak in focusing on discriminative essential information. Secondly, due to the increasing depth of the network, the model is prone to forgetting the features learned at the shallow level and receiving interference information, which eventually leads to confusion in classification[26].

In order to tackle the previously mentioned problems and improve the model's classification capability, we propose a multi-dimensional attention mechanism and feature enhancement model using ResNet50 pre-trained on the ImageNet dataset as the baseline. This model emphasizes key information from both location and channel dimensions, thereby addressing the limitations of insufficient focal information encountered in most attention methods. Furthermore, the feature enhancement module prioritizes the extraction of discriminative information from the upper layer while suppressing interference information.

III. Proposed Model

A. Baseline

ResNet50 is a well-established deep learning architecture that has showcased its success across a variety of computer vision tasks, such as scene classification, target detection, and semantic segmentation. The network itself is composed of several layers, encompassing a convolutional layer, a batch normalization layer, a ReLU activation function layer, a maximum pooling layer, and a global average pooling layer. These layers are stacked in four residual units, namely L1, L2, L3, and L4, each consisting of 3, 6, 4, and 3 residual blocks, respectively. The core of ResNet50 is its residual blocks, which are illustrated in Fig.1.

B. ECA-ResNet

The Efficient Channel Attention (ECA) mechanism is a lightweight attention method that leverages the interdependencies among feature map channels to increase the representational power of the model. By assigning weights

to the importance of each channel in the feature map through Global Average Pooling (GAP), the ECA me mechanism allows the network to focus on crucial channel features, which leads to an improvement in the network's performance. Additionally, the computational complexity of the ECA mechanism is low, which enables it to enhance the network performance without adding computational over-head. In order to improve the ResNet network's ability to emphasize channel information, the ECA mechanism was introduced into the residual block by placing it between the first BN layer and the ReLu activation function, as depicted in Fig. 1.

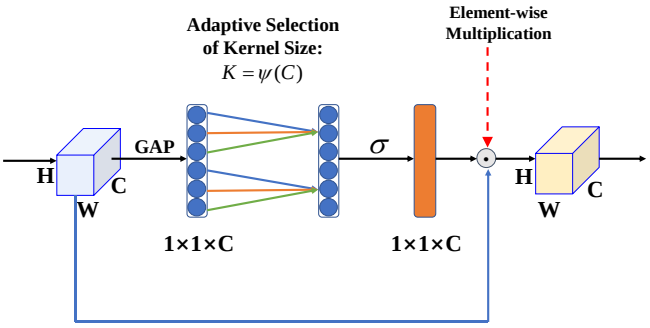

Fig. 2 Schematic diagram of the ECA module

The ECA mechanism is presented in Fig. 2. Assuming that the dimension of the input feature map is $X(H,W,C)$, where H and W are the height and width of the feature map, and C is the number of channels. The channel feature map can be obtained using GAP, and any P-th element can be expressed as follows:

$$Z_p = \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} X(i,j) \tag{1}$$

$X(i,j)$ denotes any one of its elements. In order to perform the weighting operation, a channel feature map Z must be learned first using one-dimensional convolution. This is done to calculate the weights of the channel attention. It is worth noting that ECA leverages cross-channel interaction, and the parameter K is responsible for determining the range of channels covered by the interaction. Moreover, the size of the convolutional kernel is also determined by K, which is adaptively selected based on the number of channels C.

$$K = \varphi(C) = \left| \frac{lb(C)}{2} + \frac{1}{2} \right|_{odd} \tag{2}$$

Here, $\left|[lb(C)+1]/2\right|_{odd}$ denotes the odd number closest to $[lb(C)+1]/2$. The feature map is passed through a convolution layer, and the resulting feature map is fed to the σ activation function, which outputs the attention weight A between 0 and 1. Finally, the output feature map F of the ECA module, after the channel weighting, can be expressed as:

$$F = A \odot X \tag{3}$$

Here, $\odot$ denotes the element-wise multiplication between the attention weight of each channel and the corresponding feature value of each channel. Consequently, the ECA module produces a new feature map that is weighted with significant channel information.

C. *Multidimensional Attention*

The MA module is shown in Fig. 3. For enhancing the classification performance of the ECA-ResNet50 model and efficiently extracting multidimensional characteristics, the majority of attention mechanisms have the disadvantage of capturing incomplete distinguishing features by focusing only on a single dimension of location or channel information. To address this constraint, this paper proposes a multi- dimensional attention mechanism that integrates location information with the channel information extracted by ECA, thereby improving the spatial location re-presentation of the model and focusing better on important features for classification on the feature map. Specific operations are as follows.

Step 1: The ECA introduced residual blocks were utilized to extract effective channel features, while the residual blocks of the first and last layers were fused to enhance the channel feature expression ability.

Step 2: The fused feature map is fed as input to the positional information feature grabbing module. Firstly, two convolution transformations are performed, and the second convolution layer uses an expanded convolution which can reduce the parameters of the convolution layer. This is done to avoid the influence of parameter redundancy caused by the channel attention of the upper layer on the classification effect.

Step 3: In order to make the information in the feature map fully related, it is necessary to conduct dimension splitting of positions. The input features are pooled to a height of 1 along the horizontal direction (X) and a width of 1 along the vertical direction (Y), resulting in the encoding of the two feature maps with embedded direction-specific information as two separate concern maps.

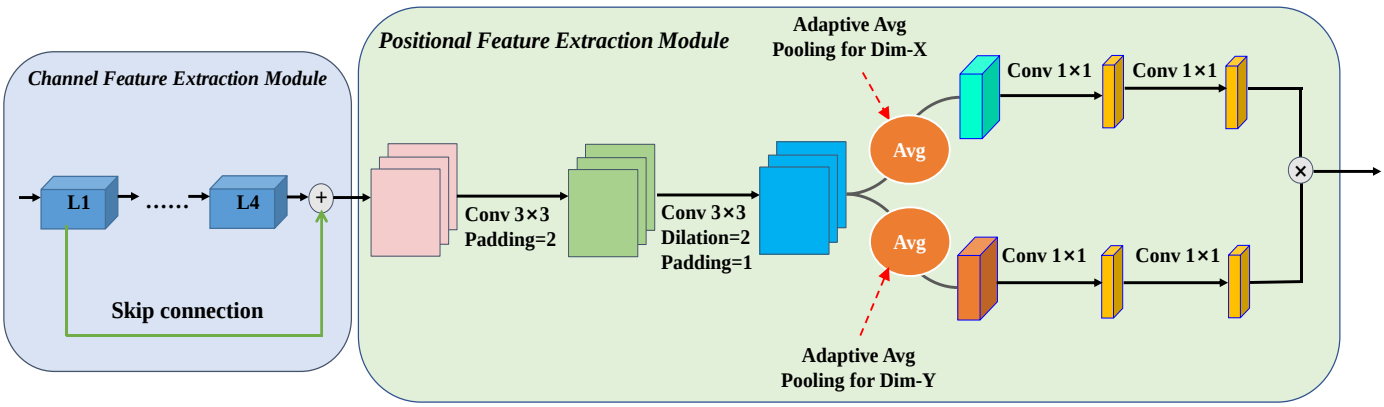

Fig. 3 Schematic diagram of the MA module

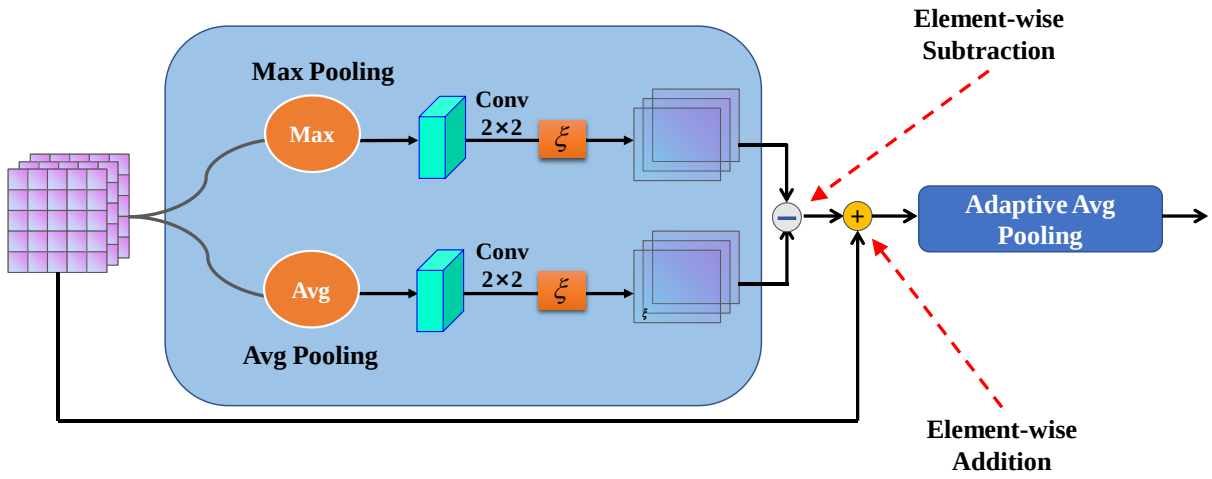

Fig. 4 Schematic diagram of the FE module

Instead of using traditional Global Average Pooling (GAP), we introduce Adaptive Average Pooling (AAP) as an alternative, which automatically calculates the parameters required for the corresponding pooling based on the size of the input and output tensor, thus improving computational efficiency and reducing in- formation loss.

$$Kernel\ size = input\_size - (output\_size) * (\frac{input\_size}{output\_size}) \tag{4}$$

Equation (4) shows the principle of AAP automatically solving for the size of the required pooling kernel, which $input\_size$ refers to the size of the tensor to be pooled and $output\_size$ refers to the output tensor size after pooling. After AAP, feature map information is obtained for two different locations.

$$Z_c^h(h) = \frac{1}{W} \sum_{0 \le i \le W} F_c(h, i) \tag{5}$$

$$Z_c^w(w) = \frac{1}{H} \sum_{0 \le j \le H} F_c(j, w) \tag{6}$$

Equations (5) and (6) are represented below, where $Z_c^h(h)$ represents the output of channel c at height h after pooling, and $Z_c^w(w)$ represents the output of the channel c at width w after pooling. Here, $F_c(h,i)$ represents any element with height h at the channel c of F, and $F_c(j,w)$ represents any element with width w at the channel c of F. In order to capture the long-distance correlation of feature maps along a spatial direction and obtain different positional information, f^h and f^w were obtained for each concern map through 1×1 convolution transformation $Conv_1$, and g^h and g^w were obtained through 1×1 convolution transformation $Conv_2$. The ReLU activation function is denoted by δ.

$$\begin{pmatrix} f^h = \delta[Conv_1(Z^h)] \\ f^w = \delta[Conv_1(Z^w)] \\ g^h = \delta[Conv_2(f^h)] \\ g^w = \delta[Conv_2(f^w)] \end{pmatrix} \tag{7}$$

Finally, Equation (8) is used to combine the features learned from the network's location information to generate a new feature map G that incorporates both channel and location information.

$$G = g^h \times g^w \tag{8}$$

D. Feature Enhancement

To enhance both the model's classification performance and generalization ability, a feature enhancement module was developed. This module is illustrated in Fig. 4.

To improve the feature map after attention has been focused on the key information in the previous layer, global average pooling and maximum pooling are applied to the features extracted from the attention module to obtain the corresponding G_{MAX} and G_{AVG} feature maps. Then, the features from each of the two directions are learned by the same convolutional transform $Conv_3$ to bring S_{MAX} and S_{AVG}. It should be noted that in order to overcome the poor robustness and weak noise immunity of the traditional ReLu activation function in deep convolution, the LeakyReLu activation function with a wider convergence range is introduced here instead of the traditional ReLu activation function, denoted by ξ.

$$\begin{pmatrix} S_{MAX} = \xi[Conv_3(G_{MAX})] \\ S_{AVG} = \xi[Conv_3(G_{AVG})] \\ S = S_{MAX} - S_{AVG} + G \end{pmatrix} \tag{9}$$

Finally, the output characteristics in both directions are subtracted to reduce interference caused by useless features. Feature fusion is then performed to increase the higher values on the feature map, thus enhancing the discriminative power of the model. The final feature map S of the feature enhancement module is obtained.

IV. Experiments

A. Datasets

Table I shows the details of the data set used in this article. The UC Merced Land Use (UCM) dataset represents a comprehensive collection of data designed to facilitate the classification of various scene features. This dataset was initially developed by researchers at the University of California (UC) Merced in the year 2004. The dataset comprises high-resolution aerial images that have been grouped into 21 categories. The images have a fixed image size of 256×256 pixels, and the entire dataset contains a total of 2100 images.

TABLE I
DATASET INFORMATION

Datasets	Number of images	Number of categories	Image size
UCM	2100	21	256×256
AID	10000	30	600×600
NWPU	31500	45	256×256

The Aerial Image Dataset (AID) is a collection of data used for analyzing aerial images, released by the Institute of Automation, Chinese Academy of Sciences. The dataset consists of 10,000 aerial images grouped into 30 categories, each image being of size 600×600 pixels, and representing various features such as buildings, farmland, roads, bridges, forests, water bodies, and more.

The NWPU-RESISC45 (NWPU) dataset, released in 2017, is a benchmark dataset for remote sensing image classification. Northwestern Polytechnical University in China collected the dataset, which comprises 31,500 images. Each image has dimensions of 256×256 pixels. The dataset covers 45 land use categories, such as farmland, airports, beaches, forests, and industrial land. This dataset has become an important tool for researchers and practitioners working in the field of remote sensing, providing a valuable resource for the development and testing of new image classification algorithms and techniques.

B. Metrics

The model's classification performance was evaluated using the overall accuracy (OA) and confusion matrix (CM).

TABLE II
SECOND-ORDER CONFUSION MATRIX

Confusion Matrix		True label	
		Positive	Negative
Predict label	Positive	TP	FP
	Negative	FN	TN

The confusion matrix is a useful tool that records the classification results and intuitively expresses the proportion of different categories misclassified into other categories. The confusion matrix is comprised of n×n matrices, the columns of the matrix are the true categories of the training samples, and the rows of the matrix are the predicted categories of the training samples. The row value in the matrix represents the classification accuracy rate between categories. True Positive (TP) indicates the positive category of the real value of the sample, and the predicted value of the model is also positive. True Negative (TN) indicates that the real value of the sample is negative, and the predicted value of the model is also negative. False Positive (FP) indicates that the real value of the sample is negative, but the predicted value of the model is positive. False Negative (FN) indicates that the true value of the sample is in the positive category, but the model predicts that it is in the negative category. Confusion matrix plays an important role in classification model evaluation Table II presents the second-order confusion matrix.

Accuracy is a critical performance metric for evaluating a model's predictive power. It refers to the proportion of correctly predicted samples to the total number of training samples in the test set. The overall accuracy is calculated using Equation (10).

$$OA = \frac{TP + TN}{TP + FN + TN + FP} \tag{10}$$

C. Experimental Parameter Settings

To minimize the influence of randomization on the final classification outcomes, we randomly partitioned the dataset and replicated the experiment five times, taking the average of the classification outcomes as the final result. To explore the generalizability of MA-FE on large-scale datasets, we created different training ratios on both the AID and NWPU datasets and conducted two sets of experiments on each dataset to ensure the accuracy of the findings. To maintain fairness, we employed the same training-test split ratio as other models on these three datasets. Specifically, for the AID dataset, the first set of experiments was divided into a 50% training set and a 50% test set, while the second set of experiments was divided into a 20% training set and an 80% test set. For the NWPU dataset, the first set of experiments employed a training-test split of 20% and 80%, while the second set of experiments was divided into a 10% training set and an 80% test set. The UCM model used an 80% and 20% training-test split. The batch size was set to 32, and we trained the model for 150 epochs to ensure convergence.

Regarding the preprocessing stage, we employed various techniques such as flipping, cropping, normalization, and shuffling of images to mitigate overfitting and ensure the experimental validity. All images were resized to 256 × 256 pixels. To facilitate the training network to reach the optimal solution, we set the learning rate to 0.0001. In the context of the present experimental setup, the utilized hardware configuration consists of an Intel i7-12700KF CPU, an NVIDIA GeForce RTX 3080(10G) graphics card, and a memory capacity of 32GB. For model construction, the PyTorch framework is employed.

D. Results and Analysis on UCM Dataset

We conducted a comparative study of 10 recently proposed methods on the UCM dataset, with a training ratio of 80%. The obtained results are presented in Table III.

TABLE III
COMPARISON OF THE OVERALL ACCURACY OF EACH METHOD ON THE UCM DATASET

Methods	Year	Accuracy(%)
GoogLeNet [3]	2017	94.31
EfficientNet [20]	2020	94.37
MobileNet [19]	2020	90.91
Coutourlet CNN [27]	2020	99.25
Skip-Connected CNN [8]	2020	98.04
DDRL-AM method [28]	2020	99.05
EfficientNetB3-Attn-2 [29]	2021	99.21
VGG_VD16 with SAFF [23]	2021	97.02
LCNN-BFF Method [30]	2021	99.29
ARCNet [31]	2021	99.12
Ours		**99.66**

It is evident from the table that our proposed MA-FE model outperforms all other methods, achieving an accuracy of 99.66%. Our model's accuracy is 0.37% higher than the LCNN-BFF method and 0.41% higher than the Coutourlet CNN method, which provides strong evidence of our model's robust classification ability.

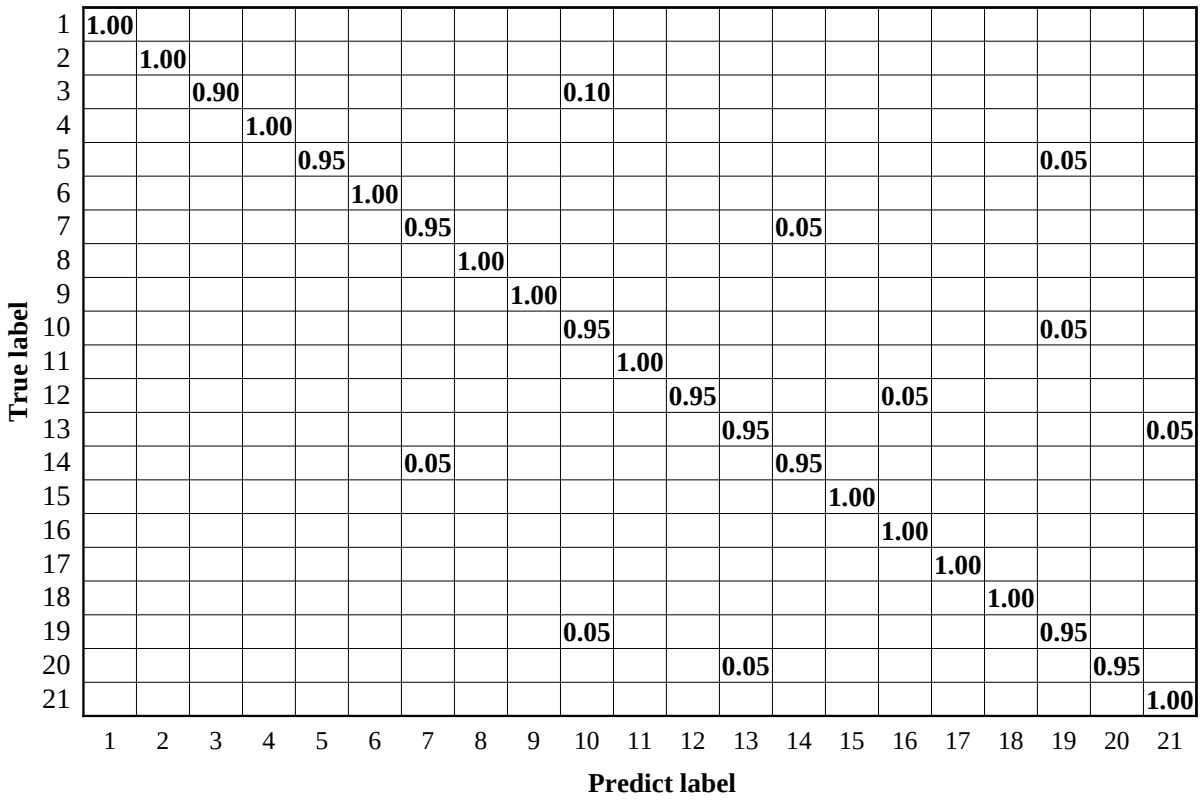

Fig. 5 Confusion matrix with a training ratio of 80% on the UCM dataset

Furthermore, Fig. 5 shows the confusion matrix for the MA-FE model on the UCM dataset, with a training ratio of 80%. The rows and columns of the matrix are numbered 1-21, representing the 21 categories of the UCM dataset (sorted alphabetically by scene category name). The figure shows that 12 of the 21 categories have an overall accuracy of 1, with only one category having an accuracy lower than 95%. Notably, the overall accuracy of categories 3 (baseball diamond) and 5 (buildings), which have intra-class diversity scenes, are 90% and 95%, respectively. Classes 10 (golf course) and 19 (sparse residential) share similar scene characteristics, resulting in an inter-class similarity, the overall accuracy of these classes is 95%, with a confusion ratio of 5%. These findings demonstrate that our model can not only differentiate between-category similarity but can also effectively distinguish within-category diversity, providing strong evidence of the efficacy of our proposed approach.

E. Results and Analysis on AID Dataset

The results, as presented in Table IV, indicate that our proposed model improves accuracy to some extent, achieving 95.68% and 93.51% accuracy for the training sets of 50% and 20%, respectively. Furthermore, our proposed method, MA-FE, outperformed VGG_VD16 with SAFF, which employs pretraining and an attention mechanism by 1.85% and 2.26%, respectively, thus demonstrating the superiority of our proposed approach.

TABLE IV
COMPARISON OF THE OVERALL ACCURACY OF EACH METHOD ON THE AID DATASET

Methods	Year	Accuracy(%)	
		50% Training Ratio	20% Training Ratio
ResNet50 [22]	2020	94.69	92.39
EfficientNet [20]	2020	88.35	86.56
MobileNet [19]	2020	90.91	88.53
Skip-Connected CNN [8]	2020	93.3	91.1
LCNN-BFF [30]	2020	94.62	91.66
VGG_VD16 with SAFF [23]	2021	93.83	90.25
ACNet [24]	2021	95.38	93.33
EfficientNetB3-Attn-2 [29]	2021	95.39	92.48
MARAA-BOVW [32]	2021	93.94	90.37
MRHNet-50 [33]	2022	95.06	91.14
Ours		**95.68**	**92.51**

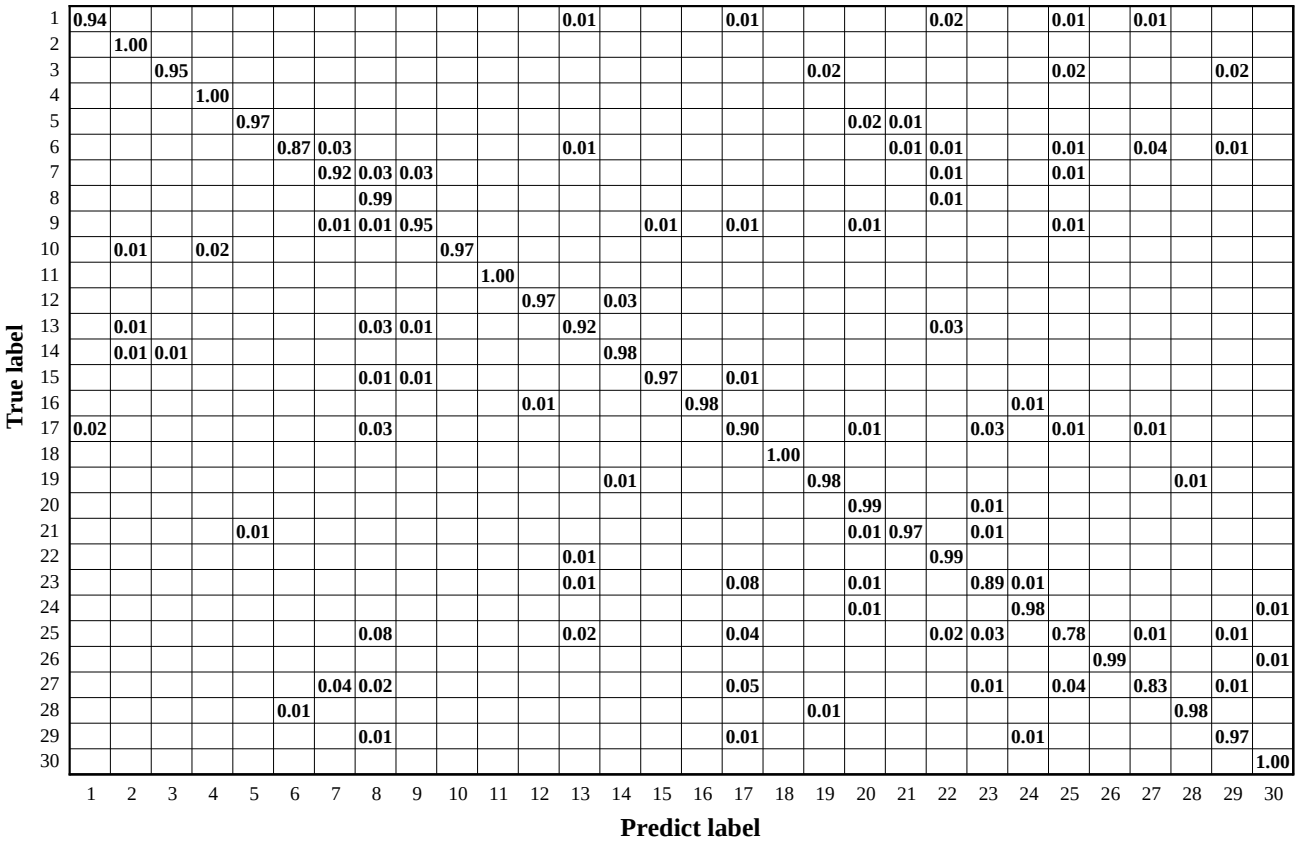

Fig. 6 Confusion matrix with a training ratio of 50% on the AID dataset

We present the confusion matrix for the AID dataset with a training ratio of 50% in Fig. 6, where the columns labeled 1-30 correspond to the 30 scene categories in the AID dataset, numbered in ascending alphabetical order. The figure shows that among the 30 categories, 17 have accuracy rates above 95%, while only one category has an accuracy rate below 80%.

Specifically, the scenes with intra-class diversity, such as 1 (Airport), 7 (Church), 8 (Commercial), and 22 (Railway Station), exhibit accuracy rates of 94%, 92%, 99%, and 99%, respectively. Furthermore, for scenes with inter-class similarity, such as 2 (Bare land) and 10 (Desert), 19 (Playground) and 28 (Stadium), our model achieves overall accuracy rates of 100%, 97%, 98%, and 98%, respectively, with a confusion ratio of only 1% between categories 19 and 28. These results demonstrate that our model can effectively differentiate both inter-class and intra-class scene diversity, highlighting the efficacy of our method.

F. Results and Analysis on NWPU Dataset

We partitioned the NWPU dataset into training ratios of 20% and 10%, respectively, and evaluated them against 12 advanced methods from recent years.

TABLE VI
COMPARISON OF THE OVERALL ACCURACY OF EACH METHOD ON THE NWPU DATASET

Methods	Year	Accuracy(%)	
		20% Training Ratio	10% Training Ratio
GoogLeNet [3]	2017	78.48	76.16
EfficientNet [20]	2020	81.83	78.57
MobileNet [19]	2020	83.26	80.32
ResNet50 [22]	2020	88.93	86.23
Skip-Connected CNN [8]	2020	87.3	84.33
Contourlet CNN [27]	2020	89.57	85.93
VGG VD16 with SAFF [23]	2021	87.86	84.38
MARAA-BOVW [32]	2021	89.76	84.82
ACNet [24]	2021	92.42	91.09
LCNN-BFF [30]	2021	91.73	86.53
MRHNet-101 [33]	2022	91.64	-
ACR-MLFF [28]	2022	92.45	90.01
Ours		**93.21**	**91.13**

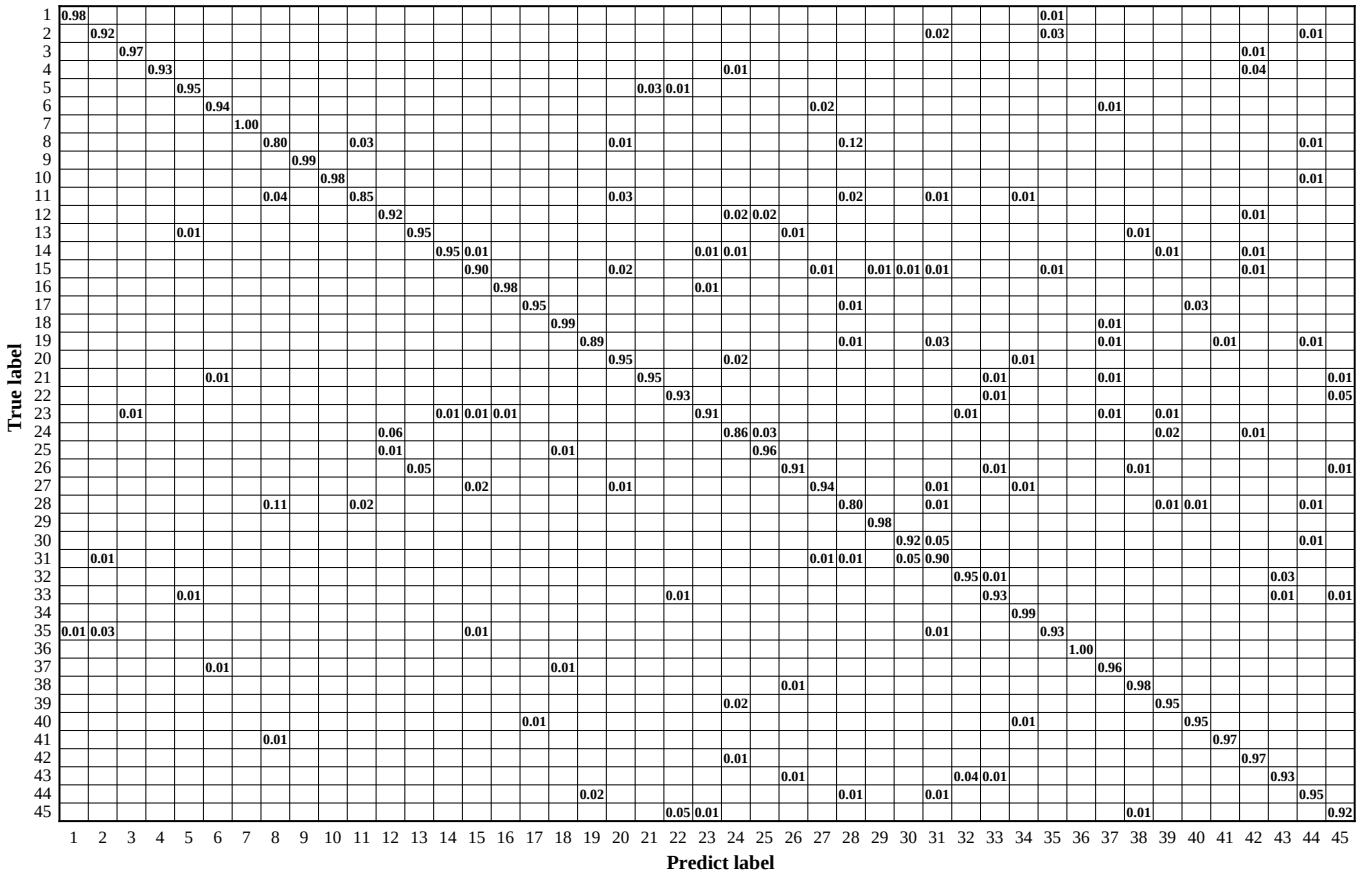

Fig. 7 Confusion matrix with a training ratio of 50% on the NWPU dataset

Table VI displays the results of our comparison. It is evident from the table that the accuracy of our proposed model improves to some extent when the training ratios are 20% and 10%, and the classification accuracy is 93.21% and 91.13%, respectively. This is 5.35% and 6.75% higher than that of VGG_VD16 with SAFF, which highlights the superior learning and generalization abilities of MA-FE on large datasets.

Fig. 7 presents the confusion matrix of the NWPU dataset with a training ratio of 20%. The labels 1-45 represent the 45 categories in the dataset, numbered in ascending alphabetical order according to the scenario category name. As can be observed from the figure, the classification accuracy of all categories is above 80%, and the overall accuracy of 40 out of 45 categories is above 90%. For the scenarios with inter-class similarity, the overall accuracy of 2 (airport) and 31 (industrial area) are 92% and 95%, respectively, with a confusion ratio of 2%. The overall accuracy of 12 (dense residential) and 24 (medium residential) are 92% and 86%, with a 2% and 6% confusion ratio, respectively. The scenarios 15 (freeway) and 35 (runway) have overall accuracy of 90% and 93%, respectively, with a confusion ratio of 1%. Additionally, categories 8 (church) and 31 (railway station) with diverse scenes within the class have an overall accuracy of 80% and 90%, respectively. These results demonstrate that our proposed model can effectively differentiate the scenes of inter-classification similarity and intra-class diversity on a large dataset, at the same time, it is also proved that the MA-FE model has good generalization ability on large data sets such as NWPU.

G. Ablation Experiment

In order to illustrate the efficacy and importance of both the MA module and the FE module, we performed experiments on three datasets, with the results showcased in Table VII.

TABLE VII
EFFECTIVENESS OF DIFFERENT MODULES ON 3 DATASETS

Datasets	Training Ratio	MA module	FE module	Accuracy(%)
UCM	80%	×	×	97.23
		√	×	99.51
		×	√	99.39
		√	√	**99.66**
AID	50%	×	×	91.88
		√	×	95.22
		×	√	93.24
		√	√	**95.98**
NWPU	20%	×	×	89.65
		√	×	93.16
		×	√	90.78
		√	√	**93.21**

The overall accuracy of the four fusion methods on the UCM dataset is 97.23%, 99.51%, 99.39%, and 99.66%, respectively. For the AID dataset, the overall accuracy is 91.88%, 95.22%, 93.24%, and 95.98%, respectively. Finally, the overall accuracy on the NWPU dataset is 89.65%, 93.16%, 90.78%, and 93.21%, respectively. These results demonstrate that the addition of the MA module allows the model to combine channel and location information,

resulting in the simultaneous capture of two-dimensional discriminative information and a significant improvement in overall accuracy. Moreover, the introduction of the FE module enhances the extracted attention feature map, reducing the value of non-discriminative features and further improving classification accuracy. By using both the FE and MA modules simultaneously, our proposed method achieves the best classification performance, further demonstrating its superiority.

V. Conclusion

In the present study, we propose a novel MA-FE model for remote sensing image scene classification. The proposed MA-FE model employs a pre-trained ResNet50 structure that is embedded with an ECA module, and combines the MA and FE modules to further enhance the classification performance of the model. Notably, the MA module, addresses the limitations of many attention models by capturing the key feature information of the two dimensions of channel and location, thereby enabling comprehensive information focus. Furthermore, the FE module carries out additional feature enhancement while suppressing other irrelevant information, thus improving the model's dis-crimination ability and classification accuracy. To evaluate the effectiveness of the proposed method, we conducted detailed experiments on three datasets, and the experimental results demonstrate the superiority of the proposed approach. From now on, our future work will focus on developing methods to enhance the generalization ability of large datasets, with the goal of further improving the accuracy of remote sensing image scene classification.

References

[1] G.-S. Xia, J. Hu, F. Hu, B. Shi, X. Bai, Y. Zhong, et al., "Aid: A Benchmark Data Set for Performance Evaluation of Aerial Scene Classification,"*IEEE Transactions on Geoscience and Remote Sensing*, vol.55, no.7, pp. 3965-3981, 2017.

[2] G. Cheng, X. Xie, J. Han, L. Guo, and G.-S. Xia, "Remote Sensing Image Scene Classification Meets Deep Learning: Challenges, Methods, Benchmarks, and Opportunities,"*IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol.13, pp. 3735-3756, 2020.

[3] C. Gong, H. Junwei, and L. Xiaoqiang, "Remote Sensing Image Scene Classification: Benchmark and State of the Art,"*Proceedings of the IEEE*, vol.105, no.10, pp. 1865-1883, 2017.

[4] F. Hu, G.-S. Xia, J. Hu, and L. Zhang, "Transferring Deep Convolutional Neural Networks for the Scene Classification of High-Resolution Remote Sensing Imagery,"*Remote Sensing*, vol.7, no.11, pp. 14680-14707, 2015.

[5] X. Bai, C. Liu, P. Ren, J. Zhou, H. Zhao, and Y. Su, "Object Classification Via Feature Fusion Based Marginalized Kernels,"*IEEE Geoscience and Remote Sensing Letters*, vol.12, no.1, pp. 8-12, 2015.

[6] X. Bai, H. Zhang, and J. Zhou, "Vhr Object Detection Based on Structural Feature Extraction and Query Expansion,"*IEEE Transactions on Geoscience and Remote Sensing*, vol.52, no.10, pp. 6508-6520, 2014.

[7] J. Xie, N. He, L. Fang, and A. Plaza, "Scale-Free Convolutional Neural Network for Remote Sensing Scene Classification,"*IEEE Transactions on Geoscience and Remote Sensing*, vol.57, no.9, pp. 6916-6928, 2019.

[8] N. He, L. Fang, S. Li, J. Plaza, and A. Plaza, "Skip-Connected Covariance Network for Remote Sensing Scene Classification,"*IEEE Transactions on Neural Networks and Learning Systems*, vol.31, no.5, pp. 1461-1474, 2020.

[9] Y. Gao, J. Shi, J. Li, and R. Wang, "Remote Sensing Scene Classification Based on High-Order Graph Convolutional Network,"*European Journal of Remote Sensing*, vol.54, no.sup1, pp. 141-155, 2021.

[10] D. G. Lowe, "Distinctive Image Features from Scale-Invariant Keypoints,"*International Journal of Computer Vision*, vol.60, pp. 91-110, 2004.

[11] T. Ojala, M. Pietikainen, and T. Maenpaa, "Multiresolution Gray-Scale and Rotation Invariant Texture Classification with Local Binary Patterns,"*IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol.24, no.7, pp. 971-987, 2002.

[12] T. H. Rassem, B. E. Khoo, N. M. Makbol, and A. A. Alsewari, "Multi-Scale Colour Completed Local Binary Patterns for Scene and Event Sport Image Categorisation,"*IAENG International Journal of Computer Science*, vol.44, no.2, pp. 197-211, 2017.

[13] N. Dalal, and B. Triggs, "Histograms of Oriented Gradients for Human Detection," *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, pp. 886-893, 2005.

[14] Y. Yang, and S. Newsam, "Bag-of-Visual-Words and Spatial Extensions for Land-Use Classification," *Proceedings of the 18th SIGSPATIAL International Conference on Advances in Geographic Information Systems*, pp. 270-279, 2010.

[15] S. Lazebnik, C. Schmid, and J. Ponce, "Beyond Bags of Features: Spatial Pyramid Matching for Recognizing Natural Scene Categories," *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, pp. 2169-2178, 2006.

[16] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet Classification with Deep Convolutional Neural Networks," *Communications of the ACM*, vol.60, no.6, pp. 84-90, 2017.

[17] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, et al., "Going Deeper with Convolutions," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1-9, 2015.

[18] A. Alzu'bi, A. Amira, and N. Ramzan, "Learning Transfer Using Deep Convolutional Features for Remote Sensing Image Retrieval,"*IAENG International Journal of Computer Science*, vol.46, no.4, pp. 637-644, 2019.

[19] H. Pan, Z. Pang, Y. Wang, Y. Wang, and L. Chen, "A New Image Recognition and Classification Method Combining Transfer Learning Algorithm and Mobilenet Model for Welding Defects,"*IEEE Access*, vol.8, pp. 119951-119960, 2020.

[20] A. M. Pour, H. Seyedarabi, S. H. A. Jahromi, and A. Javadzadeh, "Automatic Detection and Monitoring of Diabetic Retinopathy Using Efficient Convolutional Neural Networks and Contrast Limited Adaptive Histogram Equalization,"*IEEE Access*, vol.8, pp. 136668-136673, 2020.

[21] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770-778, 2016.

[22] Q. Hou, D. Zhou, and J. Feng, "Coordinate Attention for Efficient Mobile Network Design," *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13713-13722, 2021.

[23] R. Cao, L. Fang, T. Lu, and N. He, "Self-Attention-Based Deep Feature Fusion for Remote Sensing Scene Classification,"*IEEE Geoscience and Remote Sensing Letters*, vol.18, no.1, pp. 43-47, 2020.

[24] X. Tang, Q. Ma, X. Zhang, F. Liu, J. Ma, and L. Jiao, "Attention Consistent Network for Remote Sensing Scene Classification,"*IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol.14, pp. 2030-2045, 2021.

[25] Q. Wang, B. Wu, P. Zhu, P. Li, W. Zuo, and Q. Hu, "Eca-Net: Efficient Channel Attention for Deep Convolutional Neural Networks," *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11534-11542, 2020.

[26] W. Qin, W. Zhao, and M. Li, "Multi-Level Feature Representation and Multi-Layered Fusion Contrast for Few-Shot Classification," *IAENG International Journal of Computer Science*, vol.49, no.2, pp. 318-324, 2022.

[27] M. Liu, L. Jiao, X. Liu, L. Li, F. Liu, and S. Yang, "C-Cnn: Contourlet Convolutional Neural Networks,"*IEEE Transactions on Neural Networks and Learning Systems*, vol.32, no.6, pp. 2636-2649, 2020.

[28] J. Li, D. Lin, Y. Wang, G. Xu, Y. Zhang, C. Ding, et al., "Deep Discriminative Representation Learning with Attention Map for Scene Classification,"*Remote Sensing*, vol.12, no.9, 2020.

[29] H. Alhichri, A. S. Alswayed, Y. Bazi, N. Ammour, and N. A. Alajlan, "Classification of Remote Sensing Images Using Efficientnet-B3 Cnn Model with Attention,"*IEEE Access*, vol.9, pp. 14078-14094, 2021.

[30] C. Shi, T. Wang, and L. Wang, "Branch Feature Fusion Convolution Network for Remote Sensing Scene Classification,"*IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol.13, pp. 5194-5210, 2020.

[31] Q. Wang, S. Liu, J. Chanussot, and X. Li, "Scene Classification with Recurrent Attention of Vhr Remote Sensing Images,"*IEEE*

Transactions on Geoscience and Remote Sensing, vol.57, no.2, pp. 1155-1167, 2018.

[32] G. Lv, L. Dong, W. Zhang, and W. Xu, "Multi-Scale Attentive Region Adaptive Aggregation Learning for Remote Sensing Scene Classification,"*International Journal of Remote Sensing*, vol.42, no.20, pp. 7742-7776, 2021.

[33] C. Li, Y. Zhuang, W. Liu, S. Dong, H. Du, H. Chen, et al., "Effective Multiscale Residual Network with High-Order Feature Representation for Optical Remote Sensing Scene Classification,"*IEEE Geoscience and Remote Sensing Letters*, vol.19, pp. 1-5, 2021.