Progressive Feature Fusion Framework Based on Graph Convolutional Network for Remote Sensing Scene Classification

Chongyang Zhang and Bin Wang, *Senior Member, IEEE*

***Abstract*—Remote sensing (RS) scene classification plays an important role in the intelligent interpretation of RS data. Recently, convolutional neural network (CNN)-based and attention-based methods have become the mainstream of RS scene classification with impressive results. However, existing CNN-based methods do not utilize long-range information, and existing attention-based methods do not fully exploit multiscale information, although both aspects of information are essential for a comprehensive understanding of RS scene images. To overcome the above limitations, we propose a progressive feature fusion (PFF) framework based on graph convolutional network (GCN), namely PFFGCN for RS scene classification in this article, which has a strong ability to learn both multiscale and contextual (local/long-range) information in RS scene images. It mainly consists of two modules: a multilayer feature extraction module and a multiscale contextual information fusion (MCIF) module. The MFE module is utilized to extract multilevel features and global features, and the MCIF module is constructed to capture rich contextual information from multilevel features and fuse them in a progressive manner. In MCIF, GCN is adopted to explore intrinsic attributes (including the topological structure and the contextual information) hidden in each feature map. Through the PFF strategy, the graph features at each level are fused with the next-level features to reduce the semantic gap between nonadjacent features and enhance the multiscale representation of the model. Besides, grouped GCN based on channel grouping is further proposed to improve the efficiency of PFFGCN. The proposed method is extensively evaluated on various RS scene classification datasets, and the experimental results demonstrate that the proposed method outperforms current state-of-the-art methods.**

***Index Terms*—Feature fusion, graph convolutional network (GCN), graph learning, remote sensing (RS), scene classification.**

I. Introduction

THE advancement of remote sensing (RS) imaging equipment and technologies has made it much easier to obtain high-resolution RS scene images, thus providing a solid source of data support for various applications, such as environmental monitoring [1], urban planning [2], and natural hazard detection [3]. All these applications rely on accurate RS scene classification, which aims to automatically assign a semantic label to each RS scene image according to its content and has gradually become a fundamental task [4]. Since RS scene images have significant differences from natural images in various aspects such as shooting angle, spatial complexity, and image resolution, resulting in greater intra-class differences and inter-class similarities in features extracted from RS scene images, effective implementation of feature extraction is essential for accurate classification of RS scene images [5]. To address this issue, various methods have been proposed to extract discriminative features from RS scene images by supervised learning, which can be roughly divided into two categories: handcrafted feature-based methods [6], [7], [8], [9] and deep learning-based methods [4], [10], [11], [12].

Recently, with the development of artificial intelligence, deep learning-based methods have gradually become the mainstream of RS scene classification due to their excellent feature extraction capabilities. Generally, RS scene classification methods based on deep learning can be divided into two main subcategories: convolutional neural network (CNN)-based and attention-based methods.

As the most common model in deep learning, CNNs can learn rich semantic information hidden in the RS scene images with hierarchical structure [13], thus achieving impressive performance [5], [14], [15], [16], [17]. Because RS scene images differ from natural images, especially because they contain complex and detailed spatial patterns [4], context dependencies must be considered when distinguishing between different scene categories. As depicted in Fig. 1(a), in the context of the "Bridge" scene, a large number of objects related to the scene can be identified, including "Building," "River," "Car," "Boat," and more. Due to the high inter-class similarity in RS scenes, if the model only focuses on the local regions, the "Bridge" scene may be misclassified as other scenes with similar ground objects (e.g., "Park"); if these local regions can be fully related and considered entirely, the scene can be correctly classified as "Bridge." Therefore, the ideal model is expected to consider local and long-range information together to make decisions, especially long-range information. However, due to the locality of convolution operation, current

Manuscript received 16 October 2023; revised 4 December 2023; accepted 2 January 2024. Date of publication 5 January 2024; date of current version 22 January 2024. This work was supported in part by the National Natural Science Foundation of China under Grant 62371140, and in part by the National Key Research and Development Program of China under Grant 2022YFB3903404. *(Corresponding author: Bin Wang.)*

The authors are with the Key Laboratory for Information Science of Electromagnetic Waves (MoE), Fudan University, Shanghai 200433, China, and also with the Image and Intelligence Laboratory, School of Information Science and Technology, Fudan University, Shanghai 200433, China (e-mail: 21210720041@m.fudan.edu.cn; wangbin@fudan.edu.cn).

In addition, our code is available at: https://github.com/I3ab/PFFGCN.

Digital Object Identifier 10.1109/JSTARS.2024.3350129

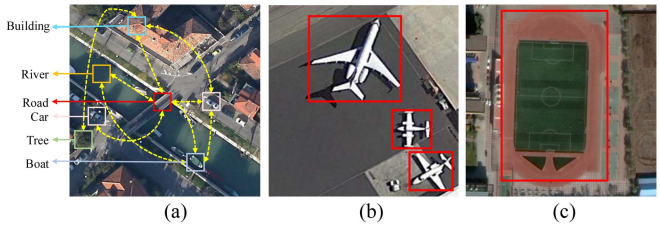


Fig. 1. Contextual (local/long-range) information and multiscale information in RS scene images. (a) Contextual information within "Bridge" category scene. (b) Different sizes of the same target object "Airplane" in a scene image. (c) A large area occupied by the target object "Playground."

CNN-based methods focus on extracting local information, but ignore the long-range relationship between local regions in RS scene images [18], [19], [20], [21].

To address the above problem, attention-based methods begin to be applied to RS scene classification [18], [22], since they have great potential to extract long-range information from RS scene images by leveraging the self-attention mechanism to learn the relationships between elements in a sequence [22]. However, this kind of methods still have some limitations: 1) They, including various vision transformers [23], [24], usually use image patches to model contextual dependencies between patches, which prevents them from extracting local information of RS scene images with complex spatial patterns and geometric structures [18]. 2) They mainly extract features at a single scale and may not have the ability of multiscale learning, which is crucial to understand RS scene images. As shown in Fig. 1(b) and (c), the target objects in different RS scene images may vary significantly in size, and even the size of the same target object in the same scene may be quite different, such as airplanes in (b), resulting in the common multiscale phenomena in RS scene images. Therefore, the ideal model is expected to possess a strong ability of multiscale information extraction and fusion to fully understand RS scene images. Besides, attention-based methods often have a high computational complexity due to the specific self-attention mechanism.

Recently, some graph-based approaches have attracted much attention, because graphs have the inherent ability to comprehensively represent data topology and geometry [25], [26], [27], [28], [29]. DFAGCN [25] introduces graph convolutional network (GCN) to reveal patch-to-patch correlations of convolutional feature maps, thus obtaining more refined features. Vision GNN (ViG) [30] is proposed as the new backbone of GCN in computer vision applications, which can be used to directly process the image data and extract image features. However, despite the good results achieved, the above GCN-related methods still have some drawbacks, such as underutilization of multiscale information from different levels, which limits performance.

To overcome the limitations of the above methods, this article proposes a progressive feature fusion (PFF) framework based on GCN, called PFFGCN, for RS scene classification. The proposed PFFGCN mainly contains a multilayer feature extraction (MFE) module and a multiscale contextual information fusion (MCIF) module. First, the MFE module is utilized to extract global features and multilevel features from RS scene images. Then, a GCN-based MCIF module with a PFF strategy is constructed and designed to fully capture the contextual information and multiscale information hidden in RS scene images. Specifically, in MCIF module, GCN is adopted to make full use of the intrinsic attribute information (including the topological structure and the contextual information) hidden in each feature map, and the PFF strategy is designed to effectively fuse the hierarchical and multiscale features in a progressive manner. By means of GCN and PFF, we can not only mine and utilize the multiscale information and contextual (local/long-range) information in RS scene images, but also significantly reduce the semantic gap between nonadjacent features, realizing that the features obtained by the MCIF module contain both rich multiscale information and contextual information. Finally, a linear classifier is simply used to achieve high-precision classification of RS scene images. Moreover, considering the high computational cost of GCN, grouped GCN block is further proposed to reduce the complexity of graph-level processing, thereby improving the overall efficiency of PFFGCN.

The main contributions of our work can be briefly summarized as follows.

1) A novel feature fusion framework, named PFFGCN, is designed for RS scene classification. This framework mainly consists of a replaceable MFE module and an MCIF module, achieving high-accuracy classification of RS scene images. Besides, the grouped GCN is further proposed to improve the efficiency of the proposed PFFGCN.
2) The MCIF module is constructed to fully fuse both multiscale information and contextual information in a progressive manner by means of the GCN blocks, significantly reducing the semantic gap between nonadjacent hierarchical features and enhancing the multiscale representation capability of the model.
3) Extensive experimental results conducted on various benchmark datasets demonstrate that the proposed PFFGCN can achieve a new state-of-the-art (SOTA) performance for RS scene classification.

The rest of this article is organized as follows: Section II briefly reviews the related works and recent progresses on RS scene classification and GCN. In Section III, the PFFGCN method is presented and described in detail. In Section IV, the extensive experiments are conducted to evaluate the performance of the proposed PFFGCN along with the ablation study and visualization. Finally, the conclusions are given in Section V.

II. Related Works

A. RS Scene Classification

The purpose of RS scene classification is to classify RS scene images into different semantic groups according to their contents. Initially, the handcrafted feature-based methods such as SIFT [7], [31], HOG [8], [32] and bag-of-visual-words (BoVW) [9], [33] are employed for RS scene classification. However, with the progress of artificial intelligence, deep learning-based methods have overtaken handcrafted feature-based methods. Recently, CNN-based and attention-based methods have become

the norm in computer vision, and RS scene classification has also embraced these methods.

1) CNN-Based Methods: Since the AlexNet [34] is proposed, CNNs have evolved rapidly in the last ten years. Some milestone works, including VGGNet [11], ResNet [10], SENet [10], and Res2Net [35], greatly facilitate the task of image classification. Based on these baseline models, accuracy of RS scene classification has been greatly improved as illustrated in [14], [15], [16], [36], and [37]. To solve the problem that CNNs pay more attention to local information while ignoring global information, RLFCNN [16] combines global and rearranged local features to realize more comprehensive representation. Different from the traditional CNNs that minimize only the cross-entropy loss, discriminative CNNs (D-CNNs) [34] apply a new discriminative objective function to optimize the training process and explicitly impose a metric learning regularization on the CNN features. SCCov [14] embeds novel skip connections and covariance pooling into the traditional CNNs to achieve a more representative feature learning.

Moreover, since RS scene images usually have information with various scales, many researchers have studied how to effectively extract the multiscale information hidden in RS scene images [5], [17], [38]. SKAL [5] utilizes a global-local two-stream architecture to produce a multiscale representation, which involves extracting global and local features from the whole image and the most significant area, respectively. MF^{2}CNet [17] designs a multiscale feature fusion covariance network with octave convolution to get multifrequency and multiscale features from RS scene images. However, although CNN-based methods achieve impressive results in RS scene classification, their ability to capture long-range information and correlations between objects in RS scene images is limited by convolution operators, and this constraint may result in a suboptimal performance in classification tasks.

2) Attention-Based Methods: With the popularity of different vision transformers like ViT [23] and Swin transformer [24], some attention-based methods have been proposed for RS scene classification to address the aforementioned challenges faced by CNNs [18], [39], [40], [41].

The existing attention-based methods for RS scene classification can be categorized into two major classes. The first class includes various variants of ViT. ViTRSIC [39] explores the impact of standard vision transformers architecture in RS scene classification. SCViT [40] considers both the detailed geometric information of the RS scene images and the contribution of the different channels contained in the classification token. The second class involves methods that integrate attention mechanisms on top of the CNN backbone, such as MBLANet [41] and EMTCAL [18]. MBLANet [41] combines a convolutional local attention module with deep residual network (ResNet-50), which can automatically perceive the key parts of the image, suppress secondary features, and extract key information in the feature map. EMTCAL [18] effectively combines the advantages of CNNs and transformers and develops an efficient multiscale transformer to explore the intrinsic contextual knowledge in RS scene images. However, although attention-based methods can address the limitations of CNNs and effectively capture long-range information [8], many of them do not possess strong multiscale learning capability [18]. Besides, the self-attention mechanism also results in high computational costs, making these methods less practical for RS scene classification.

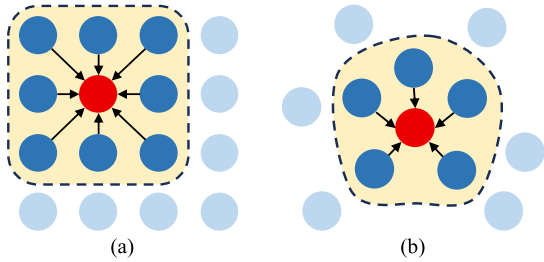


Fig. 2. Comparison of different convolutions. (a) CNN convolution. (b) Graph convolution.

B. GCNs

Recently, GCNs have become one of the most popular models because of the strong ability of contextual learning. They are usually used for point cloud classification, scene graph generation, and action recognition [42] in the field of computer vision [30]. To overcome the limitations of CNNs and transformers, GCNs begin to be applied in the RS image processing and analysis [25], [26], [27], [28], [29]. For the hyperspectral image (HSI) classification, DAGCN [26] is designed by focusing on the problems stemming from the increasing resolution of HSIs. MvRLNet [27] proposes a multiview graph learning module (MGLM) to integrate topology and spectral graph information into a unified network, capturing the latent discriminant feature response in various situations. CGE-AL [28] proposes a new class-wise graph-embedding-based active learning framework implemented by a class-wise GCN, achieving outstanding performance on the HSI classification. To improve the feature representation capability in RS scene classification, CNN-GCN [29] introduces a novel two-stream architecture that combines global-based visual features obtained by CNN and object-based location features obtained by GCN.

As shown in Fig. 2, the traditional convolution networks usually use a fixed-size kernel around the central pixel to extract features, while the graph convolution can achieve feature abstraction function in a more flexible manner by aggregating features of its neighborhood [26], which means GCNs have flexible receptive fields and have great potential in computer vision tasks. In general, graph convolution operation F can be formulated as aggregation and update operations [43] as follows:

$$\begin{aligned}\mathcal{G}' &= F(\mathcal{G},\ \mathcal{W}) \\ &= \mathrm{Update}(\mathrm{Aggregate}(\mathcal{G},\ W_{\mathrm{agg}}),\ W_{\mathrm{update}})\end{aligned} \tag{1}$$

where $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ and $\mathcal{G}' = (\mathcal{V}', \mathcal{E}')$ are the input graph and output graph, respectively, and W_{agg} and W_{update} denote the learnable parameters of the aggregation and update operations, respectively.

In GCNs, aggregation functions are utilized to extract useful information from the neighborhood of nodes, while update

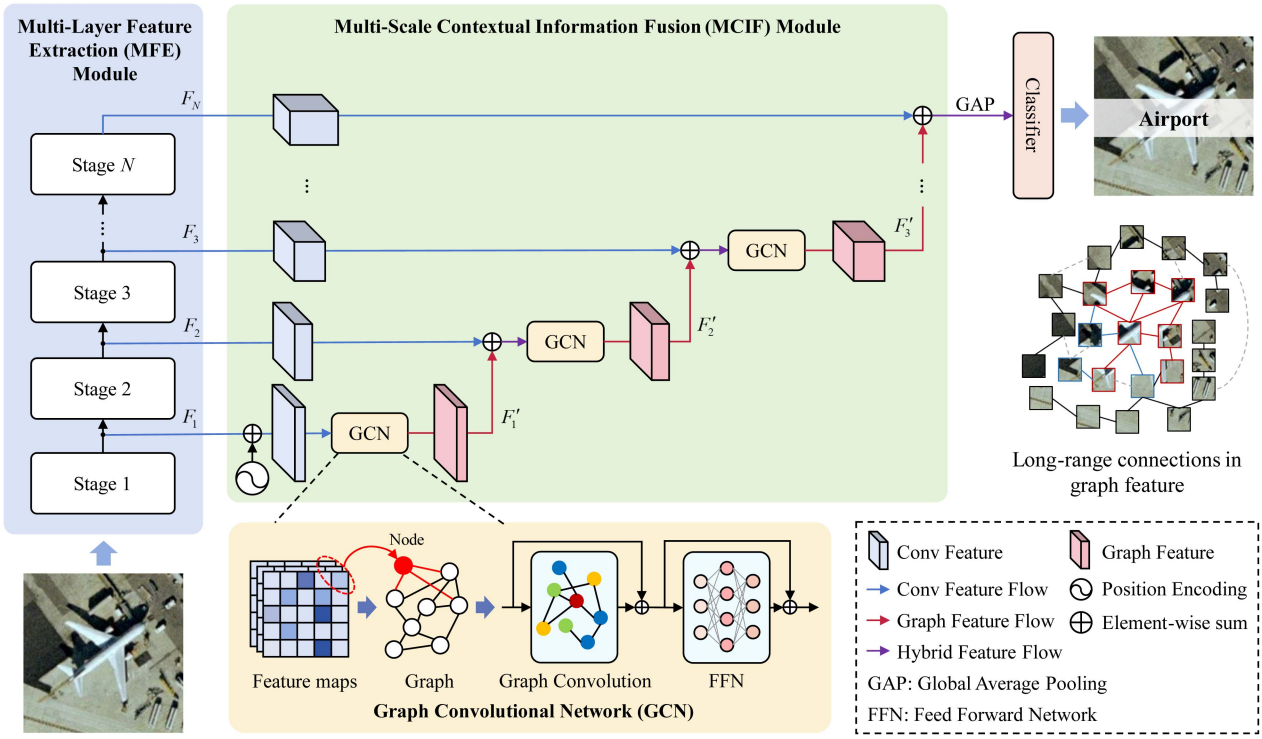


Fig. 3. Overall framework of the proposed PFFGCN, which mainly consists of an MFE module and an MCIF module.

functions perform nonlinear transforms on the aggregated information to compute new node representations. More specifically, for a node x_i, its representation x'_i is computed as follows:

$$x'_i = \phi(x_i, \rho(x_i, \mathcal{N}(x_i), W_{\text{agg}}), W_{\text{update}}) \tag{2}$$

where ρ is a node feature aggregation function and ϕ is a node feature update function, and $\mathcal{N}(x_i)$ is a set of neighbor nodes of x_i. For the sake of simplicity and efficiency, max-relative graph convolution is proposed [30], [44]

$$\rho(.) = x_i^{(\text{agg})} = \max(\{x_i - x_j | x_j \in \mathcal{N}(x_i)\}), \tag{3}$$

$$\phi(.) = x'_i = x_i^{(\text{agg})} W_{\text{update}} \tag{4}$$

where the bias term is omitted.

III. Proposed Method

Since RS scene images contain rich multiscale information and contextual (local/long-range) information, making full use of these two aspects of information can effectively improve the accuracy of RS scene classification. In view of this, we propose PFFGCN to fully exploit the discriminative information, which mainly contains an MFE module and an MCIF module, as illustrated in Fig. 3. First, the input RS scene image with the size of $H \times W \times 3$ is sent to the MFE module to obtain multilevel features. Each pixel in feature maps of a scene is viewed as a node, and an adjacency graph can be constructed by searching k-nearest neighbors. Then, the MCIF module is employed to capture the multiscale information and contextual information. Graph-level processing (conducted by GCN) on the graphs obtained from each feature map can further exploit long-range relationship between local regions in RS scene images. By progressively fusing graph features with next-level features in the MCIF module, not only can the multiscale information and contextual information in the RS scene images be fully utilized, but also the semantic gap between nonadjacent features can be reduced. Finally, a linear classifier is applied on the hybrid features obtained by the MCIF module to predict the results of query scene images. Besides, considering the high computational cost of GCN, we further propose grouped GCN to alleviate this issue.

A. MFE Module

The MFE module is employed to extract multilevel features and global vision features, consisting of multiple stages, each of which processes the output of the previous stage. Typically, the output features of each stage have different spatial scales. In the past, pretrained CNNs are usually deployed as MFE module. For example, DFAGCN [25] took a pretrained VGGNet-16 as the MFE module, while EMTCAL [18] used a pretrained ResNet-34.

As a general framework, various backbone networks can be deployed as the MFE module in PFFGCN. Considering the simplicity of method description and the feature extraction capability of the network, we chose ResNet-50 as an illustrative example of the MFE module to elaborate on the proposed PFFGCN in this section. ResNet-50 is a successful CNN model composed of four stages, with each stage containing multiple residual blocks. Different residual blocks can capture multilevel

convolutional features. In the shallow layers, convolutional features usually contain low-level details and local information, such as texture, color, and shape. As the network delves deeper, semantic information becomes increasingly enriched. Due to the complex contents in RS scene images, features from different levels can help construct more comprehensive features according to both content and semantic information, so as to facilitate the classification of RS scene images. For the convenience of description, we denote the feature obtained from the stage i as $F_i \in \mathbb{R}^{H_i \times W_i \times C_i} (i = 1, 2, \ldots, n)$. In ResNet-50, n equals 4 and the global vision feature $F_4 \in \mathbb{R}^{H_4 \times W_4 \times C_4}$ is generated by the last stage of MFE module. For simplicity, we also denote the global vision feature as G.

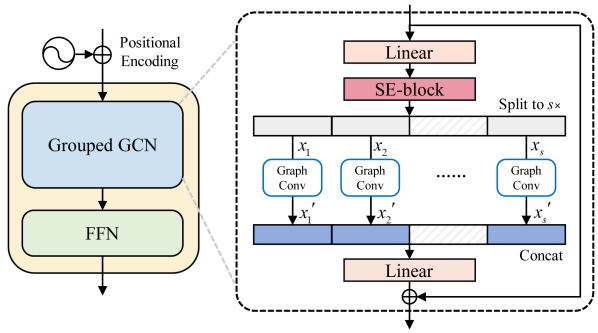


Fig. 4. Illustration of the proposed grouped GCN block.

B. MCIF Module

The MCIF module is proposed and designed to fully fuse the multiscale and contextual (local/long-range) information in RS scene images. In the MCIF module, GCN is utilized to capture the rich long-range information from the multilevel features, and the PFF strategy is designed to effectively fuse the multilevel features with various scales in a progressive manner.

1) Graph-Level Processing of Feature Maps:

a) Construction of graph: Multilevel features encompass rich local information, yet they lack long-range information that is crucial for RS scene images. To fully understand RS scene images, we employ GCN to assist in enhancing contextual information in features at different levels. We conduct graph-level processing on each of the feature maps instead of the original image because each pixel in the feature maps represents a local region and has rich local information. Taking feature map $F_1 \in \mathbb{R}^{H_1 \times W_1 \times C_1}$ as an example, each pixel can be seen as a feature vector $x_i \in \mathbb{R}^{C_1} (i = 1, 2, \ldots, H_1 W_1)$. Then, we get a set of features $X = [x_1, x_2, \ldots, x_N] \ \ (N = H_1 W_1)$. These image features can be viewed as a set of unordered nodes, which can be denoted as $\mathcal{V} = \{v_1, v_2, \ldots, v_N\}$. A graph $\mathcal{G}$ can be represented by a tuple $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where $\mathcal{E}$ is the set of all the edges. It is worth noting that the size of feature maps determines the number of nodes in each graph. To retain positional information, the learnable position encodings [23] are added to these unordered nodes.

b) Graph-level processing by GCN: Most GCNs have a fixed graph structure and only update the node features at each iteration. However, the recent work [44] points out that using dynamic graph convolution to dynamically change graph structure and neighbor nodes at each layer allows the network to obtain better graph representations, which effectively alleviates the oversmoothing problem and generates a larger receptive field. For these reasons, we recalculate edges between nodes via a K-NN function in the feature space of each GCN block to further increase the receptive field. Specifically, for each node v_i, find its k-nearest neighbors $\mathcal{N}_k(v_i)$ using the K-NN function and add an edge e_{ji} directed from v_j to v_i for all $v_j \in \mathcal{N}_k(v_i)$. Then the set of edges $\mathcal{E}$ and graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ can be updated.

As in (1)–(4), graph convolution utilize aggregation and update operations to consistently refresh the information in the graph. This processing can be denoted as $X' = \mathrm{GraphConv}(X)$. To increase the diversity of features, linear layers are applied before and after the graph convolution, which can project node features into the same domain. Besides, a nonlinear activation function is also applied after the graph convolution to avoid layer collapse. In practical application, the processing of the graph can be represented as follows:

$$X_G = \sigma(\mathrm{GraphConv}(XW_{G1}))W_{G2} + X \tag{5}$$

where $\sigma(.)$ is the nonlinear activation function, X_G represents the feature after graph-level processing, W_{G1} and W_{G2} are the weights of the first and second linear layers, respectively, and a residual connection is also utilized here to avoid vanishing gradients.

To further improve the feature transformation capability and alleviate the oversmoothing phenomenon commonly seen in GCNs, a feed forward network (FFN) consisting of a two-layer multilayer perceptron (MLP) and a residual connection is applied after graph convolution

$$\mathrm{FFN}(X_G) = \sigma(X_G W_{F1})W_{F2} + X_G \tag{6}$$

where W_{F1} and W_{F2} are the weights of the first and second linear layers, respectively.

According to the above operations and transformations, we can acquire graph-based representations of each feature map that contains rich long-range information. It is worth noting that batch normalization (BN) is applied after every linear layer or graph convolution, but it is omitted from (5) and (6) for simplicity. Furthermore, the whole graph-level processing described above can be simplified as follows:

$$\begin{aligned} Y &= \mathrm{GCN}(X) \\ &= \mathrm{FFN}(\sigma(\mathrm{GraphConv}(XW_{G1}))W_{G2} + X). \end{aligned} \tag{7}$$

Equation (7) is shown as a GCN block in Fig. 3. During experiments, it can be observed that employing GCN blocks does not introduce too many parameters, but it will cause extra computational burden, as described in Subsection IV-E.

c) Grouped GCN: To reduce the computational complexity, we further propose the grouped GCN to improve the model's efficiency. As depicted in Fig. 4, after the input feature map $X \in \mathbb{R}^{H \times W \times C}$ is processed by the first linear layer, it is split into s feature subsets along the channel dimension, denoted as $x_1, x_2, \ldots, x_s$. The number of channels of each feature subset

is 1/s of the input feature, but the spatial size is consistent with the input feature. Each feature subset x_i has a corresponding graph convolution $\mathrm{GraphConv}_i(\cdot)$. Accordingly, the number of input channels of each graph convolution also becomes 1/s of the original. All these subsets are updated in parallel and then concatenated for subsequent processing

$$x'_i = \mathrm{GraphConv}_i(x_i), i = 1, 2, \ldots, s \tag{8}$$

$$X_G = \sigma([x'_1, x'_2, x'_3, \ldots, x'_s])W_{G2} + X \tag{9}$$

where [.] represents the feature concatenation operation.

After channel division, each graph convolution possesses a reduced number of parameters and computations, which helps to diminish the model's complexity to a certain extent. Moreover, the grouping update operation allows the model to update information in parallel across multiple feature subspaces, thereby enhancing efficiency. After the grouped GCN, FFN is also utilized for feature transformation. For simplicity, the above procedure is expressed as $Y = \mathrm{GroupedGCN}(X)$. Notice that an SE-block [12] is added after the first linear layer of the grouped GCN block. This operation can yield different weights to different channels before channel splitting, enabling the model to pay more attention to useful information. By replacing the GCN blocks in PFFGCN with the grouped GCN blocks, the computation cost of graph-level processing can be significantly reduced, improving efficiency.

2) Multiscale Contextual Information Fusion:

a) PFF strategy: Different levels of features contain information of different spatial scales, and effectively fusing these features can enhance the utilization of multiscale information in RS scene images. A common strategy for multilevel feature fusion is direct feature fusion (DFF), in which features of different scales are directly combined, as employed in [25]. In DFF, features from different levels are first resized to match the size of specific features, typically the top or bottom feature. Subsequently, feature fusion operations such as concatenation or addition are performed. However, this approach suffers from the loss or degradation of feature information, which impairs the fusion effect of nonadjacent levels [45]. The semantic gap between nonadjacent hierarchical features is larger than the semantic gap between adjacent hierarchical features, especially for the bottom and top features. Therefore, directly fusing non adjacent features from different levels fails to fully leverage the multiscale information hidden in RS scene images. Moreover, the sizes vary widely between nonadjacent features, which means that a considerable number of additional parameters need to be introduced for the resize operation.

To better utilize the multiscale information in multilevel features, we propose the PFF strategy to perform feature fusion on selected features, and the comparison between DFF and PFF is shown in Fig. 5. Specifically, low-level features are first fused with the features of its next level, and the fused features are then progressively fused with higher level features until the last layer. The fusion process can be mathematically written as follows:

$$\begin{cases} F'_i = \mathrm{FF}(F_i, \mathrm{Downsampling}(F_{i-1})) \\ F'_{i+1} = \mathrm{FF}(F_{i+1}, \mathrm{Downsampling}(F'_i)) \end{cases} \tag{10}$$

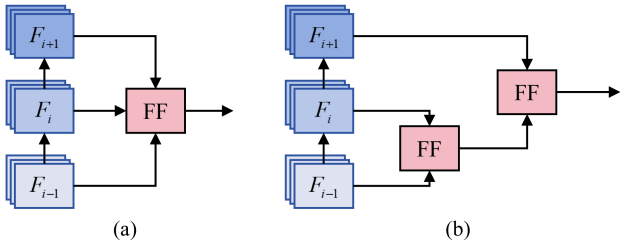


Fig. 5. Comparison of two different feature fusion strategies. Among them, F_i represents the feature obtained by the ith stage in the MFE module, "FF" denotes the feature fusion operation (e.g., concatenation or addition). (a) DFF strategy. (b) PFF strategy.

where $\mathrm{FF}(.)$ denotes the feature fusion operation, and F'_i represents the fused feature. Downsampling not only serves the purpose of resize, but also allows the convolution parameters to weight the features according to their importance.

By doing so, multilevel features with different scales can be progressively fused, closing the semantic gap between nonadjacent features. Moreover, according to the combinational explosion effect [35], the multiplexing of lower level features can yield larger receptive fields and comprehensively extract the connections between the features at different levels, enhancing the multiscale representation capability of the model.

b) Feature fusion with GCN: The GCN blocks are employed in the MCIF module to exploit the intrinsic attributes including topology and contextual information. Specifically, the MCIF module initially processes low-level features at the graph level, acquires graph features, and then fuses them with next-level convolutional features. Thus, a hybrid feature that contains graph and convolutional characteristics is obtained, as shown in Fig. 3. This process iterates progressively as described in PFF strategy. The feature paths of different levels can be represented as follows:

$$\begin{cases} F'_1 = \mathrm{GCN}(F_1) \\ F'_2 = \mathrm{GCN}(F_2 + \mathrm{Downsampling}(F'_1)) \\ F'_3 = \mathrm{GCN}(F_3 + \mathrm{Downsampling}(F'_2)) \end{cases} \tag{11}$$

where element-wise addition is used as the feature fusion operation.

Using the obtained informative graph feature F'_3, we then combine it with global vision feature G to fully understand RS scene images

$$G' = G + \mathrm{Downsampling}(F'_3). \tag{12}$$

By doing so, the long-range information contained in the graph features and the local information with various scales contained in the convolutional features can be fully matched, and the PFF between adjacent levels will mitigate the impact of semantic gaps. Consequently, the obtained hybrid feature G' contains both rich multiscale information and contextual (local/long-range) information.

In practical implementation, only 2-times downsampling is required because only adjacent features are fused. We achieve this 2-times downsampling using a 2×2 convolution with a stride

Fig. 6. Samples from UCM dataset with true labels.

of 2. Meanwhile, for each feature path, dropout is employed during training to enhance generalization.

C. Classification

For the obtained G', we first use the global average pooling operation to reduce its spatial size before classification. Next, a linear layer is used to calculate scores for different categories. Finally, the softmax function is utilized to classify RS scene images. In addition, the cross-entropy loss function is adopted to optimize the network.

IV. Experimental Results and Analysis

In this section, we evaluate the classification performance of the proposed PFFGCN on three public high-resolution RS scene datasets. First, a brief introduction to all datasets, implementation details and experimental settings are provided. Then, extensive experiments are conducted and the experimental results are reported by comparing the proposed PFFGCN with other SOTA methods on each dataset. Moreover, the results of ablation study and visualization are also reported in this section. Finally, we give an in-depth discussion on the characteristics of the proposed method.

A. Datasets and Evaluation Metrics

Three public datasets are employed to evaluate the RS scene classification performance of the proposed PFFGCN, including UC Merced Land Use dataset (UCM dataset) [9], Aerial Image dataset (AID dataset) [46], and NWPU-RESISC45 dataset (NWPU dataset) [4].

1) UCM Dataset: This dataset is collected by the Computer Vision Lab of University of California, Merced. The UCM dataset contains a total of 2100 RS scene images divided into 21 scene categories on average. Each image in this dataset consists of 256×256 pixels, and each pixel has a spatial resolution of 0.3 *m* in the RGB color space. Some examples of this benchmark dataset are shown in Fig. 6.

2) AID Dataset: This dataset is released by Wuhan University. The AID dataset contains a total of 10 000 RS scene images with a size of 600×600, and the spatial resolution varies from 0.5 to 8 *m*. The AID dataset includes 30 scene categories, with

Fig. 7. Samples from AID dataset with true labels.

Fig. 8. Samples from NWPU dataset with true labels.

the number of images in each category varying from 220 to 420. Some examples of this benchmark dataset are shown in Fig. 7.

3) NWPU Dataset: Constructed by Northwestern Polytechnical University, the NWPU dataset contains a total of 31 500 RS scene images divided into 45 scene categories on average. Each image has a size of 256×256 and spatial resolution ranging from 0.2 to 30 *m* per pixel. Some examples of this benchmark dataset are shown in Fig. 8.

For a fair comparison, the training ratios of the UCM, AID, and NWPU datasets are set to 50% and 80%, 20% and 50%, and 10% and 20%, respectively, consistent with the previous approaches [4], [9], [46], [49], [50], [51], [52], [53], [54]. Overall accuracy (OA) and confusion matrix (CM) are used to assess the classification accuracy. OA reflects the overall accuracy of a classification model, which is defined as the percentage of correctly classified images in total test images. CM is a table that accumulates the number of correctly classified and misclassified

TABLE I
ACCURACY COMPARISON BETWEEN BASELINE AND THE PROPOSED METHOD ON THREE DATASETS

Methods	UCM		AID		NWPU	
	Tr = 50%	Tr = 80%	Tr = 20%	Tr = 50%	Tr = 10%	Tr = 20%
Fine-tuned ResNet-50	98.13 ± 0.20	99.33 ± 0.39	93.74 ± 0.13	96.13 ± 0.13	91.38 ± 0.18	93.65 ± 0.23
PFFGCN (ResNet-50)	**99.04 ± 0.19**	**99.67 ± 0.21**	**95.88 ± 0.23**	**97.40 ± 0.14**	**92.91 ± 0.15**	**94.89 ± 0.12**
Fine-tuned ViG-S	98.86 ± 0.17	99.48 ± 0.23	93.13 ± 0.31	96.27 ± 0.15	91.59 ± 0.15	93.58 ± 0.29
PFFGCN (ViG-S)	**99.10 ± 0.33**	**99.76 ± 0.26**	**96.18 ± 0.13**	**97.64 ± 0.25**	**93.34 ± 0.17**	**95.22 ± 0.08**

In the case of different MFE modules, the best values are highlighted in bold.

images for each scene class and reflects them as a percentage. Besides, the total number of parameters and the floating-point operations (FLOPs) are used to evaluate the efficiency of the model.

B. Implementation Details and Experimental Setup

In our experiments, two common but different types of visual backbone networks, CNN-based ResNet-50 [10] and GCN-based ViG-S [30], are employed as the MFE module of PFFGCN to verify the generalization capability of the entire framework. They are initialized by pretrained parameters (using ImageNet dataset [34]). For RS scene classification, existing works have demonstrated that using a model pretrained on a large-scale dataset, such as ImageNet has better performance than training from scratch [47]. The rest of our model is initialized randomly.

For all GCNs, ℓ_2 distance is used to measure the distance between nodes in the feature space, and GELU [48] is used as a nonlinear activation function. For the implementation of K-NN, an MLP with BN and ReLU is adopted to complete the update function in (4). When using ResNet-50 as the MFE module, we only add the learnable positional encoding to each node before the first GCN block, because each level of information, including positional information, is passed backward progressively and does not need to be added repeatedly. However, when using ViG-S as the MFE module, we do not add positional encoding because ViG-S itself already contains it.

All experiments are implemented using PyTorch framework and executed on the Ubuntu 18.04 operating system. To speed up the training process, we utilize a GeForce RTX 3090 with 24G memory. Choose the Adam algorithm as the optimizer to train our model 50 epochs and set the batch size to 16. The learning rate is initialized to 0.00003, and a cosine decay learning rate scheduler with a linear warm-up is adopted. To meet the size requirements, the images in all three datasets are resized to 224×224. In addition, horizontal flips, vertical flips, and random rotations are used for data argumentation. For all networks, the dropout rate is set to 0.3.

To obtain reliable experimental results, we repeat all the experiments five times by randomly selecting training and test samples. Finally, the average classification results and standard deviations of these five runs are reported. In the report, to distinguish PFFGCN using the GCN block and the grouped GCN block, we refer to the models using the two different GCN blocks for graph-level processing as PFFGCN-v1 and PFFGCN-v2, respectively. In the following analysis of this section, PFFGCN defaults to PFFGCN-v1.

TABLE II
EFFICIENCY COMPARISON WITH CLASSICAL METHODS

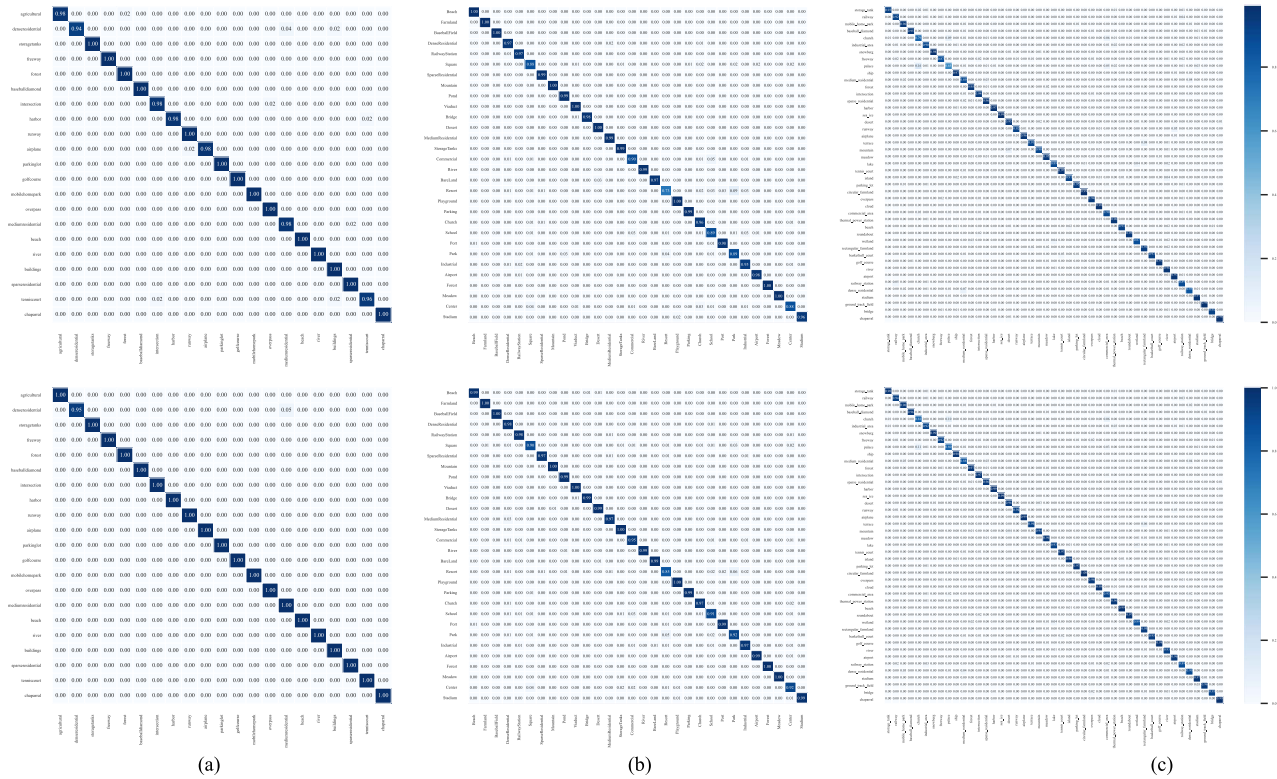
Methods	Publication	Parameters	FLOPs
VGG-16 [11]	ICLR2015	134.38M	15.47G
ResNet-50 [10]	CVPR2016	23.57M	4.13G
ViT-Base [23]	ICLR2021	86.42M	16.86G
Swin-Base [24]	ICCV2021	87.70M	15.17G
PFFGCN-v1 (ResNet-50)	Ours	40.28M	10.93G
PFFGCN-v2 (ResNet-50)	Ours	39.28M	8.16G

C. Performance of the PFFGCN Method

From Table I, it can be found that the OA of PFFGCN (ResNet-50) has been improved by 0.91% and 0.34%, 2.14% and 1.27%, and 1.53% and 1.24% compared to the baseline method (ResNet-50) on UCM, AID, and NWPU datasets, respectively. Compared with ViG-S, the OA improvement obtained is 0.24% and 0.28%, 3.05% and 1.37%, and 1.75% and 1.64% on these three datasets, respectively. The above results show that PFFGCN can be effective for various MFE modules. Moreover, on almost every dataset, the less training data used, the more significant the improvement, suggesting that our method also helps to enhance generalization.

Besides the impressive OA results, the corresponding CMs further confirm the proposed PFFGCN's superior performance. Taking PFFGCN based on ResNet-50 as an example, the CMs on the three datasets are shown in Fig. 9(a)–(c). Fig. 9(a) shows that our PFFGCN achieves amazing accuracy ($\geq$98%) in most scene classes in UCM dataset, many of which even reach 100% such as "Storage Tanks," "Freeway," and "River." From Fig. 9(b), the proposed PFFGCN yields impressive results (95%) in most of the scene categories in AID dataset. As shown in Fig. 9(c), although the NWPU dataset is more challenging, our PFFGCN still achieves competitive results ($\geq$90%) on most classes. Besides, the categories of "Church" and "Palace" in NWPU dataset are easily confused with each other, as they both have similar monolithic objects, such as magnificent roofs and greenery, which are similar in texture structure and color characteristics. For both categories, PFFGCN achieves acceptable results ($\geq$65%) at both training ratios of 10% and 20%.

In terms of the model's efficiency, taking ResNet-50 based PFFGCN as an example, the comparison between our PFFGCN and other classical CNN and transformer methods is shown in Table II. The total parameters and computation complexity of PFFGCN are smaller than those of VGG-16, ViT-Base, and Swin-Base, but larger than those of ResNet-50. Compared with PFFGCN-v1, PFFGCN-v2 achieves a FLOPs improvement of



(a) (b) (c)

Fig. 9. Confusion matrices of the proposed PFFGCN based on ResNet-50 on three public datasets with different training ratios. (a) UCM dataset, Tr = 50% (top) and Tr = 80% (bottom). (b) AID dataset, Tr = 20% (top) and Tr = 50% (bottom). (c) NWPU dataset, Tr = 10% (top) and Tr = 20% (bottom).

2.77G. This indicates that PFFGCN has a moderate amount of storage and computation, posing no significant computational burden.

D. Comparison With SOTA Methods

To comprehensively evaluate the classification performance of the proposed PFFGCN in RS Scene classification and verify the great potential of GCNs in the field of RS, we compare our method with existing SOTA methods. We divide these methods into four categories: handcrafted feature-based methods, CNN-based methods, attention-based methods, and GCN-based methods, and all results of the compared methods are shown in Table III. Among them, DFAGCN [25] is a representative GCN-based method used for comparison, which applies GCN to further complete feature aggregation and achieve excellent results. All results for the comparison methods were from the primary literature, but some methods have not been tested on certain datasets, so we omit these results in our report.

1) Experiments on UCM dataset: The amount of data in the UCM dataset is small, and many methods have reached saturation results on this dataset. It is clear from Table III that both the CNN-based methods and the attention-based methods consistently achieve impressive accuracy, and the proposed PFFGCN also achieves excellent accuracy on the UCM dataset. When the training ratio is 50%, compared with the best CNN-based method (i.e., MF^{2}CNet) and the best attention-based method (i.e., SCViT), the OA improvement obtained by PFFGCN with ResNet-50 is 0.28% and 0.14%, respectively. When the training ratio is 80%, the OA result of PFFGCN (ViG-S) is comparable to the best outcome (i.e., MGSNet), which is 1.28% higher than DFAGCN.

2) Experiments on AID dataset: As can be seen from Table III, PFFGCN achieves the best accuracy among all these methods. When the training ratio is 20%, PFFGCN using ViG reaches 96.18% and the OA improvement is 0.62% compared with the suboptimal method (i.e., SCViT). Compared with the best attention-based method (i.e., SCViT) and another GCN-based method (i.e., DFAGCN), the OA improvement obtained by PFFGCN (ViG-S) is 0.61% and 2.76%, respectively, when the training ratio equals 50%. And PFFGCN with ResNet-50 is also better than the vast majority of the previous methods. Experimental results show that our PFFGCN can understand RS scenes more comprehensively and achieve excellent results.

3) Experiments on NWPU dataset: Compared to the UCM and AID datasets, the NWPU dataset is more challenging and has more data. The high inter-class similarity and intra-class diversity of NWPU dataset easily results in misclassification. According to Table III, the proposed PFFGCN outperforms SOTA methods and achieves the best performance. At the training ratio of 10% and 20%, the OA of PFFGCN (ResNet-50) reaches 92.91% and 94.89%, respectively. On basis of ViG-S, PFFGCN achieves higher accuracy of 93.34% and 95.22%, respectively.

Moreover, both ResNet-50 and ViG-S based PFFGCN-v2 can also achieve competitive results on all three datasets. These

TABLE III
COMPARISON OF OA AND STANDARD DEVIATIONS (%) OF STATE-OF-THE-ART METHODS ON THREE PUBLIC DATASETS

Type	Methods	Publication	UCM		AID		NWPU	
			Tr = 50%	Tr = 80%	Tr = 20%	Tr = 50%	Tr = 10%	Tr = 20%
Handcrafted Feature-Based	BoVW (SIFT) [46]	TGRS2017	71.90 ± 0.79	74.12 ± 3.30	61.40 ± 0.41	67.65 ± 0.49	–	–
	BoVW (LBP) [46]	TGRS2017	73.13 ± 1.50	77.12 ± 1.93	56.73 ± 0.54	64.25 ± 0.55	–	–
	salCLM (eSIFT) [6]	JSTARS2017	91.88 ± 1.06	93.62 ± 0.85	85.58 ± 0.83	88.41 ± 0.63	–	–
	salM^{3}LBP-CLM [6]	JSTARS2017	94.21 ± 0.75	95.75 ± 0.80	86.92 ± 0.35	89.76 ± 0.45	–	–
	BoVW [4]	PROC2017	–	–	–	–	41.72 ± 0.21	44.97 ± 0.28
	BoVW + SPM [4]	PROC2017	–	–	–	–	27.83 ± 0.61	32.96 ± 0.47
	LLC [4]	PROC2017	–	–	–	–	38.81 ± 0.23	40.03 ± 0.34
CNN-Based	CaffeNet [46]	TGRS2017	93.98 ± 0.67	95.02 ± 0.81	86.86 ± 0.47	89.53 ± 0.31	–	–
	VGG-VD-16 [46]	TGRS2017	94.14 ± 0.69	95.21 ± 1.20	86.59 ± 0.29	89.64 ± 0.36	–	–
	D-CNN [36]	TGRS2018	–	98.93 ± 0.10	90.82 ± 0.16	96.89 ± 0.10	89.22 ± 0.50	91.89 ± 0.22
	VGG-16-CapsNet [49]	RS2019	95.33 ± 0.18	98.81 ± 0.22	91.63 ± 0.19	94.74 ± 0.17	85.08 ± 0.13	89.18 ± 0.14
	SCCov [14]	TNNLS2020	–	99.05 ± 0.25	93.12 ± 0.25	96.10 ± 0.16	89.30 ± 0.35	92.10 ± 0.25
	LSENet [50]	TIP2021	98.53 ± 0.37	99.78 ± 0.18	94.41 ± 0.16	96.36 ± 0.19	92.23 ± 0.14	93.34 ± 0.15
	CSDS [51]	JSTARS2021	98.48 ± 0.21	99.52 ± 0.13	94.29 ± 0.35	96.70 ± 0.14	91.64 ± 0.16	93.59 ± 0.21
	MF^{2}CNet [17]	TGRS2022	98.76 ± 0.18	99.52 ± 0.25	95.54 ± 0.17	97.02 ± 0.28	92.07 ± 0.22	93.85 ± 0.27
	T-CNN [52]	TGRS2022	–	99.33 ± 0.11	94.55 ± 0.27	96.27 ± 0.23	90.25 ± 0.14	93.05 ± 0.12
	HDTFF-Net [53]	JSTARS2023	–	–	–	97.46 ± 0.16	–	94.47 ± 0.26
	MGSNet [54]	TGRS2023	–	**99.76 ± 0.14**	95.46 ± 0.21	97.18 ± 0.16	92.40 ± 0.16	94.57 ± 0.12
Attention-Based	ViT-B 32 [23]	ICLR2021	97.83 ± 0.16	98.95 ± 0.24	93.74 ± 0.27	95.84 ± 0.29	90.05 ± 0.29	92.61 ± 0.14
	Swin-T [24]	ICCV2021	–	99.46 ± 0.11	94.56 ± 0.14	96.92 ± 0.12	90.84 ± 0.09	93.18 ± 0.15
	T2T-ViT-12 [55]	ICCV2021	97.79 ± 0.20	99.10 ± 0.32	94.39 ± 0.22	96.29 ± 0.24	90.62 ± 0.18	93.19 ± 0.10
	PVT-V2-B0 [56]	CVM2022	97.94 ± 0.44	98.86 ± 0.38	93.52 ± 0.35	96.27 ± 0.14	89.72 ± 0.16	92.95 ± 0.09
	EMTCAL [18]	TGRS2022	98.67 ± 0.16	99.57 ± 0.28	94.69 ± 0.14	96.41 ± 0.23	91.63 ± 0.19	93.65 ± 0.12
	SCViT [40]	TGRS2022	98.90 ± 0.19	99.57 ± 0.31	95.56 ± 0.17	96.98 ± 0.16	92.72 ± 0.04	94.66 ± 0.10
GCN-Based	DFAGCN [25]	TNNLS2022	–	98.48 ± 0.42	–	94.88 ± 0.27	–	89.29 ± 0.28
	PFFGCN-v1 (ResNet-50)	Ours	**99.04 ± 0.19**	99.67 ± 0.21	**95.88 ± 0.23**	97.40 ± 0.14	**92.91 ± 0.15**	**94.89 ± 0.12**
	PFFGCN-v2 (ResNet-50)	Ours	98.61 ± 0.54	99.33 ± 0.57	**95.80 ± 0.20**	97.16 ± 0.26	**92.93 ± 0.37**	**94.88 ± 0.14**
	PFFGCN-v1 (ViG-S)	Ours	**99.10 ± 0.33**	**99.76 ± 0.26**	**96.18 ± 0.13**	**97.64 ± 0.25**	**93.34 ± 0.17**	**95.22 ± 0.08**
	PFFGCN-v2 (ViG-S)	Ours	98.91 ± 0.14	99.62 ± 0.12	**96.05 ± 0.20**	97.39 ± 0.09	**93.11 ± 0.33**	**95.10 ± 0.10**

The SOTA values are marked in bold.

positive results show that the proposed method can make full use of the multiscale and contextual information in RS scene images.

E. Ablation Study

To analyze the influence of different hyperparameters in GCN and the contribution of different components in our PFFGCN, we conduct experiments on AID and NWPU datasets with different training ratios in this section based on ResNet-50.

1) Analysis of GCN: The number of neighbor nodes k plays an important role in GCN, which can directly control the aggregation range when constructing graph and updating information. Too few neighbors will degrade information exchange, while too many neighbors will result in oversmoothing and extra computation. To explore the impact of k, we tune k from 3 to 15 for all GCNs in PFFGCN. Moreover, we also study the impact of channel grouping on the model's efficiency. The experimental results are summarized in Tables IV and V, where the total numbers of parameters and the FLOPs of entire PFFGCN and a single GCN block (corresponding to $F_2 \in \mathbb{R}^{28\times28\times512}$) are listed.

When use ResNet-50 as the MFE module, the corresponding PFFGCN has three GCN blocks at different levels. In order to explore the parameter sensitivity of different levels of GCN, we set three different parameter strategies for k: uniform (applying identical parameter setting for different levels of GCN), increasing (where k gradually becomes larger as the level going deeper), and decreasing (where k gradually decreases as the level going deeper). From Table IV, setting identical k for all GCNs can achieve the best results in most cases on both the AID and NWPU datasets, and also is convenient to adjust for the practical applications. Therefore, in our implementation, the same parameter settings are used for different levels of GCN to facilitate ease of use.

It can be found from Table IV that the performance of whole PFFGCN including the accuracy and efficiency varies with the number of neighbors of GCN. When k increases from 3 to 6, the overall accuracy of the model improves. However, when k is greater than 6, the accuracy does not always improve, but instead causes extra computation, which indicates that a larger aggregation range is not always better and selecting an appropriate k value in GCN is important for RS scene classification. As can be seen from Table IV, the best trade-off between overall accuracy and efficiency can be achieved when setting k to 6 for all GCNs, and therefore, this setting is chosen for all experiments.

When $k = 6$, we conduct experiments using three different numbers of feature subsets, i.e., $s = 2$, 4, and 8, respectively.

TABLE IV
EXPERIMENTAL RESULTS OF DIFFERENT PARAMETER SETTINGS

Strategy	k			Entire PRMs	Entire FLOPs	AID		NWPU	
						Tr = 20%	Tr = 50%	Tr = 10%	Tr = 20%
Unified	3	3	3	40.28M	9.08G	95.67 ± 0.15	97.28 ± 0.19	92.68 ± 0.16	94.74 ± 0.17
	6	6	6	40.28M	10.93G	**95.88 ± 0.23**	**97.40 ± 0.14**	92.91 ± 0.15	**94.89 ± 0.12**
	9	9	9	40.28M	12.78G	95.55 ± 0.23	97.33 ± 0.17	92.90 ± 0.14	94.71 ± 0.18
	12	12	12	40.28M	14.63G	95.88 ± 0.26	97.26 ± 0.07	**92.98 ± 0.19**	94.72 ± 0.17
	15	15	15	40.28M	16.48G	95.61 ± 0.21	97.24 ± 0.14	92.86 ± 0.36	94.77 ± 0.11
Increase	3	6	9	40.28M	10.93G	95.65 ± 0.25	**97.40 ± 0.11**	92.89 ± 0.12	94.67 ± 0.20
	6	9	12	40.28M	12.78G	95.72 ± 0.18	97.26 ± 0.12	92.87 ± 0.12	94.82 ± 0.33
	9	12	15	40.28M	14.63G	95.44 ± 0.14	97.26 ± 0.31	92.91 ± 0.26	94.67 ± 0.09
	3	9	15	40.28M	12.78G	95.43 ± 0.37	97.16 ± 0.43	92.85 ± 0.14	94.64 ± 0.12
Decrease	9	6	3	40.28M	10.93G	95.59 ± 0.31	97.25 ± 0.24	92.92 ± 0.10	94.75 ± 0.21
	12	9	6	40.28M	12.78G	95.65 ± 0.20	97.22 ± 0.17	92.90 ± 0.27	94.83 ± 0.15
	15	12	9	40.28M	14.63G	95.71 ± 0.14	97.15 ± 0.29	92.88 ± 0.21	94.84 ± 0.17
	15	9	3	40.28M	12.78G	95.67 ± 0.31	97.24 ± 0.14	92.71 ± 0.31	94.72 ± 0.04

Entire PRMs denotes the number of total parameters of the entire PFFGCN.
The best values are marked in bold.

TABLE V
EXPERIMENTAL RESULTS OF DIFFERENT GCN BLOCKS

Operation	k	s	Block PRMs	Block FLOPs	Entire PRMs	Entire FLOPs	AID		NWPU	
							Tr = 20%	Tr = 50%	Tr = 10%	Tr = 20%
GCN	6	1	1.09M	1.85G	40.28M	10.93G	95.88 ± 0.23	97.40 ± 0.14	92.91 ± 0.15	94.89 ± 0.12
GGCN	6	2	0.95M	1.24G	39.59M	9.08G	95.72 ± 0.24	97.14 ± 0.08	**93.02 ± 0.12**	**94.91 ± 0.10**
	6	4	0.89M	0.93G	39.25M	8.16G	**95.80 ± 0.20**	**97.16 ± 0.26**	92.93 ± 0.37	94.88 ± 0.14
	6	8	0.86M	0.77G	39.07M	7.69G	95.58 ± 0.29	96.97 ± 0.16	92.82 ± 0.18	94.69 ± 0.14

Block PRMs denotes the number of total parameters of a single GCN block.
The best values are marked in bold.

From Table V, as s increases, the total number of parameters and FLOPs of grouped GCN block decrease. When $s = 4$, the FLOPs are only half of what they are when $s = 1$. The classification accuracy is similar when $s = 2$ and $s = 4$. When $s = 4$, the model exhibits an overall better performance on the AID dataset. When $s = 2$, the model performs better on the NWPU dataset and even achieves an astonishing 93.02% OA when the training ratio is 10%. However, when s increases to 8, the performance starts to significantly decline. This is because the reduction in parameters results in a decrease in generalization. Due to the trade-off between the model's complexity and accuracy, we choose $s = 4$ as the default parameter of PFFGCN-v2 and report results in Subsection IV-D. Moreover, as shown in Tables II and III, replacing the GCN blocks with the grouped GCN blocks ($s = 4$) reduces the FLOPs of the entire model by 25%, and the classification accuracy is only slightly reduced.

2) Ablation of Each Component: We design ablation experiments to analyze the contributions of different components in PFFGCN. As mentioned in Section III, our PFFGCN mainly consists of an MFE module and an MCIF module. The MCIF module is mainly composed of GCN blocks and PFF strategy. The ablation experiments are conducted to evaluate the effect of each component in PFFGCN. Also, to compare the PFF strategy with the DFF strategy and the corresponding models are constructed. Totally, the following five models are constructed for comparison:

1) Model 1 (Fine-tuned ResNet): MFE.
2) Model 2 (Fine-tuned ResNet with DFF): MFE + DFF.
3) Model 3 (Fine-tuned ResNet with PFF): MFE + PFF.
4) Model 4 (PFFGCN with DFF): MFE + DFF + GCN.
5) Model 5 (PFFGCN): MFE + PFF + GCN.

The results of ablation experiments are shown in Table VI. It can be observed that Model 1 has the lowest average OA, meaning that DFF, PFF, and GCN should all be valid, but a more detailed comparison is needed.

a) Evaluation of PFF: By comparing the experimental results of the Model 1, Model 2, and Model 3, it can be found that the average OAs of the Model 2 and Model 3 are both higher than the Model 1, which indicates multilevel feature fusion can contribute positively to RS scene classification. Meanwhile, the performance of the Model 3 is significantly superior to that of the Model 2. This indicates that PFF can better integrate multilevel features than DFF, effectively alleviating semantic gap between different levels. Moreover, PFF has fewer parameters and less computation than DFF. For the AID dataset under training ratio 20% and the NWPU dataset under 10% training ratio, the OA improvement obtained by PFF is significant, reaching 1.67% and 1.28%, respectively. The results above demonstrate that features from different levels are valuable for enhancing the model's understanding of RS scene images. It is also proved that PFF is an effective and efficient multilevel feature fusion method, which can narrow the semantic gap and facilitate the construction of comprehensive semantic representation.

b) Evaluation of GCN: Comparisons between Model 2 and Model 4, as well as between Model 3 and Model 5, all reveal that incorporating GCN can further improve accuracy

TABLE VI
RESULTS OF ABLATION EXPERIMENTS FOR DFF, PFF, AND GCN

Model	PRMs	FLOPs	MFE	MCIF			AID		NWPU	
				DFF	PFF	GCN	Tr = 20%	Tr = 50%	Tr = 10%	Tr = 20%
1	23.57M	4.13G	✓				93.74 ± 0.13	96.13 ± 0.13	91.38 ± 0.18	93.65 ± 0.23
2	82.30M	7.01G	✓	✓			94.84 ± 0.17	96.80 ± 0.17	92.35 ± 0.15	94.37 ± 0.14
3	34.59M	5.37G	✓		✓		95.41 ± 0.10	97.06 ± 0.25	92.59 ± 0.07	94.62 ± 0.14
4	87.99M	12.57G	✓	✓		✓	95.10 ± 0.13	97.05 ± 0.19	92.48 ± 0.19	94.60 ± 0.20
5	40.28M	10.93G	✓		✓	✓	**95.88 ± 0.23**	**97.40 ± 0.14**	**92.91 ± 0.15**	**94.89 ± 0.12**

The best values are marked in bold.

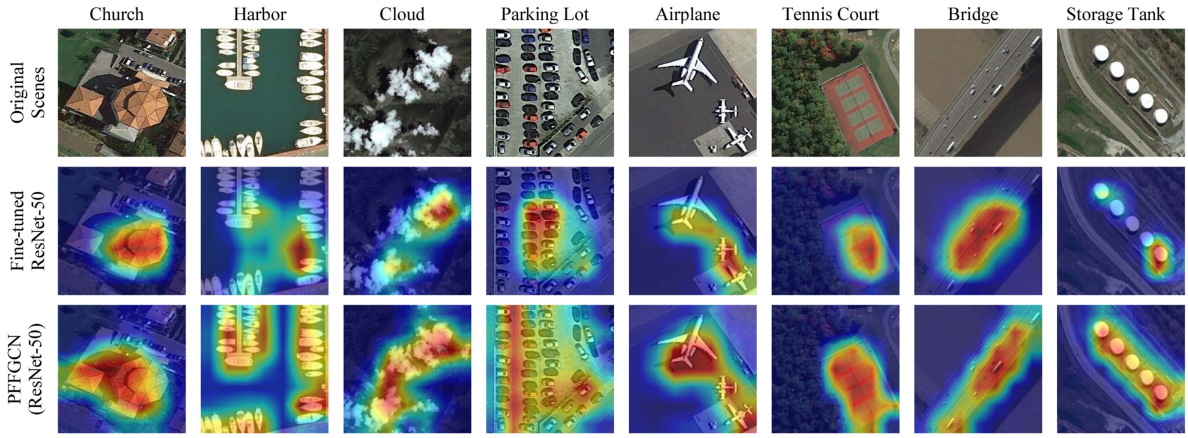


Fig. 10. Visualization using Grad-CAM on NWPU data set. Original scenes with labels are given in the first row. The CAMs generated by fine-tuned ResNet-50 and PFFGCN (ResNet-50) are shown in the second row and the third row, respectively.

for RS scene classification. Taking Model 5 as an example, for the AID dataset, the OA improvement obtained by GCN is 0.43% and 0.36% when the training ratios equals 20% and 50%, respectively. For the NWPU dataset, the OA improvement is 0.32% and 0.27%. By combining PFF strategy and GCNs, the model can not only learn a more comprehensive multiscale representation but also further explore the long-range relationship between different local regions in RS scene images. Therefore, GCN is significant for PFFGCN since it can further extract long-range information from RS scene images. In addition, from the analysis of GCN, it can also be seen that the number of neighbor nodes k in GCN can be adjusted to make the model have different performances, and appropriate s can be selected according to different task requirements.

Based on the above analysis, we leverage the unique strengths of PFF strategy and GCN as a whole MCIF module to develop an integrated solution for RS scene classification, yielding impressive results.

F. Visualization

Regions that the model pays attention to on RS scene images can be analyzed intuitively by heat maps. To intuitively analyze the multiscale ability and the understanding of different categories of proposed PFFGCN, we use Grad-CAM [57] to carry out class activation mapping (CAM) visualization. Grad-CAM is a popular visualization method, which utilizes gradients to compute the importance of spatial locations and makes it easy for us to understand how a model learns an image. We choose eight RS scene images with variable objects from the NWPU dataset to realize CAM and the results are shown in Fig. 10. Finetuned ResNet-50 and proposed PFFGCN are chosen for comparison. The regions display in red mean that these regions are of interest to the model. To further validate the multiscale ability of PFFGCN, visualization using Grad-CAM on objects of different sizes within the same category is shown in Fig. 11.

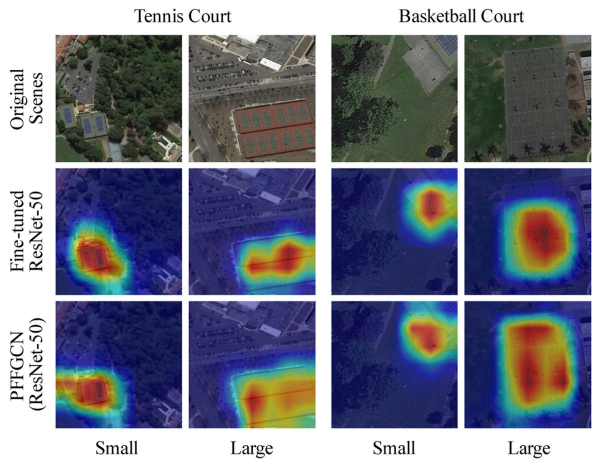


Fig. 11. Visualization using Grad-CAM on objects of different sizes in the RS scene images of categories "Tennis Court" and "Basketball Court."

As shown in Fig. 10, although the CAM results based on baseline (ResNet-50) are correct, the coverage of some objects is not comprehensive enough, such as "Church," "Harbor," and "Cloud," which indicates the deficiency in multiscale capability.

Due to the powerful multiscale ability, the activation maps of PFFGCN tend to cover the whole object in different scenes like "Airplane," "Church," and "Tennis Court." Fig. 11 also shows the PFFGCN based CAM results have more concentrated and comprehensive activation maps for small and large objects, which further demonstrates the multiscale capability of the proposed method.

In addition, for RS scene images with dispersed objects, finetuned ResNet-50 can only focus on local regions, such as "Harbor," "Parking Lot," and "Storage Tank." This indicates that the baseline (ResNet-50) lacks the ability to represent long-range information effectively. Benefiting from GCN, PFFGCN can effectively grasp a broader range of contextual information in these scenes and distinguish between different categories.

G. Discussions

According to the above experimental results, we can summarize the advantages and disadvantages of the proposed method as follows.

1) *Strong representation capability of multiscale and contextual information:* Benefiting from GCN and the PFF strategy in the MCIF module, the proposed PFFGCN significantly improves the classification accuracy of the baseline models and achieves new SOTA results on all three datasets. In addition to the excellent accuracy, the CAMs vividly demonstrate that our PFFGCN can accurately and comprehensively understand information at various scales and the intrinsic attributes in RS scene images.
2) *Acceptable amount of storage and computation:* The total parameters and FLOPs of PFFGCN are all at moderate and acceptable levels compared with other classical methods, suggesting that PFFGCN can serve as a very useful model in practice.
   It should be noted that the introduction of graph-level processing makes the model more complex and affects its operational efficiency to some extent. Although we have proposed the grouped GCN to alleviate this problem, the efficiency improvement is still limited. The high computational cost of GCN itself may still pose challenges for applications that require real-time processing or computing resource-constrained devices (such as user terminals). In the future work, we may use knowledge distillation and implement lightweight GCNs to make the model more efficient.
3) *Easy adjustment of the hyperparameters:* In our PFFGCN, only a few hyperparameters need to be tuned. Among them, tuning k can control the aggregation range of GCN to adapt the characteristics of different datasets, while adjusting s can reduce the model's computational complexity, thus enhancing the model's efficiency. Our PFFGCN used the same hyperparameter setting on all three datasets, and achieved good experimental results, which also shows that our method is not sensitive to the setting of these hyperparameters within a certain range.
4) *End-to-end training and replaceable MFE module:* It is easy to see that the training process of the proposed PFFGCN is end-to-end, and the MFE module can be replaced with different general visual backbone networks, which will be convenient and beneficial for the practical and widespread application of our method.

Besides, cross-dataset training plays a crucial role in enhancing the model's adaptability to new data and tasks, thereby improving its generalization capabilities and overall performance. However, in our current work, our primary focus is on achieving high classification accuracy for RS scene classification tasks. In the future work, we plan to conduct a more comprehensive study on the interaction between various backbone networks and the proposed PFFGCN and investigate the domain adaptability performance of the model on different datasets, so as to further improve the cross-domain ability and generalization of the proposed model.

V. Conclusion

This article has presented a novel GCN-based framework (named PFFGCN) for RS scene classification, which mainly consists of the MFE module and the MCIF module. The MFE module is employed to extract multilevel and local/global features. Using both PFF and GCN in the MCIF module, the proposed PFFGCN exhibits a strong representation capability of multiscale and contextual (local/long-range) information in RS scene images. More importantly, the proposed framework can be applied to different backbone networks and exhibits excellent generalization. The experimental results on three widely used datasets have shown that the proposed PFFGCN can achieve an excellent performance for the task of RS scene classification.

Despite promising results achieved by the proposed method, there is still room for further improvement. In the future, our goal is to further improve the efficiency and cross-domain capabilities of the model, as well as the generalization performance, so as to make it more practical and applicable.

References

[1] X. Huang, D. Wen, J. Li, and R. Qin, "multilevel monitoring of subtle urban changes for the megacities of China using high-resolution multi-view satellite imagery," *Remote Sens. Environ.*, vol. 196, pp. 56–75, Jul. 2017.
[2] Z. Y. Lv, W. Shi, X. Zhang, and J. A. Benediktsson, "Landslide inventory mapping from bitemporal high-resolution remote sensing images using change detection and multiscale segmentation," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 11, no. 5, pp. 1520–1532, May 2018.
[3] N. Longbotham, C. Chaapel, L. Bleiler, C. Padwick, W. J. Emery, and F. Pacifici, "Very high resolution multiangle urban classification analysis," *IEEE Trans. Geosci. Remote Sens.*, vol. 50, no. 4, pp. 1155–1170, Apr. 2012.
[4] 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.
[5] Q. Wang, W. Huang, Z. Xiong, and X. Li, "Looking closer at the scene: Multiscale representation learning for remote sensing image scene classification," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 33, no. 4, pp. 1414–1428, Apr. 2022.
[6] 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.
[7] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. J. Comput. Vis.*, vol. 60, no. 2, pp. 91–110, 2004.
[8] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, 2005, vol. 1, pp. 886–893.

[9] 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.

[10] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 770–778.

[11] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. Int. Conf. Learn. Representations*, 2015, pp. 1–13.

[12] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 7132–7141.

[13] Z. Shao, L. Wang, Z. Wang, W. Du, and W. Wu, "Saliency-aware convolution neural network for ship detection in surveillance video," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 30, no. 3, pp. 781–794, Mar. 2020.

[14] N. He, L. Fang, S. Li, J. Plaza, and A. Plaza, "Skip-connected covariance network for remote sensing scene classification," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 31, no. 5, pp. 1461–1474, May 2020.

[15] C. Peng, Y. Li, L. Jiao, and R. Shang, "Efficient convolutional neural architecture search for remote sensing image scene classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 7, pp. 6092–6105, Jul. 2021.

[16] Y. Yuan, J. Fang, X. Lu, and Y. Feng, "Remote sensing image scene classification using rearranged local features," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 3, pp. 1779–1792, Mar. 2019.

[17] L. Bai, Q. Liu, C. Li, Z. Ye, M. Hui, and X. Jia, "Remote sensing image scene classification using multiscale feature fusion covariance network with octave convolution," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5620214.

[18] 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 Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5626915.

[19] S. Mei, J. Ji, Y. Geng, Z. Zhang, X. Li, and Q. Du, "Unsupervised spatial-spectral feature learning by 3D convolutional autoencoder for hyperspectral classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 9, pp. 6808–6820, Sep. 2019.

[20] S. Mei, X. Li, X. Liu, H. Cai, and Q. Du, "Hyperspectral image classification using attention-based bidirectional long short-term memory network," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2021, Art. no. 5509612.

[21] X. Tang, X. Zhang, F. Liu, and L. Jiao, "Unsupervised deep feature learning for remote sensing image retrieval," *Remote Sens.*, vol. 10, no. 8, 2018, Art. no. 1243.

[22] Y. Zhou, F. Wang, J. Zhao, R. Yao, S. Chen, and H. Ma, "Spatial-temporal based multihead self-attention for remote sensing image change detection," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 10, pp. 6615–6626, Oct. 2022.

[23] A. Dosovitskiy, L. Beyer, and A. Kolesnikov, "An image is worth 16x16 words: Transformers for image recognition at scale," in *Proc. Int. Conf. Learn. Representations*, 2021.

[24] Z. Liu, Y. Lin, and Y. Cao, "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 9992–10002.

[25] 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 Trans. Neural Netw. Learn. Syst.*, vol. 33, no. 10, pp. 5751–5765, Oct. 2021.

[26] J. Bai, B. Ding, Z. Xiao, L. Jiao, H. Chen, and A. C. Regan, "Hyperspectral image classification based on deep attention graph convolutional network," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5504316.

[27] X. Cheng et al., "Multi-view graph convolutional network with spectral component decompose for remote sensing images classification," *IEEE Trans. Circuits Syst. Video Technol.*, to be published, doi: 10.1109/TCSVT.2022.3227172.

[28] X. Liao, B. Tu, J. Li, and A. Plaza, "Class-wise graph embedding-based active learning for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5522813.

[29] J. Liang, Y. Deng, and D. Zeng, "A deep neural network combined CNN and GCN for remote sensing scene classification," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 13, pp. 4325–4338, 2020.

[30] K. Han, Y. Wang, J. Guo, and E. Wu, "Vision GNN: An image is worth graph of nodes," in *Proc. Adv. Neural Inf. Process. Syst.*, 2022, vol. 35, pp. 8291–8303.

[31] Y. Yang and S. Newsam, "Comparing SIFT descriptors and gabor texture features for classification of remote sensed imagery," in *Proc. IEEE 15th Int. Conf. Image Process.*, 2008, pp. 1852–1855.

[32] G. Cheng, P. Zhou, X. Yao, C. Yao, Y. Zhang, and J. Han, "Object detection in VHR optical remote sensing images via learning rotation- invariant HOG feature," in *Proc. 4th Int. Workshop Earth Observ. Remote Sens. Appl.*, 2016, pp. 433–436.

[33] 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.

[34] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2012, pp. 1097–1105.

[35] S. H. Gao, M.-M. Cheng, K. Zhao, X.-Y. Zhang, M.-H. Yang, and P. Torr, "Res2net: A new multi-scale backbone architecture," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 2, pp. 652–662, Feb. 2021.

[36] G. Cheng, C. Yang, X. Yao, L. Guo, and J. Han, "When deep learning meets metric learning: Remote sensing image scene classification via learning discriminative CNNs," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 5, pp. 2811–2821, May 2018.

[37] X. Lu, H. Sun, and X. Zheng, "A feature aggregation convolutional neural network for remote sensing scene classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 10, pp. 7894–7906, Oct. 2019.

[38] E. Li, J. Xia, P. Du, C. Lin, and A. Samat, "Integrating multilayer features of convolutional neural networks for remote sensing scene classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 10, pp. 5653–5665, Oct. 2017.

[39] Y. Bazi, L. Bashmal, M. M. A. Rahhal, R. A. Dayil, and N. A. Ajlan, "Vision transformers for remote sensing image classification," *Remote Sens.*, vol. 13, no. 3, 2021, Art. no. 516.

[40] 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 Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 4409512.

[41] 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 Trans. Image Process.*, vol. 31, pp. 99–109, 2022.

[42] Y.-F. Song, Z. Zhang, C. Shan, and L. Wang, "Richly activated graph convolutional network for robust skeleton-based action recognition," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 31, no. 5, pp. 1915–1925, May 2021.

[43] C. Wu, X.-J. Wu, and J. Kittler, "Graph2Net: Perceptually-enriched graph learning for skeleton-based action recognition," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 4, pp. 2120–2132, Apr. 2022.

[44] G. Li, M. Müller, A. Thabet, and B. Ghanemet, "DeepGCNs: Can GCNs go as deep as CNNs?," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2019, pp. 9266–9275.

[45] G. Yang, J. Lei, and Z. Zhu, "AFPN: Asymptotic feature pyramid network for object detection," 2023, *arXiv:2306.15988*.

[46] 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.

[47] O. A. B. Penatti, K. Nogueira, and J. A. dos Santos, "Do deep features generalize from everyday objects to remote sensing and aerial scenes domains?," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops*, 2015, pp. 44–51.

[48] H. Dan and K. Gimpel, "Gaussian error linear units (gelus)," 2016, *arXiv:1606.08415*.

[49] 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.

[50] Q. Bi, K. Qin, H. Zhang, and G.-S. Xia, "Local semantic enhanced ConvNet for aerial scene recognition," *IEEE Trans. Image Process.*, vol. 30, pp. 6498–6511, 2021.

[51] 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 J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 10484–10499, 2021.

[52] W. Wang, Y. Chen, and P. Ghamisi, "Transferring CNN with adaptive learning for remote sensing scene classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5533918.

[53] W. Song, Y. Cong, S. Zhang, Y. Wu, and P. Zhang, "HDTFF-Net: Hierarchical deep texture features fusion network for high-resolution remote sensing scene classification," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 16, pp. 7327–7342, 2023.

[54] 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.

[55] L. Yuan et al., "Tokens-to-token vit: Training vision transformers from scratch on imagenet," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2021, pp. 558–567.
[56] W. Wang, E. Xie, and X. Li, "Pvt v2: Improved baselines with pyramid vision transformer," *Comput. Vis. Media*, vol. 8, no. 3, pp. 415–424, 2022.
[57] R. R. Selvaraju, M. Cogswell, and A. Das, "Grad-CAM: Visual explanations from deep networks via gradient-based localization," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 618–626.

Chongyang Zhang received the B.S. degree in measurement and control technology and instrument from the University of Electronic Science and Technology of China, Chengdu, China, in 2021. He is currently working toward the M.S. degree in circuits and systems with the School of Information Science and Technology, Fudan University, Shanghai, China.

His research interests include remote sensing scene classification, machine learning, and pattern recognition.

Bin Wang (Senior Member, IEEE) received the B.S. degree in electronic engineering and the M.S. degree in communication and electronic systems from Xidian University, Xi'an, China, in 1985 and 1988, respectively, and the Ph.D. degree in system science from Kobe University, Kobe, Japan, in 1999.

After his graduation in 1988, he was with Xidian University as a Teacher. From 1999 to 2000, he was with the Communications Research Laboratory, Ministry of Posts and Telecommunications, Kobe, as a Research Fellow, working on magnetoencephalography signal processing and its application for brain science. Then, as a Senior Supervisor, he was with the Department of Etching, Tokyo Electron AT Ltd., Tokyo, Japan, from 2000 to 2002, dealing with the development of advanced process control systems for etching semiconductor equipment. Since September 2002, he has been with the Department of Electronic Engineering, Fudan University, Shanghai, China, where he is currently a Full Professor and Leader of the Image and Intelligence Laboratory. He has authored and coauthored more than 150 scientific papers in important domestic and international periodicals. He is the holder of several patents. His main research interests include multispectral/hyperspectral image analysis, automatic target/object detection and recognition, pattern recognition, signal detection and estimation, and machine learning.

Dr. Wang is an Associate Editor of the IEEE JOURNAL OF SELECTED TOPICS IN APPLIED EARTH OBSERVATIONS AND REMOTE SENSING.