Large Kernel Separable Mixed ConvNet for Remote Sensing Scene Classification

Keqian Zhang, Tengfei Cui, Wei Wu, Xueke Zheng, and Gang Cheng

***Abstract*—Among tasks related to intelligent interpretation of remote sensing data, scene classification mainly focuses on the holistic information of the entire scene. Compared with pixel-level or object-based tasks, it involves a richer semantic context, making it more challenging. With the rapid advancement of deep learning, convolutional neural networks (CNNs) have found widespread applications across various domains, and some work has introduced them into scene classification tasks. However, traditional convolution operations involve sliding small convolutional kernels across an image, primarily focusing on local details within a small receptive field. To achieve better modeling of the entire image, the smaller receptive field limits the ability of convolution operation to capture features over a broader range. To this end, we introduce large kernel CNNs into the scene classification task to expand the receptive field of the mode, which allows us to capture comprehensive nonlocal information while still acquiring rich local details. However, in addition to encoding spatial association, the effective information within the feature maps is also strongly channel related. Therefore, to fully model this channel dependency, a novel channel separation and mixing module has been designed to realize feature correlation in the channel dimension. The combination of them forms a large kernel separable mixed ConvNet, enabling the model to capture effective dependencies of feature maps in both spatial and channel dimensions, thus achieving enhanced feature expression. Extensive experiments conducted on three datasets have also validated the effectiveness of the proposed method.**

***Index Terms*—Channel separation and mixing, large kernel convolution, remote sensing, scene classification.**

I. Introduction

As A basic and challenging task in remote sensing community, remote sensing scene classification receives growing attention, which aims to understand the semantic content in scene images and assign corresponding labels, and has been widely applied to geological survey [1], urban planning [2], [3], [4], [5], disaster monitoring [6], and other fields [7], [8], [9]. Recently, with the development of remote sensing data acquisition technologies [10], [11], [12], more and more available data have emerged, providing sufficient data support for research in this direction, so a lot of research work has been proposed to better understand the corresponding scene information. These methods can be roughly divided into three main categories according to different means of feature extraction: handcrafted-feature-based methods [13], unsupervised-feature-learning-based methods [14], [15], and deep-feature-learning-based methods [16], [17], [18].

Traditional scene classification predominantly relies on handcrafted features. These methodologies center around leveraging substantial engineering expertise and domain-specific knowledge to design various human-engineered features, such as color, shape, spatial relations, and spectral characteristics. Among these, color histograms have been employed [19] for image classification utilizing the enhanced color structure code, augmented by grid and vector analyses. Cheng et al. [20] devised a pragmatic rotation-invariant framework grounded in an ensemble of partial detectors. This framework facilitates the detection of objects or recurring spatial patterns across a predefined range of orientations. Aptoula [21] proposed the utilization of global morphological texture descriptors for tasks within remote sensing image processing. This initiative delves into the viability of multiscale texture descriptors. In [22], a more efficient sparse model was trained through a coarse-to-fine framework, using an unsupervised hidden layer autoencoder to detect redundancies. Then, a supervised single hidden layer neural network was used to train fine sparse and activation vectors, improving classification performance on the UC Merced land use (UCM) dataset.

Although the above methods based on handcrafted features have made progress on the scene classification problem; however, human participation in the feature design process significantly affects the feature expressiveness and effectiveness for classification tasks, especially when the scene becomes more complex, the representational power of handcrafted features becomes limited or even deficient [23], [24]. To remedy this constraint, the pursuit of automatic feature learning from images emerges as a more suitable approach. Recently, unsupervised-feature-learning-based methods have become an attractive alternative to handcrafted-feature-based methods, yielding noteworthy advancements in remote sensing scene classification endeavors. By acquiring features from images

Manuscript received 24 October 2023; revised 26 November 2023 and 23 December 2023; accepted 10 January 2024. Date of publication 15 January 2024; date of current version 12 February 2024. This work was supported in part by the National Key R&D Program Projects under Grant 2016YFC08033103, in part by the National Natural Science Foundation of China Grant Program under Grant 41001226, in part by the China Postdoctoral Science Foundation under Grant 2015M582831, and in part by the Special Funding for Basic Scientific Research Business Expenses of Henan Universities under Grant NSFRF180329. *(Corresponding author: Gang Cheng.)*

Keqian Zhang, Xueke Zheng, and Gang Cheng are with the College of Surveying and Land Information Engineering, Henan Polytechnic University, Jiaozuo 454000, China (e-mail: keqianzhang@home.hpu.edu.cn; 311705000216@home.hpu.edu.cn; chenggang@hpu.edu.cn).

Tengfei Cui is with the Metropolitan College, Boston University, Boston, MA 02215 USA (e-mail: tecui@bu.edu).

Wei Wu is with the College of Geomatics and Geoinformation, Guilin University of Technology, Guilin 541004, China (e-mail: wu_wei@glut.edu.cn).

Digital Object Identifier 10.1109/JSTARS.2024.3353796

via learning, as opposed to manual design, the potential to obtain more discriminative features arises, which aligns more fittingly with the current problem. Classical unsupervised-feature-learning-based methods include and are not limited to principal component analysis [25], [26], [27], [28], [29], [30], [31], k-means clustering [32], [33], [34], [35], [36], [37], sparse coding [38], [39], [40], [41], [42], and autoencoders [43], [44]. Chaib et al. [45] employed scale-invariant feature transformation and robust feature operators to extract local features from satellite imagery. Subsequently, a sparse principal component analysis was harnessed to assimilate category-specific information and facilitate comprehensive classification. The widely recognized bag-of-visual-words paradigm found application in remote sensing scene classification assignments, driven by its straightforwardness and efficacy [46], [47], [48]. Cheriyadat [14] employed sparse coding to acquire a suite of bias functions from images. Subsequently, these bias functions were leveraged to encode low-level features, generating novel sparse representations. Nonetheless, the lack of label information limits the further development of the methods, which hinders to the improvement of classification performance.

At present, most advanced methods usually rely on deep learning to obtain good feature representations. Compared with the previous method, deep-feature-learning-based methods can not only automatically extract the abstract features contained in the image, but also can better modify the model through the deep structure of the neural network and label information, so as to obtain a more powerful feature representation [49], [50], [51], [52]. Yu and Liu [53] leveraged a pair of pretrained convolutional neural networks (CNNs) to acquire profound features from original and processed images. Subsequently, an extreme learning machine was employed to classify the amalgamated features. In addition, an improved pretraining AlexNet was proposed [54], which combined the scale pooling–spatial pyramid pooling [55] and side supervision to solve the problem of overfitting and effectively use multiscale information to represent the semantic features of the scene, achieving good performance. Anwer et al. [56] substantiated that a CNN trained on mapped coded images incorporating explicit LBP-based texture information offers supplementary insights to the deep model. They further investigated the influence of diverse network fusion architectures on classification outcomes. Although the methods mentioned above have made significant progress with the robust feature extraction and information representation capabilities of CNNs, some challenges remain to be solved. First, most of these approaches employ small convolution kernels for feature extraction. This practice hampers the comprehensive modeling of spatial correlations across entire scene images during the feature extraction process. Currently, some work has tried to use large kernel convolution to improve the ability of CNN to capture long-range dependencies, and some progress has been made. However, larger convolution kernels introduce more parameters and increase the difficulty of model optimization [57], [58]. Therefore, how to construct a remote sensing image scene classification framework based on large kernel convolution is worthy of study. Second, although the large kernel convolution can encode spatial association, its independence in the channel dimension makes it insufficient to model channel correlation, so it is difficult to ensure that the extracted features are sufficient enough for the classification task.

To solve the above problems, this article introduces a novel large kernel separable mixed ConvNet (LSMNet) to capture spatial and channel relationships within scene images. The primary contributions of this study are delineated as follows.

1) To make up for the limited receptive field of traditional small convolution kernels and the difficulty of training large convolution kernels, a deepwise large kernel convolution is introduced. On one hand, it captures nonlocal information under a larger receptive field, on the other hand, it reduces the parameters of large kernels. When combined with conventional small kernel convolutions, it enables the acquisition of global information and local detail information simultaneously in spatial dimension.
2) The designed channel separation and mixing module is used to capture interchannel dependency relationships, thereby addressing the limitations of the aforementioned convolutional operations in the spectral dimension. The combination of them provides a more comprehensive modeling of both the spatial and spectral features of the data.
3) By combining the above modules, we propose an LSMNet, which better captures holistic information in diverse scenes, reduces the optimization challenges associated with applying large kernel convolutions in scene classification models, and models the feature correlations between channels. Excellent experimental results on multiple datasets also validate the effectiveness of the proposed method.

II. Proposed Method

A. Overall Structure of Proposed Method

The schematic representation of the proposed LSMNet framework is depicted in Fig. 1. Given a remote sensing scene image denoted as $I \in \mathbb{R}^{C\times H\times W}$, where H, W, and C, respectively, signify the image's height, width, and channel count. In the case of a color image, C is equal to 3. The initial step of LSMNet involves a preliminary feature mapping conducted by a feature extraction module, yielding the resultant extracted features denoted as $Z_0 \in \mathbb{R}^{D_0\times H_0\times W_0}$. Subsequently, a feature fusing block (FFB), comprising two large kernel separation mixing layers and a Fused-MBConv [59], is devised. This block operates on the input Z_0 to yield a more intricate feature representation denoted as $Z \in \mathbb{R}^{D_1\times H_1\times W_1}$. Following the execution of several FFBs, an enriched feature encoding is attained, with the detailed configuration depicted in Fig. 1(a). The large kernel separation mixing layer incorporates large kernel deepwise convolution and a channel separation and mixing module to undertake feature modeling across spatial and channel dimensions, while Fused-MBConv is used to selectively enhance the local details of features. Finally, the enriched feature expression is sent to the classifier to obtain the classification result.

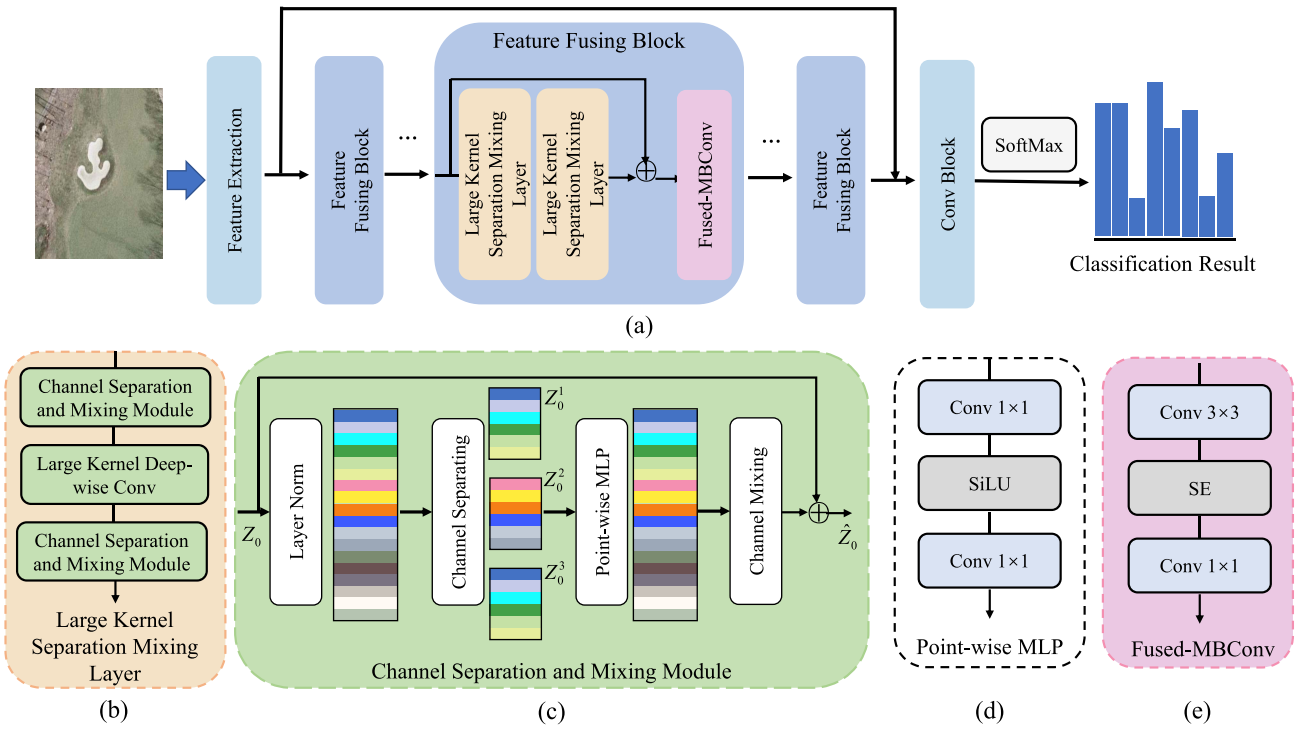


Fig. 1. (a) Schematic illustration of the proposed LSMNet. (b) Composition of large kernel separation mixing layer. (c) Schematic illustration of channel separation and mixing module. (d) Structural composition of pointwise MLP. (e) Structural composition of Fused-MBConv.

B. Channel Separation and Mixing Module

A channel separation and mixing module is employed to capture feature dependencies within the channel dimension more effectively. This module encompasses channel separating, pointwise multilayer perceptron (pointwise MLP), and channel mixing, as depicted in Fig. 1(c). More specifically, the input feature Z_0 undergoes an initial partitioning into three distinct feature subsets: Z_0^1, Z_0^2, and Z_0^3. Subsequently, a pointwise MLP executes channel modeling across these three separate feature groups, facilitating the amalgamation of features within the channel dimension. Ultimately, a channel mixing operation facilitates information interchange across the concatenated features. This process can be formally expressed as follows:

$$\begin{aligned} Z_0^1, Z_0^2, Z_0^3 &= \text{Split}(\text{LayerNorm}(Z_0)) \\ \hat{Z}_0^1, \hat{Z}_0^2, \hat{Z}_0^3 &= W_1(\sigma(W_0(Z_0^1, Z_0^2, Z_0^3))) \\ \hat{Z}_0 &= \text{Shuffle}([\hat{Z}_0^1, \hat{Z}_0^2, \hat{Z}_0^3]) + Z_0 \end{aligned} \tag{1}$$

where σ is SiLU nonlinearity function [60], W_0 and W_1 are pointwise convolutions, and Split$(\cdot)$ and Shuffle$(\cdot)$ represent the splitting and shuffling operations on the channel dimension. The shuffling fuser is repeated and arranged before and after the large kernel deepwise convolution to learn visual representations.

C. Large Kernel Deepwise Convolution

Traditional CNNs extract features by stacking multiple small convolutional kernels, with the sizes of these kernels predominantly focused around 3×3, 5×5, or 7×7 dimensions [61], [62]. Nevertheless, several investigations [63] have proved that

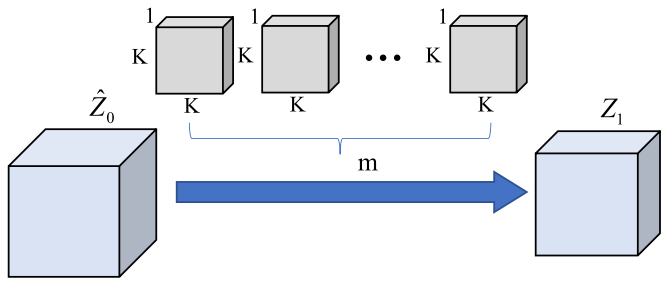


Fig. 2. Framework of large kernel deepwise convolution.

employing convolution with large kernels is more effective in capturing long-range dependencies than stacking small kernels, so as to better model the spatial context relationship in the images. However, incorporating large kernel convolution inevitably intensifies the challenge of model training. To solve the above problem, a large kernel deepwise convolution is integrated into the large kernel separation mixing layer (as illustrated in Fig. 2). This operation not only expands the receptive field effectively, enabling the encoding of comprehensive and accurate spatial structural information but also significantly reduces the number of parameter, thereby mitigating challenges associated with training large kernels. More specifically, the feature $\hat{Z}_0$, derived from the channel separation and mixing module, is passed through a convolutional layer comprised of m convolution kernels, whose size is $K \times K \times 1$, resulting in the evolved feature Z_1. After the processing of channel separation and mixing module and large kernel deepwise convolution, the

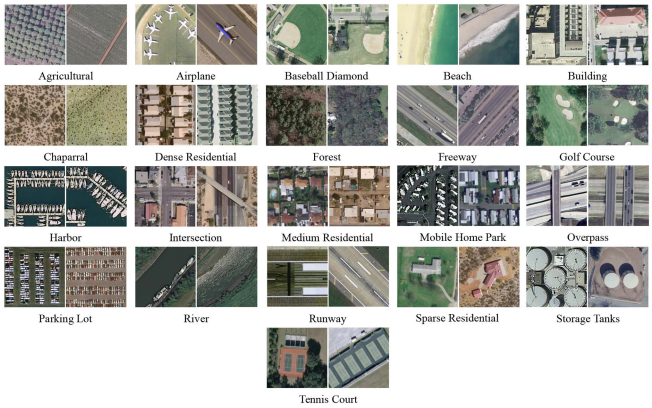


Fig. 3. Sample images of the UCM dataset. Two images of each class are exhibited.

Fig. 4. Sample images of the AID. Two images of each class are exhibited.

original feature maps are feature associated in both channel and spatial dimensions. This solves the previously mentioned problems of small receptive fields and strong channel independence, providing better discriminative features for the subsequent classification.

Due to the localized correlations present in remote sensing scene images, the FFB supported by large-kernel convolutions cannot fully utilize spatial local features. Furthermore, it requires enhanced capability to spatially model the features, aiming to attain improved performance in scene classification. To this end, a convolutional block that captures local spatial information is integrated into the model to enhance local connectivity. Specifically, after two large kernel separation mixing layers, a Fused-MBConv [its specific structure is shown in Fig. 1(e)] is introduced, which consists of a convolution with a size of 3×3, a squeeze-and-excitation (SE) layer [64], and another convolution with a size of 1×1. The first convolution is employed to capture local detailed features, then utilizing an SE layer for spatial attention, enabling the model to focus on significant regions. Subsequently, a 1×1 convolution is applied to restore the channel dimensions of the feature map.

III. Experiments

A. Datasets Descriptions and Evaluation Metrics

1) Datasets Descriptions: To test the effectiveness of the proposed method, extensive experiments are conducted on three public datasets, including the UCM dataset, the aerial image dataset (AID), and the NWPU-RESISC45 (NWPU) dataset.

1) As one of the widely recognized datasets in scene classification tasks, the UCM dataset [65] consists of 2100 images with a spatial resolution of 0.3 m, including 21 scene categories, including golf course, overpass, river, runway, etc., which is shown in Fig. 3. This dataset is derived from aerial orthophotos downloaded from the United States Geological Survey, and its size is 256×256 pixels.

2) Fig. 4 shows that the AID [66] from Wuhan University contains 10 000 images covering 30 scene categories, such as airports, bare ground, beaches, and centers. Compared with

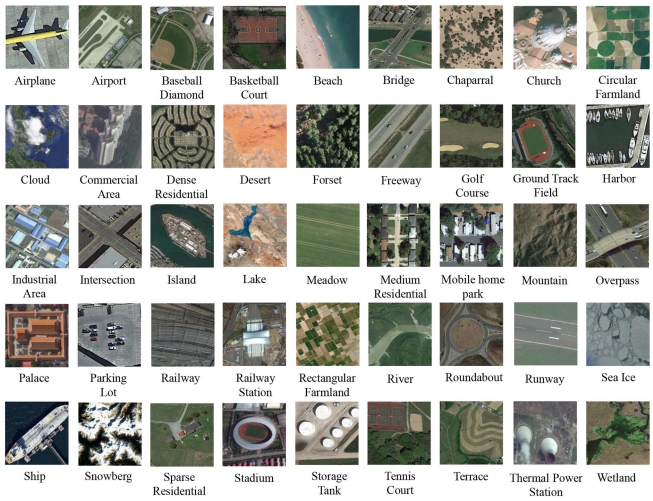


Fig. 5. Sample images of the NWPU dataset. Two images of each class are exhibited.

the UCM dataset, it faces more significant challenges in the scene classification task because the images of the AID come from multiple remote sensing sensors, and the spatial resolution ranges from 0.5 to 8 m, resulting in substantial intraclass diversity. Moreover, the number of samples in each category varies from 220 to 420, a significant difference in sample quantities among classes. Consequently, the AID presents a higher level of classification difficulty.

3) The NWPU dataset [47], created by Northwestern Polytechnical University and demonstrated in Fig. 5, is a large-scale scene classification dataset known for its rich scene diversity and instance variability. This dataset encompasses 45 categories: airplane, beach, bridge, church, and more. There are 700 images in each class with the size of 256×256 pixels. Compared with other datasets, first, it showcases a vast range of spatial resolutions, spanning from 0.2 to 30 m. Moreover, its scene diversity and the total number of images significantly surpass other datasets. Finally, it demonstrates pronounced variations

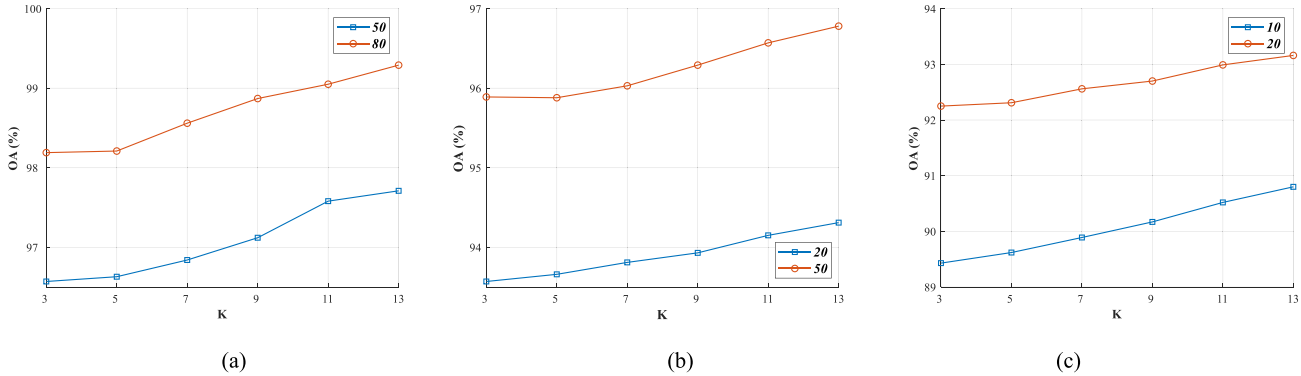


Fig. 6. Influence of the size of K on the classification accuracy. (a) AID. (b) UCM datasets. (c) NWPU dataset.

in translation, viewpoint, object pose, lighting, background, and occlusion. Therefore, this dataset is recognized as a very challenging dataset.

2) Evaluation Metrics: To establish the superiority of the proposed method in terms of classification performance compared with other state-of-the-art methods, we incorporate both quantitative and qualitative analyses during the experimental phase.

1) In terms of quantitative assessment, we utilize a widely recognized classification metric—overall accuracy (OA)—to demonstrate the classification results of each method. OA is the ratio of correctly classified samples to the total number of samples within the dataset.
2) In qualitative analysis, we visualize the confusion matrix, enhancing the intuitive understanding of the proposed method's classification results across diverse datasets. Specifically, the columns of the confusion matrix delineate the model's predictive outcomes, while the rows illustrate the true sample distribution in the dataset. Consequently, the diagonal cells of the confusion matrix represent the percentage of correctly identified samples.

B. Analysis of Experimental Parameters and Computational Consumption

As discussed in Section II-C, the size of the large kernel convolution is a critical parameter that governs the receptive field and the efficiency of feature extraction within the corresponding convolution operation. To scrutinize the effect of varying K on classification performance, we set K to [3, 5, 7, 9, 11] and perform pertinent experiments across three datasets. The graphs in Fig. 6 illustrate that as K grows from smaller to larger values, classification accuracy correspondingly improves across all three datasets. This trend arises because a larger K equips the corresponding large kernel deepwise convolution with an expanded receptive field, facilitating the capture of extensive long-range dependencies throughout the scene. The subsequent Fused-MBConv compensates for the lack of locally detailed features. As a result, for subsequent experimentation, we opt to set the K size to 13.

TABLE I
CLASSIFICATION PERFORMANCE OF DIFFERENT METHODS ON THE UCM DATASET

Method	80% Training Ratio	50% Training Ratio
CaffeNet [66]	95.02±0.81	93.98±0.67
GoogLeNet [66]	94.31±0.89	92.70±0.60
VGG-16 [66]	95.21±1.20	94.14±0.69
salM^{3}LBP-CLM [67]	95.75±0.80	94.21±0.75
TEX-Net-LF [56]	96.62±0.49	95.89±0.37
Fusion by addition [16]	97.42±1.79	/
VGG-16-CapsNet [68]	98.81±0.22	95.33±0.18
Inception-v3-CapsNet [68]	99.02±0.24	97.58±0.16
GBNet+global feature [69]	98.57±0.48	97.05±0.19
MSA-Network [70]	98.96±0.21	97.80±0.33
PVT-V2-B0 [71]	98.86±0.38	**97.94±0.44**
LSMNet (ours)	**99.29±0.15**	97.71±0.14

Bold values represents the best values of the experimental results.

To demonstrate the running time and the memory usage of the model, relevant statistics are performed on the UCM dataset. The training time is 46.21 s and the testing time is 7.08 s. The floating-point operations and parameters of the model are 9143.87 and 2.15 M, respectively (the unit M represents 1×10^6). From the obtained results, the model runs efficiently and requires a moderate amount of memory, which is sufficient for subsequent applications.

C. Comparison With State-of-The-Art Methods

In this section, to validate the classification effectiveness of LSMNet, some state-of-the-art methods are selected for comparison. To this end, we introduce two widely used classification metrics: OA and confusion matrix. The analysis of qualitative indicators are reported in Tables I–III, while the specific classification results on each dataset are presented in Figs. 7–9.

1) Quantitative Indicators Evaluation: To evaluate the classification performance of the proposed method, Table I gives the comparative evaluation with several state-of-the-art classification methods on the UCM dataset. It can be concluded from the table that the proposed LSMNet achieves the highest OA values of 99.29% under 80% training proportions, which is at least 0.33% higher than other methods, indicating that the proposed

TABLE II
CLASSIFICATION PERFORMANCE OF DIFFERENT METHODS ON THE AID

Method	50% Training Ratio	20% Training Ratio
CaffeNet [66]	89.53±0.31	86.86±0.47
GoogLeNet [66]	86.39±0.55	83.44±0.40
VGG-16 [66]	89.64±0.36	86.59±0.29
salM^{3}LBP-CLM [67]	89.76±0.45	86.92±0.35
TEX-Net-LF [56]	92.96±0.18	90.87±0.11
Fusion by addition [16]	91.87±0.36	/
VGG-16-CapsNet [68]	94.74±0.17	91.63±0.19
Inception-v3-CapsNet [68]	96.32±0.12	93.79±0.13
GBNet+global feature [69]	98.48±0.12	92.20±0.23
MSA-Network [70]	96.01±0.43	93.53±0.21
PVT-V2-B0 [71]	96.27±0.14	93.52±0.35
LSMNet (ours)	**96.78±0.16**	**94.31±0.10**

Bold values represents the best values of the experimental results.

TABLE III
CLASSIFICATION PERFORMANCE OF DIFFERENT METHODS ON THE NWPU DATASET

Method	20% Training Ratio	10% Training Ratio
AlexNet [47]	79.85±0.13	76.69±0.21
GoogLeNet [47]	78.48±0.26	76.19±0.38
VGG-16 [47]	79.79±0.15	76.47±0.18
Two-Stream Fusion [53]	83.16±0.18	80.22±0.22
BoCF [47]	84.32±0.17	82.65±0.34
Triple Networks [72]	92.33±0.20	/
VGG-16-CapsNet [68]	89.18±0.14	85.08±0.13
Inception-v3-CapsNet [68]	92.60±0.11	89.03±0.21
ACR-MLFF [73]	92.45±0.20	90.01±0.33
ViT-B-32 [74]	92.61±0.14	90.05±0.29
PVT-V2-B0 [71]	92.95±0.09	89.72±0.16
LSMNet (ours)	**93.16±0.13**	**90.80±0.15**

Bold values represents the best values of the experimental results.

LSMNet effectively models the spatial and channel associations, and improves the feature expression ability.

The proposed approach is also tested on the AID to show its effectiveness. Table II lists the classification performance of LSMNet and other comparative methods. It can be seen that the proposed method achieves an OA of 96.78% and 94.31% when using 50% and 20% training samples, respectively, achieving the best performance compared with other methods. In this dataset, the progress made by LSMNet is more obvious, which proves that the proposed method can achieve better performance with the blessing of more data.

Table III gives the classification performance comparison between the proposed method and existing methods on the more challenging NWPU dataset. When the proportion of training samples is 10%, LSMNet has achieved at least 0.75% accuracy improvement, which is a remarkable classification result. When 20% of the images are selected as the training set, it has achieved 0.21% OA improvement compared with the suboptimal method. The excellent performance of the proposed method further verifies its applicability and robustness in complex scenarios.

2) Confusion Matrix Display: To show the detailed classification results of the proposed method in each category more concretely, the corresponding confusion matrix is calculated and visualized. A total of six confusion matrices are drawn according to the experimental results obtained from different proportions of training data on three datasets.

Fig. 7 shows the confusion matrices of the classification results obtained by LSMNet on the UCM dataset when the training ratios are 80% and 50%. It can be seen from the figure that even when the training set accounted for 50%, there are 17 categories with an accuracy exceeding 99.5%. When the training samples increase, the classification performance is further improved.

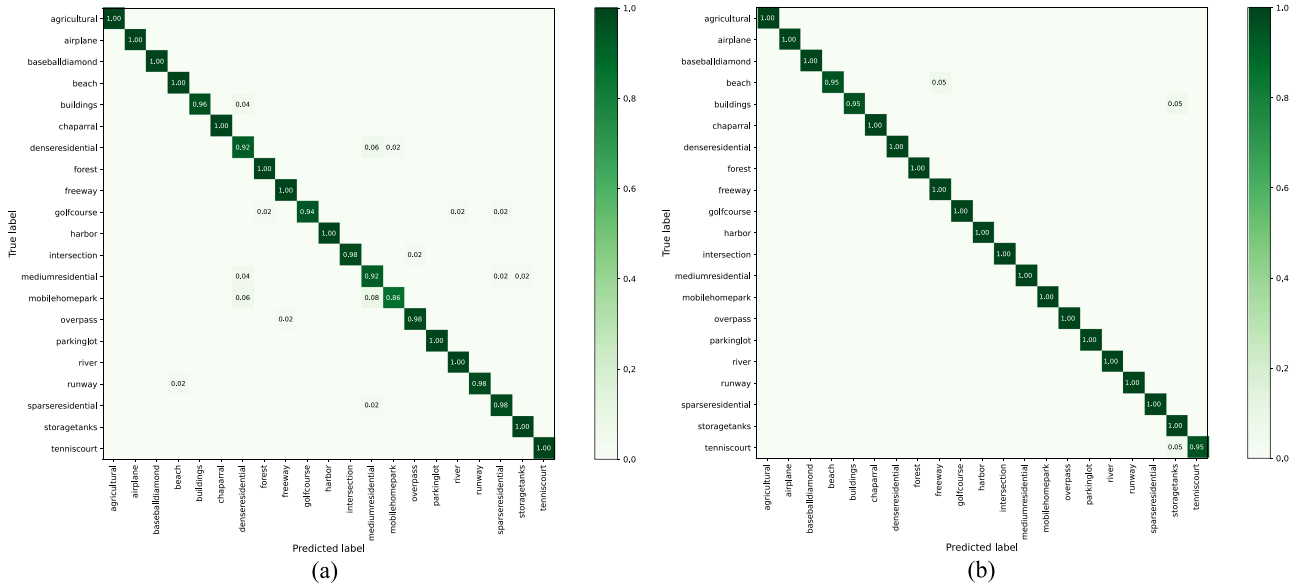
Fig. 8 shows the classification results on AID. When the proportion of training samples is 20%, only five categories' accuracy is lower than 90%. On the one hand, the similarity of samples between different categories increased, and on the other hand, the difference of samples within a class has also been reflected. With the increase of training sample, the classification accuracy of *center*, *park*, *school*, and *square* all exceeded 90%. This is because with the increase of sample number, LSMNet can better capture space and channel information and improve the discrimination of classification features.

Fig. 9 gives the confusion matrix generated from the classification results by LSMNet on the NWPU dataset with the training ratios of 20% and 10%. It can be seen from the figure that the classification accuracy of most categories exceeds 90% in both cases, which proves that the proposed method can achieve satisfactory results even in challenging and complex scenes.

D. Ablation Studies

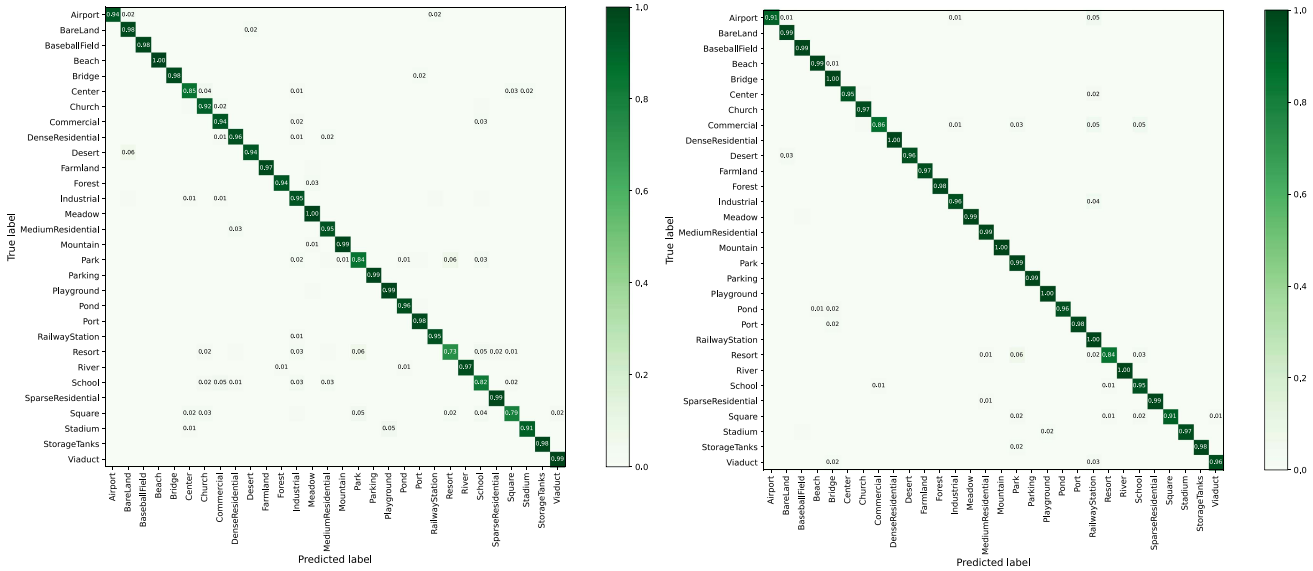
To verify the role of channel separation and mixing module, large kernel deepwise convolution, and Fused-MBConv in this article, this section conducts a series of ablation experiments to explore the importance of various modules for network performance improvement. The specific experimental results are listed in Table IV.

The variant 1 ("without LDC") verifies the impact of large convolution kernels on experimental performance by replacing large kernel deepwise convolution with traditional small convolution kernels (3×3 and 5×5). By comparing "Full model" and variant 1 ("without LDC"), the introduction of large kernel deepwise convolution enhances the capture of nonlocal overall information in the scene, and reduces the optimization difficulty of the large convolution kernel model, so that the model can still obtain better classification performance with limited training samples. The comparison between "Full model" and variant 2 ("without CSM") demonstrates that the interaction between feature channels can be better achieved and feature representation ability can be improved after the feature dependency relationship between channels is fully modeled. The improvement of performance proves that the independence of large kernel convolution in the channel dimension makes it insufficient to model channel correlation, and the designed channel separation and mixing module makes up for this limitation well. Finally, in the absence of Fused-MBConv (without FMBConv), the lack of local detailed features diminishes the model's ability to capture spatial local features, decreasing classification accuracy. By combining large kernel deepwise convolution, channel separation and mixing module and Fused-MBConv, the proposed LSMNet effectively obtains long-range dependencies and local



(a) (b)

Fig. 7. Confusion matrices of the proposed method on the UCM dataset by fixing the training ratios to 50% and 80%: (a) 50% and (b) 80%.



(a) (b)

Fig. 8. Confusion matrices for the proposed method on the AID by fixing the training ratio to 20% and 50%: (a) 20% and (b) 50%.

TABLE IV
ABLATION STUDIES FOR THE PROPOSED LSMNET ON THREE DATASETS

No.	OA (%)	Full model	Without LDC	Without CSM	Without FMBC
#1	UCM(50%)	97.71	97.25	97.33	97.52
#2	UCM(80%)	99.29	98.76	99.01	99.05
#3	AID(20%)	94.31	93.89	94.12	94.15
#4	AID(50%)	96.78	96.23	96.54	96.49
#5	NWPU (10%)	90.80	90.17	90.56	90.61
#6	NWPU (20%)	93.16	92.61	92.93	92.97

CSM: channel separation and mixing module, LDC: large kernel deep-wise convolution, and FMBC: Fused-MBConv.

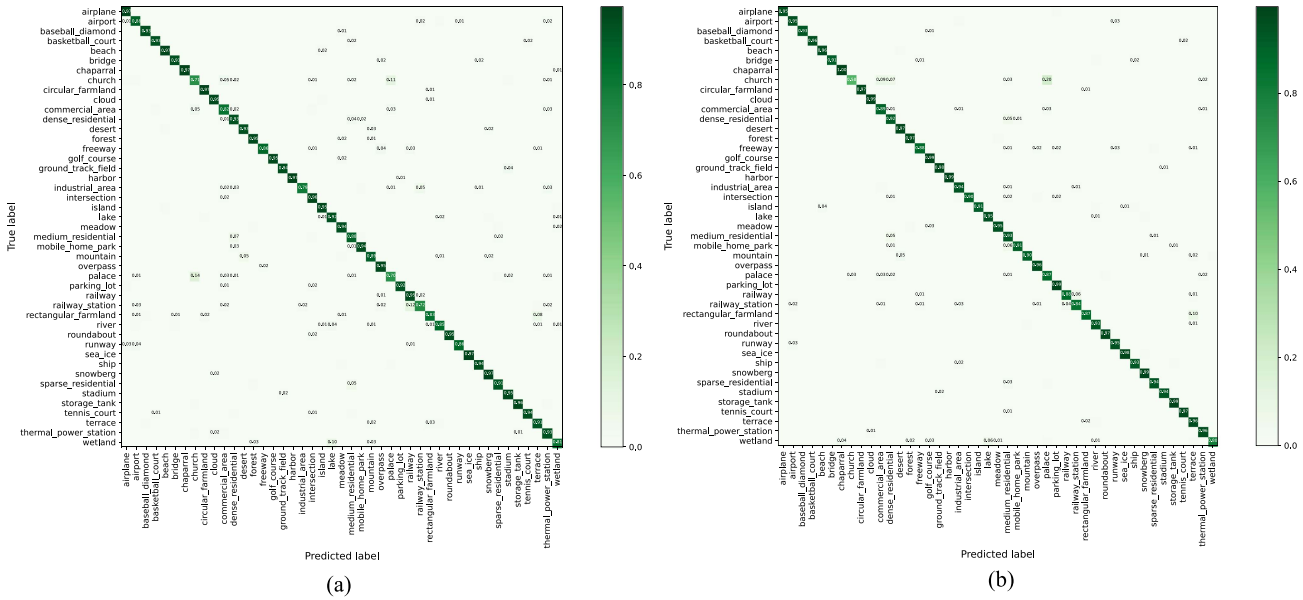


Fig. 9. Confusion matrices for the proposed method on the NWPU dataset by fixing the training ratio to 10% and 20%: (a) 10% and (b) 20%.

details, while guaranteeing channel correlations, enabling it to achieve state-of-the-art performance in multiple datasets.

IV. Conclusion

In this article, a novel LSMNet is proposed and applied to the task of remote sensing scene classification. First, to address the limitations of traditional small convolutional kernels with restricted receptive fields and the high training complexity associated with large convolutional kernels, a large kernel depthwise convolution is introduced to comprehensively understand the overall information of scene images. Combined with conventional convolutions, this approach enables information extraction at different receptive fields, enriching the spatial dimension of feature expressiveness. Furthermore, to extract interactions and dependencies among channel dimensions, a novel channel separation and mixing module is designed to model channel dependencies. The combination of them enables the acquisition of both spatial and channel dependencies within the feature maps, resulting in enhanced representational capabilities in the final classification features. Extensive experimental results on three datasets also prove the superiority of the proposed method from the perspective of quantitative analysis and visual analysis.

References

[1] H. Jiang et al., "A survey on deep learning-based change detection from high-resolution remote sensing images," *Remote Sens.*, vol. 14, no. 7, 2022, Art. no. 1552.

[2] X. Liu et al., "Classifying urban land use by integrating remote sensing and social media data," *Int. J. Geographical Inf. Sci.*, vol. 31, no. 8, pp. 1675–1696, 2017.

[3] W. Zhao, Y. Bo, J. Chen, D. Tiede, T. Blaschke, and W. J. Emery, "Exploring semantic elements for urban scene recognition: Deep integration of high-resolution imagery and OpenStreetMap (OSM)," *ISPRS J. Photogrammetry Remote Sens.*, vol. 151, pp. 237–250, 2019.

[4] Y. Su, Y. Zhong, Q. Zhu, and J. Zhao, "Urban scene understanding based on semantic and socioeconomic features: From high-resolution remote sensing imagery to multi-source geographic datasets," *ISPRS J. Photogrammetry Remote Sens.*, vol. 179, pp. 50–65, 2021.

[5] R. Yang et al., "Identifying urban wetlands through remote sensing scene classification using deep learning: A case study of Shenzhen, China," *ISPRS Int. J. Geo- Inf.*, vol. 11, no. 2, 2022, Art. no. 131.

[6] M. Ragab, "Leveraging mayfly optimization with deep learning for secure remote sensing scene image classification," *Comput. Elect. Eng.*, vol. 108, 2023, Art. no. 108672.

[7] J. Wang, F. Gao, J. Dong, and Q. Du, "Adaptive dropblock-enhanced generative adversarial networks for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 6, pp. 5040–5053, Jun. 2021.

[8] J. Yao, B. Zhang, C. Li, D. Hong, and J. Chanussot, "Extended vision transformer (ExViT) for land use and land cover classification: A multimodal deep learning framework," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5514415, doi: 10.1109/TGRS.2023.3284671.

[9] D. Hong et al., "Cross-city matters: A multimodal remote sensing benchmark dataset for cross-city semantic segmentation using high-resolution domain adaptation networks," *Remote Sens. Environ.*, vol. 299, 2023, Art. no. 113856.

[10] J. Wang, W. Li, Y. Wang, R. Tao, and Q. Du, "Representation-enhanced status replay network for multisource remote-sensing image classification," *IEEE Trans. Neural Netw. Learn. Syst.*, early access, Jun. 28, 2023, doi: 10.1109/TNNLS.2023.3286422.

[11] J. Wang, W. Li, Y. Gao, M. Zhang, R. Tao, and Q. Du, "Hyperspectral and SAR image classification via multiscale interactive fusion network," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 34, no. 12, pp. 10823–10837, Dec. 2022.

[12] M. Zhang, W. Li, X. Zhao, H. Liu, R. Tao, and Q. Du, "Morphological transformation and spatial-logical aggregation for tree species classification using hyperspectral imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5501212, doi: 10.1109/TGRS.2022.3233847.

[13] V. Bisot, R. Serizel, S. Essid, and G. Richard, "Feature learning with matrix factorization applied to acoustic scene classification," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 25, no. 6, pp. 1216–1229, Jun. 2017.

[14] A. M. Cheriyadat, "Unsupervised feature learning for aerial scene classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 52, no. 1, pp. 439–451, Jan. 2013.

[15] F. Zhang, B. Du, and L. Zhang, "Saliency-guided unsupervised feature learning for scene classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 53, no. 4, pp. 2175–2184, Apr. 2014.

[16] S. Chaib, H. Liu, Y. Gu, and H. Yao, "Deep feature fusion for VHR remote sensing scene classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 8, pp. 4775–4784, Aug. 2017.

[17] 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 J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 13, pp. 3735–3756, 2020, doi: 10.1109/JSTARS.2020.3005403.

[18] L. Sun, J. Pan, and J. Tang, "ShuffleMixer: An efficient ConvNet for image super-resolution," *Adv. Neural Inf. Process. Syst.*, vol. 35, pp. 17314–17326, 2022.

[19] H. Li, H. Gu, Y. Han, and J. Yang, "Object-oriented classification of high-resolution remote sensing imagery based on an improved colour structure code and a support vector machine," *Int. J. Remote Sens.*, vol. 31, no. 6, pp. 1453–1470, 2010.

[20] G. Cheng, J. Han, P. Zhou, and L. Guo, "Multi-class geospatial object detection and geographic image classification based on collection of part detectors," *ISPRS J. Photogrammetry Remote Sens.*, vol. 98, pp. 119–132, 2014.

[21] E. Aptoula, "Remote sensing image retrieval with global morphological texture descriptors," *IEEE Trans. Geosci. Remote Sens.*, vol. 52, no. 5, pp. 3023–3034, May 2013.

[22] G. Cheng, J. Han, L. Guo, and T. Liu, "Learning coarse-to-fine sparselets for efficient object detection and scene classification," in *Proc. IEEE Conf. Comput. Vis. pattern Recognit.*, 2015, pp. 1173–1181.

[23] Y. Bazi, M. M. Al Rahhal, H. Alhichri, and N. Alajlan, "Simple yet effective fine-tuning of deep CNNs using an auxiliary classification loss for remote sensing scene classification," *Remote Sens.*, vol. 11, no. 24, 2019, Art. no. 2908.

[24] A. Raza, H. Huo, S. Sirajuddin, and T. Fang, "Diverse capsules network combining multiconvolutional layers for remote sensing image scene classification," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 13, pp. 5297–5313, 2020, doi: 10.1109/JSTARS.2020.3021045.

[25] J. Luo and M. Boutell, "Natural scene classification using overcomplete ICA," *Pattern Recognit.*, vol. 38, no. 10, pp. 1507–1519 , 2005.

[26] J. Wang, C. Luo, H. Huang, H. Zhao, and S. Wang, "Transferring pre-trained deep CNNs for remote scene classification with general features learned from linear PCA network," *Remote Sens.*, vol. 9, no. 3, 2017, Art. no. 225.

[27] A. Sinha, S. Banerji, and C. Liu, "New color GPHOG descriptors for object and scene image classification," *Mach. Vis. Appl.*, vol. 25, pp. 361–375, 2014.

[28] J. Li et al., "The study of scene classification in the multisensor remote sensing image fusion," *Math. Problems Eng.*, vol. 2013, 2013.

[29] K. Hotta, "Local autocorrelation of similarities with subspaces for shift invariant scene classification," *Pattern Recognit.*, vol. 44, no. 4, pp. 794–799, 2011.

[30] L. Dong, J. Su, and E. Izquierdo, "Scene-oriented hierarchical classification of blurry and noisy images," *IEEE Trans. Image Process.*, vol. 21, no. 5, pp. 2534–2545, May 2012.

[31] K. Hotta, "Local co-occurrence features in subspace obtained by KPCA of local blob visual words for scene classification," *Pattern Recognit.*, vol. 45, no. 10, pp. 3687–3694, 2012.

[32] A. Bolovinou, I. Pratikakis, and S. Perantonis, "Bag of spatio-visual words for context inference in scene classification," *Pattern Recognit.*, vol. 46, no. 3, pp. 1039–1053, 2013.

[33] X. Zhu and Z. Yang, "Multi-scale spatial concatenations of local features in natural scenes and scene classification," *PLoS One*, vol. 8, no. 9, 2013, Art. no. e76393.

[34] L. Zhou, Z. Zhou, and D. Hu, "Scene classification using a multi-resolution bag-of-features model," *Pattern Recognit.*, vol. 46, no. 1, pp. 424–433, 2013.

[35] F. Hu, G.-S. Xia, Z. Wang, X. Huang, L. Zhang, and H. Sun, "Unsupervised feature learning via spectral clustering of multidimensional patches for remotely sensed scene classification," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 8, no. 5, pp. 2015–2030, May 2015.

[36] Y. Zhong, Q. Zhu, and L. Zhang, "Scene classification based on the multifeature fusion probabilistic topic model for high spatial resolution remote sensing imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 53, no. 11, pp. 6207–6222, Nov. 2015.

[37] M. A. Kaljahi, S. Palaiahnakote, M. H. Anisi, M. Y. I. Idris, M. Blumenstein, and M. K. Khan, "A scene image classification technique for a ubiquitous visual surveillance system," *Multimedia Tools Appl.*, vol. 78, pp. 5791–5818, 2019.

[38] G. Sheng, W. Yang, T. Xu, and H. Sun, "High-resolution satellite scene classification using a sparse coding based multiple feature combination," *Int. J. Remote Sens.*, vol. 33, no. 8, pp. 2395–2412, 2012.

[39] Q. Kunlun, Z. Xiaochun, W. Baiyan, and W. Huayi, "Sparse coding-based correlation model for land-use scene classification in high-resolution remote-sensing images," *J. Appl. Remote Sens.*, vol. 10, no. 4, pp. 042005–042005, 2016.

[40] A. Qayyum et al., "Scene classification for aerial images based on CNN using sparse coding technique," *Int. J. Remote Sens.*, vol. 38, no. 8–10, pp. 2662–2685, 2017.

[41] X.-H. Han and Y.-W. Chen, "Generalized aggregation of sparse coded multi-spectra for satellite scene classification," *ISPRS Int. J. Geo- Inf.*, vol. 6, no. 6, 2017, Art. no. 175.

[42] B. Gajjar, H. Mewada, and A. Patani, "Sparse coded spatial pyramid matching and multi-kernel integrated SVM for non-linear scene classification," *J. Elect. Eng.*, vol. 72, no. 6, pp. 374–380, 2021.

[43] E. Li, P. Du, A. Samat, Y. Meng, and M. Che, "Mid-level feature representation via sparse autoencoder for remotely sensed scene classification," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 10, no. 3, pp. 1068–1081, Mar. 2017.

[44] E. Othman, Y. Bazi, N. Alajlan, H. Alhichri, and F. Melgani, "Using convolutional features and a sparse autoencoder for land-use scene classification," *Int. J. Remote Sens.*, vol. 37, no. 10, pp. 2149–2167, 2016.

[45] S. Chaib, Y. Gu, and H. Yao, "An informative feature selection method based on sparse PCA for VHR scene classification," *IEEE Geosci. Remote Sens. Lett.*, vol. 13, no. 2, pp. 147–151, Feb. 2016.

[46] Q. Zhu, Y. Zhong, B. Zhao, G.-S. Xia, and L. Zhang, "Bag-of-visual-words scene classifier with local and global features for high spatial resolution remote sensing imagery," *IEEE Geosci. Remote Sens. Lett.*, vol. 13, no. 6, pp. 747–751, Jun. 2016.

[47] G. Cheng, J. Han, and X. Lu, "Remote sensing image scene classification: Benchmark and state of the art," *Proc. IEEE*, vol. 105, no. 10, pp. 1865–1883, Oct. 2017.

[48] K. Amiri, M. Farah, and U. M. Leloglu, "BOVSG: Bag of visual subgraphs for remote sensing scene classification," *Int. J. Remote Sens.*, vol. 41, no. 5, pp. 1986–2003, 2020.

[49] J. Wang, W. Li, M. Zhang, R. Tao, and J. Chanussot, "Remote-sensing scene classification via multistage self-guided separation network," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5615312, doi: 10.1109/TGRS.2023.3295797.

[50] J. Yao, D. Hong, H. Wang, H. Liu, and J. Chanussot, "UCSL: Toward unsupervised common subspace learning for cross-modal image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5514212, doi: 10.1109/TGRS.2023.3282951.

[51] J. Yao et al., "Semi-active convolutional neural networks for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5537915, doi: 10.1109/TGRS.2022.3206208.

[52] D. Hong et al., "SpectralGPT: Spectral foundation model," 2023, *arXiv:2311.07113*.

[53] Y. Yu et al., "A two-stream deep fusion framework for high-resolution aerial scene classification," *Comput. Intell. Neurosci.*, vol. 2018, 2018.

[54] X. Han, Y. Zhong, L. Cao, and L. Zhang, "Pre-trained AlexNet architecture with pyramid pooling and supervision for high spatial resolution remote sensing image scene classification," *Remote Sens.*, vol. 9, no. 8, 2017, Art. no. 848.

[55] P. S. Yee, K. M. Lim, and C. P. Lee, "DeepScene: Scene classification via convolutional neural network with spatial pyramid pooling," *Expert Syst. with Appl.*, vol. 193, 2022, Art. no. 116382.

[56] R. M. Anwer, F. S. Khan, J. Van De Weijer, M. Molinier, and J. Laaksonen, "Binary patterns encoded convolutional neural networks for texture recognition and remote sensing scene classification," *ISPRS J. Photogrammetry Remote Sens.*, vol. 138, pp. 74–85, 2018.

[57] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. Alemi, "Inception-v4, inception-ResNet and the impact of residual connections on learning," *Proc. AAAI Conf. Artif. Intell.*, vol. 31, no. 1, 2017.

[58] M.-H. Guo, C.-Z. Lu, Q. Hou, Z. Liu, M.-M. Cheng, and S.-M. Hu, "SegNeXt: Rethinking convolutional attention design for semantic segmentation," *Adv. Neural Inf. Process. Syst.*, vol. 35, pp. 1140–1156, 2022.

[59] M. Tan and Q. Le, "Efficientnetv2: Smaller models and faster training," in *Proc. Int. Conf. Mach. Learn.*, 2021, pp. 10096–10106.

[60] S. Elfwing, E. Uchibe, and K. Doya, "Sigmoid-weighted linear units for neural network function approximation in reinforcement learning," *Neural Netw.*, vol. 107, pp. 3–11, 2018.

[61] D. Yu, Q. Xu, H. Guo, C. Zhao, Y. Lin, and D. Li, "An efficient and lightweight convolutional neural network for remote sensing image scene classification," *Sensors*, vol. 20, no. 7, 2020, Art. no. 1999.

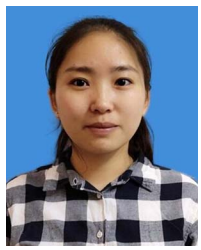
[62] Q. Bi, K. Qin, Z. Li, H. Zhang, K. Xu, and G.-S. Xia, "A multiple-instance densely-connected ConvNet for aerial scene classification," *IEEE Trans. Image Process.*, vol. 29, pp. 4911–4926, 2020, doi: 10.1109/TIP.2020.2975718.
[63] X. Ding, X. Zhang, J. Han, and G. Ding, "Scaling up your kernels to 31x31: Revisiting large kernel design in CNNs," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 11963–11975.
[64] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 7132–7141.
[65] Y. Yang and S. Newsam, "Bag-of-visual-words and spatial extensions for land-use classification," in *Proc. 18th SIGSPATIAL Int. Conf. Adv. Geographic Inf. Syst.*, 2010, pp. 270–279.
[66] G.-S. Xia et al., "AID: A benchmark data set for performance evaluation of aerial scene classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 7, pp. 3965–3981, Jul. 2017.
[67] X. Bian, C. Chen, L. Tian, and Q. Du, "Fusing local and global features for high-resolution scene classification," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 10, no. 6, pp. 2889–2901, Jun. 2017.
[68] W. Zhang, P. Tang, and L. Zhao, "Remote sensing image scene classification using CNN-CapsNet," *Remote Sens.*, vol. 11, no. 5, 2019, Art. no. 494.
[69] H. Sun, S. Li, X. Zheng, and X. Lu, "Remote sensing scene classification by gated bidirectional network," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 1, pp. 82–96, Jan. 2020.
[70] G. Zhang et al., "A multiscale attention network for remote sensing scene images classification," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 9530–9545, 2021, doi: 10.1109/JSTARS.2021.3109661.
[71] W. Wang et al., "Pvt v2: Improved baselines with pyramid vision transformer," *Comput. Vis. Media*, vol. 8, no. 3, pp. 415–424, 2022.
[72] Y. Liu and C. Huang, "Scene classification via triplet networks," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 11, no. 1, pp. 220–237, Jan. 2018.
[73] X. Wang, L. Duan, A. Shi, and H. Zhou, "Multilevel feature fusion networks with adaptive channel dimensionality reduction for remote sensing scene classification," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, 2021, Art. no. 8010205, doi: 10.1109/LGRS.2021.3070016.
[74] A. Dosovitskiy et al., "An image is worth 16x16 words: Transformers for image recognition at scale," 2021, *arXiv:2010.11929*.

Keqian Zhang received the B.S. degree in surveying and mapping engineering from Shenyang Urban Construction University, Shenyang, China, in 2021, and is currently working toward the M.S. degree in resources and environment from Henan Polytechnic University.

He has authored or coauthored one academic paper, and granted one invention patent and two software copyrights. His research interests include remote sensing image scene classification and pattern recognition.

Tengfei Cui received the M.S. degree in economics from Boston University, Boston, MA, USA, in 2021.

He has authored or coauthored several academic papers where his research focuses on computer vision.



Wei Wu received the B.S. degree in surveying and mapping engineering from Shenyang Urban Construction University, Shenyang, China, in 2021, and is currently working toward the M.S. degree in resources and environment from the Guilin University of Technology.

She has authored or coauthored one academic paper, and granted one invention patent and two software copyrights. Her research interests include image classification and target detection.

Ms. Wu attended and presented a paper at the *International Conference on Data Science and Network Security* in 2023.

Xueke Zheng received the B.S. degree in remote sensing science and technology from Henan Polytechnic University, Jiaozuo, China, in 2021, and is currently working toward the M.S. degree in resources and environment from Henan Polytechnic University.

His research interests include remote sensing image classification and soil moisture inversion.

Gang Cheng received the Ph.D. degree in cartography and geoinformation engineering from Wuhan University, Wuhan, China, in 2008.

He has presided over and participated in several projects, such as 973 Preliminary Research Special Project, the National Key R&D Program, the National Natural Science Foundation of China, the Humanities and Social Sciences Foundation of the Ministry of Education, Innovative Science and Technology Team of Henan Province, Science and Technology Tackling Project of Henan Province, and China Postdoctoral Fund. He has authored or coauthored more than 40 academic papers, been authorized to issue more than 20 patents/software copyrights, and edited three books.

Dr. Cheng was the recipient of more than ten awards at provincial and ministerial levels, such as the National Award for Progress in Surveying and Mapping Science and Technology, the Award for Progress in Geographic Information Science and Technology, the Science and Technology Award of the Coal Industry, and the Science and Technology Progress Award of Henan Province.