Large Kernel Sparse ConvNet Weighted by Multi-Frequency Attention for Remote Sensing Scene Understanding

Junjie Wang, Wei Li, Mengmeng Zhang, Jocelyn Chanussot

Large Kernel Sparse ConvNet weighted by Multi-frequency Attention for Remote Sensing Scene Understanding

Junjie Wang, Wei Li, *Senior Member, IEEE*, Mengmeng Zhang, Jocelyn Chanussot, *Fellow, IEEE*

***Abstract*—Remote sensing scene understanding is a highly challenging task, and has gradually emerged as a research hotspot in the field of intelligent interpretation of remote sensing data. Recently, the use of convolutional neural networks (CNNs) has been proven to be a fruitful advancement. However, with the emergence of visual transformers (ViTs), the limitations of traditional small convolutional kernels in directly capturing a large receptive field have posed significant challenges to their dominant role. Additionally, the fixed neuron connections between different convolutional layers have weakened the practicality and adaptability of the models. Furthermore, the global average pooling also leads to the loss of effective information in the acquired features. In this work, a Large kernel Sparse ConvNet weighted by Multi-frequency Attention (LSCNet) is proposed. Firstly, unlike traditional convolutional neural networks, it utilizes two parallel rectangular convolutional kernels to approximate a large kernel, achieving comparable or even better results than ViTs-based methods. Secondly, an adaptive sparse optimization strategy is employed to dynamically optimize the fixed neuron connections between different convolutional layers, achieving a favorable connectivity pattern for capturing abstract features more accurately. Lastly, a novel multi-frequency attention (MFA) module is used to replace global average pooling (GAP), so as to preserve more useful information while weighting the recognition features, thereby enhancing the discriminative and learning abilities of the model. In the conducted experiments, LSCNet achieves the best recognition results on three well-known remote sensing aerial datasets when compared to the state-of-the-art methods (including ViTs-based methods).**

***Index Terms*—Remote sensing, scene understanding, large kernel convolution, adaptive sparse optimization, multi-frequency attention.**

I. Introduction

REMOTE sensing scene understanding is a vital yet difficult task in the field of intelligent interpretation of remote sensing data. It aims to capture high-level semantic information from images and precisely assign corresponding class labels to them. It has applications in various military and civilian domains, including natural disaster detection, weapon guidance, traffic supervision, and land cover monitoring [1]–[5]. In recent years, the advancement in remote sensing technology has increased the level of data abstraction from pixels to objects and ultimately to scenes [6]–[8]. In order to keep pace with these advancements, numerous researchers have dedicated their efforts over the past few decades to address the challenges and achieve scene-level image understanding [9]–[11]. In this task, effective feature extraction plays a crucial role, and based on the means of feature extraction, existing scene understanding works can be roughly divided into three directions: methods using low-level visual features, methods relying on mid-level visual representations, and methods based on high-level visual information [12] [13].

Early works for scene understanding mainly used low-level visual features and focused on designing various handcrafted features, such as color, texture, shape, spatial, and spectral information. For instance, one of the most straightforward yet useful visual characteristics for scene understanding tasks is the color histogram feature [14]. It is not only easy to implement but also exhibits translation and rotation invariance. Additionally, texture descriptors analyze the structural characteristics of an image by computing relative differences in local regions, which facilitates the recognition of textural scene images [15]. The GIST descriptor provides a global representation by computing the spatial distribution of local feature detector outputs in subregions, allowing for the representation of scale and orientation information in scene images [13], [16]–[19]. On the other hand, the methods based on scale-invariant feature transform (SIFT) utilize gradient information around key points in the scene for feature description [20]. These methods work well in scenes with homogenous spatial distributions using low-level visual features. However, they struggle to describe scenes with high spatial disparity and heterogeneous distributions [21]. In contrast to methods that use low-level visual features, methods relying on mid-level visual representations aim to utilize basic functions for feature encoding and take a series of low-level features or raw pixel values as input. Among them, the well-known bag-of-visual-words (BoVW) model is one of the most popular mid-level feature-based methods [22]. It learns a vocabulary of visual words by performing k-means clustering on local features. In addition, methods such as principal component analysis (PCA) [23], sparse coding [24], and autoencoders [25] are also typical approaches that rely on mid-level visual representations.

With the popularity of deep learning methods, remarkable achievements have been made in various application domains, including image classification, object detection, and semantic segmentation, and the feature representation of images has

This paper is supported by the National Key R&D Program of China (Grant No. 2021YFB3900502). (Corresponding author: Wei. Li)

J. Wang, W. Li, and M. Zhang are with the School of Information and Electronics, Beijing Institute of Technology, and Beijing Kay Laboratory of Fractional Signals and Systems, Beijing 100081, China (e-mail: junjiewang@bit.edu.cn, liwei089@ieee.org, mengmengzhang@bit.edu.cn).

Jocelyn Chanussot is with the LJK, CNRS, INRIA, Grenoble INP, University of Grenoble Alpes, 38000 Grenoble, France, and also with the Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing 100094, China (e-mail: jocelyn.chanussot@gipsa-lab.grenoble-inp.fr).

entered a new stage [26]–[30]. Unlike low-level and mid-level features, deep learning models can extract more abstract, effective, and discriminative high-level visual features without the need for extensive engineering efforts or domain expertise. Among all deep learning models, convolutional neural networks (CNNs) have demonstrated an excellent application for scene understanding tasks and have achieved satisfying performance. He *et al.* [31] incorporated skip connections and covariance pooling into CNNs, combining multi-resolution feature maps and leveraging second-order information within these feature maps. Zhang *et al.* [32] introduced dilated convolutions and channel attention to the network so as to generate more discriminative features, resulting in a lightweight and efficient network architecture. To refine the abstract features that the VGG-Net collected, a discriminant correlation analysis (DCA) [33] was utilized to fuse different features. Xu *et al.* [34] proposed a feature aggregation network that consists of two streams, a discriminative feature stream and a general feature stream, which are integrated using a weighted fusion method. Fang *et al.* [35] introduced frequency domain features into the traditional spatial domain structure, obtaining more advanced feature representations through joint learning.

In addition to CNNs, a new deep learning model based on visual transformers (ViTs) has been proposed and widely applied in various domains [36]. The rise of ViTs can be partly attributed to their ability to capture a larger receptive field. Compared to CNNs that perform convolution operations on a small sliding window, ViTs utilize global or local attention with larger window sizes, enabling each layer to capture a larger receptive field. As a result, some studies have started to explore the integration of ViTs into remote sensing scene understanding tasks. In [37], a remote sensing scene classification method based on vision transformers was proposed. It divided the image into small patches and transformed them into sequences, after which the generated sequences were fed into a multi-head attention layer to generate the final representation. Ma *et al.* [38] proposed a feature learning module to simultaneously explore homogeneous and heterogeneous features in remote sensing scenes. In [39], a model called efficient multiscale transformer and cross-level attention learning (EMTCAL) was proposed, and the multi-level feature extraction modules and context information extraction modules were used to obtain rich perceptual information, combined with a developed cross-level attention model to aggregate and explore the correlation between multi-level features. However, the core conclusion of the original ViTs is that when there is enough data for training, ViTs outperform CNNs, surpassing the limitations of the lack of inductive biases, and achieving better performance in downstream tasks. However, when the training dataset is not large enough, the performance of ViTs is usually worse than equivalently sized CNNs. Inspired by this, some researchers have attempted to design advanced pure CNN architectures and equip them with larger convolutional kernels to obtain a larger receptive field. For example, Ding *et al.* [40] extended the size of the convolutional kernel to 31x31, successfully achieving results comparable to methods based on ViTs. However, simply using large kernels makes the training process very challenging [41], thus requiring a novel approach and strategy to enlarge the convolutional receptive field while mitigating training difficulty.

In summary, existing research has made efforts from various aspects, such as feature extraction and network architecture design, to address the challenges of aerial scene understanding in practical applications. However, as remote sensing technology develops and the growing demand for practical applications, there are still some unresolved issues. 1) In remote sensing image processing tasks, the acquisition process of training samples requires significant manpower and material resources, making the process time-consuming. As a result, the number of available sample is usually limited. This poses a challenge for ViTs to achieve satisfactory results under such circumstances. On the other hand, traditional convolution operations with small kernel sizes hinder the acquisition of a large receptive field, thus limiting the further improvement of model performance. 2) Convolutional neural networks (CNNs) have successfully reduced network parameters through their advantage of sparse connections. However, the fixedness of sparse connections between different convolutional layers limits the further evolution of the model, weakening its practicality and adaptability. 3) Channel attention, which assigns different weights to different channels of feature maps, has become a popular and important tool in deep learning models. Obtaining the weight coefficients of channels is a crucial step in this process [42]–[44]. The commonly used global average pooling (GAP) has been a standard choice for obtaining these coefficients due to its simplicity and efficiency. However, GAP only utilizes a small portion of the obtained features, leading to a significant loss of potentially useful information and thereby reducing the discriminative power and learning capacity of the model.

To address the aforementioned issues, this paper proposes a novel Large kernel Sparse ConvNet (LSCNet) weighted by Multi-frequency attention for aerial scene understanding. Firstly, to tackle the problem of small receptive fields in traditional convolutional kernels, large-kernel convolution is introduced into the model. Specifically, two parallel rectangular kernels are utilized to approximate a large kernel, achieving comparable or even superior results compared to ViTs-based methods. This decomposition strikes a balance between capturing long-range dependencies and extracting local detailed features. Subsequently, an adaptive sparse optimization strategy is proposed to dynamically adjust sparse connections during the training process. This strategy allows the model to gradually evolve its sparse structure, leading to better performance. Finally, a multi-frequency attention (MFA) module is designed to allocate weights to different feature channels. In contrast to previous methods, it compresses and encodes channel information to explore as much potentially useful information as possible.

In conclusion, the key contributions are listed below.

1) Aiming at expanding the receptive field of the convolutional kernel, a large kernel convolution is introduced into the model. By replacing small convolutional kernels with two parallel rectangular kernels, the receptive field is expanded while preserving the ability to capture

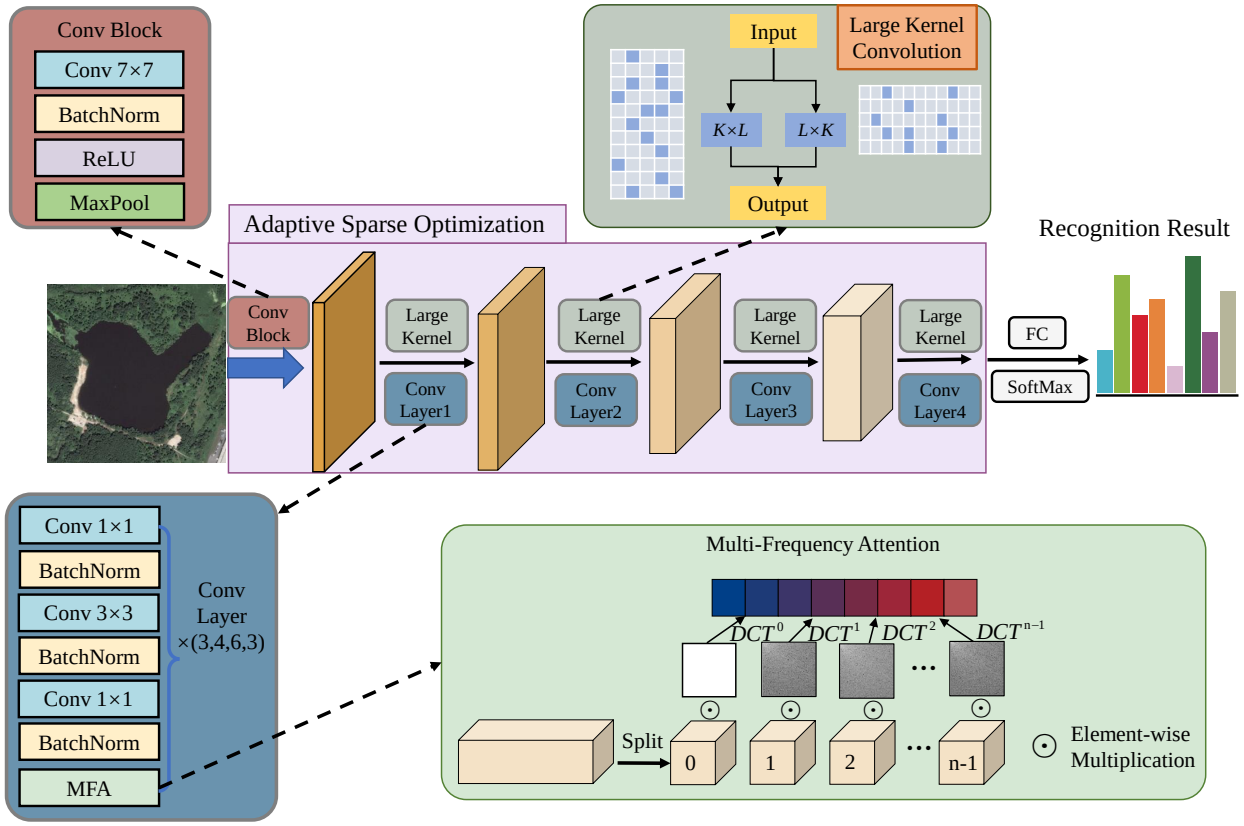


Fig. 1. Schematic illustration of the proposed LSCNet. First, the input image undergoes a *Conv Block* for basic feature extraction, followed by four *Conv Layers* that include the multi-frequency attention module to focus on important information. Next, large kernel convolutions are applied to achieve feature extraction and learning with a large receptive field. The entire network utilizes an adaptive sparse optimization strategy to dynamically adjust the connections, ensuring that the model can evolve into a better pattern. Finally, the results are obtained through a classifier.

local detailed features, thus achieving comparable performance to ViTs-based methods.

2) To adjust the fixed sparse connections between layers, an adaptive sparse optimization strategy is proposed, which adaptively adjusts the fixed connections between original layers through a dynamic "prune-and-grow" scheme. By pruning the least important weights and adding new ones, the network connections are gradually optimized towards a favorable pattern, thereby encouraging finer capturing of local features.
3) To obtain effective channel attention weights, a novel multi-frequency attention (MFA) module is designed. In contrast to traditional methods, this paper regards the channel attention representation as a compression process. By incorporating frequency domain analysis, it provides a fresh perspective that preserves more useful information while weighting the final recognition features, thus enhancing the discriminative power and learning capability of the model.

The rest of the paper is organized as follows. Section II provides a detailed description of the proposed LSCNet, including how two parallel rectangular small convolutional kernels are used to approximate a large kernel, how the connections between layers are adaptively pruned, and how more frequency components are utilized to achieve more effective attention. The efficiency of the suggested modules and the function of each module are tested through detailed experiments and discussions in Section III. Finally, section IV draws the conclusion.

II. Proposed LSCNet Framework

The proposed LSCNet framework is illustrated in Fig.1. Firstly, the scene images are fed into a feature extraction network consisting of *Conv Block* and *Conv Layer*, where large-kernel convolutions are employed to obtain a larger receptive field, and each *Conv Layer* contains multiple convolution and regularization operations. Subsequently, the multi-frequency attention module is added to the *Conv Layer* to weigh important information on the obtained feature vectors to improve the model's attention to crucial information. Throughout the training process, an adaptive sparse optimization strategy is applied to dynamically adjust the sparse connections between different layers, enabling the network to evolve to a more optimal state. By combining these modules, the understanding and learning capabilities of the model are enhanced, achieving state-of-the-art aerial scene understanding.

A. Large Kernel Convolution

Although ViTs have achieved remarkable performance in various fields, their reliance on a large amount of training data

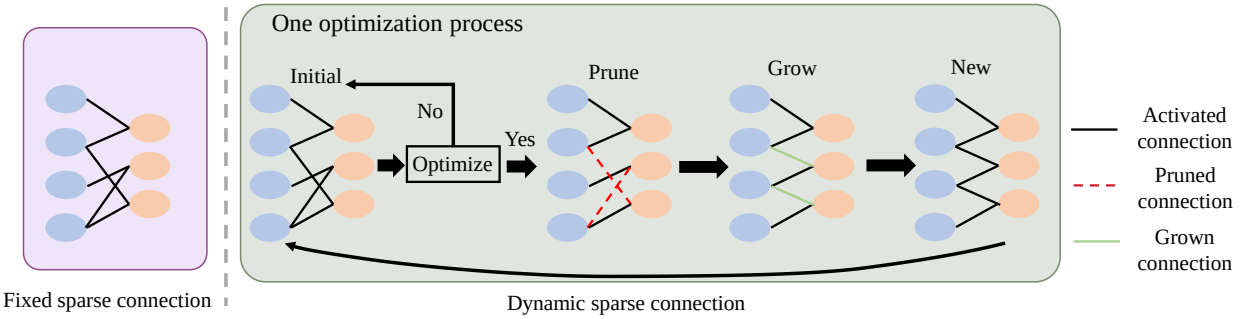


Fig. 2. Dynamic sparse connection enables the training of neural networks with initially sparse components (sparse kernels) from scratch. Throughout the training process, it adaptively fine-tunes the sparse weights by removing the least significant weights while growing new ones. This dynamic procedure gradually optimizes the sparse kernels into an effective pattern, thereby facilitating a more comprehensive extraction of local features.

limits their application in remote sensing, so it is natural to think whether it is possible to obtain larger receptive fields by using convolutions with larger kernels, thus achieving performance comparable to or even better than ViTs. Here, a $K \times K$ large-kernel convolution is replaced with two parallel rectangular convolutions, with sizes of $K \times L$ and $L \times K$, where K is equal to the length and width of the feature map, and $L < K$. The benefit of this approach lies in its ability to balance the performance of the convolutional kernels in capturing global dependencies and extracting local detailed features (the longer side is used to capture global dependencies, while the shorter side is used to extract local detailed features). Some previous works attempted to utilize two parallel or stacked complementary $K \times 1$ and $1 \times K$ convolutional kernels [45] [46]. However, the length limitation of the shorter side to 1 affected its effectiveness in extracting features along the corresponding dimension. To address this limitation, the proposed large kernel convolution increases the length of the shorter side to L (the specific value of L will be discussed in Section III-B 1), thereby improving the training stability and memory scalability of the large kernel convolution.

B. Adaptive Sparse Optimization

The popularity of Convolutional Neural Networks (CNNs) attributes to their powerful representational capabilities and the reduction of the number of parameters, among which the sparse connection between different layers is also one of the important reasons for the reduction of the number of parameters. However, these fixed sparse connections limit the further evolution of the model. As for now, related work has been proposed to obtain a new sparse connection network, including methods such as dropout and model compression. However, the stochastic nature of these sparse techniques affects the stability of the model [47] [48]. Therefore, an adaptive sparse optimization strategy is developed to adjust the sparse connections in the model by pruning the least important weights and adding new ones, which is executed once for every 100 batches of data fed into the model by the dataloader.

Firstly, neural network pruning is formulated as an optimization problem. Given a dataset $D = (x_i, y_i)_{i=1}^{num}$, and a desired dense level k. neural network pruning can be written as the following constrained optimization problem:

$$\begin{aligned} min\mathcal{L}(w; D) &= min\frac{1}{num}\sum_{i=1}^{num} l(w; (x_i, y_i)) \\ w \in R^m, \quad &\|w\|_0 \leq k. \end{aligned} \tag{1}$$

where $l(\cdot)$ is the loss function used in the model, w is the set of parameters of the neural network, m is the number of parameters, and $\|\cdot\|_0$ is the L_0 normalization.

However, to adaptively prune the neural network, a criterion is designed to measure the importance of each connection. Firstly, an auxiliary indicator variable c is introduced to represent the connections between parameters w. Now, given a dense level k, Eq. 1 is reformulated as:

$$\begin{aligned} min\mathcal{L}(c \odot w; D) &= min\frac{1}{num}\sum_{i=1}^{num} l(c \odot w; (x_i, y_i)) \\ w \in R^m, \quad c \in \{0,1\}^m, \quad &\|c\|_0 \leq k. \end{aligned} \tag{2}$$

where $\odot$ represents the Hadamard product. Different from Eq. 1, the key idea here is that since the connection weights w and the existence of connections c have been separated, the importance of each connection can be determined by measuring its impact on the loss function. For example, the value of c_j represents whether connection j exists ($c_j = 1$) or not ($c_j = 0$) in the network. Therefore, the impact of connection j on the model loss can be measured by changing the value of c_j while keeping other values unchanged,

$$\triangle\mathcal{L}_j(w; D) = \mathcal{L}(\mathbf{1} \odot w; D) - \mathcal{L}((\mathbf{1} - e_j) \odot w; D), \tag{3}$$

where e_j is the indicator vector for connection j (it has a value of 1 only at index j, and 0 everywhere else) and $\mathbf{1}$ is the vector of dimension m. However, computing $\triangle\mathcal{L}_j$ for each connection is computationally expensive, and since c is binary, $\mathcal{L}$ is non-differentiable with respect to c. Therefore, by relaxing the binary constraint on c, $\triangle\mathcal{L}_j$ can be approximated by the derivative of $\mathcal{L}$ with respect to c_j, denoted as $d_j(w; D)$,

$$\begin{aligned} \triangle \mathcal{L}_j(w;D) \approx d_j(w;D) &= \frac{\partial \mathcal{L}(c \odot w;D)}{\partial c_j}\Big|_{c=1} \\ = lim_{\delta \to 0} &\frac{\mathcal{L}(c \odot w;D) - \mathcal{L}((c-\delta e_j) \odot w;D)}{\delta}\Big|_{c=1} \end{aligned} \tag{4}$$

Returning to the original objective, the goal is to identify important connections in the neural network and remove unimportant ones, thereby achieving model pruning. For this purpose, the magnitude of the derivative d_j is used as an evaluation metric (if the magnitude of the derivative is high, it indicates that connection j has a significant impact on the loss and should be retained to ensure the learning of w_j). Based on these assumptions, the importance of a connection is defined as the normalized magnitude of the derivative,

$$s_j = \frac{|d_j(w;D)|}{\sum_{k=1}^{m} |d_k(w;D)|} \tag{5}$$

After computing the importance of the connections, only the top-k dense level. As a result, the target variable c is updated accordingly,

$$c_j = \mathit{1}[s_j - \tilde{s}_k \geq 0], \quad j \in 1...m \tag{6}$$

where $\tilde{s}_k$ is the kth largest element in the vector s and $\mathit{1}[\cdot]$ is the indicator function. At this point, neural network pruning has been completed, and the next step is to randomly grow the same number of connections (as shown in Fig. 2) to achieve prune-and-grow scheme. The weights of the new connections in the random growth process are randomly initialized. By doing so, the model can adaptively adjust the sparse weights, gradually evolving the connections in the model toward a better pattern.

C. Multi-frequency Attention Module

The commonly used channel attention modules typically include a Global Average Pooling (GAP) operation to assign a scalar weight to each channel. The initial operation is to average the information at all positions in the spatial dimension into a single value. This is done because the final weights act on the entire channel, and thus, it is essential to calculate the weights based on the overall channel information. Additionally, the aim is to leverage inter-channel correlations rather than spatial distribution correlations. Employing Global Average Pooling (GAP) to suppress spatial distribution information enables more accurate weight computation. However, due to the simplicity of GAP, it might be difficult to successfully extract complicated information from a variety of inputs, leading to a loss of significant information. Therefore, a multi-frequency attention (MFA) module is designed, which treats the acquisition of scalar weights as a compression process while preserving the overall representation capability of the channels. Specifically, a discrete cosine transform (DCT) is applied to compress the channels, followed by the utilization of multiple frequency components to achieve channel attention.

Before proceeding to the detailed method introduction, some necessary content review as well as detailed derivations are first performed, including a revisiting of the DCT and the representation flaws of the GAP. Typically, a 2D DCT is expressed as:

$$\begin{aligned} F_{DCT}(u,v) = a_0 c(u,v) &\sum_{x=0}^{K-1}\sum_{y=0}^{K-1} F(x,y) \\ cos&\frac{(2x+1)u\pi}{2K} cos\frac{(2y+1)v\pi}{2K} \\ u,v = 0,1 \ldots K-1, a_0 &= \frac{2}{K}, \\ c(u,v) = \begin{cases} 1/2 & u=v=0 \\ 1/\sqrt{2} & uv=0, u \neq v \\ 1 & uv>0 \end{cases} & \end{aligned} \tag{7}$$

where F is the input. While GAP can be viewed as a special case of 2D DCT, where its result is proportional to the lowest frequency unit of the 2D DCT. Assuming that the variables u and v in Eq. 7 are set to 0:

$$\begin{aligned} F_{DCT}(0,0) &= \frac{a_0}{2}\sum_{x=0}^{K-1}\sum_{y=0}^{K-1} F(x,y) \\ cos&\frac{(2x+1)0\pi}{2K} cos\frac{(2y+1)0\pi}{2K} \\ &= \frac{a_0}{2}\sum_{x=0}^{K-1}\sum_{y=0}^{K-1} F(x,y) \\ &= \frac{a_0}{2} GAP(F) KK \end{aligned} \tag{8}$$

From the above equations, GAP only utilizes the lowest frequency information in the frequency domain, disregarding a significant amount of potentially useful information, resulting in information loss. Therefore, addressing this issue, the proposed multi-frequency attention module leverages more information from the transformed 2D DCT, including the lowest frequency unit, to achieve a more comprehensive channel attention mechanism.

First, the input $F \in R^{C \times K \times K}$ is divided into n parts along the channel dimension, represented as [F^0, ..., F^{n-1}], where $F^i \in R^{C' \times K \times K}, C' = \frac{C}{n}$. For each part, a 2D DCT transformation is applied, and the resulting outputs are used for channel attention,

$$\begin{aligned} T^i &= 2DDCT^{u_i,v_i}(F^i) \\ &= \alpha \sum_{x=0}^{K-1}\sum_{y=0}^{K-1} F^i(x,y) cos\frac{(2x+1)u^i\pi}{2K} cos\frac{(2y+1)v^i\pi}{2K} \end{aligned} \tag{9}$$

where $\alpha = a_0 c(u^i, v^i)$ represents a constant normalization factor, $[u_i, v_i]$ are the frequency component 2D indices corresponding to F^i, and $T^i \in R^{C'}$ is the C' dimensional vector after the compression. The frequency component indices $[u_i, v_i]$ for each part F^i are chosen by applying "zig-zag" scanning, which is a method of content selection, where it starts from the top-left corner and proceeds to select frequencies along a diagonal pattern [49]. After obtaining the DCT transformation vectors for each part, they are concatenated and passed through

a sigmoid function to obtain the corresponding attention vector A,

$$A = sigmoid(fc(cat(T^0, T^1, ..., T^{n-1}))) \tag{10}$$

The above equations demonstrate that the proposed MFA module improves upon the original GAP by utilizing multiple frequency components, enriching the compressed channel representation. After obtaining the attention vector A, each channel of input F is scaled by the corresponding attention value:

$$\hat{F} = a_l F_l, \quad l \in 0, 1..C-1 \tag{11}$$

where $\hat{F}$ is the output of attention mechanism, a_l is the l-th element of attention vector A, and F_l is the l-th channel of input.

Fig. 3. Sample images of the UCM dataset: two images of each class are exhibited. (Semantic category $\sim$ Number of samples in this category)

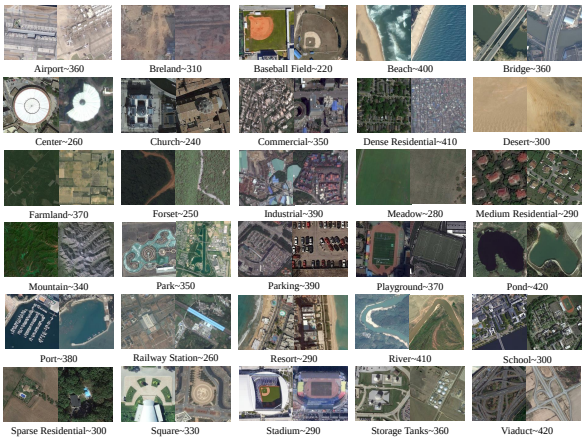


Fig. 4. Sample images of the AID dataset: two images of each class are exhibited. (Semantic category $\sim$ Number of samples in this category)

III. Experimental Results and Discussions

A. Datasets Descriptions and Evaluation Metrics

1) Datasets Descriptions: (1) The UCM dataset [62] consists of 2100 scene images, which are divided into 21 land-use classes, including airplane, forest, freeway, overpass, etc. Each class comprises 100 RGB images with the size of 256×256

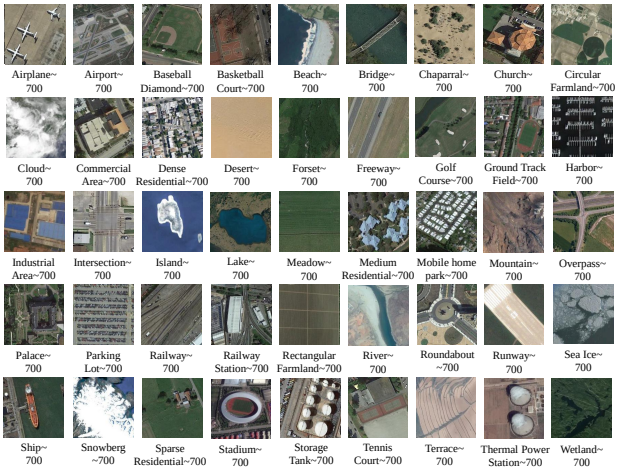


Fig. 5. Sample images of the NWPU-RESISC45 dataset: one image of each class are exhibited. (Semantic category $\sim$ Number of samples in this category)

TABLE I
Comparison of overall accuracy and standard deviations (%) of state-of-the-art methods on UCM dataset with the training ratio of 80%

Type	Method	Publication Year	Training ratio (80%)
△	BoVW(LBP) [13]	TGRS2017	77.12±1.93
	BoVW(SIFT) [13]	TGRS2017	74.12±3.30
	salM^{3}LBP-CLM [50]	JSTARS2017	95.75±0.80
	salCLM(eSIFT) [50]	JSTARS2017	94.52±0.79
□	Two-Fusion [51]	CIN2018	98.02±1.03
	CCPNet [52]	RS2018	97.52±0.97
	SCCov [31]	TNNLS2019	99.05±0.25
	ARCNet-VGG [53]	TGRS2019	99.12±0.40
	GBNet [54]	TGRS2020	98.57±0.48
	MG-CAP [55]	TIP2020	99.00±0.10
	SEMSDNet [56]	JSTARS2021	99.41±0.14
	CSDS [57]	JSTARS2021	99.52±0.13
	T-CNN [58]	TGRS2022	99.33±0.11
	DFAGCN [59]	TNNLS2022	98.48±0.42
◊	ViT-B-16 [36]	ICLR2021	99.28±0.23
	T2T-ViT-12 [60]	ICCV2021	99.10±0.30
	EMTCAL [39]	TGRS2022	99.57±0.28
	SCViT [61]	TGRS2022	99.14±0.27
Ours	LSCNet		**99.81±0.19**

△:Methods using Low-level Visual Features □:Convolution-Based Methods
◊:Vision Transformer-Based Methods

pixels and a spatial resolution of 0.3 meters per pixel. This dataset is extracted from aerial orthophotos downloaded from the United States Geological Survey (USGS), and has been extensively used for tasks such as remote sensing scene recognition and retrieval. Fig. 3 shows some samples of this benchmark dataset. (2) The AID dataset is extracted by Wuhan University from Google Earth images [13]. Fig. 4 illustrates some images of each class in this dataset. Compared to datasets with images from a single source (such as the UCM dataset), the AID dataset presents more challenges in scene recognition due to the diverse remote sensing sensors used in Google Earth imagery. The dataset consists of 10,000 RGB images across 30 classes, including field, meadow, medium residential, mountain, and more. Between 220 and 420 im-

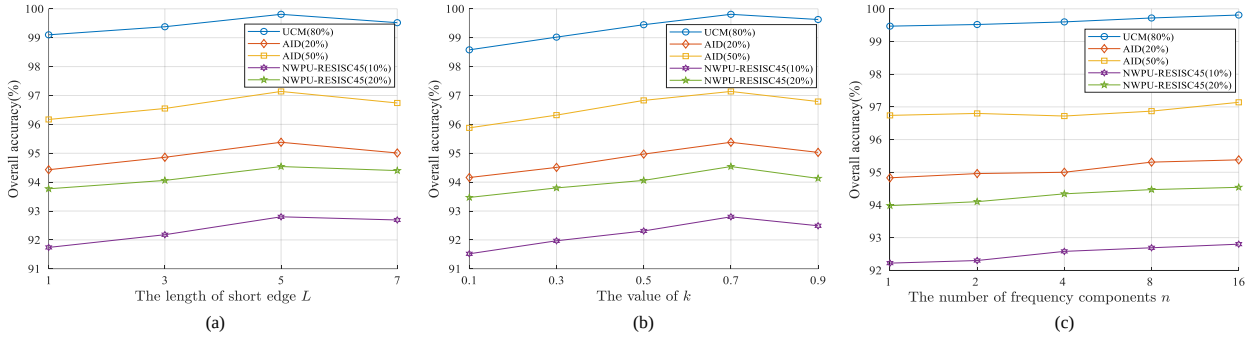


Fig. 6. Analysis of parameters contained in the model. (a) Analysis of short edge length in large kernel convolutions. (b) Analysis of dense level. (c) Analysis of the number of frequency components.

ages are included in each scene class, which is a significant variation in the amount of example images. The collection contains images with resolutions that vary from 0.5 to 8 meters, each with a fixed size of 600×600 pixels. (3) The NWPU-RESISC45 dataset is a large-scale scene recognition dataset with rich image variation and diversity created by Northwestern Polytechnical University [18]. The dataset contains 31,500 RGB images, covering 45 scene classes, including beach, church, cloud, desert, river, etc. Each class consists of 700 images with a fixed size of 256×256 pixels. The spatial resolution of the images ranges from approximately 0.2 meters to 30 meters. Compared to other datasets, the NWPU-RESISC45 dataset is characterized by its large scale in terms of the number of scenes and total images. It also exhibits significant variations in translation, spatial resolution, viewpoint, object pose, lighting, background, and occlusion. Furthermore, it possesses high intra-class diversity and inter-class similarity, making it a challenging dataset for scene recognition task. Fig. 5 exhibits some images of this challenging dataset.

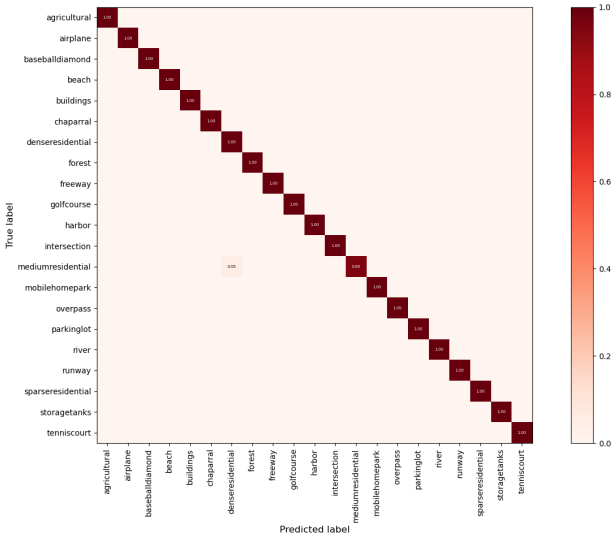


Fig. 7. Confusion matrix (CM) on UCM dataset with 80% of the dataset for training and 20% for testing.

2) Evaluation Metrics and Experimental Setting: To demonstrate the superiority of the proposed method compared to other state-of-the-art methods, a widely used quantitative analysis metric - overall accuracy (OA), is introduced. OA refers to the ratio of correctly classified samples to the total number of samples in the dataset. Additionally, to provide a more intuitive representation of the specific recognition results of the proposed method on different datasets, a confusion matrix is introduced to visualize intra-class recognition and inter-class confusion. Specifically, the columns of the confusion matrix describe the prediction of the model, and the sum of each column denotes the number of samples predicted as that class. The total of each row represents the actual number of samples for that class, whereas each row itself depicts the actual distribution of the data. Therefore, the cells on the diagonal represent the proportion of correctly recognized samples, while the other cells represent cases of misrecognition.

To demonstrate the performance of this method, a series of experiments on the three datasets were conducted to evaluate the results. All the experiments were run on a GPU of NVIDIA RTX 3070 with 8 GB RAM. For a fair comparison, the frameworks were all based on Pytorch. The Adam optimizer is used to train the network parameters. The learning rate was set to 0.0001. The epoch was set to 500. The best training model on the validation set is used to verify the test set. The average values with standard deviation were obtained from the results of 5 repeated experiments.

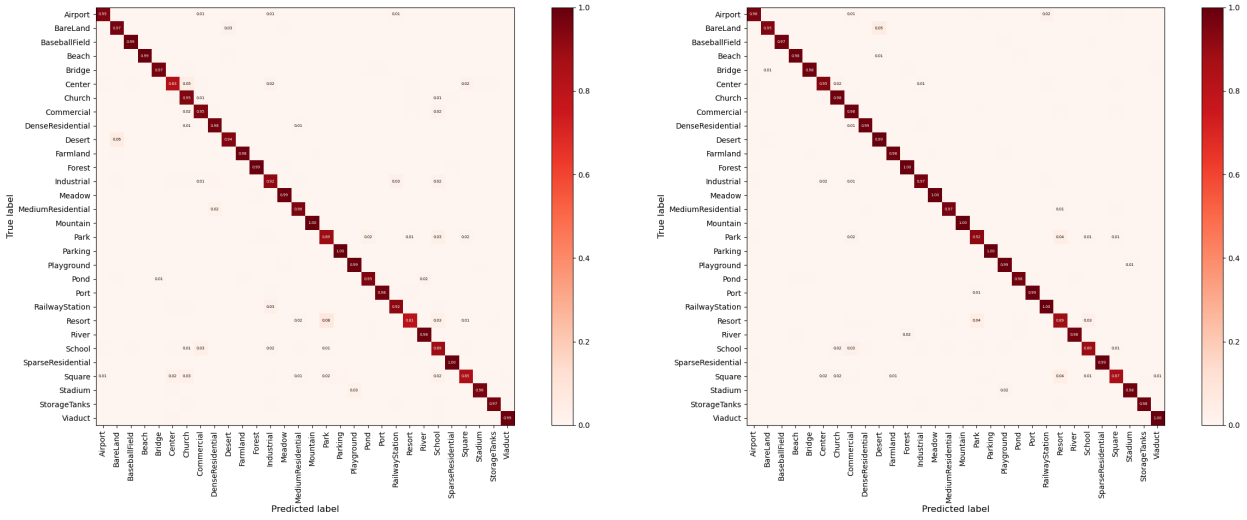
B. Parameter Analysis

1) Analysis of short edge length in large kernel convolutions: The length of the short edge L in the large kernel convolution, mentioned in Section II-A, is an important parameter that controls the receptive field and feature extraction effectiveness along the corresponding dimension. To ascertain how various quantities of L affect the outcomes of the experiment, L is set to [1, 3, 5, 7], and its performance on three datasets is reported. From Fig.6 (a), it can be observed that the overall accuracy shows improvement as the length of the short side increases from 1 to 5, and the accuracy has not improved or even decreased after that. This is because with the increase of the length of the short edge, the receptive field of the model is expanded, allowing it to capture more local features. However, once a certain scale is reached, further expansion introduces

TABLE II
COMPARISON OF OVERALL ACCURACY AND STANDARD DEVIATIONS (%) OF STATE-OF-THE-ART METHODS ON AID DATASET WITH THE TRAINING RATIO OF 20% AND 50%

Type	Method	Publication Year	Training ratios	
			20%	50%
△	BoVW(LBP)) [13]	TGRS2017	56.73±0.54	64.25±0.55
	BoVW(SIFT) [13]	TGRS2017	61.40±0.41	67.65±0.49
	salM^{3}LBP-CLM [50]	JSTARS2017	86.92±0.35	89.76±0.45
	salCLM(eSIFT) [50]	JSTARS2017	85.58±0.83	88.41±0.63
□	Two-Fusion [51]	CIN2018	92.32±0.41	94.58±0.25
	GCFs+LOFs [63]	RS2018	92.48±0.38	96.85±0.23
	SCCov [31]	TNNLS2019	93.12±0.25	96.10±0.16
	ARCNet-VGG [53]	TGRS2019	88.75±0.40	93.10±0.55
	GBNet [54]	TGRS2020	92.20±0.23	95.48±0.12
	MG-CAP [55]	TIP2020	93.34±0.18	96.12±0.12
	CSDS [57]	JSTARS2021	94.29±0.35	96.70±0.14
	PSGAN [64]	TGRS2022	89.47±0.34	92.67±0.55
	T-CNN [58]	TGRS2022	94.55±0.27	96.27±0.23
◊	ViT-B-16 [36]	ICLR2021	93.81±0.21	96.08±0.14
	T2T-ViT-12 [60]	ICCV2021	94.39±0.22	96.29±0.24
	EMTCAL [39]	TGRS2022	94.69±0.14	96.72±0.23
	SCViT [61]	TGRS2022	95.31±0.11	96.72±0.16
Ours	LSCNet		**95.38±0.15**	**97.14±0.14**

△:Methods using Low-level Visual Features □:Convolution-Based Methods ◊:Vision Transformer-Based Methods



(a) (b)

Fig. 8. Confusion matrix (CM) on AID dataset. (a) 20% of the dataset for training and the rest for testing. (b) 50% of the dataset for training and the rest for testing.

additional noise and complicates model training. Therefore, in the subsequent experimental settings, the length of the short edge in the large kernel convolution is set to 5.

2) Analysis of dense level: As mentioned in Section II-B, in the process of adaptive sparse optimization, the importance of each connection is determined based on its impact on the loss function. Therefore, it is necessary to set a dense level k, only the top k connections are retained, and the remaining connections are pruned. In this subsection, corresponding experiments are conducted to select the optimal value for k. k is set to [0.1, 0.3, 0.5, 0.7, 0.9], and its optimal value was determined based on performance on three datasets. It can be seen from Fig.6 (b) that the optimal performance is achieved when k is set to 0.7. When k is small, only a few connections are retained while a large number of connections are pruned and randomly regrown, resulting in difficulties in model optimization. On the other hand, when k is large, most connections remain unchanged, preventing the model from fully evolving into a better pattern. Therefore, in the subsequent experimental settings, the dense value k is set to 0.7.

TABLE III
COMPARISON OF OVERALL ACCURACY AND STANDARD DEVIATIONS (%) OF STATE-OF-THE-ART METHODS ON NWPU-RESISC45 DATASET WITH THE TRAINING RATIO OF 10% AND 20%

Type	Method	Publication Year	Training ratios	
			10%	20%
△	BoVW [18]	RPOC2017	41.72±0.21	44.79±0.28
	BoVW+SPM [18]	RPOC2017	27.83±0.61	32.96±0.47
	LLC [18]	RPOC2017	38.81±0.23	40.03±0.34
□	Fine-tuned VGG-16 [18]	RPOC2017	87.15±0.45	90.36±0.18
	Two-Fusion [51]	CIN2018	80.22±0.22	83.16±0.18
	CNN-CapsNet [12]	RS2019	89.03±0.21	92.60±0.11
	SCCov [31]	TNNLS2019	89.30±0.35	92.10±0.25
	MF^{2}Net [10]	GRSL2020	90.17±0.25	92.73±0.21
	MG-CAP [55]	TIP2020	90.83±0.12	92.95±0.11
	SEMSDNet [56]	JSTARS2021	91.68±0.39	93.89±0.63
	CSDS [57]	JSTARS2021	91.64±0.16	93.59±0.21
	PSGAN [64]	TGRS2022	84.72±0.72	88.47±0.56
	T-CNN [58]	TGRS2022	90.25±0.14	93.05±0.12
◊	ViT-B-16 [36]	ICLR2021	90.96±0.08	93.36±0.17
	T2T-ViT-12 [60]	ICCV2021	90.62±0.18	93.19±0.10
	EMTCAL [39]	TGRS2022	91.63±0.19	93.65±0.12
	SCViT [61]	TGRS2022	92.65±0.20	94.24±0.16
Ours	LSCNet		**92.80±0.14**	**94.54±0.19**

△:Methods using Low-level Visual Features □:Convolution-Based Methods ◊:Vision Transformer-Based Methods

3) Analysis of the number of frequency components: In the multi-frequency attention module, n represents the number of frequency components to be selected in the DCT transformation. To examine the impact of different values of n on the learning ability of the model, experiments are conducted in this section with n set to [1, 2, 4, 8, 16], and the results on three datasets are shown in Fig.6 (c). Firstly, all experiments utilizing multiple frequency components showed a significant performance gain compared to using only GAP (equivalent to $n = 1$). This validates the importance of utilizing multiple frequency components to enhance channel attention. Secondly, among the remaining choices, the experiment achieves optimal results when n is set to 16. Therefore, in the subsequent settings, n is set to 16.

C. Comparison with State-of-the-art Methods

1) Results on the UCM dataset: To validate the recognition performance of the proposed method, Table I presents a comparative evaluation of LSCNet and several other representative recognition methods on the UCM dataset with 80% of the samples as the training set and the remaining samples as the test set. The results in the table reveal that methods using low-level visual features make it difficult to achieve better experimental results due to the fixity of their feature extraction process. The convolution-based methods not only obtain the abstract features automatically but also understand the scene information with the assistance of labeling information, achieving satisfactory classification results. The vision transformer-based approaches also achieve good performance because they capture the long-range dependencies in the scene and better model the global information in the image. The proposed method expands the receptive field of the model through large kernel convolutions and strengthens the preservation of channel attention information, achieving the best recognition performance under improved sparse connections. When the training data accounts for 80%, LSCNet achieves the highest overall accuracy (OA) of 99.81% among all methods, showing a significant improvement of at least 4.06% compared to methods using low-level visual features. It also demonstrates notable performance gains compared to deep learning methods based on CNNs and ViTs, indicating that LSCNet not only inherits the powerful feature extraction capabilities of CNNs but also captures long-range dependencies comparable to ViTs.

Fig.7 shows the confusion matrix generated by LSCNet based on the recognition results when the training rate is 80%. Among the 21 scene classes, only *medium residential* is misclassified as *dense residential*, while all other classes achieve 100% recognition accuracy. This misrecognition occurs due to the high similarity between two scenes, such as the similarity in building structures and background. It becomes challenging to distinguish them accurately, leading to confusion and affecting the experimental results.

2) Results on the AID dataset: Table II lists the comparison results between LSCNet and other state-of-the-art methods on the AID dataset, where two columns of results represent using 20% and 50% samples for training, and the remaining samples as the test set. Compared to methods using low-level visual features, LSCNet leverages deep network structures to obtain richer feature representations, enhancing the representation and discriminative power of recognition features. It achieves accuracy improvements of 8.46% and 7.38% at training rates of 20% and 50%, respectively. Compared to methods based on CNNs, LSCNet achieves improvements of 0.83% and 0.44% at different training rates by utilizing larger convolutional kernels to obtain a larger receptive field, while enhancing the rationality of network connections and the effectiveness of

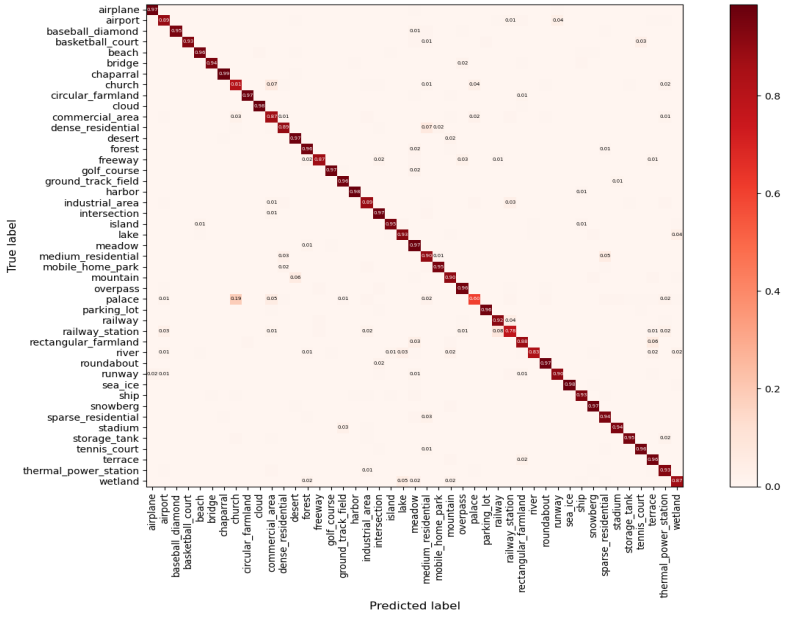


Fig. 9. Confusion matrix (CM) on NWPU-RESISC45 dataset with 10% of the dataset for training and the rest for testing

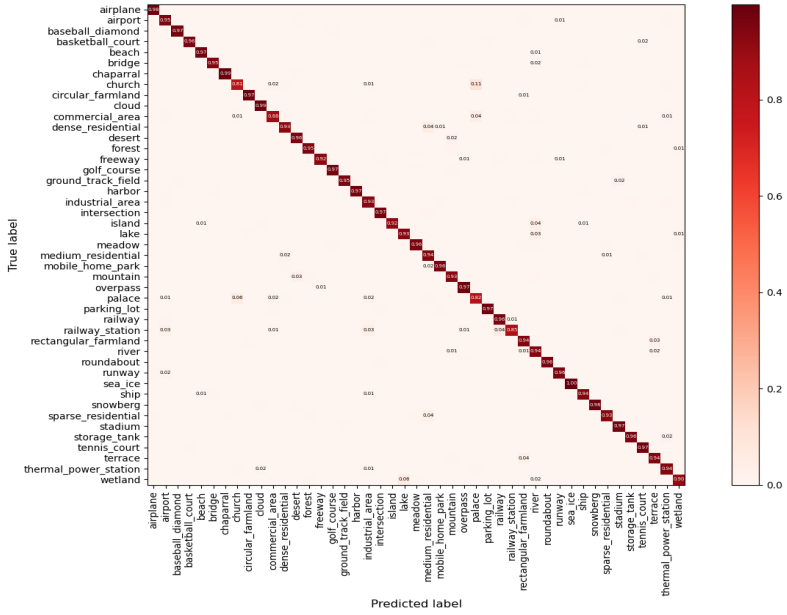


Fig. 10. Confusion matrix (CM) on NWPU-RESISC45 dataset with 20% of the dataset for training and the rest for testing.

channel attention. In comparison to methods based on ViTs, LSCNet retains the ability of CNNs to capture local details while bridging the gap in the receptive field between CNNs and ViTs, achieving comparable or even superior performance than the relevant methods.

Fig.8 shows the confusion matrices on the AID dataset under different training rates. Among the 30 classes, only five (under 20% training rate) and three (under 50% training rate) classes have recognition accuracy below 90%, while the recognition accuracy of mountain, parking and other classes has reached 100%. Classes with small inter-class differences, such as *sparse residential*, *medium residential*, and *dense residential*, are also accurately classified with accuracy exceeding 99%. However, classes like *resort*, *school*, and *square* have relatively lower recognition accuracy compared to other classes. For example, *resort* is often misclassified as *park*, as they have similar spatial distribution, including buildings and lakes. Nevertheless, LSCNet still achieves satisfactory recognition performance.

3) Results on the NWPU-RESISC45 dataset: Compared to the previous two datasets, the NWPU-RESISC45 dataset is richer in terms of scene classes and the number of images,

TABLE IV
ABLATION STUDIES FOR THE PROPOSED LSCNET ON THREE DATASETS.(SKC: SMALL KERNEL CONVOLUTION; LKC: LARGE KERNEL CONVOLUTION; ASO: ADAPTIVE SPARSE OPTIMIZATION; FSC: FIXED SPARSE CONNECTION; MAF: MULTI-FREQUENCY ATTENTION MODULE)

Variant	Convolutional module			Layer connection method		Attention module		OA (%) on different dataset				
	SKC	LKC	$K \times 1 + 1 \times K$	ASO	FSC	MAF	SENet(using GAP)	UCM (80%)	AID (20%)	AID (50%)	NWPU-RESISC45 (10%)	NWPU-RESISC45 (20%)
1		✓		✓		✓		**99.81**	**95.38**	**97.14**	**92.80**	**94.54**
2			✓	✓		✓		99.10	94.82	96.60	92.05	93.86
3	✓			✓		✓		98.74	94.38	96.29	91.53	93.47
4		✓			✓	✓		99.40	94.89	96.65	92.11	94.06
5		✓		✓			✓	99.47	94.83	96.74	92.22	93.98
6		✓			✓		✓	98.93	94.40	96.17	91.75	93.66
7	✓			✓			✓	98.61	94.10	95.96	91.47	93.13
8	✓				✓	✓		98.58	93.97	95.73	91.20	93.18

making it more challenging. Table III lists a comparison of LSCNet and existing state-of-the-art methods in terms of recognition performance on this dataset. Similarly, the experimental accuracies of the high-level visual information-based methods (convolution-based and vision transformer-based) are generally better than methods based on low-level visual features. The convolution-based and vision transformer-based methods have their own strengths and weaknesses, achieving better feature extraction in local and global features, respectively. With 10% and 20% of the samples respectively chosen as the training set, and the remaining samples are used for testing, the proposed method outperforms other comparative methods. Compared to the second-best model trained under 10% and 20% training rates, LSCNet achieved an OA improvement of 0.15% and 0.30% respectively. These results demonstrate a significant advancement over methods using low-level visual features, highlighting the exceptional recognition performance of LSCNet. Moreover, they further confirm the effectiveness of large kernel convolution, adaptive sparse optimization, and multi-frequency attention module in enhancing the experimental results.

Fig.9 and Fig.10 illustrate two confusion matrices on the NWPU-RESISC45 dataset with training rates of 10% and 20%. In both cases, more than 90% accuracy is achieved for 34 and 41 out of 45 classes, respectively. This demonstrates not only the outstanding performance but also the balanced recognition ability of the model. In the two sets of experiments, *palace* and *church* have exhibited relatively poor recognition performance compared to other classes, which can be attributed to the similarity in the distribution of images, leading to confusion by the classifier. However, for most classes, including *overpass* and *intersection*, which are similar classes, LSCNet still achieves excellent recognition results.

D. Ablation Studies

Aiming at the limitations of traditional CNNs methods on the receptive field, this paper proposes a large kernel convolution to expand receptive fields. Additionally, to overcome the issue of limited model generalization caused by fixed connections between layers, an adaptive sparse optimization strategy is designed. Furthermore, through further analysis of channel attention weights, the utilization of useful information is enhanced by employing a multi-frequency attention module. To validate their contributions to scene understanding tasks, a series of ablation experiments were conducted in this subsection to examine the effects of different modules on performance improvement. The specific experimental results are listed in Table IV. Firstly, to individually assess the superiority of each module compared to traditional methods, #2 - #5 demonstrate the recognition accuracy under different variants (#1 represents the result of the proposed method). The comparison between #1 and #2, #3 demonstrates the advantages of large kernel convolution over small kernel convolution. Here, $K \times 1 + 1 \times K$ represents the utilization of two rectangular convolution kernels with a short side length of 1. The comparison of their results further verifies the advantages of rectangular convolution kernels with a longer short side, as mentioned earlier. When the length is 1, it fails to capture local features in the other direction, leading to incomplete feature extraction. On the other hand, the comparison between #1 and #4 showcases the benefits of the adaptive sparse optimization strategy, which continuously optimizes the connection between layers through the pruning-growth process, and improves the fitting ability of the model. And the comparison between #1 and #5 illustrates the improvement in channel attention achieved by multiple frequency components. GAP, which corresponds to utilizing the lowest frequency component, neglects the useful information inherent in other components. Secondly, the comparison between #1 and #6, #7, #8 demonstrates the impact of solely using large kernel convolution, adaptive sparse optimization, and multi-frequency attention modules on experimental results. The significant improvement of #1 compared to the other three variants also confirms the complementarity among these three components. The combination of these three components can effectively enhance the understanding and learning abilities of the modal, resulting in LSCNet achieving state-of-the-art performance across all three datasets.

E. Running time and Memory Requirement

To demonstrate the execution time and the resources (memory usage) of the model, relevant statistics are performed on

the UCM dataset. The training time of the model is calculated with 32 images within a round of training, while the test time is obtained using the entire test set (420 images). The training time is 0.413s and the testing time is 0.825s. In addition, the memory requirement of the model is 10.69 MB. From the obtained results, the model runs efficiently and requires a moderate amount of memory, which is sufficient for subsequent applications.

IV. Conclusions

In this paper, a novel Large kernel Sparse ConvNet weighted by Multi-frequency Attention is proposed for remote sensing scene understanding. Firstly, to address the limited receptive field of traditional small kernel convolutions, two parallel rectangular kernels are utilized to approximate a large kernel, enabling a larger receptive field. The long and short sides of the rectangular kernel capture long-range dependencies and local detailed information, respectively. Additionally, an adaptive sparse optimization strategy is introduced to modify the fixed sparse connections in the network, allowing the model to evolve into a better recognition pattern. Lastly, instead of simple global average pooling, multiple components in the frequency domain are used to obtain more reasonable and effective channel attention weights, improving the performance of the modal. Extensive experiments on three publicly available datasets demonstrate the effectiveness and superiority of LSCNet from both quantitative and qualitative analyses, showcasing its applicability in scene understanding tasks.

References

[1] X. Yao, J. Han, G. Cheng, X. Qian, and L. Guo, "Semantic annotation of high-resolution satellite images via weakly supervised learning," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 54, no. 6, pp. 3660–3671, 2016.

[2] Y. Gao, M. Zhang, W. Li, X. Song, X. Jiang, and Y. Ma, "Adversarial complementary learning for multisource remote sensing classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 61, pp. 1–13, 2023.

[3] L. Gómez-Chova, D. Tuia, G. Moser, and G. Camps-Valls, "Multimodal classification of remote sensing images: A review and future directions," *Proceedings of the IEEE*, vol. 103, no. 9, pp. 1560–1584, 2015.

[4] J. Wang, W. Li, Y. Gao, M. Zhang, R. Tao, and Q. Du, "Hyperspectral and sar image classification via multiscale interactive fusion network," *IEEE Transactions on Neural Networks and Learning Systems*, pp. 1–15, 2022.

[5] X. Huang, H. Liu, and L. Zhang, "Spatiotemporal detection and analysis of urban villages in mega city regions of china using high-resolution remotely sensed imagery," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 53, no. 7, pp. 3639–3657, 2015.

[6] J. Wang, F. Gao, J. Dong, S. Zhang, and Q. Du, "Change detection from synthetic aperture radar images via graph-based knowledge supplement network," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 15, pp. 1823–1836, 2022.

[7] W. Li, J. Wang, Y. Gao, M. Zhang, R. Tao, and B. Zhang, "Graph-feature-enhanced selective assignment network for hyperspectral and multispectral data classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–14, 2022.

[8] Y. Gao, M. Zhang, J. Wang, and W. Li, "Cross-scale mixing attention for multisource remote sensing data fusion and classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 61, pp. 1–15, 2023.

[9] W. Li, Y. Gao, M. Zhang, R. Tao, and Q. Du, "Asymmetric feature fusion network for hyperspectral and sar image classification," *IEEE Transactions on Neural Networks and Learning Systems*, pp. 1–14, 2022.

[10] K. Xu, H. Huang, Y. Li, and G. Shi, "Multilayer feature fusion network for scene classification in remote sensing," *IEEE Geoscience and Remote Sensing Letters*, vol. 17, no. 11, pp. 1894–1898, 2020.

[11] J. Wang, W. Li, M. Zhang, R. Tao, and J. Chanussot, "Remote sensing scene classification via multi-stage self-guided separation network," *IEEE Transactions on Geoscience and Remote Sensing*, 2023.

[12] W. Zhang, P. Tang, and L. Zhao, "Remote sensing image scene classification using cnn-capsnet," *Remote Sensing*, vol. 11, no. 5, p. 494, 2019.

[13] G.-S. Xia, J. Hu, F. Hu, B. Shi, X. Bai, Y. Zhong, L. Zhang, and X. Lu, "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.

[14] J. A. dos Santos, O. A. B. Penatti, and R. da S. Torres, "Evaluating the potential of texture and color descriptors for remote sensing image retrieval and classification," in *Proceedings of the International Conference on Computer Vision Theory and Applications - Volume 2: VISAPP, (VISIGRAPP 2010)*, INSTICC. SciTePress, 2010, pp. 203–208.

[15] 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.

[16] J. Yin, H. Li, and X. Jia, "Crater detection based on gist features," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 8, no. 1, pp. 23–29, 2015.

[17] A. Oliva and A. Torralba, "Modeling the shape of the scene: A holistic representation of the spatial envelope," *International journal of computer vision*, vol. 42, pp. 145–175, 2001.

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

[19] S.-B. Chen, Q.-S. Wei, W.-Z. Wang, J. Tang, B. Luo, and Z.-Y. Wang, "Remote sensing scene classification via multi-branch local attention network," *IEEE Transactions on Image Processing*, vol. 31, pp. 99–109, 2022.

[20] Y. Ke and R. Sukthankar, "Pca-sift: a more distinctive representation for local image descriptors," in *Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2004. CVPR 2004.*, vol. 2, 2004, pp. II–II.

[21] Y. Yang and S. Newsam, "Geographic image retrieval using local invariant features," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 51, no. 2, pp. 818–832, 2013.

[22] L. Zhao, P. Tang, and L. Huo, "A 2-d wavelet decomposition-based bag-of-visual-words model for land-use scene classification," *International Journal of Remote Sensing*, vol. 35, no. 6, pp. 2296–2310, 2014.

[23] H. Abdi and L. J. Williams, "Principal component analysis," *Wiley Interdisciplinary Reviews: Computational Statistics*, vol. 2, no. 4, pp. 433–459, 2010.

[24] B. A. Olshausen and D. J. Field, "Sparse coding with an overcomplete basis set: A strategy employed by v1?" *Vision Research*, vol. 37, no. 23, pp. 3311–3325, 1997.

[25] G. E. Hinton and R. R. Salakhutdinov, "Reducing the dimensionality of data with neural networks," *Science*, vol. 313, no. 5786, pp. 504–507, 2006.

[26] X. Zhang, W. Li, C. Gao, Y. Yang, and K. Chang, "Hyperspectral pathology image classification using dimension-driven multi-path attention residual network," *Expert Systems with Applications*, p. 120615, 2023.

[27] J. Wang, F. Gao, J. Dong, and Q. Du, "Adaptive dropblock-enhanced generative adversarial networks for hyperspectral image classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 59, no. 6, pp. 5040–5053, 2021.

[28] J. Wang, M. Zhang, W. Li, and R. Tao, "A multistage information complementary fusion network based on flexible-mixup for hsi-x image classification," *IEEE Transactions on Neural Networks and Learning Systems*, 2023.

[29] J. Wang, W. Li, Y. Wang, R. Tao, and Q. Du, "Representation-enhanced status replay network for multisource remote-sensing image classification," *IEEE Transactions on Neural Networks and Learning Systems*, 2023.

[30] Y. Zhou, H. Liu, F. Ma, Z. Pan, and F. Zhang, "A sidelobe-aware small ship detection network for synthetic aperture radar imagery," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 61, pp. 1–16, 2023.

[31] 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.

[32] B. Zhang, Y. Zhang, and S. Wang, "A lightweight and discriminative model for remote sensing scene classification with multidilation pooling

module," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 12, no. 8, pp. 2636–2653, 2019.
[33] S. Chaib, H. Liu, Y. Gu, and H. Yao, "Deep feature fusion for vhr remote sensing scene classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 55, no. 8, pp. 4775–4784, 2017.
[34] K. Xu, H. Huang, P. Deng, and G. Shi, "Two-stream feature aggregation deep neural network for scene classification of remote sensing images," *Information Sciences*, vol. 539, pp. 250–268, 2020.
[35] J. Fang, Y. Yuan, X. Lu, and Y. Feng, "Robust space–frequency joint representation for remote sensing image scene classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 57, no. 10, pp. 7492–7502, 2019.
[36] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.
[37] Y. Bazi, L. Bashmal, M. M. A. Rahhal, R. A. Dayil, and N. A. Ajlan, "Vision transformers for remote sensing image classification," *Remote Sensing*, vol. 13, no. 3, p. 516, 2021.
[38] J. Ma, M. Li, X. Tang, X. Zhang, F. Liu, and L. Jiao, "Homo–heterogenous transformer learning framework for rs scene classification," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 15, pp. 2223–2239, 2022.
[39] X. Tang, M. Li, J. Ma, X. Zhang, F. Liu, and L. Jiao, "Emtcal: Efficient multiscale transformer and cross-level attention learning for remote sensing scene classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–15, 2022.
[40] X. Ding, X. Zhang, J. Han, and G. Ding, "Scaling up your kernels to 31x31: Revisiting large kernel design in cnns," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 11 963–11 975.
[41] S. Liu, T. Chen, X. Chen, X. Chen, Q. Xiao, B. Wu, M. Pechenizkiy, D. Mocanu, and Z. Wang, "More convnets in the 2020s: Scaling up kernels beyond 51x51 using sparsity," *arXiv preprint arXiv:2207.03620*, 2022.
[42] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 7132–7141.
[43] Z. Gao, J. Xie, Q. Wang, and P. Li, "Global second-order pooling convolutional networks," in *Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition*, 2019, pp. 3024–3033.
[44] Q. Wang, B. Wu, P. Zhu, P. Li, W. Zuo, and Q. Hu, "Eca-net: Efficient channel attention for deep convolutional neural networks," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 11 534–11 542.
[45] 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," *Advances in Neural Information Processing Systems*, vol. 35, pp. 1140–1156, 2022.
[46] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. Alemi, "Inception-v4, inception-resnet and the impact of residual connections on learning," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 31, no. 1, 2017.
[47] P. Baldi and P. J. Sadowski, "Understanding dropout," *Advances in neural information processing systems*, vol. 26, 2013.
[48] J. Ngiam, Z. Chen, S. Bhaskar, P. Koh, and A. Ng, "Sparse filtering," *Advances in neural information processing systems*, vol. 24, 2011.
[49] M. S. Al-Ani and F. H. Awad, "The jpeg image compression algorithm," *International Journal of Advances in Engineering & Technology*, vol. 6, no. 3, pp. 1055–1062, 2013.
[50] X. Bian, C. Chen, L. Tian, and Q. Du, "Fusing local and global features for high-resolution scene classification," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 10, no. 6, pp. 2889–2901, 2017.
[51] Y. Yu and F. Liu, "A two-stream deep fusion framework for high-resolution aerial scene classification," *Computational Intelligence and Neuroscience*, vol. 2018, 2018.
[52] K. Qi, Q. Guan, C. Yang, F. Peng, S. Shen, and H. Wu, "Concentric circle pooling in deep convolutional networks for remote sensing scene classification," *Remote Sensing*, vol. 10, no. 6, p. 934, 2018.
[53] 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, 2019.
[54] H. Sun, S. Li, X. Zheng, and X. Lu, "Remote sensing scene classification by gated bidirectional network," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 58, no. 1, pp. 82–96, 2020.
[55] S. Wang, Y. Guan, and L. Shao, "Multi-granularity canonical appearance pooling for remote sensing scene classification," *IEEE Transactions on Image Processing*, vol. 29, pp. 5396–5407, 2020.
[56] T. Tian, L. Li, W. Chen, and H. Zhou, "Semsdnet: A multiscale dense network with attention for remote sensing scene classification," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 14, pp. 5501–5514, 2021.
[57] X. Wang, L. Yuan, H. Xu, and X. Wen, "Csds: End-to-end aerial scenes classification with depthwise separable convolution and an attention mechanism," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 14, pp. 10 484–10 499, 2021.
[58] W. Wang, Y. Chen, and P. Ghamisi, "Transferring cnn with adaptive learning for remote sensing scene classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–18, 2022.
[59] K. Xu, H. Huang, P. Deng, and Y. Li, "Deep feature aggregation framework driven by graph convolutional network for scene classification in remote sensing," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 33, no. 10, pp. 5751–5765, 2022.
[60] L. Yuan, Y. Chen, T. Wang, W. Yu, Y. Shi, Z.-H. Jiang, F. E. Tay, J. Feng, and S. Yan, "Tokens-to-token vit: Training vision transformers from scratch on imagenet," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 558–567.
[61] P. Lv, W. Wu, Y. Zhong, F. Du, and L. Zhang, "Scvit: A spatial-channel feature preserving vision transformer for remote sensing image scene classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–12, 2022.
[62] Y. Yang and S. Newsam, "Bag-of-visual-words and spatial extensions for land-use classification," in *Proceedings of the 18th SIGSPATIAL International Conference on Advances in Geographic Information Systems*. Association for Computing Machinery, 2010, p. 270–279.
[63] D. Zeng, S. Chen, B. Chen, and S. Li, "Improving remote sensing scene classification by integrating global-context and local-object features," *Remote Sensing*, vol. 10, no. 5, p. 734, 2018.
[64] G. Cheng, X. Sun, K. Li, L. Guo, and J. Han, "Perturbation-seeking generative adversarial networks: A defense framework for remote sensing image scene classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–11, 2022.