remote sensing

MDPI

Article

Multi-Scale Graph-Based Feature Fusion for Few-Shot Remote Sensing Image Scene Classification

Nan Jiang, Haowen Shi and Jie Geng *

School of Electronics and Information, Northwestern Polytechnical University, Xi'an 710072, China
* Correspondence: gengjie@nwpu.edu.cn

Abstract: Remote sensing image scene classification has drawn extensive attention for its wide application in various scenarios. Scene classification in many practical cases faces the challenge of few-shot conditions. The major difficulty of few-shot remote sensing image scene classification is how to extract effective features from insufficient labeled data. To solve these issues, a multi-scale graph-based feature fusion (MGFF) model is proposed for few-shot remote sensing image scene classification. In the MGFF model, a graph-based feature construction model is developed to transform traditional image features into graph-based features, which aims to effectively represent the spatial relations among images. Then, a graph-based feature fusion model is proposed to integrate graph-based features of multiple scales, which aims to enhance sample discrimination based on different scale information. Experimental results on two public remote sensing datasets prove that the MGFF model can achieve superior accuracy than other few-shot scene classification approaches.

Keywords: few-shot learning; graph-based feature; multi-scale feature fusion; remote sensing image scene classification

Citation: Jiang, N.; Shi, H.; Geng J. Multi-Scale Graph-Based Feature Fusion for Few-Shot Remote Sensing Image Scene Classification. *Remote Sens.* **2022**, *14*, 5550. https://doi.org/10.3390/rs14215550

Academic Editor: Javier Marcello

Received: 9 September 2022
Accepted: 1 November 2022
Published: 3 November 2022

1. Introduction

Scene classification of a remote sensing image is a significant assignment, which has attracted significant attention in various applications. Recent years have witnessed the continuous improvement of imaging technology, then the resolution of remote sensing images has gradually increased, and more useful information can be captured from remote sensing images, including extensive land cover information, such as terrain, mountain, as well as water. For different scenes, remote sensing images are processed differently. It is of great significance to assign semantic labels to remote sensing images [1,2], which is helpful for the management and analysis the remote sensing data in a unified manner. Therefore, scene classification aims at classifying remote sensing images based on similar scene characteristics by extracted features [3]. Currently, scene classification technology has been broadly applied in geological exploration, planning management, environmental monitoring, object detection, and other fields [4,5]. The methods for remote sensing image scene classification consist of three types according to visual features, which are methods derived from low-level visual features, middle-level visual representations, and high-level visual information [6].

The classification methods based on low-level visual features obtain the structure, edge, and other basic information from images, where these features are extracted based on traditional hand-crafted methods. For example, scale-invariant feature transform (SIFT) [7] is used to obtain low-level visual features, which has been frequently adopted for the scene classification of remote sensing images. The spatial envelope feature algorithm [8] adopts global feature information to classify remote sensing images, which can represent the overall information of images macroscopically. Nevertheless, scene classification derived from the global feature information is difficult to handle complex scenarios of remote sensing images. Furthermore, low-level visual features contain less semantic information, which is

challenging to obtain satisfactory results for scene classification with high complexity and non-uniform spaces.

The classification approaches derived from middle-level visual information encode the local features extracted by traditional manual approaches, and further optimize the feature representation with more recognition ability. Among middle-level feature based methods, bag-of-words model (BoVW) [9] is one of the most classic methods. BoVW model uses SIFT to extract local features from images, and then encodes local features into global features using visual word bags. Thus, BoVW combines various underlying visual features to enhance the performance of recognition ability of the model itself [10]. Many subsequent methods [11,12] that use multiple features to deal with the scene classification are developed based on BoVW. Although mid-level feature based methods achieve superior results than low-level feature based methods, they also face many shortcomings. In particular, the methods of re-representing local features will ignore the relationship among each local feature, which weaken the classification effect.

As for classification approaches derived from high-level visual information, they are mainly employing deep neural networks, which can learn more abstract semantic information from remote sensing images. Recent years have witnessed that the deep learning technology has greatly improve the performances in numerous visual assignments, and various convolutional neural networks, such as AlexNet [13], VGG-VD-16 [14], and GoogLeNet [15] have been proposed for image classification task. Penatti et al. [16] first adopted convolutional neural network to deal with scene classification of remote sensing image, which is proved to achieve superb results. Polarimetric-feature-driven deep CNN [17] is proposed to solve PolSAR classification with limited training samples. A convolutional neural network is used in the task of image scene classification, which is usually divided into three training methods. The first is complete training, which is conducted under the condition that all weight parameters of the network are initialized randomly, and good performance is obtained [18,19]. The second is to directly use the trained model as a feature extractor, and then use the extracted features in the scene classification task through algorithm processing [20]. The third is to transfer the trained parameters to the model to achieve better results [21]. Compared with approaches derived from low-level information, as well as middle-level information, classification derived from high-level visual features can completely express the deeper abstract semantic information of images, which is capable for enhancing the performance of scene classification.

Generally, classification methods based on deep learning technology require extensive labeled samples for network training. If the training samples are insufficient, deep model will face the problem of overfitting, which may result in a sharp decrease in the performance on the test set. As for most practical applications, it is quite challenging to collect extensive labeled samples, and collection of annotated samples often consumes a lot of time. Therefore, scene classification of remote sensing images in a few-shot case is of great significance. Inspired by human cognition of novel objects based on prior knowledge, few-shot learning has been developed to learn policies for recognition and classification with quite limited labeled samples [22,23]. At present, few-shot learning methods are mainly derived from meta learning, as well as transfer learning.

In contrast with optical images, remote sensing images possess the distinctions of large size, abundant data, and rich land cover information [24]. Since remote sensing images capture complex ground scenes, performance of scene classification will be affected by environmental factors and other factors [25]. Specifically, if there exist distinguished intra-class differences, remote sensing images of the same category may be classified into different categories, where intra-class variability arises because similar scenes contain large differences in underlying feature space. Furthermore, remote sensing images of different types may be classified into the same scene because they contain similar backgrounds or spatial structures, where this phenomenon is called as inter-class similarity. These issues, mentioned above, will reduce the efficacy for scene classification. In addition, traditional scene classification methods generally utilize deep features of the top layer for

classification, but the front layer features still contain useful information. Such information may be lost during convolution or other operations. Therefore, the extraction of multi-scale feature is beneficial to expand the information from remote sensing images. How to utilize multi-scale features in a few-shot case is also an issue that needs to be resolved.

For few-shot scene classification of remote sensing images, a multi-scale graph-based feature fusion (MGFF) model is proposed, which aims at excavating effective features in few-shot cases. In our approach, we firstly pre-train a feature extractor to learn multi-scale features of remote sensing images, and then a transposed convolution is applied to correct the extracted multi-scale features. After that, image features are transformed into graph-based features based on the construction module. Finally, a graph-based feature fusion model is proposed to enhance sample discrimination based on different scale features. The major contributions of this paper are summarized as follows:

- A graph-based feature learning model is developed to learn features from remote sensing images firstly, which enables to effectively express the spatial relations among remote sensing images. It is able to take advantage of relation information for scene classification, which is beneficial to few-shot scene classification.
- A graph-based feature fusion model is proposed, which can integrate graph-based features of multiple scales. It is able to enhance sample discrimination based on different scale features, which integrates more abundant and effective semantic information. The proposed model can take full advantage of image features to improve few-shot classification accuracies, which reduces the influence of inconsistent semantic information.
- Experimental results on two public remote sensing data illustrate that the proposed MGFF yield an improvement of classification accuracy about 2–10% contrast to other advanced methods, which proves the efficacy of our MGFF model.

The remainder of this paper is organized as follows. Section 2 presents the related work. Section 3 describes the MGFF model in detail. The results and discussions are, respectively, shown in Sections 4 and 5. In the end, conclusions are summarized in Section 6.

2. Related Works

In this section, related work related to remote sensing image scene classification, few-shot learning, and graph learning will be introduced in detail.

2.1. Remote Sensing Image Scene Classification

Although remote sensing image scene classification consists of three categories according to different visual features, classification derived from high-level information utilizes the powerful learning ability of deep neural network [26,27], which can excavate discriminative information from remote sensing images. Deeper abstract semantic information enables to boost the classification performance. Therefore, methods derived from deep learning are widely adopted to deal with the scene classification of remote sensing image [28–30].

In deep learning based methods, a convolutional neural network (CNN) has been broadly adopted for scene classification [31,32]. Nogueira et al. [33] applied a CNN to extract features and used linear SVM for scene classification. Cheng et al. [34] proposed a discriminative CNN to deal with the problem of misclassification due to factors of intra-class dissimilarity and inter-class similarity. Wang et al. [35] proposed an attention loop CNN model for scene classification, which can extract high-level feature information and guarantee effective discarding of non-critical information. Tang et al. [36] proposed class-level prototype guided multi-scale feature learning method, which enables to make a distinction from different semantics with limited labeled samples. Rafael Pires et al. [37] developed a transfer learning method to learn the effect of a CNN in dealing with the scene classification task, which demonstrates the feasibility of transfer learning from natural images to remote sensing images. Zeng et al. [38] proposed specific contrastive learning model, which enables to boost the scene classification accuracy with limited supervised

samples. Although the above works make an effect on boosting the performance of remote sensing image scene classification, they all focus on extracting high-level visual features from remote sensing images, but low-level features and middle-level features also contain useful information. Therefore, multi-scale feature fusion should be considered to learn abundant information from different scale features.

2.2. Few-Shot Learning

In recent years, few-shot learning has drawn great attention, which aims to enhance the learning capacity on the novel categories with fewer labeled samples [39,40]. Few-shot learning has also been developed for remote sensing image scene classification, where various few-shot learning approaches have been widely applied [41,42]. The follow-up work is to design a more suitable few-shot learning method for remote sensing image scene classification. Alajaji et al. [43] proposed a prototype network, which combines with a pre-trained SqueezeNet to obtain better prototype features of each category. Jiang et al. [44] proposed a multi-scale metric learning (MSML) method, which extracts multi-scale features and learns the relations among samples for few-shot classification. Due to the issues of intra-class difference, as well as inter-class similarity in remote sensing images, prototype-based few-shot learning models ignore the verification of prototype features, which reduce the scene classification performance. To overcome these issues, Cheng et al. [45] proposed a Siamese prototype network (SPNet), which develops prototype self calibration and mutual calibration to optimize sample features. Zeng et al. [46] proposed an iterative distribution learning network, where three sub-modules are developed to improve the discrimination of features. In the network, the similarity distribution learning model is adopted to calculate the relationship of different instances, the label matching model is utilized as few-shot classifier derived from prior knowledge, and the attention-based feature calibration model is developed to optimize the sample features and yield the final features for the next iteration.

Meta learning is also adopted to few-shot scene classification. In [47], a meta learning based method named RS-MetaNet is proposed to learn a metric space through a series of assignments, which is suitable for scene classification with limited label data. Li et al. [48] proposed a discriminative learning based adaptive matching network (DLA-MatchNet), which is a few-shot learning method that adds a matcher into the feature extractor. In [49], a life-long few-shot learning model is proposed for few-shot classification, which realizes knowledge transferring from one dataset to a novel dataset.

2.3. Graph Learning

Graph learning aims to study how to apply deep neural networks on graph-based data, and has accomplished various classification tasks. In recent years, graph learning [50] has been adopted to few-shot learning tasks and realized effective results, where a series of graph neural networks (GNNs) have been proposed [51,52]. GNNs can utilize the available information of similar samples for classification.

Several recent studies have been proposed to enable the neural network to handle the graph structures [53]. In a graph convolutional network (GCN) [54], a graph convolution operator is introduced, which can be cascaded to obtain a deep learning architecture. A graph attention network (GAT) [55] extends the GCN model by adding a learnable attention kernel, which allows to assign different weights to the set of adjacent nodes. A simplified graph convolutional network (SGC) [56] optimizes the network structure by simplifying the non-linear transformation function to the single linear transformation function, which achieves superior results with fewer parameters. A graph-based embedding smoothing network (GES-Net) [57] adopts an unsupervised non-parametric regularizer, severed as embedding smoothing, which can obtain superb results for few-shot remote sensing image scene classification. Although the above works all use graph learning to solve classification tasks, the fusion of graph features may result in missing of useful information and reducing the stability of deep network. Therefore, effective utilization of a graph-based neural

network needs to be studied to fully excavate the image features for enhancing few-shot scene classification performance.

3. Methodology

The proposed multi-scale graph-based feature fusion model is shown in Figure 1. In our framework, a pre-trained feature extractor is firstly adopted to extract multi-scale image features. Furthermore, a transposed convolution operation is applied to correct the extracted multi-scale features. After that, image features are transformed into graph-based features based on the construction module. Finally, a graph-based feature fusion model is proposed to enhance sample discrimination based on different scale features, which is utilized for scene classification.

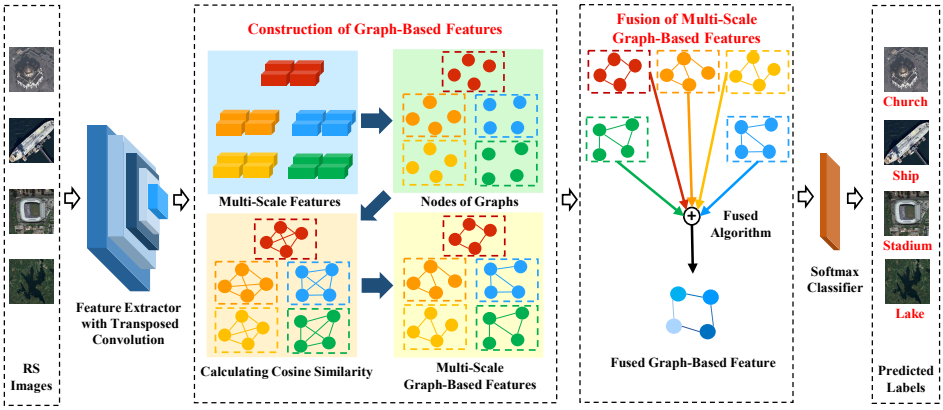


Figure 1. The framework of proposed multi-scale graph-based feature fusion model.

3.1. Problem Formulation

The problem with few-shot remote sensing image scene classification is in utilizing an extremely limited number of labeled data for training a classifier. As for few-shot learning tasks, all the samples can be divided into two distinct datasets, named D_{base} and D_{novel}, where there is no overlap in categories. In D_{base}, each category includes abundant labeled data, which can be adopted for training and validation. However, in D_{novel}, there are only extremely limited number of labeled data of each class, which is applied for testing.

During the training and validation process, we split D_{base} into four different parts without overlap in categories, namely, $S_{train} = \{x_i, y_i\}$, $Q_{train} = \{x_j, y_j\}$, $S_{valid} = \{x_m, y_m\}$, $Q_{valid} = \{x_n, y_n\}$, in which x_i denotes the ith sample, and y_i denotes corresponding label. S_{train} and Q_{train} are applied during the training process. S_{valid} and Q_{valid} are adopted during the validation process.

In the training process, the parameters of pre-trained feature extractor are updated on S_{train}, and the evaluation of the classification module is conducted on Q_{train}. When the evaluation results tend to be stable, it can be considered that the feature extractor has been well trained. In the testing process, we can regard the classification as a M-way Z-shot K-query task, where Z labeled samples of each class ($M \times Z$ samples in total) are applied as the prior knowledge, and $M \times K$ unlabeled samples are utilized for prediction. We split D_{Novel} into two parts, which are $S_{test} = \{x_p, y_p\}(p = 1, 2, \ldots, M \times Z)$, and $Q_{test} = \{x_q, y_q\}(q = 1, 2, \ldots, M \times K)$. The number of labeled samples is extremely small, where the shot number Z is set to 1 or 5 in the experiments.

3.2. Extraction of Multi-Scale Features

To extract multi-scale features from images, ResNet-12 is adopted as the backbone in our framework. ResNet-12 contains four residual convolution blocks and one average pooling layer. The structure of the backbone is shown in Figure 2. Each residual convolution block is composed of three convolution layers, batch normalization layers and ReLU layers, alternately.

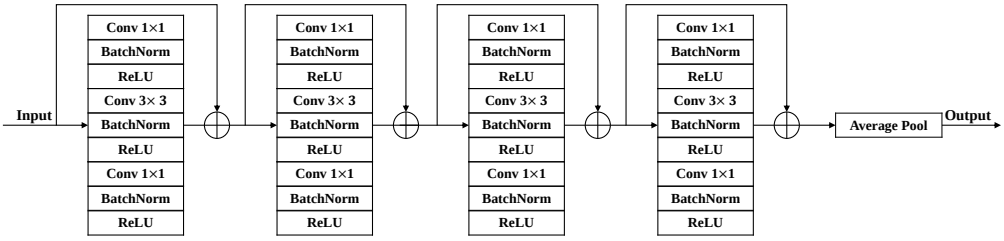


Figure 2. The structure of the backbone.

Considering the sample number of the training set is insufficient, data augmentation is utilized to extend pseudo training samples, where random rotation is applied to generate pseudo images. Thus, original images and generated images are used for deep network training. The loss function of backbone optimization is defined as follows:

$$L_t = (1 - \lambda) \cdot L(I) + \lambda \cdot L(R) \tag{1}$$

where L_t stands for the total loss, $L(\cdot)$ denotes cross-entropy loss function, $L(I)$ represents the loss of original images, $L(R)$ stands for the loss of the generated images, and λ is a hyperparameter in order to balance the two types of loss.

Considering the scarcity of samples, we prefer to improve our deep network from the perspective of sample feature enhancement. As for traditional classification methods, only the top layer features are utilized for final classification. However, the low-level features as well as mid-level features from deep neural network also contain useful information, which may be weakened or even disappeared during the convolution and pooling operations. Therefore, multi-scale feature fusion is beneficial to expand the information of image features.

As mentioned above, the feature extractor used in this paper can be divided into five stages, namely, four convolution blocks and the last one average pooling block. Thus, we can combine these five stages to obtain multi-scale features. Considering that multi-scale fusion requires the consistency of dimensions, a transposed convolution operation is adopted to revise features of each stage to a certain extent. The structure of the extraction of multi-scale features is shown in Figure 3.

Therefore, for a remote sensing image $\mathbf{X_k}$, the extraction of multi-scale features can be defined as follows:

$$\mathbf{F_k} = \{T_i[P_i(\mathbf{X_k})]\}_{i=1}^{5} \tag{2}$$

where $\mathbf{F_k}$ denotes the extracted multi-scale features of the sample $\mathbf{X_k}$, $T_i[\cdot]$ represents the ith transposed convolution layer, $P_i(\cdot)$ stands for features output from the ith stage of the hidden layer, and the input image passes through the 1th, 2th, ..., and 5th stage of the hidden layers.

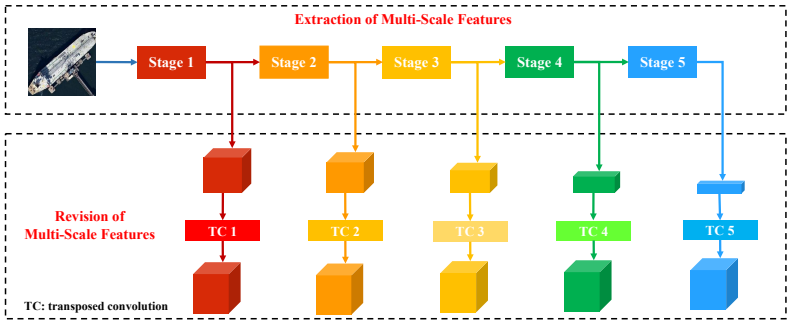


Figure 3. The structure of the extraction of multi-scale features.

3.3. Construction of Graph-Based Features

In order to obtain more effective expression of the spatial information among remote sensing images, a construction model of graph-based feature is proposed based on the KNN algorithm, which can construct graph-based features to reflect relation information.

Assume the image features have the size of $N \times C \times H \times W$, where N expresses the batch size, C denotes the channel number of the feature, H and W, respectively, represent the height and the width of the feature. Graph-based features mainly include two types of features, one is node features $\mathbf{V_i}$ and the other is edge features $\mathbf{E_i}$. In this paper, the feature of a whole graph is defined as follows:

$$\mathbf{G_i} = \{\mathbf{E_i}, \mathbf{V_i}\} \tag{3}$$

where i represents the ith graph constructed by the ith scale image features, and $i \in [1, 5]$.

Node features are essentially the denoted image features. More specifically, features of a node come from the features of an image obtained by the feature extractor. In our settings, batch norm is applied, which means N (batch size) images are imported at one time. Therefore, node features of a graph include features of N images. In our proposed model, features of different scales are considered, and all the node features in a graph are acquired at the same scale. Then the node features can be written as follows:

$$\mathbf{V_i} = \left\{\mathbf{F^T_{1,i}}, \mathbf{F^T_{2,i}}, \ldots, \mathbf{F^T_{N,i}}\right\} \tag{4}$$

where $\mathbf{V_i} \in R^{N \times d}$, and $\mathbf{F_{k,i}} \in R^{1 \times d}$ represents the image features of the kth sample on the ith scale.

At the same time, we perform matrix transformation on the image features. The dimensions of features are transferred to $N \times T \times W$, where T is equal to $C \times H$, which will not cause loss of information and aims to simplify the later calculation and programming.

The edge features reflect the degree of similarity among nodes, which are represented by an adjacent matrix $\mathbf{E_i} \in R^{N \times N}$. In order to obtain the adjacent matrix, cosine similarity between any two nodes is calculated firstly. For a certain node m, the cosine similarity can be calculated as follows:

$$\mathbf{R_{m,i}} = [cos(\mathbf{F_{m,i}}, \mathbf{F_{n,i}})] \tag{5}$$

where n denotes the sample index, and $n \in [1, N]$. Theoretically, we can calculate the distance between any two nodes. In order to reduce the model complexity, neighbor nodes are searched by KNN algorithm, where the most similar nodes are considered. For a certain node, the top k most similar nodes of the same class are represented, where the cosine similarity denotes the edge connection between them. Other nodes with much lower similarities are set to 0, which stands for no edge connection. Thus, the modified cosine similarity results with KNN algorithm can be denoted as $\hat{\mathbf{R}}_{\mathbf{m,i}}$.

Based on cosine similarity results, the adjacency matrix can be defined as follows:

$$\mathbf{R_i}[m, n] = \begin{cases} \hat{\mathbf{R}}_{\mathbf{m,i}}[1, n] & \text{if } m \neq n \\ 0 & \text{else} \end{cases} \tag{6}$$

Then, the adjacency matrix is normalized to yield the final adjacency matrix $\mathbf{E_i}$. The normalization process is represented as follows:

$$\mathbf{E_i} = \mathbf{U_i^{-\frac{1}{2}}} \mathbf{R_i} \mathbf{U_i^{\frac{1}{2}}} \tag{7}$$

where $\mathbf{U_i}$ stands for the degree diagonal matrix, and $\mathbf{U_i}[m, n] = \sum_n \mathbf{R_i}[m, n]$.

3.4. Fusion of Multi-Scale Graph-Based Features

To integrate different scale features, a fusion model of multi-scale graph-based features is developed to obtain more abundant and effective semantic information. The fusion model

is considered from two perspectives, which are node features and edge features. With regard to node features, which are essentially image features, we can consider that features obtained from the deeper layers contain more abstract information, and, thus, the fusion weight should be larger for the deeper layers. At the same time, features from low-layer network include abundant detailed information, which should not be negligible. For scene classification, the proportion of low-layer features is relatively small, and, thus, the fusion weights of them are relatively lower.

As for edge features, they are constructed based on cosine similarity among node features. The edge feature can reflect the spatial relation between two nodes. Only when the cosine similarity of two nodes reaches a certain level, there can be a non-zero edge feature expression. The images of the same scene have the characteristics of high similarity information, even if there are some individual differences. Drawing on this assumption, we believe that after repeated training, features with high similarity extracted by the model are the key to classification and are worth retaining for feature fusion. The most straightforward way is to directly multiply the edge features and the node features, which can ignore the features of nodes that are not connected by edges.

Denoting the fused graph as $\mathbf{G_f}$, which includes the fused node features $\mathbf{V_f}$ and the fused edge features $\mathbf{E_f}$. It can be defined as follows:

$$\mathbf{G_f} = \{\mathbf{E_f}, \mathbf{V_f}\} \tag{8}$$

Based on the above analysis, the node features of five graphs with different scales are fused, where the fusion process is defined as follows:

$$\mathbf{V_f} = \sum_{i=1}^{4} \alpha_i \mathbf{E_i}^{\beta_i} \mathbf{V_i} + \alpha_5 (\mathbf{I} + \mathbf{E_5})^{\beta_5} \mathbf{V_5} \tag{9}$$

where $\mathbf{I}$ represents the identity matrix, and α and β stand for the hyper-parameters. In our framework, features of the fifth scale are obtained by the deepest layer, which means they contain the most effective information. $\mathbf{I}$ is to enhance features of the fifth scale and fuse more effective information. α is to weight graph-based features obtained from different scales to balance the importance of features from different levels. Here, features obtained from the deeper layer are corresponding to larger weight. Therefore, we set the rule that $\alpha_1 \leq \alpha_2 \leq \alpha_3 \leq \alpha_4 \leq \alpha_5$. β is the degree to balance the neighbor information. The smaller the β, the stronger the degree will be expressed.

After the node features are fused, the fused edge features are also constructed using the same method as above. The initial matrix based on the fused node features is generated as $\mathbf{R_f}$, and the corresponding degree diagonal matrix is denoted as $\mathbf{U_f}$. Thus, the standardized fused edge features are defined as follows

$$\mathbf{E_f} = \mathbf{U_f}^{-\frac{1}{2}} \mathbf{R_f} \mathbf{U_f}^{\frac{1}{2}} \tag{10}$$

Finally, fused graph-based features of five scales are utilized for few-shot classification, where the logistic regression classifier is trained by S_{train} and Q_{train}, and updated by S_{test}. Labels of samples from Q_{test} are predicted to calculate classification accuracy.

4. Results

In this section, two public remote sensing datasets are applied to verify the effectiveness of the MGFF model. The datasets, parameter setting and experimental comparisons are presented in detail.

4.1. Datasets

The evaluation of MGFF model is executed on two public available datasets, including NWPU-RESISC45 and WHU-RS19. These two datasets are depicted in Figure 4. NWPU-RESISC45 dataset [58] is a public dataset for remote sensing image scene classification,

which was created by Northwestern Polytechnical University. This dataset consists of 31,500 images with 45 categories, including airplane, chaparral, dense residential, forest, rectangular farmland, and so on. Each type consists of 700 images with 256 × 256 pixels. In the evaluation experiments, 45 classes are divided into 25, 10, and 10 classes to, respectively, serve as the training set, validation set, as well as testing set. The details of data partitioning are reported in Table 1.

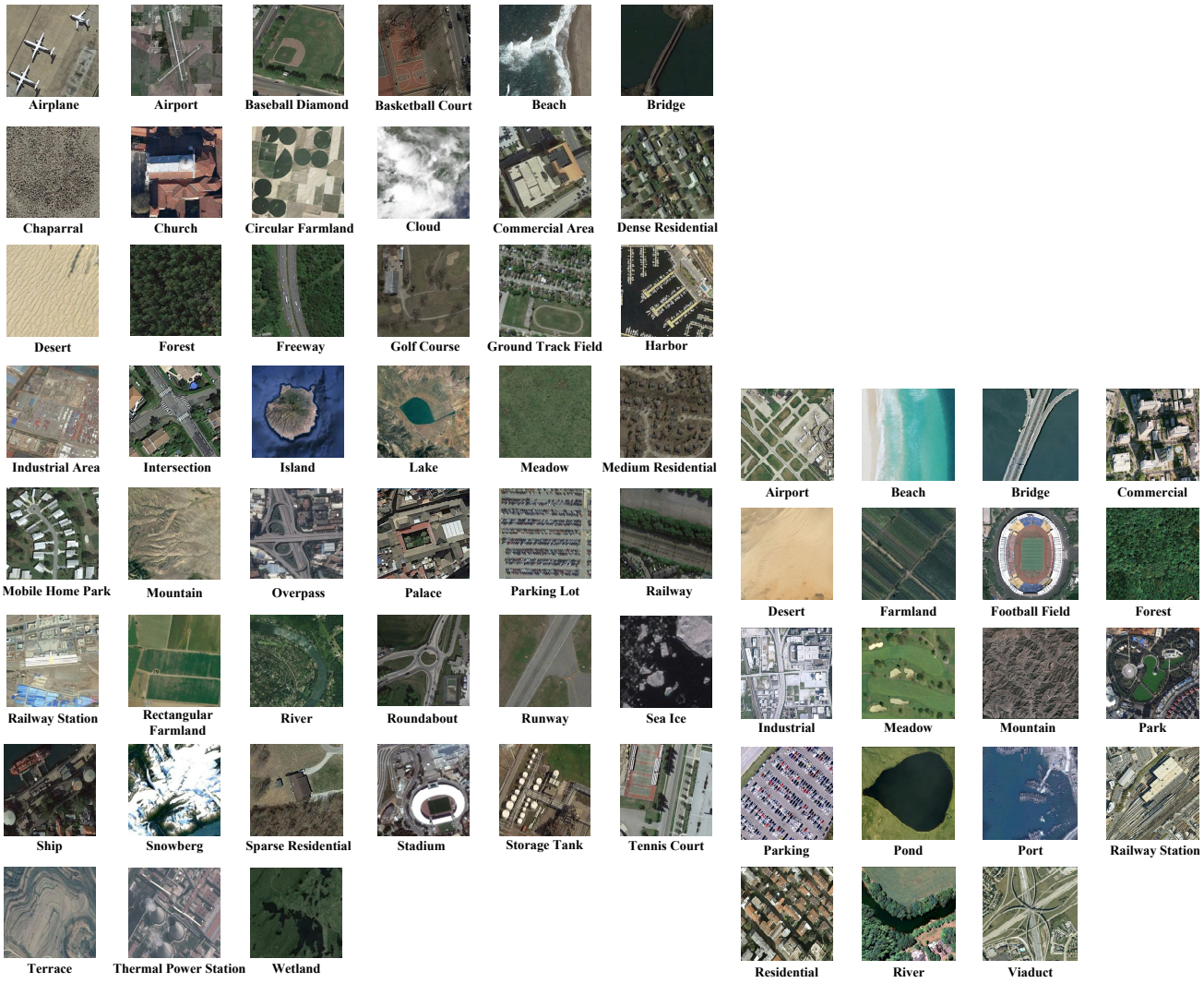


Figure 4. Datasets utilized in the experiments, (**a**) images of NWPU-RESISC45 dataset, (**b**) images of WHU-RS19 dataset.

WHU-RS19 dataset [59] is a benchmark for remote sensing image scene classification, which was established by Wuhan University. The dataset consists of 19 categories, including mountain, beach, park, commercial, farmland, railway station, desert, meadow, football field, industrial, forest, pond, parking lot, river, residential viaduct, port, bridge, and airport. Each category contains 50 or more images with 256 × 256 pixels. In the evaluation experiments, 19 categories are divided into 9, 5, and 5 categories to serve as the training set, validation set, as well as testing set, respectively, which is shown in Table 1.

Table 1. Details of NWPU-RESISC45 dataset and WHU-RS19 dataset.

Datasets	Training	Validation	Testing
NWPU-RESISC45	Sea ice; Beach; Rectangular farmland; Mountain; Stadium; Cloud;Railway; Ship; Desert; Forest; Island; Baseball Diamond; Lake; Meadow; Snowberg; Airplane; Palace; Ground field;Harbor; Bridge; Chaparral; Church; Wetland; Mobile home park;	Storage tank; Power station; Runway; Sparse residential; Terrace; Railway station; Tennis Court; Overpass; Commerical area; Industrial area;	Mid residential; River; Intersection; Dense residential; Parking lot; Golf course; Circle farmland ; Airport; Freeway; Basketball court;
WHU-RS19	Park; Residential; Airport; Football field; Meadow; Desert; Parking lot; Bridge; Industrial;	Farmland; Railway station; Port; Forest; Beach;	Viaduct; Mountain; Pond; Commerical; River;

4.2. Experimental Settings

In our experiments, ResNet-12 is adopted as the backbone in our framework for feature extraction. In the pre-training stage, the weight factor λ of loss function is set to 0.5, and SGD optimizer is utilized as the optimization method. The learning rate is set to 0.001 and the batch size is set to 64. In order to construct graph-based features, the number of nodes in a graph should also be consistent with the batch size, which is set to 64. The settings of experiment parameters are summarized in Table 2. Other hyper-parameters are selected based on the experimental results, which have been discussed in the following subsection.

Table 2. The settings of experiment parameters.

Parameters	Values
λ	0.5
learning rate	0.001
batch size	64

4.3. Comparisons with the State-of-the-Art Approaches

To verify the efficacy of the MGFF model, several state-of-the-art approaches for the same image scene classification task are applied for comparisons, including SCL-MLNet [60], Meta-SGD [61], TPN [62], Relation Net [41], MAML [63], DLA-MatchNet [48], and GES-Net [57]. In the field of few-shot scene classification of remote sensing images, the most concerned indicator is the overall classification accuracy. In order to compare fairly with other methods, we adopt the overall accuracy (ACC) as the evaluation index. In the experiments, classification results of 5-way 1-shot case and 5-way 5-shot case of each method are presented in Tables 3 and 4, respectively.

It can be clearly seen from Table 3 that our MGFF model produces the best accuracies under conditions of the two cases on NWPU-RESISC45 data, surpassing the GES-Net

model by 4.26% and 0.97%, respectively. Compared with DLA-MatchNet, MGFF model has 6.29% and 1.61% improvements in 5-way 1-shot case and 5-way 5-shot case, respectively. Furthermore, the proposed MGFF is more advanced in terms of accuracy compared to the remaining other compared methods. It is verified that our proposed multi-scale graph-based feature fusion model is able to yield superb performance on NWPU-RESISC45 data.

Experimental comparisons conducted on WHU-RS19 dataset are shown in Table 4. Our proposed MGFF model presents the best results in cases of 5-way 1-shot and 5-way 5-shot. It exceeds the GES-Net model by 0.64% and 2.49%, respectively. Compared with DLA-MatchNet, MGFF model has 8.21% and 4.97% improvements under the two conditions. In addition, MGFF yields better few-shot classification performance contrast to other advanced approaches.

The above results illustrate that the proposed MGFF model is capable of effectively utilizing the limited information from a few samples, and through multi-scale graph-based feature fusion, information of the low-level features can be greatly reconciled with that of the high-level features. At the same time, our model excavates the spatial relations by construction of graph-based features from image features, so as to obtain more effective information than traditional methods. Therefore, the proposed MGFF is able to effectively ameliorate the accuracy of remote sensing image scene classification with limited labeled data.

Table 3. Classification accuracies of 5-way 1-shot and 5-way 5-shot on the NWPU-RESISC45 dataset.

Method	5-Way 1-Shot	5-Way 5-Shot
SCL-MLNet [60]	62.21 ± 1.12	80.86 ± 0.76
Meta-SGD [61]	60.69 ± 0.72	75.72 ± 0.49
TPN [62]	66.52 ± 0.76	78.47 ± 0.64
Relation Network [41]	66.41 ± 0.48	78.53 ± 0.41
MAML [63]	47.32 ± 0.10	63.03 ± 0.55
DLA-MatchNet [48]	68.80 ± 0.70	81.63 ± 0.46
GES-Net [57]	70.83 ± 0.85	82.27 ± 0.55
MGFF (Ours)	75.09 ± 0.94	83.24 ± 0.65

Table 4. Classification accuracies of 5-way 1-shot and 5-way 5-shot on the WHU-RS19 dataset.

Method	5-Way 1-Shot	5-Way 5-Shot
SCL-MLNet [60]	63.36 ± 0.88	77.62 ± 0.81
Meta-SGD [61]	51.59 ± 0.92	63.95 ± 0.87
TPN [62]	59.24 ± 0.86	71.43 ± 0.67
Relation Network [41]	60.88 ± 0.42	79.76 ± 0.67
MAML [63]	51.06 ± 0.21	65.83 ± 0.17
DLA-MatchNet [48]	68.27 ± 1.83	79.89 ± 0.33
GES-Net [57]	75.84 ± 0.78	82.37 ± 0.38
MGFF (Ours)	76.48 ± 0.96	84.86 ± 0.76

In order to comprehensively present the classification effect of the proposed method, precision (PRE) and F_1 score (F_1) are also adopted as the evaluation indicators. The few-shot classification results of the proposed model are shown in the Table 5. It can be found that under two conditions, PRE and F_1 on the two datasets achieve excellent results close to ACC, which shows the effectiveness and stability of the proposed model.

Table 5. Classification results of our proposed model on two datasets.

Indicators	NWPU-RESISC45		WHU-RS19	
	5-Way 1-Shot	5-Way 5-Shot	5-Way 1-Shot	5-Way 5-Shot
PRE	75.19 ± 0.96	83.75 ± 1.09	76.08 ± 1.16	84.97 ± 0.59
F_1	73.80 ± 0.62	82.85 ± 0.93	74.09 ± 1.08	83.52 ± 0.64
ACC	75.09 ± 0.94	83.24 ± 0.65	76.48 ± 0.96	84.86 ± 0.76

Moreover, similar to other literature [46,57], the confusion matrix is utilized to evaluate the performance of few-shot remote sensing image scene classification. Figure 5 shows the confusion matrices of the proposed model with different conditions on two datasets. It can be seen that performance of 5-way 5-shot is superior to that of 5-way 1-shot.

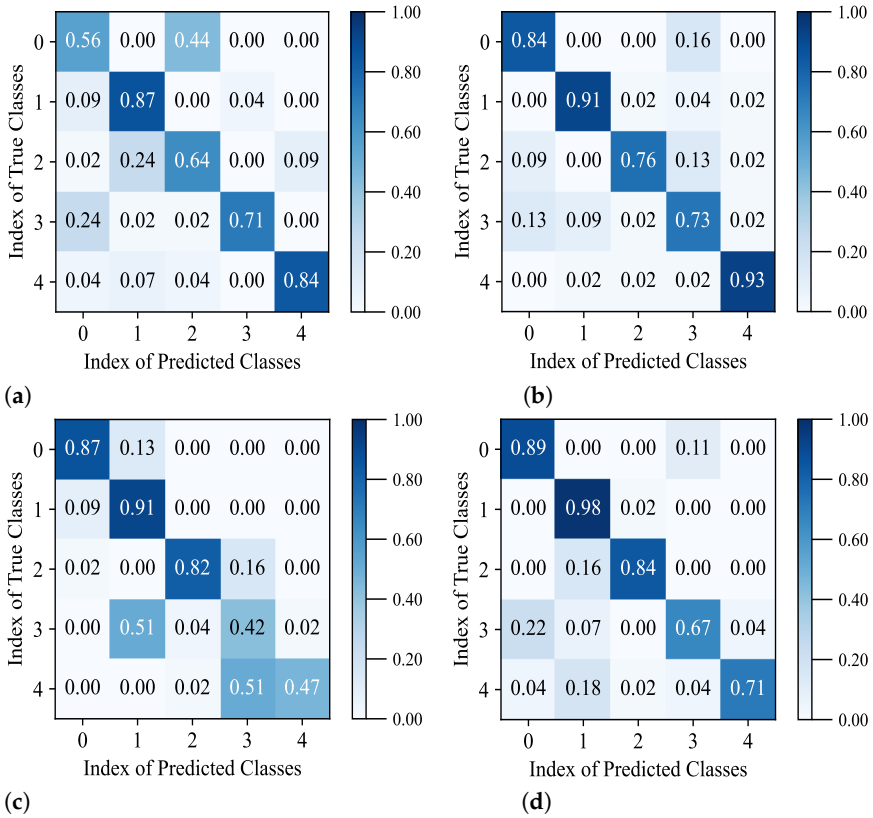


Figure 5. Confusion matrices of our proposed model with different conditions on two datasets. (**a**) the confusion matrix on the NWPU-RESISC45 of 5-way 1-shot case, (**b**) the confusion matrix on the NWPU-RESISC45 of 5-way 5-shot case, (**c**) the confusion matrix on the WHU-RS19 of 5-way 1-shot case, and (**d**) the confusion matrix on the WHU-RS19 of 5-way 5-shot case.

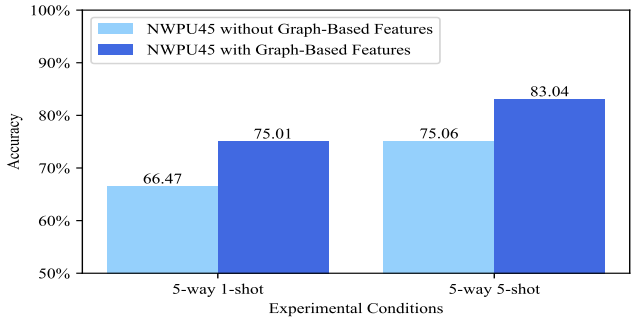
5. Discussions

To further clearly discuss and illustrate the effectiveness of MGFF model, three kinds of ablation experiments are conducted, which are the effect of graph-based features, the effect of multi-scale feature fusion strategy and the effect of parameters.

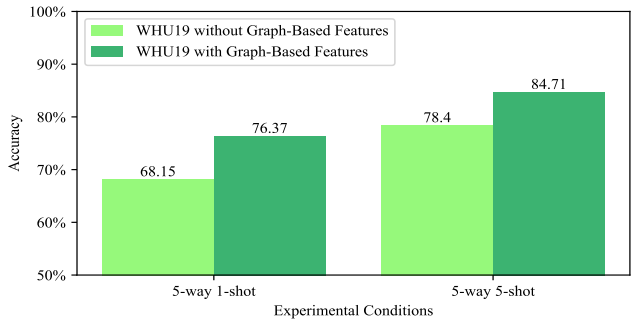
5.1. Effect of Graph-Based Features

In our proposed model, in order to extract the spatial relation information, we construct graph-based features from the image features. To validate the effectiveness of this module, few-shot classification with or without graph-based features is conducted. In the comparisons, we perform linear weighted fusion of image features obtained from the

feature extractor and transposed convolution layers. The compared results on NWPU-RESISC45 and WHU-RS19 are shown in Figure 6.



(**a**)



(**b**)

Figure 6. Comparisons of different types of features on two datasets. (**a**) shows the test results on the NWPU-RESISC45 dataset with different types of features, (**b**) shows the test results on the WHU-RS19 dataset with different types of features.

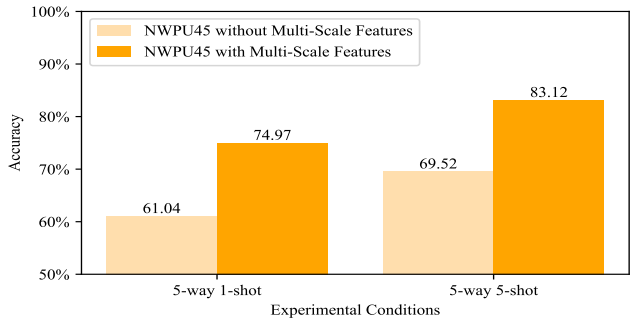
It can be clearly seen from Figure 6, accuracy of the image feature fusion model is obviously lower than that of the graph-based feature fusion model. On the NWPU-RESISC45 dataset, under the two conditions, the accuracy results of the image feature fusion model are 66.47% and 75.06%, respectively, which are 8.54% and 7.98% lower than that of the graph feature fusion model. On the WHU-RS19 data, under the two conditions, the accuracy results of the image feature fusion model are 68.15% and 78.40%, respectively, which are 8.22% and 6.31% lower than that of the graph-based feature fusion model. It can be concluded that graph-based features reflect more effective information than image features, which have a significant effect on improving the accuracy of few-shot remote sensing image scene classification.

5.2. Effect of Multi-Scale Feature Fusion Strategy

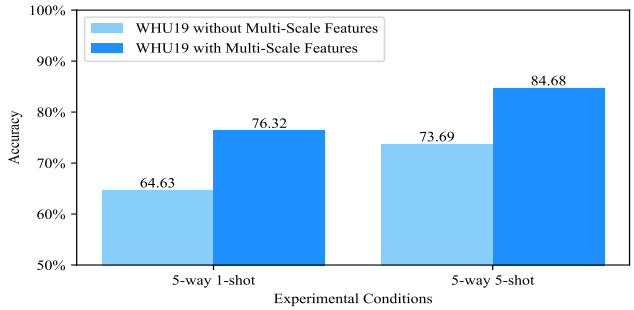
In MGFF model, a multi-scale graph-based feature fusion model is introduced to integrate high-level graph-based features with low-level graph-based features, where features of five different scales are fused. To illustrate the effectiveness of this fusion module, we conduct experiments on single-scale feature. In the compared model, the top-level features are obtained after five stages of the feature extractor, which is utilized for training and testing. Compared results obtained on NWPU-RESISC45 and WHU-RS19 are presented in the Figure 7.

It can be clearly seen that without multi-scale feature fusion, the accuracy drops significantly. On the NWPU-RESISC45 dataset, under the two conditions, the accuracy results of the single-scale feature are 61.04% and 69.52%, respectively, which are 13.93% and 13.60% lower than that of the multi-scale feature fusion model. On the WHU-RS19 data, under the two conditions, the accuracies of the single-scale feature are 64.63% and 73.69%, respectively, which are 11.69% and 10.99% lower than that of the multi-scale feature fusion

model, respectively. The compared results show that the multi-scale feature fusion model can make the features expressing specific information that the single-scale feature do not have, which is able to provide more effective knowledge for scene classification. Therefore, the developed multi-scale feature fusion strategy makes a great effect on boosting the accuracy of few-shot remote sensing scene classification.



(**a**)



(**b**)

Figure 7. Comparisons of different scales of features on two datasets. (**a**) shows the test results on the NWPU-RESISC45 dataset with different scales of features, (**b**) shows the test results on the WHU-RS19 dataset with different scales of features.

5.3. Discussions of Parameters

In MGFF model, the parameters have great influence on the classification results. Thus, it is of great significance to analyze the parameters. There are mainly three types of hyperparameters in the proposed model, including parameters of graph-based feature construction model, parameters of multi-scale graph feature fusion, and parameters of the test condition.

In our graph-based feature construction model, the hyperparameter k determines the number of connections between each node and other nodes in a single graph, which affects the expressiveness of edge features. So as to specifically illustrate the influence caused by the hyperparameter k to the final classification performance, we select a number of different values to conduct experiments under the condition of $N = 64$. The relevant experimental results are presented in the Figure 8, where k is changed from 1 to 14. It can be seen that the settings of different k will cause obvious differences in the classification accuracies, and the maximum difference of the accuracy changing can even reach about 6%. Furthermore, the classification performance will be achieved the best when k is set to 6.

In our multi-scale graph-based feature fusion model, the fusion is conducted based on Equation (9), where there are multiple hyperparameters. These hyperparameters have a direct and important impact on the fusion of features. In order to explore the degree of their influence, we set four different groups of values for α_1 to α_5, including A1($\alpha_1 = 0.25$, $\alpha_2 = 0.25$, $\alpha_3 = 0.25$, $\alpha_4 = 0.25$, $\alpha_5 = 0.25$), A2($\alpha_1 = 0.20$, $\alpha_2 = 0.20$, $\alpha_3 = 0.20$, $\alpha_4 = 0.20$, $\alpha_5 = 0.25$), A3($\alpha_1 = 0.05$, $\alpha_2 = 0.05$, $\alpha_3 = 0.05$, $\alpha_4 = 0.05$, $\alpha_5 = 0.25$), and A4($\alpha_1 = 0.05$, $\alpha_2 = 0.10$, $\alpha_3 = 0.15$, $\alpha_4 = 0.20$, $\alpha_5 = 0.25$). In addition, for β_1 to β_5, we assume that they are all equal and set

12 different values ranging from 0.5 to 3. Experimental results on NWPU-RESISC45 and WHU-RS19 are presented in Figure 9.

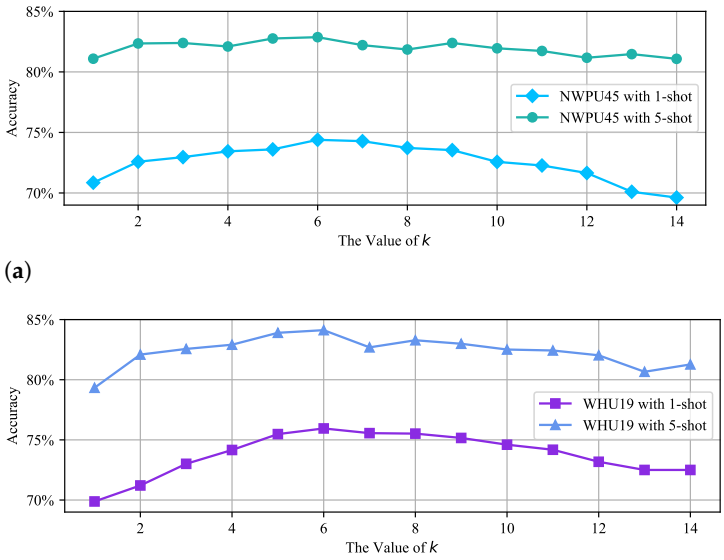


Figure 8. Analysis with different values of k on two datasets. (**a**) shows the test results on the NWPU-RESISC45 dataset with different values of k, (**b**) shows the test results on the WHU-RS19 dataset with different values of k.

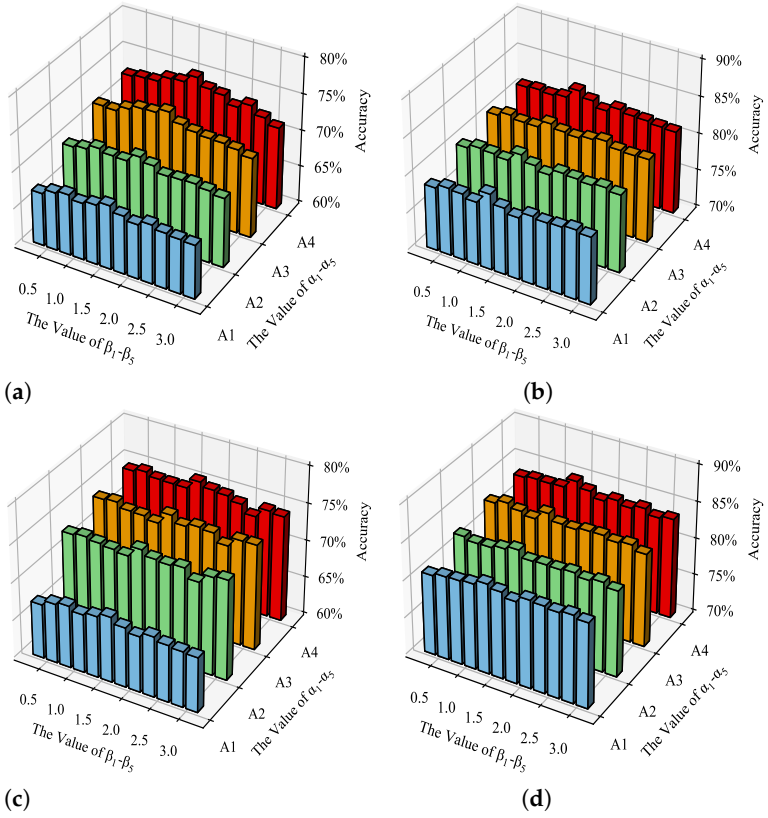


Figure 9. Accuracy with different values of α_i and β_i on two datasets. (**a**,**b**) the results on the NWPU-RESISC45 under the condition of 5-way 1-shot and 5-way 5-shot, respectively. (**c**,**d**) the results on the WHU-RS19 under the condition of 5-way 1-shot and 5-way 5-shot, respectively.

It can be clearly seen from the Figure 9 that α_i and β_i make a great effect on the performance of the model. With different hyperparameter settings, the maximum impact on the accuracy can reach about 8%. For α_1 to α_5, it is obvious that the weight setting of group A4 is optimal compared to the weight setting of other groups. The results verify that different hyperparameters should be set according to the scales of the graph-based features. For β_1 to β_5, it is obvious the optimal parameter setting is 1.5. The results also show that a certain degree of exponentiation is beneficial to improve the classification results.

During the testing, the parameters of the test conditions can illustrate the accuracy of MGFF model on different cases. So as to explore the effect of the test conditions on the few-shot classification results, different numbers of shot are set for analysis, where the results on the two datasets are shown in the Figure 10.

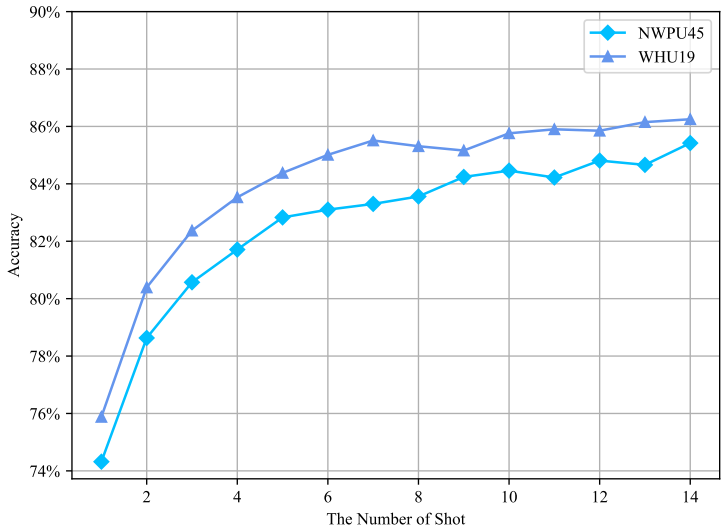


Figure 10. Accuracy with different numbers of shot on two datasets.

From Figure 10, as the shot number increases, the accuracy of the proposed model on two datasets also increases. The growth rate is much faster when the shot number is lower than 5. When the shot number is greater than 5, the growth rate of the accuracy is gradually slowing down. Therefore, it is more appropriate for us to choose the number of 1 and 5 as the main test conditions, which, respectively, reflect the few-shot learning ability of MGFF model in the most extreme condition.

6. Conclusions

In this paper, a multi-scale graph-based feature fusion (MGFF) model is proposed for few-shot remote sensing image scene classification. In MGFF model, the graph-based feature learning is proved to take advantage of relation information for scene classification. Moreover, the graph-based feature fusion model is proposed to integrate graph-based features of multiple scales, which is verified to excavate more abundant semantic information and enhance sample discrimination. Experimental results on two public remote sensing datasets illustrate that the proposed MGFF method can achieve superior accuracy than other advanced few-shot scene classification approaches.

As for few-shot remote sensing image scene classification, there are still some issues worthy of further research. In the remote sensing data, there are still some noise data in the labeled samples. Therefore, it is an important research direction to establish a classification model with strong robustness that overcomes the influence of noise samples.

Author Contributions: Conceptualization, N.J. and J.G.; methodology, N.J. and J.G.; software, N.J. and H.S.; validation, N.J.; formal analysis, J.G.; investigation, N.J. and H.S.; data curation, N.J.; writing—original draft preparation, N.J., H.S. and J.G.; writing—review and editing, N.J., H.S. and J.G.; supervision, J.G. All authors have read and agreed to the published version of the manuscript.

Funding: This research was supported by National Natural Science Foundation of China under Grant 61901376, Project funded by China Postdoctoral Science Foundation under Grant 2021TQ0271 and Grant 2021M700110, and the national undergraduate innovation and entrepreneurship training program.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The NWPU-RESISC45 dataset can be obtained from [58]. The WHU-RS19 dataset can be obtained from [59].

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

References

1. Yao, X.; Han, J.; Cheng, G.; Qian, X.; Guo, L. Semantic annotation of high-resolution satellite images via weakly supervised learning. *IEEE Trans. Geosci. Remote Sens.* **2016**, *54*, 3660–3671. [CrossRef]
2. Cui, Z.; Yang, W.; Chen, L.; Li, H. MKN: Metakernel networks for few shot remote sensing scene classification. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 1–11. [CrossRef]
3. Zhao, B.; Zhong, Y.; Xia, G.S.; Zhang, L. Dirichlet-derived multiple topic scene classification model for high spatial resolution remote sensing imagery. *IEEE Trans. Geosci. Remote Sens.* **2015**, *54*, 2108–2123. [CrossRef]
4. Huang, X.; Wang, Y. Investigating the effects of 3D urban morphology on the surface urban heat island effect in urban functional zones by using high-resolution remote sensing data: A case study of Wuhan, Central China. *ISPRS J. Photogramm. Remote Sens.* **2019**, *152*, 119–131. [CrossRef]
5. Chen, S.W.; Cui, X.C.; Wang, X.S.; Xiao, S.P. Speckle-free SAR image ship detection. *IEEE Trans. Image Process.* **2021**, *30*, 5969–5983. [CrossRef]
6. Xia, G.S.; Hu, J.; Hu, F.; Shi, B.; Bai, X.; Zhong, Y.; Zhang, L.; Lu, X. AID: A benchmark data set for performance evaluation of aerial scene classification. *IEEE Trans. Geosci. Remote Sens.* **2017**, *55*, 3965–3981. [CrossRef]
7. Lowe, D.G. Distinctive image features from scale-invariant keypoints. *Int. J. Comput. Vis.* **2004**, *60*, 91–110. [CrossRef]
8. Oliva, A.; Torralba, A. Modeling the shape of the scene: A holistic representation of the spatial envelope. *Int. J. Comput. Vis.* **2001**, *42*, 145–175. [CrossRef]
9. Yang, Y.; Newsam, S. Bag-of-visual-words and spatial extensions for land-use classification. In Proceedings of the 18th SIGSPATIAL International Conference on Advances in Geographic Information Systems, Beijing, China, 2–5 November 2010; pp. 270–279.
10. Shao, W.; Yang, W.; Xia, G.S. Extreme value theory-based calibration for the fusion of multiple features in high-resolution satellite scene classification. *Int. J. Remote Sens.* **2013**, *34*, 8588–8602. [CrossRef]
11. Negrel, R.; Picard, D.; Gosselin, P.H. Evaluation of second-order visual features for land-use classification. In Proceedings of the 2014 12th International Workshop on Content-Based Multimedia Indexing (CBMI), Klagenfurt, Austria, 18–20 June 2014; IEEE: Piscataway, NJ, USA, 2014; pp. 1–5.
12. Zhao, L.; Tang, P.; Huo, L. A 2-D wavelet decomposition-based bag-of-visual-words model for land-use scene classification. *Int. J. Remote Sens.* **2014**, *35*, 2296–2310. [CrossRef]
13. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. Imagenet classification with deep convolutional neural networks. *Adv. Neural Inf. Process. Syst.* **2012**, *25*. [CrossRef]
14. Simonyan, K.; Zisserman, A. Very deep convolutional networks for large-scale image recognition. *arXiv* **2014**, arXiv:1409.1556.
15. Szegedy, C.; Liu, W.; Jia, Y.; Sermanet, P.; Reed, S.; Anguelov, D.; Erhan, D.; Vanhoucke, V.; Rabinovich, A. Going deeper with convolutions. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 1–9.
16. Penatti, O.A.; Nogueira, K.; Dos Santos, J.A. Do deep features generalize from everyday objects to remote sensing and aerial scenes domains? In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, Boston, MA, USA, 7–12 June 2015; pp. 44–51.
17. Chen, S.W.; Tao, C.S. PolSAR image classification using polarimetric-feature-driven deep convolutional neural network. *IEEE Geosci. Remote Sens. Lett.* **2018**, *15*, 627–631. [CrossRef]
18. Shawky, O.A.; Hagag, A.; El-Dahshan, E.S.A.; Ismail, M.A. Remote sensing image scene classification using CNN-MLP with data augmentation. *Optik* **2020**, *221*, 165356. [CrossRef]
19. Zhang, T.; Liang, J.; Ding, B. Acoustic scene classification using deep CNN with fine-resolution feature. *Expert Syst. Appl.* **2020**, *143*, 113067. [CrossRef]
20. Khan, A.; Chefranov, A.; Demirel, H. Image scene geometry recognition using low-level features fusion at multi-layer deep CNN. *Neurocomputing* **2021**, *440*, 111–126. [CrossRef]
21. Mcilwaine, B.; Casado, M.R. JellyNet: The convolutional neural network jellyfish bloom detector. *Int. J. Appl. Earth Obs. Geoinf.* **2021**, *97*, 102279. [CrossRef]
22. Gidaris, S.; Bursuc, A.; Komodakis, N.; Pérez, P.; Cord, M. Boosting few-shot visual learning with self-supervision. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Korea, 27 October–2 November 2019; pp. 8059–8068.

23. Chu, W.H.; Li, Y.J.; Chang, J.C.; Wang, Y.C.F. Spot and learn: A maximum-entropy patch sampler for few-shot image classification. In Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 6251–6260.
24. Cheng, G.; Han, J.; Guo, L.; Liu, Z.; Bu, S.; Ren, J. Effective and efficient midlevel visual elements-oriented land-use classification using VHR remote sensing images. *IEEE Trans. Geosci. Remote Sens.* **2015**, *53*, 4238–4249. [CrossRef]
25. Chen, S.W. SAR image speckle filtering with context covariance matrix formulation and similarity test. *IEEE Trans. Image Process.* **2020**, *29*, 6641–6654. [CrossRef]
26. Wang, S.; Wang, X.; Zhang, L.; Zhong, Y. Auto-AD: Autonomous hyperspectral anomaly detection network based on fully convolutional autoencoder. *IEEE Trans. Geosci. Remote Sens.* **2021**, *60*, 1–14. [CrossRef]
27. Zhu, S.; Du, B.; Zhang, L.; Li, X. Attention-based multiscale residual adaptation network for cross-scene classification. *IEEE Trans. Geosci. Remote Sens.* **2021**, *60*, 1–15. [CrossRef]
28. Xu, C.; Zhu, G.; Shu, J. A lightweight and robust lie group-convolutional neural networks joint representation for remote sensing scene classification. *IEEE Trans. Geosci. Remote Sens.* **2021**, *60*, 1–15. [CrossRef]
29. Wang, X.; Wang, S.; Ning, C.; Zhou, H. Enhanced feature pyramid network with deep semantic embedding for remote sensing scene classification. *IEEE Trans. Geosci. Remote Sens.* **2021**, *59*, 7918–7932. [CrossRef]
30. Lu, X.; Zheng, X.; Yuan, Y. Remote sensing scene classification by unsupervised representation learning. *IEEE Trans. Geosci. Remote Sens.* **2017**, *55*, 5148–5157. [CrossRef]
31. Cheng, G.; Xie, X.; Han, J.; Guo, L.; Xia, G.S. Remote sensing image scene classification meets deep learning: Challenges, methods, benchmarks, and opportunities. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2020**, *13*, 3735–3756. [CrossRef]
32. Lu, X.; Gong, T.; Zheng, X. Multisource compensation network for remote sensing cross-domain scene classification. *IEEE Trans. Geosci. Remote Sens.* **2019**, *58*, 2504–2515. [CrossRef]
33. Nogueira, K.; Penatti, O.A.; Dos Santos, J.A. Towards better exploiting convolutional neural networks for remote sensing scene classification. *Pattern Recognit.* **2017**, *61*, 539–556. [CrossRef]
34. Cheng, G.; Yang, C.; Yao, X.; Guo, L.; Han, J. When deep learning meets metric learning: Remote sensing image scene classification via learning discriminative CNNs. *IEEE Trans. Geosci. Remote Sens.* **2018**, *56*, 2811–2821. [CrossRef]
35. Wang, Q.; Liu, S.; Chanussot, J.; Li, X. Scene classification with recurrent attention of VHR remote sensing images. *IEEE Trans. Geosci. Remote Sens.* **2018**, *57*, 1155–1167. [CrossRef]
36. Tang, X.; Lin, W.; Ma, J.; Zhang, X.; Liu, F.; Jiao, L. Class-Level Prototype Guided Multiscale Feature Learning for Remote Sensing Scene Classification With Limited Labels. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 1–15. [CrossRef]
37. Pires de Lima, R.; Marfurt, K. Convolutional neural network for remote-sensing scene classification: Transfer learning analysis. *Remote Sens.* **2019**, *12*, 86. [CrossRef]
38. Zeng, Q.; Geng, J. Task-specific contrastive learning for few-shot remote sensing image scene classification. *ISPRS J. Photogramm. Remote Sens.* **2022**, *191*, 143–154. [CrossRef]
39. Vinyals, O.; Blundell, C.; Lillicrap, T.; Wierstra, D. Matching networks for one shot learning. *Adv. Neural Inf. Process. Syst.* **2016**, *29*, 3637–3645.
40. Snell, J.; Swersky, K.; Zemel, R. Prototypical networks for few-shot learning. *Adv. Neural Inf. Process. Syst.* **2017**, *30*, 4080–4090.
41. Sung, F.; Yang, Y.; Zhang, L.; Xiang, T.; Torr, P.H.; Hospedales, T.M. Learning to compare: Relation network for few-shot learning. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 1199–1208.
42. Zhang, X.; Qiang, Y.; Sung, F.; Yang, Y.; Hospedales, T. RelationNet2: Deep comparison network for few-shot learning. In Proceedings of the 2020 International Joint Conference on Neural Networks (IJCNN), Glasgow, UK, 19–24 July 2020; IEEE: Piscataway, NJ, USA, 2020; pp. 1–8.
43. Alajaji, D.; Alhichri, H.S.; Ammour, N.; Alajlan, N. Few-shot learning for remote sensing scene classification. In Proceedings of the 2020 Mediterranean and Middle-East Geoscience and Remote Sensing Symposium (M2GARSS), Tunis, Tunisia, 9–11 March 2020; IEEE: Piscataway, NJ, USA, 2020; pp. 81–84.
44. Jiang, W.; Huang, K.; Geng, J.; Deng, X. Multi-scale metric learning for few-shot learning. *IEEE Trans. Circuits Syst. Video Technol.* **2020**, *31*, 1091–1102. [CrossRef]
45. Cheng, G.; Cai, L.; Lang, C.; Yao, X.; Chen, J.; Guo, L.; Han, J. SPNet: Siamese-prototype network for few-shot remote sensing image scene classification. *IEEE Trans. Geosci. Remote Sens.* **2021**, *60*, 1–11. [CrossRef]
46. Zeng, Q.; Geng, J.; Jiang, W.; Huang, K.; Wang, Z. Idln: Iterative distribution learning network for few-shot remote sensing image scene classification. *IEEE Geosci. Remote Sens. Lett.* **2021**, *19*, 1–5. [CrossRef]
47. Li, H.; Cui, Z.; Zhu, Z.; Chen, L.; Zhu, J.; Huang, H.; Tao, C. RS-MetaNet: Deep meta metric learning for few-shot remote sensing scene classification. *arXiv* **2020**, arXiv:2009.13364.
48. Li, L.; Han, J.; Yao, X.; Cheng, G.; Guo, L. DLA-MatchNet for few-shot remote sensing image scene classification. *IEEE Trans. Geosci. Remote Sens.* **2020**, *59*, 7844–7853. [CrossRef]
49. Zhai, M.; Liu, H.; Sun, F. Lifelong learning for scene recognition in remote sensing images. *IEEE Geosci. Remote Sens. Lett.* **2019**, *16*, 1472–1476. [CrossRef]
50. Koch, G.; Zemel, R.; Salakhutdinov, R. Siamese neural networks for one-shot image recognition. In Proceedings of the ICML Deep Learning Workshop, Lille, France, 6–11 July 2015; Volume 2.

51. Kim, J.; Kim, T.; Kim, S.; Yoo, C.D. Edge-labeling graph neural network for few-shot learning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 16–17 June 2019; pp. 11–20.
52. Gidaris, S.; Komodakis, N. Generating classification weights with gnn denoising autoencoders for few-shot learning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 21–30.
53. Hamilton, W.L.; Ying, R.; Leskovec, J. Representation learning on graphs: Methods and applications. *arXiv* **2017**, arXiv:1709.05584.
54. Kipf, T.N.; Welling, M. Semi-supervised classification with graph convolutional networks. *arXiv* **2016** arXiv:1609.02907.
55. Veličković, P.; Cucurull, G.; Casanova, A.; Romero, A.; Lio, P.; Bengio, Y. Graph attention networks. *arXiv* **2017**, arXiv:1710.10903.
56. Wu, F.; Souza, A.; Zhang, T.; Fifty, C.; Yu, T.; Weinberger, K. Simplifying graph convolutional networks. In Proceedings of the International Conference on Machine Learning, PMLR, Long Beach, CA, USA, 9–15 June 2019; pp. 6861–6871.
57. Yuan, Z.; Huang, W.; Tang, C.; Yang, A.; Luo, X. Graph-Based Embedding Smoothing Network for Few-Shot Scene Classification of Remote Sensing Images. *Remote Sens.* **2022**, *14*, 1161. [CrossRef]
58. Cheng, G.; Han, J.; Lu, X. Remote Sensing Image Scene Classification: Benchmark and State of the Art. *Proc. IEEE* **2017**, *105*, 1865–1883. [CrossRef]
59. Sheng, G.; Yang, W.; Xu, T.; Sun, H. High-resolution satellite scene classification using a sparse coding based multiple feature combination. *Int. J. Remote Sens.* **2012**, *33*, 2395–2412. [CrossRef]
60. Li, X.; Shi, D.; Diao, X.; Xu, H. SCL-MLNet: Boosting Few-Shot Remote Sensing Scene Classification via Self-Supervised Contrastive Learning. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 1–12. [CrossRef]
61. Li, Z.; Zhou, F.; Chen, F.; Li, H. Meta-sgd: Learning to learn quickly for few-shot learning. *arXiv* **2017**, arXiv:1707.09835.
62. Liu, Y.; Lee, J.; Park, M.; Kim, S.; Yang, E.; Hwang, S.J.; Yang, Y. Learning to propagate labels: Transductive propagation network for few-shot learning. *arXiv* **2018**, arXiv:1805.10002.
63. Finn, C.; Abbeel, P.; Levine, S. Model-agnostic meta-learning for fast adaptation of deep networks. In Proceedings of the International Conference on Machine Learning, PMLR, Sydney, Australia, 6–11 August 2017; pp. 1126–1135.