Few-Shot Remote Sensing Scene Classification via Subspace Based on Multiscale Feature Learning

Anyong Qin, Fuyang Chen, Qiang Li, Tiecheng Song, Yu Zhao, and Chenqiang Gao

***Abstract*—Because of the challenges associated with the difficulty of accurately labeling the remote sensing (RS) scene images and the need to identify new scene classes, few-shot learning has shown significant advantages in addressing the remote sensing scene classification (RSSC) tasks, leading to a growing interest. However, due to the scale variations of targets and irrelevant complex background in scene images, the current few-shot methods exist the following problems: the problem of the extraction capability of feature extractor in the few-shot mechanism; the problem of the separability of few-shot RS scene images classifier. To solve the above problems, an approach, called few-shot RSSC via subspace based on multiscale feature learning is introduced in this work. We first design a multiscale feature learning technique to address scale variations of the targets in the scene images. Concretely, different branches are utilized to learn scene features at various scales. The self-attention mechanism is embedded in each branch to incorporate the understanding of the global information in the different scale features. After that, a multiscale feature fusion operation, incorporating channel attention, will be devised to effectively merge the different scale features, so as to obtain a more precise feature representation of RS scene images. Furthermore, the subspace is utilized to capture the shared characteristics of each category, to reduce the impact of the complex irrelevant backgrounds in the scene images. The results of our experiments conducted on the public available RS scene datasets demonstrate the strong competitiveness of our approach.**

***Index Terms*—Attention mechanism, few-shot learning, multiscale, remote sensing scene classification (RSSC), subspace classifier.**

I. Introduction

REMOTE sensing scene classification (RSSC) has aroused growing interest as a crucial technology for accurately interpreting the remote sensing (RS) images. The goal of RSSC is to assign a distinct category label to each RS image according to its content, enabling differentiation of various scenes [1]. It has wide applications in diverse fields such as urban planning [2], disaster detection [3], geospatial object detection [4], land cover classification [5], etc., [6], [7], [8].

Recently, machine learning, which has been extensively explored by many researchers, has become the popular approach for solving RS tasks [9], [10], [11]. In the fields of RSSC, deep learning techniques utilizing convolutional neural networks (CNNs) to extract image features have demonstrated promising outcomes. Penatti et al. [12] were the pioneer to employ the CNNs for the RSSC tasks in 2015. Subsequently, a succession of the CNN-based methodologies have appeared, and all exhibiting better classification performance [7], [13]. While the deep learning techniques have indicated competitive results for RSSC tasks, these techniques still demand a substantial amount of annotated scenes for model training [14], [15]. If the quantity of labeled scene images is insufficient, overfitting may arise, leading to a decrease in classification performance [16]. Currently, due to the rapid advancements in RS technology, acquiring a vast quantity of scene images is no longer difficult. However, labeling these scene images is highly challenging, needing substantial financial investment and expertise in the corresponding field. Therefore, traditional deep learning frameworks exhibit significant limitations due to the lack of annotated scene images.

Generative adversarial networks (GANs) generate additional scene images to alleviate the problem of insufficient annotated RS scene images, but the authenticity of generated RS scene images is low and the GAN models are complex [17], [18]. Furthermore, both deep learning methods and GANs face a common problem: they can only classify categories seen during the training process and cannot identify new unknown categories [19]. However, the real world is open and the new scene categories will continue to appear, thus classifying the scene images of these new categories is a problem that the above methods cannot solve [1], [20]. So the insufficient labeled scene images and the need to identify new scene categories both pose new challenges for the RSSC. A method that can classify the new category samples with limited number of labeled scene images carry significant importance for the RSSC task [1], [20].

Under this background, few-shot learning (FSL) techniques receive the attention of researchers, and have a wide range of applications in many fields such as image classification [21], [22], detection [23], and segmentation [24], [25], [26], [27]. FSL classification, just needing a small quantity of annotated

Manuscript received 5 May 2024; revised 17 June 2024; accepted 20 July 2024. Date of publication 24 July 2024; date of current version 5 August 2024. This work was supported in part by the National Key R&D Program of China under Grant 2022YFA1004100, in part by the National Natural Science Foundation of China under Grant 62201111, in part by the Science and Technology Research Program of Chongqing Municipal Education Commission under Grant KJQN202100646, in part by China Postdoctoral Science Foundation under Grant 2024T171109 and 2022MD713684, and in part by the Special Funding for Chongqing Postdoctoral Research Project under Grant 2022CQBSHTB3086. *(Corresponding author: Yu Zhao.)*

Anyong Qin, Fuyang Chen, Qiang Li, Tiecheng Song, and Yu Zhao were with the School of Communications and Information Engineering, Chongqing University of Posts and Telecommunications, Chongqing 400065, China (e-mail: qinay@cqupt.edu.cn; zhaoyu@cqupt.edu.cn).

Chenqiang Gao was with the School of Intelligent Systems Engineering, Sun Yat-sen University, Shenzhen 518107, China (e-mail: gaochq6@mail.sysu.edu.cn).

Digital Object Identifier 10.1109/JSTARS.2024.3432895

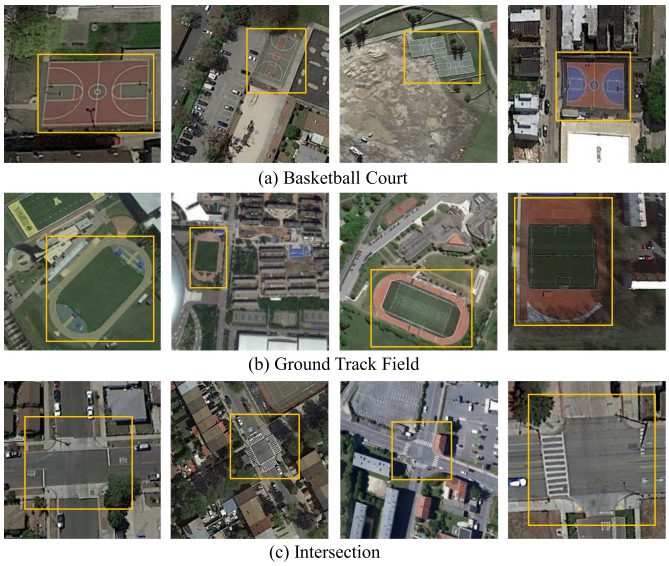


Fig. 1. Schematic diagram of the target objects scale variations and category-unrelated objects. In each image, the portion marked with a yellow box is the target object, and the unboxed portion are the category-unrelated objects. (a) Basketball Court. (b) Ground Track Field. (c) Intersection.

data, enable models to acquire prior knowledge from base class data to categorize new class data [19], [28], [29]. Currently, FSL techniques have indicated superior capability in the fields of natural image classification. However, since RS scene images are often acquired from the air by remote sensing sensors (e.g., satellites, aircrafts, or other drones), it makes the scene images be taken at a longer distance and with a larger range of viewpoints. As a result, compared with the natural images captured from a horizontal perspective and closer distance, the target objects in remote sensing scene images of the same category show large-scale variations, as shown in Fig. 1. Moreover, the large range of viewpoints makes remote sensing scene images to contain many irrelevant and complex backgrounds in addition to the target objects. As shown in Fig. 1(a), the images with the category "Basketball Court" contain other irrelevant objects, such as vegetation, roads, and houses. Fig. 1(b) and (c) also contain other ground objects in addition to ground track field and intersection. These characteristics make it more difficult to improve the extraction capability of feature extractor in few-shot mechanism and the separability of few-shot RS scene images classifier.

The few-shot feature learning methods for natural images face the challenge of large-scale variations of target objects when processing remote sensing scene images. The scale variation leads to the fact that existing feature extraction methods in few-shot framework cannot effectively capture the key information of objects at different scales [30]. While these easily neglected objects may be the key areas directly related to the class of the RS images, which can result in failure to obtain critical information of the target objects, further affecting the accuracy of the RS scene image feature representation [31], [32], [33]. Therefore, a feature extractor that can adapt to multiscale target objects in scene images is crucial for the few-shot remote sensing scene classification (FSRSSC) task.

Aiming at the extraction capability problem of the feature extractor in the FSRSSC framework, this study proposes a multiscale feature learning approach to promote feature extraction network. The proposed feature learning method not only captures the detailed and global information of the RS scene images under different receptive fields, but also considers the important influence of the global information in the scene images on feature learning. By incorporating the self-attention mechanisms into the branches of different scales, and understanding the global information in the scene images at different scales, the important features of the RS scene images will be enhanced. Furthermore, a feature fusion module based on channel attention is employed to highlight the semantic features of key channels in a complementary manner at both global and local levels, achieving effective fusion of information at different scales and obtaining more discriminative representations of the scene images.

In addition, the irrelevant and complex background in remote sensing scene images poses a challenge for the separability of few-shot RS scene images classifiers. The existing prototype networks (ProtoNet) [34] and its various variants that compute class prototypes by averaging the support features of each class have received widespread attention. Subsequently, according to the distance between the query samples and each class prototype, ProtoNet can achieve the ability to classify all query samples. However, the above methods fail to consider the adverse effects of the irrelevant complex background presented in RS scene images [35], [36], [37], [38]. As a result, the information of these irrelevant background is finally preserved by the prototype classifiers, diminishing the representation capability of the class prototype.

Compared to the ProtoNet that directly averages the feature representations, the subspace-based approach approximates the class representation as a set of linearly independent basis vectors, i.e., feature subspace. This subspace is abstracted from the feature space by singular value decomposition (SVD) [39]. The feature vectors obtained by SVD can better fit the subspace of the scene images. The most important category information extracted by our proposed multiscale feature learning from the scene images will be extracted into the leading eigenvectors, while other unrelated background information will be discarded [40], [41]. The leading eigenvectors with the category information can form a more accurate subspace. Therefore, our few-shot classifier takes advantage of class subspace as metric benchmark, and can capture the common representation of each scene category, thus mitigating the effect of irrelevant and complex backgrounds.

In summary, to solve the extraction capability problem of feature extractor and the separability problem of few-shot classifier for the FSRSSC task, this work proposes an FSRSSC approach via subspace based on multiscale feature learning. The primary contributions of the work are summarized as follows:

1) To handle the problem of scale variations for the scene images, different scale feature extraction branches based

on self-attention mechanisms are proposed to emphasize the crucial features in the whole RS scene image.
2) To enhance the semantic information of important channels within each scale feature, the proposed multiscale feature fusion module based on channel attention can combine the global and local features, to improve feature accuracy in the FSRSSC task.
3) To mitigate the adverse impact of irrelevant complex background in RS scene images on classification accuracy, a few-shot classifier utilizing subspace as the metric is proposed to fully learn the common representation for each scene class.

II. Related Work

RSSC is to use a learning algorithm to determine the class of scene image. This section will delve into the advancements made by researchers in the areas of RSSC and FSL.

A. Remote Sensing Scene Classification

Early on, RSSC relied heavily on manually designed features for classification. Some representative handcrafted features include color histograms, texture descriptors, generalized search trees (GIST) features, histogram of oriented gradient (HOG) features, and scale invariant feature transform (SIFT) features [42]. Approaches based on the handcrafted features offer stability, strong interpretability, and effectiveness in handling small-scale data. However, the handcrafted features suffer from subjectivity as they are often designed for specific tasks and datasets, that makes the generalization to other datasets and tasks challenging. Furthermore, the classification performance of methods based on handcrafted features tends to degrade when dealing with large-scale datasets and complex scene datasets. Therefore, due to these drawbacks, approaches based on handcrafted features are no longer meeting current RSSC tasks.

Deep learning techniques, especially the CNNs, have been widely applied in RSSC task, yielding favorable results. Compared to traditional classification approaches, CNNs can offer end-to-end processing and have the ability to learn high-level semantic features, which are unattainable by handcrafted features. Penatti et al. [12] first proposed using CNNs for RSSC task, demonstrating that pretrained CNNs can recognize natural objects and generalize well to scene images. Chaib et al. [43] introduced a feature refinement approach using conspicuity relation analysis, which enhances the classification effect by correcting the original features. Liu et al. [44] presented a multiscale CNN framework for addressing varying target scales in RS scene images. Lu et al. [45] designed a feature aggregation network that integrates image-related information using semantic label, enhancing the classification performance. Wang et al.[32] proposed a global–local dual-path network structure to learn multiscale feature representations of RS scene images, improving classification accuracy by capturing more representative features. Sun et al. [46] introduced a gated bidirectional network for RS scene, which eliminates interference and aggregates information from different CNN layers. He et al. [47] designed a skip-connected covariance network with fewer parameters, enabling better feature representation through feature combination and higher order information extraction from feature maps.

Despite the satisfactory classification results obtained using these CNN-based RSSC approaches, they typically require large amounts of labeled scene images for training, leading to poor performance when the labeled scene images per category are only a few (such as one or five). The insufficient labeled scene images can result in incomplete extraction of critical features from small target, which are often overlooked but directly related to the semantic class of scene images. Moreover, the existence of target objects scale variations in the scene images further affects the accuracy of extracting important features. In addition, all these approaches lack the ability to recognize new RS scene categories, which is unavoidable in real-world applications.

B. Few-Shot Learning

The goal of FSL differs from that of supervised learning, such as CNNs. In FSL, the aim is not to train a model to recognize data within the training set and then generalize to the test set. Instead, the goal is to enable the model to learn for itself and acquire prior knowledge, which can then be used to classify new data categories with only a small number of labeled data [19]. The idea of FSL is particularly useful in addressing the limitations of deep learning techniques in RSSC task [1], [35], [36], [37].

Measure-based FSL approaches first extract image features using a predesigned feature extractor and then calculate the distance between query and support images using a distance metric. The distance metric can be nonparametric, such as Euclidean distance or cosine distance, or parametric and learnable [48]. Matching networks (MatchingNet) use separate feature encoders for the query and support sets, mapping each to a corresponding vector representation. The distance between the query image and each support image representation is then calculated for classification [49]. Prototypical networks (ProtoNet) improve the MatchingNet by comparing query images with a class prototype instead of individual support images, which is computed based on the support set information [34].

However, RS scene images often contain complex background information (such as many category-unrelated objects). Prototype classifiers may retain excess irrelevant information from the scene images, then leading to poor classification performance due to the adverse effects of the complex background information.

III. Methodology

This study proposes a novel FSRSSC approach based on subspace and multiscale feature learning (MSFL), as shown in Fig. 2. We propose different scale feature extraction branches based on self-attention mechanisms to address the challenge of large-scale variations in RS scene images. To obtain more precise representations for the RS scene images, we further propose a multiscale feature fusion module based on channel attention, which can join the local and global feature representation of the scene images. In addition, for the FSRSSC, we also construct a few-shot classifier that makes use of feature subspace as

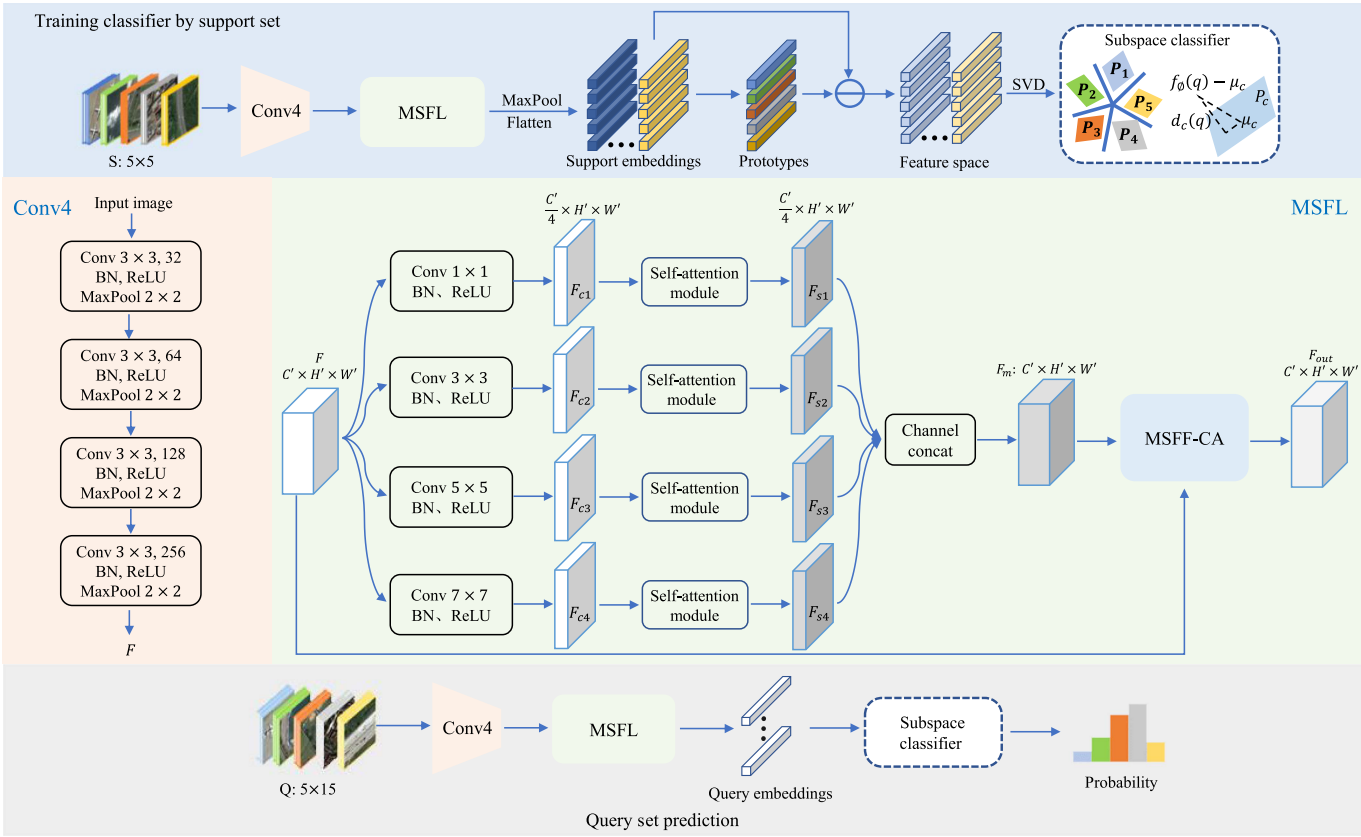


Fig. 2. Architecture of the multiscale feature learning-based network for FSRSSC via subspace. S (5×5: 5-way 5-shot) and Q (5×15: 5-way 15-query images), respectively, denote the support set and query set per task. Conv4 denotes the feature extractor including four convolutional blocks. The MSFL represents the proposed network of multiscale feature learning. The MSFF-CA represents the multiscale feature fusion scheme based on channel attention.

benchmark to alleviate adverse the impact of unrelated complex backgrounds in RS images.

A. Problem Definition

In this study, according to the same category divisions as in [35], [36], and [50], we partition each RS scene dataset D into three sets, including testing set D_{test}, training set D_{train}, and validation set D_{val}. The scene categories in D_{train}, D_{val}, and D_{test} form the class sets C_{train}, C_{val}, and C_{test}, respectively. Specifically, C_{train}, C_{val}, and C_{test} are distinct and nonoverlapping, that is, $C_{\text{train}} \cap C_{\text{val}} = \emptyset$, $C_{\text{train}} \cap C_{\text{test}} = \emptyset$, $C_{\text{val}} \cap C_{\text{test}} = \emptyset$, $C_{\text{train}} \cup C_{\text{val}} \cup C_{\text{test}} = C_{\text{total}}$. In this work, the episode-based strategy is adopted to train, validate, and test. Hence, it is necessary to generate a sequence of tasks. The tasks constructed from training set, validation set, and test set are called training tasks, validation tasks, and test tasks, respectively. Taking the construction of a training task as an example, firstly, N distinct categories are randomly selected from the set C_{train}. Then, we select $K + M$ scene images per category, with K images designated for support set and remaining M scene images for query set. It is important to note that for each task, the remote sensing scene images in the query set will not appear in the support set. This task setup is referred to as an N-way K-shot task. Specifically, each task consists of a support set $S = \{(x_{1,1}, y_{1,1}), (x_{1,2}, y_{1,2}), \ldots, (x_{N,K}, y_{N,K})\}$, and query set $Q = \{q_1, q_2, \ldots, q_{N\times M}\}$, where $x_{i,j}$ represents the jth scene image being from the ith class, $y_{i,j} \in \{0, 1, 2, \ldots, N-1\}$ denotes label of support image $x_{i,j}$, and q_l denotes the lth query image. Validation tasks and test tasks are built in the same way.

B. Method Overview

This work proposes a FSRSSC network that utilizes multiscale feature learning and subspace. Our framework is displayed in Fig. 2.

Our feature learning method retains the first four convolution blocks (Conv4) of the Conv5, and then replaces the fifth convolution block in Conv5 with multiscale feature learning. First, Conv4 is employed to perform the initial feature extraction of the RS scene images. Second, four different scale feature learning branches based on self-attention mechanisms are designed. Specifically, each branch performs convolution operations with convolution kernels of different sizes, and incorporates the self-attention mechanism to merge global information into the features with different scales, aiming to enhance feature representations of important regions in RS scene images. Subsequently, a channel attention-based multiscale feature fusion scheme is devised to effectively integrate information

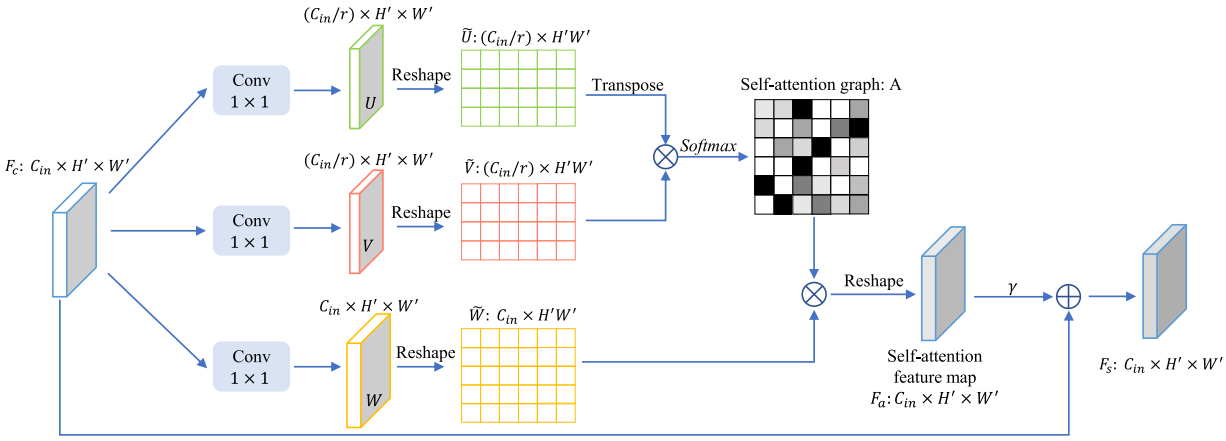


Fig. 3. Self-attention module.

from different scales, resulting in accurate feature embedding for each scene image. Finally, a subspace classifier is used for the few-shot scene image classification.

C. Multiscale Feature Learning

The proposed multiscale feature learning takes the feature map obtained from Conv4 as its input, as illustrated in Fig. 2. Conv4 is comprised of four convolution blocks, each of which consists of 3×3 convolution, a normalized, a ReLU, and a 2×2 max pooling layer. About 32, 64, 128, and 256 are the output channels of the four convolution blocks, respectively [35]. The input RS scene image is denoted by the symbol $x_i \in \Re^{C \times H \times W}$, where C represents the number of the RS image channels and the size of RS scene images represents as $H \times W$. Through the four convolution blocks (Conv4), a feature map can be obtained

$$F = g_\varphi(x_i; \varphi) \tag{1}$$

where $F \in \Re^{C' \times H' \times W'}$, g_φ represents the Conv4, φ are the parameters of the network, and x_i represents the input image.

Based on the feature map F mentioned above, we employ the proposed multiscale feature learning (MSFL) method to acquire a precise representation of the RS scene image. This method mainly consists of different scale feature extraction branches based on self-attention mechanisms, as well as a channel attention-based multiscale feature fusion module (MSFF-CA). The following is a detailed introduction to the proposed multiscale feature learning method.

1) Scale Branch Based on Self-Attention Mechanism: Each branch continues to convolve the feature map with convolutional kernels of different sizes, ensuring that each branch can capture the information from different scales. In addition, a self-attention module is embedded within each branch, aiming to catch the relationships among various regions within the RS scene images by incorporating an understanding of the global information into each scale feature. So, the important regional information from each scale feature could be emphasized.

Specifically, each branch comprises a convolutional, a normalized, and a ReLU layer. For target objects of different scales in the scene images, we also use commonly used convolution kernels (3×3, 5×5, and 7×7) to capture the key feature information. Although a 9×9 convolution kernel can provide a larger receptive field, stacking multiple smaller convolution kernels can achieve similar effects with fewer parameters and computation. In addition, the main function of the 1×1 convolution kernel is to achieve interaction and information integration between different channels. Moreover, in our experiments, the size of feature map obtained by the Conv4 (3×3) is $256 \times 8 \times 8$, that is not suitable for 9×9 convolution kernel. So, we set four scale branches and adopt 1×1 Conv, 3×3 Conv, 5×5 Conv, and 7×7 Conv, respectively.

As shown in Fig. 2, in order to effectively fuse multiscale features, the number of output channels for each scale branch is set as $C'/4$. The intermediate feature maps obtained by the above convolution in each branch are represented as F_{c1}, F_{c2}, F_{c3}, and F_{c4} (both $\in \Re^{(C'/4) \times H' \times W'}$). Subsequently, the self-attention module used in each branch is shown in Fig. 3. Here, to introduce the self-attention mechanism, the input to this module is denoted as $F_c \in \Re^{C_{in} \times H' \times W'}$.

First, the feature map F_c will be separately converted into two distinct feature spaces U and V, which is achieved by the pointwise convolution, i.e., 1×1 convolution, with input channels $C_{in} = C'/4$ and output channels C_{in}/r ($r = 8$ is a scale factor, which is used to reduce the number of channels [51].). Then, U and V are reshaped along the spatial dimensions to obtain $\widetilde{U}$ and $\widetilde{V}$. Then the relationships between positions in the space can be obtained

$$R = \widetilde{U}^T \widetilde{V} \tag{2}$$

where $R \in \Re^{H'W' \times H'W'}$, R_{ij} denotes correlation between the ith pixel position and the jth pixel position in the feature space, the ith row of $\widetilde{U}^T \in \Re^{H'W' \times (C_{in}/r)}$ denotes values of all channels at the ith pixel position in the space; the jth column of $\widetilde{V} \in \Re^{(C_{in}/r) \times H'W'}$ denotes values of all channels at the jth pixel position in the space.

The attention map can be calculated according to the formulas

$$A = \text{Softmax}(R) \tag{3}$$

$$A_{ij} = \frac{\exp(R_{ij})}{\sum_{j=1}^{H'W'} \exp(R_{ij})} \tag{4}$$

where Softmax$(\cdot)$ represents rowwise normalization, A_{ij} represents the attention weight of the jth pixel position on the ith pixel position, $\sum_{j=1}^{H'W'} A_{ij} = 1$.

In the same way, the feature map F_c is also converted into the feature spaces W by the pointwise convolution, i.e., 1×1 convolution, with input channels C_{in} and output channels C_{in}. Then, W is reshaped along the spatial dimensions to obtain $\widetilde{W}$. After that self-attention feature map can be computed via

$$F_a = \text{reshape}(\widetilde{W} A^T) \tag{5}$$

where reshape$(\cdot)$ denotes dimension transformation, reshape$(\cdot)$: $\Re^{C_{in} \times H'W'}$ to $\Re^{C_{in} \times H' \times W'}$, the ith column of $\widetilde{W} \in \Re^{C_{in} \times H'W'}$ denotes values of all channels at the ith pixel position in the space.

Finally, the output of the self-attention module is represented as

$$F_s = \gamma F_a + F_c \tag{6}$$

where $F_s \in \Re^{C_{in} \times H' \times W'}$, γ represents the learnable scale factor, F_a represents the self-attention feature map, and its input is represented as F_c.

In the four different scale feature learning branches, the self-attention module maintains a consistent structure, but the parameters are not shared. Through the above operations, four feature maps with different scale information are obtained, denoted as F_{s1}, F_{s2}, F_{s3}, and F_{s4} (both $\in \Re^{C_{in} \times H' \times W'}$). Then, we concat F_{s1}, F_{s2}, F_{s3}, and F_{s4} along channel dimension to gain multiscale representation F_m, $F_m \in \Re^{C' \times H' \times W'}$.

2) Feature Fusion Based on Channel Attention: To prevent the information loss during the extraction of different scale features and enhance the semantic information of important channels in each scale feature map, we adopt a channel attention-based multiscale feature fusion module to effectively combine the feature maps F and F_m, as shown in Fig. 4. Here, F represents the output of Conv4, and F_m represents the result of concatenating different scale feature maps.

The feature fusion module first adds F and F_m by the elementwise. Then, we design two branches from the perspectives of global information and local information. One branch first obtains global information through global average pooling, and then uses pointwise convolution to dynamically learn the channel attention of global features. The other branch directly uses pointwise convolution to dynamically learn the channel attention of local features. The processes of these two branches can be represented by the following formulas:

$$G = BN(\text{Conv}_2(ReLU(BN(\text{Conv}_1(GAP(F + F_m)))))) \tag{7}$$

$$L = BN(\text{Conv}_2(ReLU(BN(\text{Conv}_1(F + F_m))))) \tag{8}$$

where G represents the channel attention weights for global features, L represents the channel attention weights for local features, $GAP(\cdot)$ stands for the global average pooling operation, $\text{Conv}_1(\cdot)$ and $\text{Conv}_2(\cdot)$ represents pointwise convolution

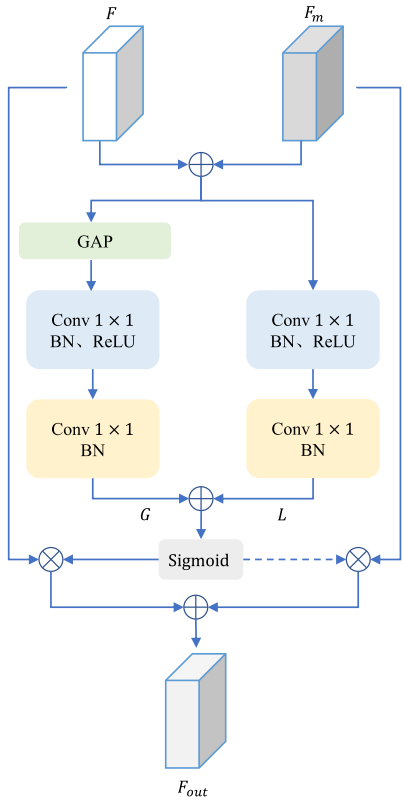


Fig. 4. Multiscale feature fusion module based on channel attention.

with different output channel numbers. $\text{Conv}_1(\cdot)$ reduces the number of channels and then $\text{Conv}_2(\cdot)$ restore it to the original size. The pointwise convolution in the two branches is the same, but they do not share parameters.

Based on the channel attention weights for global features and local features, the output of the multiscale feature fusion is

$$F_{\text{out}} = F \otimes \text{Sigmoid}(G \oplus L) + F_m \otimes (1 - \text{Sigmoid}(G \oplus L)) \tag{9}$$

where F_{out} also represents the output of the multiscale feature learning method, F represents the output of Conv4, F_m represents the result of concatenating different scale feature maps, $\otimes$ represents elementwise multiplication, $\oplus$ represents elementwise addition. Since G and L have different dimensions, the broadcast operation on G is required when performing $\oplus$.

The multiscale feature fusion based on channel attention combines global and local information for adaptive feature fusion, which can enhance the understanding of semantic scenes and thus learn more significant image features. In order to shrink the dimension of the feature map, a max pooling layer is employed following the multiscale feature learning method. Then, the pooled features are flattened into vectors form as the embedded features corresponding to the RS image, which can be represented by

$$Z = \text{Flatten}(\text{MaxPool}(F_{\text{out}})) \tag{10}$$

where Flatten$(\cdot)$ denotes flattening the feature map into a vector and MaxPool$(\cdot)$ represents a 2×2 max pooling operation.

D. Subspace Classifier

In order to improve the separability of classifier for the FSRSSC task, our few-shot classifier that takes advantage of the high-order information within the representation space is proposed to capture the commonality feature of each class, reducing the adverse effects of irrelevant complex backgrounds. Given an *N*-way K-shot task for RS scene image classification, the feature representations for the support RS images and query RS images can be obtained through formulas

$$Z_{c,i}^{sup} = E(x_{c,i}; \varphi, \psi) \tag{11}$$

$$Z_{l}^{que} = E(q_l; \varphi, \psi) \tag{12}$$

where $E(\cdot)$ represents the above feature extraction network, mainly consisting of the Conv4 and multiscale feature learning, with network parameters denoted as φ and ψ, $Z_{c,i}^{\text{sup}}$ represents the feature embedding of the i-support scene image ($x_{c,i}$) from the c-class, Z_{l}^{que} represents the feature embedding of the l-query scene image (q_l).

Subsequently, by leveraging feature representations learned from the support scene images for each category within a given task, the corresponding subspace can be obtained, enabling the construction of the subspace classifier. For instance, let us consider the cth class, we will give the procedure for obtaining its subspace in detail.

First, we compute the prototype of the cth class according to (13). Each category in a task corresponds to a prototype

$$\mu_c = \frac{1}{K}\sum_{i=1}^{K} Z_{c,i}^{\text{sup}} \tag{13}$$

where the number of the cth class support images is expressed as K. According to the feature embedding and prototype representation of the cth class, we reconstruct the feature space as shown

$$\bar{X}_c = \left[Z_{c,1}^{\text{sup}} - \mu_c, Z_{c,2}^{\text{sup}} - \mu_c, \ldots, Z_{c,K}^{\text{sup}} - \mu_c\right] \tag{14}$$

where $Z_{c,i}^{\text{sup}}$ represents the feature embedding of the i-support scene image ($x_{c,i}$) from the c-class and μ_c represents the prototype of the cth class.

Then, the matrix $\mathcal{B}_c$ that possesses orthogonal basis characteristics (i.e., $\mathcal{B}_c\mathcal{B}_c^T = I$) can be obtained through SVD on $\bar{X}_c$. We truncate the matrix B_c to acquire the feature subspace $\mathcal{P}_c$ of the cth class. After that, the feature subspaces of all classes within the given task are computed in the same way. We can finally classify the query scene images based on their projected distance to each class subspace.

Like support scene images, we also need to use the obtained the cth class prototype (μ_c) to construct the feature space of each query scene image (q_l), that is $Z_{l}^{\text{que}} - \mu_c$ (Z_{l}^{que} is the feature embedding of query scene image q_l). So, the projection vector of $Z_{l}^{\text{que}} - \mu_c$ on feature subspace of the cth class is $\mathcal{M}_c(Z_{l}^{\text{que}} - \mu_c)$, where $\mathcal{M}_c = \mathcal{P}_c\mathcal{P}_c^{\mathcal{T}}$. Then, according to the vector triangle rule, the vector perpendicular to the corresponding subspace can be expressed as $(Z_{l}^{\text{que}} - \mu_c) - \mathcal{M}_c(Z_{l}^{\text{que}} - \mu_c) = (I - \mathcal{M}_c)(Z_{l}^{\text{que}} - \mu_c)$, where I represents unit matrix. Finally, the projected distance of the query scene image q_l to feature subspace of the cth class is calculated by

$$d_c(q_l) = ||(\mathcal{I} - \mathcal{M}_c)(Z_{l}^{\text{que}} - \mu_c)||_2 \tag{15}$$

where $||.||_2$ represents L2 Norm.

According to (15), we calculate projected distances between query scene image q_l to all class subspaces. Then, the probability that query scene image q_l belongs to the cth class is computed by the softmax function, as shown in

$$\mathfrak{p}(y = c|q_l) = \frac{\exp(-d_c(q_l))}{\sum_{c'}\exp(-d_{c'}(q_l))} \tag{16}$$

where c' represents the set of classes within a task. Finally, the query scene image can be assigned to the corresponding class with maximum predicted probability.

E. Loss Function

In this work, to compute the classification loss more accurately for the FSRSSC task, the cross-entropy loss function is adopted. All subsequent experiments are based on this loss function, and the formula is expressed as

$$\mathfrak{L} = -\frac{1}{NM}\sum_{q}\log(\mathfrak{p}_{c,q}) \tag{17}$$

where N represents the number of classes and M represents the number of query samples per class in a task, and the classification probability that the qth query image belongs to its true category (c) is $p_{c,q}$.

IV. Experimental Results

For three RS benchmark datasets, we perform some necessary experiments to demonstrate the effectiveness of the proposed approach. Initially, we present an overview of the three RS benchmark datasets, as well as the details of experimental implementation. Subsequently, we compare with the existing approaches and analyze the experimental results. Finally, we perform ablation studies to substantiate the contribution of the core components in our method.

A. Dataset Description

In our study, we employ the NWPU-RESISC45 [42], UC Merced (UCM) [52], and AID [53] datasets to assess the effectiveness of our method. In order to make a fair comparison with other approaches, the data partitioning scheme used in [35], [36], and [50] is still adopted. Table I shows the details of three RS scene datasets used for the FSRSSC task. Specifically, we utilize the scene images from the training categories for model training, scene images from the validation categories for model selection, and scene images from testing categories to measure the classification performance of the proposed approach.

The NWPU-RESISC45 [42] dataset that was provided by Northwestern Polytechnical University (NWPU) consists of 45 scene classes with a total of 31 500 images. Each class includes 700 RGB images with 256 × 256 pixels. The spatial resolution of most of the images ranges from 0.2 to 30 m per pixel. As shown in Table I, we adopt 25 categories, 10 categories, and 10

TABLE I
THREE RS SCENE DATASETS USED FOR THE FSRSSC TASK

Datasets	Training classes	Validation classes	Testing classes
NWPU-RESISC45	Railway, Ship, Baseball diamond, Stadium, Chaparral, Rectangular farmland, Mountain, Palace, Meadow, Wetland, Lake, Island, Beach, Mobile home park, Airplane, Desert, Sparse residential, Sea ice, Roundabout, Bridge, Church, Harbor, Cloud, Freeway, Golf course	Terrace, Thermal power station, Industrial area, Railway station, Runway, Overpass, Snowberg. Tennis court, Commercial area, Storage tank	Parking lot, Intersection, Medium residential, Airport, Forest, Circular farmland, River, Ground track feld, Dense residential, Basketball court
UCM	Parking lot, Agricultural, Baseball diamond, Harbor, Overpass, Chaparral, Freeway, Buildings, Dense residential, Medium residential	Intersection, Storage tanks, Airplane, Forest, Runway	River, Tennis court, Golf course, sparse residential, Mobile home park, Beach
AID	Park, Beach, Center, Commercial, Desert, Stadium, Square, Parking, Bridge, Port, Playground, Railway Station, Resort, Farmland, Mountain, Storage Tank	Church,Bare Land,Pond Baseball Field, Meadow, River, School	Airport, Dense Residential, Industrial, Viaduct, Medium Residential, Forest, Sparse Residential

categories for training, validation, and testing, respectively [35], [36].

The UC Merced (UCM) [52] dataset, created by the UC Merced Computer Vision Laboratory, consists of 21 scene classes with a total of 2100 images. Each class includes 100 RGB images with 256 × 256 pixels and a spatial resolution of 0.3 m per pixel. As shown in Table I, the number of categories for the training, validation, and testing are 10, 5, and 6, respectively [35], [36].

The AID [53] dataset was created by Huazhong University of Science and Technology and Wuhan University. The dataset has 30 scene classes with a total of 10 000 RGB images. Each class contains 220∼420 images with 600 × 600 pixels. The spatial resolution varies from 0.5 to 8 m. As shown in Table I, the training set includes 16 classes, the validation set includes 7 classes, and the testing set includes the remaining 7 classes [50].

B. Implementation Details

This section conducted experiments on the NWPU-RESISC45, UCM, and AID datasets, and each task was set as 5-way 1-shot and 5-way 5-shot, respectively. For the 1-shot classification task, one RS scene image per class is selected to constitute support set. However, an additional scene image is generated through image flipping for classifier construction in the experiments. For the 5-shot classification task, five support RS images per class are selected. In each task setting, a query set consisting of 15 different RS scene images is randomly chosen for each class, ensuring they are not included in the support set. The performance of our method is assessed by computing the average classification accuracy across 600 randomly generated test tasks, accompanied by a 95% confidence interval.

All experiments were implemented using CUDA 11.4 in the PyTorch framework [54]. All RS scene images are resized to a standard dimension of 128×128 pixels for input. Adam optimizer [55] is employed for model parameter optimization, initially with a learning rate of 0.001. At five-shot, the learning rate decays every 50 epochs with a decay factor of 0.1, and at one-shot, the learning rate decays every 100 epochs with a decay factor of 0.5. The scale factor r used to reduce the number of channels in the self-attention module is set to 8 [51].

TABLE II
FEW-SHOT CLASSIFICATION ACCURACY (%) WITH 95% CONFIDENCE INTERVAL OF VARIOUS APPROACHES

Approach	NWPU-RESISC45	
	1-shot	5-shot
MAML* [56]	48.40±0.82	62.90±0.69
MetaSGD* [57]	60.63±0.90	75.75±0.65
LLSR* [63]	51.43	72.90
ProtoNet* [34]	40.33±0.18	63.82±0.56
MatchingNet* [49]	37.61	47.10
RelationNet* [48]	66.43±0.73	78.35±0.51
DeepEMD* [58]	64.39±0.84	78.01±0.56
FRN [59]	64.98±0.42	81.65±0.25
MCL-Katz [60]	63.30	80.78
GLIML [61]	66.86±0.68	78.91±0.45
DLA-MatchNet* [35]	68.80±0.70	81.63±0.46
RS-MetaNet* [64]	64.07±0.90	79.62±0.65
SPNet* [36]	67.84±0.87	83.94±0.50
SCL-MLNet* [37]	62.21±1.12	80.86±0.76
SPFSR* [62]	65.97±1.22	80.72±0.79
MPCLNet* [65]	55.94±0.04	76.24±0.12
HProtoNet* [66]	66.41±0.87	82.71±0.41
MSoPNet* [67]	67.05±0.80	82.02±0.46
Ours	**69.48±0.85**	**84.00±0.46**

C. Experimental Results and Discussions

In order to validate the effectiveness of our proposed approach, we select several FSL approaches for comparison. The comparison approaches include: the approaches for few-shot natural image classification such as MAML [56], Meta-SGD [57], ProtoNet [34], MatchingNet [49], RelationNet [48], DeepEMD [58], FRN [59], MCL-Katz [60], GLIML [61], SPFSR [62], and the approaches for FSRSSC task such as LLSR [63], RS-MetaNet [64], DLA-MatchNet [35], SPNet [36], SCL-MLNet [37], MPCLNet [65], HProtoNet [66], and MSoP-Net [67]. Next, detailed descriptions of experimental results for three RS datasets are provided, including 1-shot and 5-shot tasks. Note that, for MAML, Meta-SGD, LLSR, ProtoNet, MatchingNet, RelationNet, DeepEMD, DLA-MatchNet, RS-MetaNet, we refer to the experimental results in [50]; for FRN, MCL-Katz, GLIML, and SPFSR, their results were obtained by ourselves; for SPNet, SCL-MLNet, MPCLNet, HProtoNet, and MSoPNet, we reported the results as their original publications. In Tables II

TABLE III
FEW-SHOT CLASSIFICATION ACCURACY (%) WITH 95% CONFIDENCE INTERVAL OF VARIOUS APPROACHES

Approach	UCM	
	1-shot	5-shot
MAML* [56]	48.86±0.74	60.78±0.62
MetaSGD* [57]	50.52±2.61	60.82±2.00
LLSR* [63]	39.47	57.40
ProtoNet* [34]	52.27±0.20	69.86±0.15
MatchingNet* [49]	34.70	52.71
RelationNet* [48]	48.08±1.67	61.88±0.50
DeepEMD* [58]	58.47±0.76	70.42±0.58
FRN [59]	50.89±0.37	68.34±0.30
MCL-Katz [60]	50.73	68.95
GLIML [61]	56.41±0.62	70.40±0.41
DLA-MatchNet* [35]	53.76±0.62	63.01±0.51
RS-MetaNet* [64]	49.68±0.71	67.53±0.59
SPNet* [36]	57.64±0.73	73.52±0.51
SCL-MLNet* [37]	51.37±0.79	68.09±0.92
SPFSR* [62]	55.40±1.11	71.38±0.77
MPCLNet* [65]	56.46±0.21	**76.57±0.07**
HProtoNet* [66]	57.51±0.95	75.08±0.29
MSoPNet* [67]	54.27±0.60	69.77±0.38
Ours	**59.05±0.84**	76.34±0.51

TABLE IV
FEW-SHOT CLASSIFICATION ACCURACY (%) WITH 95% CONFIDENCE INTERVAL OF VARIOUS APPROACHES

Approach	AID	
	1-shot	5-shot
MAML* [56]	43.20±0.77	60.37±0.75
MetaSGD* [57]	45.01±0.98	62.58±0.80
LLSR* [63]	45.18	61.76
ProtoNet* [34]	54.32±0.86	67.80±0.64
MatchingNet* [49]	33.87	50.40
RelationNet* [48]	54.62±0.80	68.80±0.66
DeepEMD* [58]	61.04±0.77	74.51±0.55
FRN [59]	**62.29±0.37**	79.33±0.24
MCL-Katz [60]	55.28	75.66
GLIML [61]	61.28±0.61	79.56±0.41
DLA-MatchNet* [35]	61.99±0.94	75.03±0.67
RS-MetaNet* [64]	58.51±0.84	73.76±0.69
SCL-MLNet* [37]	59.46±0.96	76.31±0.68
SPFSR* [62]	60.01±1.09	75.40±0.76
MPCLNet* [65]	60.61±0.43	76.78±0.08
HProtoNet* [66]	59.78±0.58	75.87±0.35
Ours	61.62±0.75	**81.32±0.45**

–IV, the methods that refer to the results of other experiments are marked with *. The results of unmarked methods were obtained by ourselves. The best and second-best results are highlighted in bold and underline, respectively.

The comparison results of the NWPU-RESISC45 dataset are shown in Table II. Our approach attains an overall accuracy of 69.48% (1-shot) and 84.00% (5-shot), respectively. These results demonstrate that the classification performance of our proposed method surpasses that of all other comparison methods.

Table III records the experimental results of different methods on UCM dataset. Our approach stands out with the highest performance in the 1-shot settings. In the 5-shot scenario, our approach achieves a classification accuracy of 76.34% and the MPCLNet [65] achieves the highest performance (76.57%). Because MPCLNet [65] can not only learn the multiscale and rotation-invariant information of scene images at the same time,

TABLE V
ALGORITHMIC EFFICIENCY OF VARIOUS APPROACHES UNDER FIVE-WAY SCENARIO COUNTED ON THE NWPU-RESISC45 DATASET

Approach	Params (M)	1-shot			5-shot		
		Time (ms)	FLOPs (G)	Memory (G)	Time (ms)	FLOPs (G)	Memory (G)
FRN	**0.98**	47	24.00	**4.35**	56	29.65	**4.35**
GLIML	12.63	63	282.05	6.24	76	352.57	6.55
MCL-Katz	**0.98**	**31**	**22.59**	7.26	**39**	**28.24**	8.10
SPFSR	10.32	542	104	5.37	608	130	6.19
MPCLNet*	45.01	-	-	-	-	-	-
MSoPNet*	2.10	-	-	-	-	-	-
Ours	1.85	47	26.91	5.95	41	33.63	6.46

but also use three different loss functions to fully capture the diverse land covers within remote sensing scenes, compact intraclass samples, and separate interclass samples, respectively.

The performance evaluations of distinct approaches on the AID dataset are presented in Table IV. However, given that no AID dataset experiments were reported in [36] and [67], so the two methods is excluded from the current analysis. In the 5-shot scenario, our approach achieves a classification accuracy of 81.32%, which outperforms all other comparison methods. For the 1-shot scenario, although our method does not achieve the best classification performance, it still achieves a classification accuracy of 61.62%. We analyzed the reasons for this result as follows. In the experiments of proposed method, we uniformly resized the scene images to 128×128 as inputs. Since the image size of the NWPU-RESISC45 and UCM datasets are all 256×256 pixels, while the image size of the AID dataset is 600×600 pixels, the operation of resize caused more feature loss to the images in AID dataset. As a result, the feature extraction ability of the proposed multiscale feature extraction network on the AID dataset is negatively affected. In addition, under 1-shot scenario, there is only one labeled image per category. If the feature extraction for the image is inaccurate, it will affect the representativeness of the class metric, leading to lower classification results.

In addition, we compared the algorithmic efficiency of our method with some comparative methods reproduced by ourselves, such as FRN, MCL-Katz, GLIML and SPFSR. The parameter counts of MPCLNet and MSoPNet are directly adopted from corresponding work. For other inaccessible models, we cannot analyze their efficiency, and these methods basically use resnet-12 as backbone, with parameter counts more than 12 M. As can be seen from Table V , parameter count (Params) and floating-point operations (FLOPs) of our method are only higher than FRN and MCL-Katz, and our inference time (Time) is only higher than MCL-Katz. Overall, our method can achieve better classification results and has higher efficiency. In addition, since SVD cannot be completed with only one sample in the 1-shot case, additional scene images are obtained by flipping, that causes the inference time of our method in the 1-shot case to be longer than that of 5-shot case.

Finally, to provide a comprehensive assessment, we select several failure cases to discuss and analyze, as shown in Fig. 5. We can find that the misclassified RS scene images present the following common features: scene images of different categories

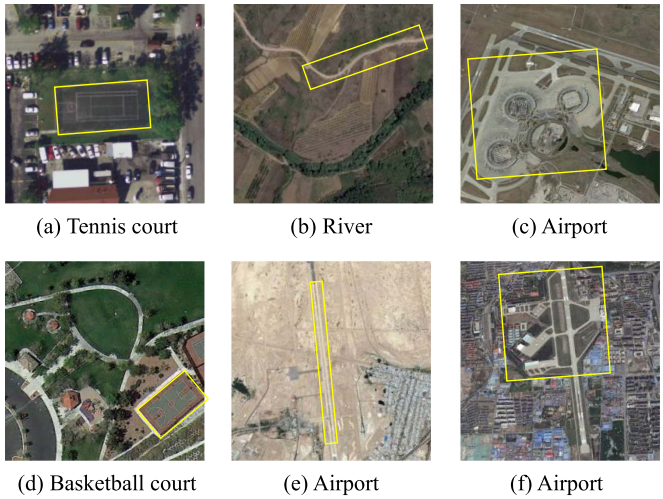


Fig. 5. Some failure cases of our proposed method. (a) Tennis court. (b) River. (c) Airport. (d) Basketball court. (e) Airport. (f) Airport.

show high similarities from background to target objects, such as Fig. 5(a) and (d) as well as Fig. 5(b) and (e), while target objects in scene images of the same category show great differences, such as Fig. 5(c) and (f). Especially in the 1-shot case, the above characteristics increase the difficulty of classifying such scene images. Fortunately, there are many strategies available, such as contrastive learning loss [65], which can alleviate these problems to a certain extent. However, this is not the focus of this work, so we only consider the general cross-entropy loss function without adding additional loss.

In summary, for three public RS scene datasets, our proposed method demonstrates relatively good overall classification performance, especially in the case of 5-shot scenario. Based on the result analyses, it is evident that our approach is more suitable and effective for the task of FSRSSC. Specifically, the multiscale feature learning-based feature extraction technique effectively handles the large-scale variations of target objects in RS images, learning more accurate feature representations of the RS scene images. Then few-shot scene images classification performance can be further improved by the subspace classifier, by capturing the shared characteristics of each class effectively and minimizing the impact of irrelevant complex backgrounds in the RS scenes.

D. Ablation Study

Next, we will respectively investigate the effectiveness of multiscale feature learning, the effectiveness of different parts within the multiscale feature learning method, the effectiveness of self-attention mechanism position, the effectiveness of proposed subspace, as well as effectiveness of multiscale feature learning for different few-shot classifiers.

1) Effectiveness of Multiscale Feature Learning: Because this work replaces the last convolutional block in Conv5 with our multiscale feature learning module, we conducted an experiment to assess the performance of the proposed feature extractor technique compared to Conv5. Except for the difference in feature extraction, all other aspects remain unchanged. Table VI shows the experimental results of three RS scene datasets under 1-shot and 5-shot sets.

As can be seen from Table VI, across three RS scene datasets, compared to using Conv5 for feature extraction, the proposed multiscale feature extraction method can significantly improve the classification accuracy for the 1-shot and 5-shot sets. Specifically, for NWPU-RESISC45 dataset, the classification accuracy of our multiscale learning method for the 1-shot and 5-shot tasks are 69.48% and 84.00%, respectively, which are higher by 6.51% and 2.68% compared with Conv5. These experimental results provide strong evidence for the efficacy of our multiscale feature extraction method. Because the designed multiscale feature learning module can adapt well to the scale variations of target objects in the scene images, emphasizing key regions and channels information at different scales to obtain more representative multiscale scene image features, thereby enhancing the feature representation.

In addition, we conduct a comprehensive analysis of the effectiveness of our multiscale feature learning via the confusion matrix. We calculate the confusion matrices when using the proposed feature extraction network and Conv5, respectively. These confusion matrices clearly record the classification situation of scene images within each category. Figs. 6 and 7 illustrate the confusion matrices in the case of 1-shot and 5-shot for NWPU-RESISC45 dataset, respectively. Each figure presents two confusion matrices corresponding to different feature extraction network settings. From the confusion matrices, compared to Conv5, the classification accuracies on ground track field, intersection, and basketball court are significantly improved when using our proposed feature extraction network in 1-shot and 5-shot scenarios. In particular, for 1-shot, the classification accuracies of our method on ground track field, intersection, and basketball court are 78.20%, 73.22%, and 51.74%, respectively, making increases of 15.54%, 14.11%, and 10.41% over Conv5. For 5-shot, the classification accuracies of our method on the above three categories are 90.50%, 86.62%, and 72.20%, representing improvements of 9.63%, 6.42%, and 9.73% over Conv5. Upon analyzing all test images, it becomes evident that the scale variations of target objects are more pronounced in RS images belonging to these three categories compared to other classes. This observation further supports the effectiveness of our proposed multiscale feature learning approach.

Finally, we further explore the classification performance and method efficiency of Conv5s with different convolutional kernel sizes (1×1, 3×3, 5×5, 7×7) and our multiscale feature extraction network on the NWPU-RESISC45 dataset. We calculate the number of parameters (Params), memory usage (Memory), floating point numbers (FLOPs), and inference time (Time). From Table VII, for the Conv5, as the convolutional kernel size increases, the Params, FLOPs, and memory usage gradually increase, and the inference time remains essentially unchanged. Compared with 1×1 and 3×3 convolution kernels, 5×5 and 7×7 convolution kernels will achieve better classification performance. Moreover, when the convolution kernel becomes 7×7, the number of parameters will become larger (5.33 M). For example, the parameter count of a Conv5s (9×9) will be larger than that of ResNet12 (12 M). Therefore, when designing

TABLE VI
ABLATION STUDY OF THE EFFECTIVENESS OF OUR MULTISCALE FEATURE LEARNING

Feature Extraction	NWPU-RESISC45		UCM		AID	
	1-shot	5-shot	1-shot	5-shot	1-shot	5-shot
Conv5	62.97±0.85	81.32±0.53	56.99±0.76	74.40±0.50	57.96±0.77	78.59±0.52
Multiscale (Ours)	**69.48±0.85**	**84.00±0.46**	**59.05±0.75**	**76.34±0.51**	**61.62±0.75**	**81.32±0.45**

Bold values represent the best results in a set of ablation experiments.

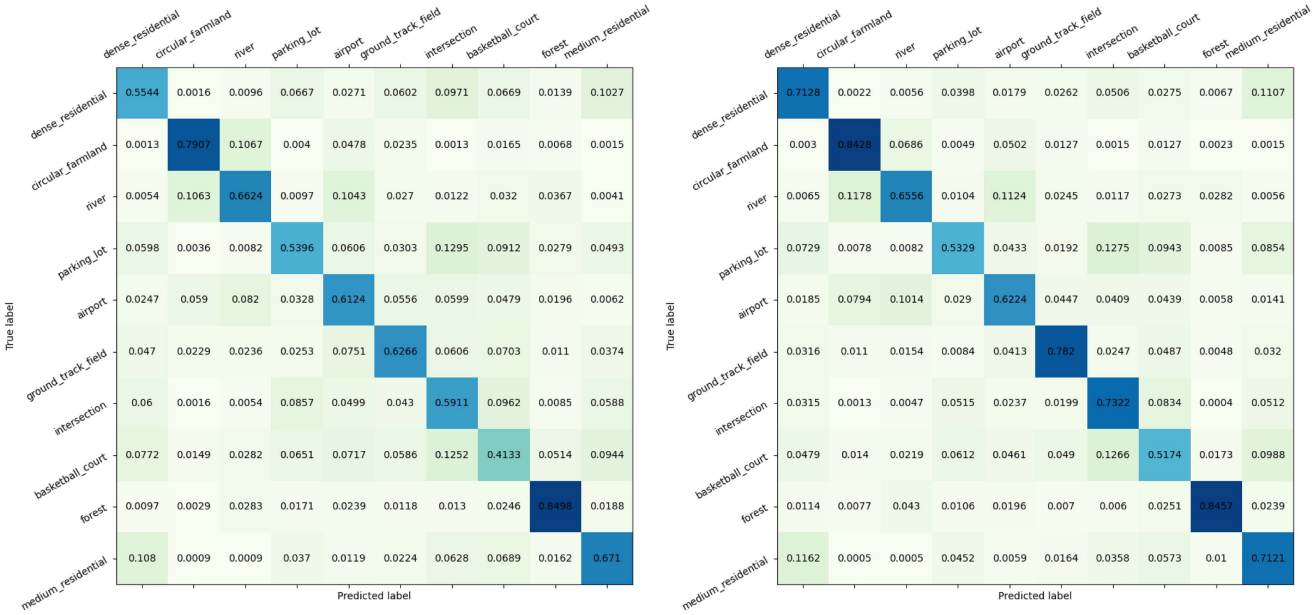


Fig. 6. Confusion matrices under different feature extraction schemes in the case of 1-shot. Left: Confusion matrix under the Conv5 network; Right: Confusion matrix under our proposed feature extraction network.

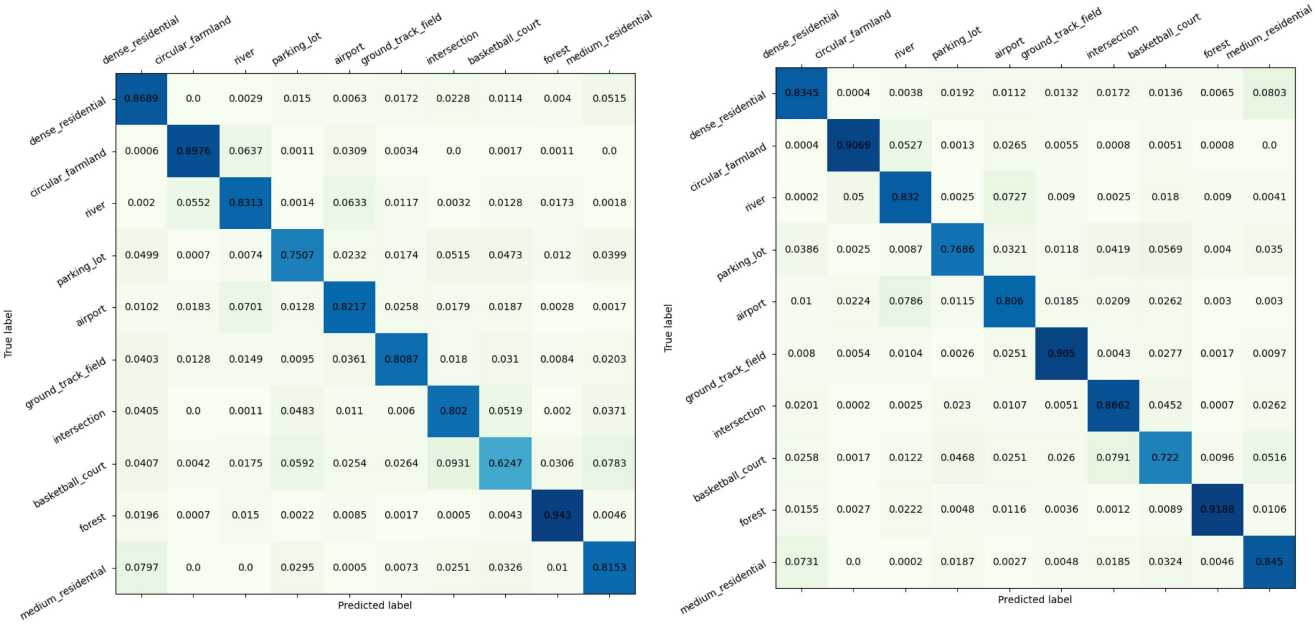


Fig. 7. Confusion matrices under different feature extraction schemes in the case of 5-shot. Left: Confusion matrix under the Conv5 network; Right: Confusion matrix under our proposed feature extraction network.

TABLE VII
FEW-SHOT CLASSIFICATION ACCURACY (%) AND ALGORITHMIC EFFICIENCY OF DIFFERENT CONV5S AND MULTISCALE FEATURE EXTRACTION

Feature extractor	Params (M)	1-shot				5-shot			
		Acc	Time (ms)	FLOPs (G)	Memory (G)	Acc	Time (ms)	FLOPs (G)	Memory (G)
Conv5 (1×1)	**0.11**	57.47±0.82	**38**	**2.79**	**5.82**	74.82±0.61	33	**3.49**	**6.35**
Conv5 (3×3)	0.98	62.97±0.85	**38**	22.59	5.85	81.32±0.52	**32**	28.24	6.38
Conv5 (5×5)	2.72	62.77±0.87	**38**	62.19	6.06	82.86±0.49	33	77.73	6.39
Conv5 (7×7)	5.33	64.52±0.87	39	121.58	7.23	82.33±0.49	34	151.97	7.56
Multiscale (Ours)	1.85	**69.48±0.85**	47	26.91	5.95	**84.00±0.46**	41	33.63	6.46

Bold values represent the best results in a set of ablation experiments.

TABLE VIII
ABLATION STUDY OF DIFFERENT PARTS IN THE MULTISCALE FEATURE LEARNING METHOD

Different Parts			NWPU-RESISC45		UCM		AID	
Multiscale Branchs	Self-attention	Channel Attention	1-shot	5-shot	1-shot	5-shot	1-shot	5-shot
✓	×	×	65.21±0.86	82.93±0.50	57.35±0.78	75.43±0.48	59.08±0.79	79.02±0.48
✓	×	✓	66.82±0.85	83.81±0.49	57.34±0.72	75.57±0.48	58.77±0.74	79.75±0.51
✓	✓	×	66.70±0.87	83.37±0.49	57.54±0.75	75.51±0.49	59.82±0.81	79.04±0.48
✓	✓	✓	**69.48±0.85**	**84.00±0.46**	**59.05±0.75**	**76.34±0.51**	**61.62±0.75**	**81.32±0.45**

Bold values represent the best results in a set of ablation experiments.

CNN, it does not stack large convolution kernels, so Conv5s (9×9) is not be adopted in our work.

The Params of our proposed multiscale feature extraction network is between the Params of Conv5 (3×3) and the Params of Conv5 (5×5). The memory usage of our method does not differ much from Conv5 (3×3) and Conv5 (5×5), and the FLOPs is higher than that of Conv5 (3×3) but much lower than that of Conv5 (5×5). Although inference time of our multiscale method is a little higher than that of all Conv5, our classification accuracy is significantly higher than theirs.

2) Effectiveness of Different Parts in the Multiscale Feature Learning: The proposed multiscale feature learning method consists of four different scale feature learning branches based on self-attention mechanisms, also including a channel attention-based multiscale feature fusion module. In this part, different experiments are conducted on the various parts of this feature learning method, to study their impacts on the few-shot scene image classification accuracy, respectively. The different experimental settings are as follows: the first one represents the only use of four scale feature learning branches without incorporating the self-attention mechanisms, and replacing the channel attention-based multiscale feature fusion with a simple addition; the second one represents the use of four different scale feature learning branches without utilizing self-attention mechanisms, and finally adopting the channel attention-based feature fusion method; the third one represents that the feature extractor network includes the four scale branches based on self-attention mechanisms, but only adopts the addition-based multiscale feature fusion; the fourth one includes the four scale feature learning branches based on the self-attention mechanisms, also and the channel attention-based feature fusion method. Table VIII presents the results of these ablation experiments.

Through analyzing the results of whether to include the self-attention mechanisms and channel attention in Table VIII, it is found that after the multiscale learning branches, respectively, adding the self-attention mechanisms and channel attention both can improve the few-shot scene images classification performance to a certain extent, which suggests that the different scale feature learning branches combined with the self-attention mechanisms can adapt to the large-scale variations in the scene images, and then focus on the key regional features under the corresponding scale. By enhancing the semantic information of important channels in each scale feature map, the channel attention-based feature fusion can also improve the representation capability of the scene images. Furthermore, by comparing the results of both incorporating the self-attention mechanisms and the channel attention, it is evident that the multiscale branches based on self-attention mechanisms and the channel attention-based multiscale feature fusion module complement each other, leading to further improve the feature representation scene images.

3) Effectiveness the Self-Attention Mechanism Position: The proposed multiscale feature learning module incorporates self-attention mechanisms into each branch to integrate an understanding of global information under different scales, thereby emphasizing the important areas features at each scale. In order to validate the effectiveness of this approach, two sets of experiments are conducted in this part regarding the position of self-attention mechanism. In one set of experiments, the self-attention mechanism is placed after concatenating each scale feature. That means each branch only performs the convolution operations, and the learned features from each branch are simply concatenated, finally self-attention mechanism is applied to the concatenated feature vectors. In this case, the position of self-attention mechanism is marked as "after concatenating." In another set of experiments, the self-attention mechanism is added to feature learning branches at four different scales separately, which is the structure used in this work. In this case, the position of the self-attention mechanism is marked as "after convolution.'. Apart from the above differences, all other settings in the two sets of experiments remain the same. Table IX records the classification results of the two sets of experiments.

TABLE IX
ABLATION STUDY OF THE LOCATION OF SELF-ATTENTION MECHANISM

Self-attention Mechanism After	NWPU-RESISC45		UCM		AID	
	1-shot	5-shot	1-shot	5-shot	1-shot	5-shot
Concatenating	66.29±0.85	83.17±0.48	57.73±0.77	75.98±0.50	58.70±0.76	79.50±0.49
Convolution (Ours)	**69.48±0.85**	**84.00±0.46**	**59.05±0.75**	**76.34±0.51**	**61.62±0.75**	**81.32±0.45**

Bold values represent the best results in a set of ablation experiments.

TABLE X
FEW-SHOT CLASSIFICATION ACCURACY WITH DIFFERENT COMBINATIONS OF FEATURE EXTRACTORS AND CLASSIFIERS

Feature Extractor	Few-Shot Classifier	NWPU-RESISC45		UCM		AID	
		1-shot	5-shot	1-shot	5-shot	1-shot	5-shot
Conv5	Prototype	60.46±0.88	78.73±0.56	53.18±0.81	72.58±0.52	52.22±0.73	68.02±0.60
	Subspace (Ours)	**62.97±0.85**	**81.32±0.53**	**56.99±0.76**	**74.40±0.50**	**57.96±0.77**	**78.59±0.52**
Multiscale (Ours)	Prototype	62.98±0.84	81.54±0.51	55.79±0.74	75.53±0.50	54.19±0.77	67.73±0.56
	Subspace (Ours)	**69.48±0.85**	**84.00±0.46**	**59.05±0.75**	**76.34±0.51**	**61.62±0.75**	**81.32±0.45**

Bold values represent the best results in a set of ablation experiments.

From Table IX, it can be observed that, compared to applying self-attention after concatenating features from each scale, the proposed multiscale feature learning module embedded the self-attention mechanism in each branch separately, can yield better classification performance on the three scene image datasets. Specifically, the classification accuracy is improved by 1.32% to 3.19% in the one-shot scenario and by 0.36% to 1.82% in the five-shot scenario. Applying the self-attention mechanism to each scale feature is able to better integrate the global information at the corresponding scale without introducing other scale features. Finally, a more accurate scene image feature at that scale is obtained by focusing on the important areas. Conversely, merely concatenating the feature vectors from each scale leads to offsetting each other, and then applying the self-attention mechanism to the overall concatenating features will fail to focus on the important areas in the scene image.

4) Effectiveness of Multiscale Feature Learning Method for Different Few-Shot Classifiers: In the FSRSSC task, the final classifier will be constructed based on the features extracted from the support set RS scene images. More accurate feature representations are beneficial for obtaining a few-shot scene image classifier with better classification performance. Next, we verify whether the proposed multiscale feature extraction method has a positive effect on the construction of different few-shot classifiers, further demonstrating the effectiveness and wide applicability of the proposed multiscale feature learning method in this work. We conducted a series of experiments on the NWPU-RESISC45, UCM, as well as the AID datasets by combination of our multiscale feature extractors with different few-shot classifiers (such as prototype and our subspace classifiers). Table X presents the experimental results.

When using the prototype classifier [34], besides the 5-shot scenario of the AID dataset, our proposed multiscale feature learning method can achieve higher classification accuracy than when using Conv5 as the feature extractor. When using the subspace classifier, for all 1-shot and 5-shot scenarios, our proposed multiscale feature learning approach can achieve better classification performance on three datasets compared with Conv5. These experimental results indicate: (a) under the prototype and subspace classifiers, adopting the proposed multiscale feature learning method can enhance the classification performance of the RS images, demonstrating that our multiscale feature learning method can effectively deal with the scale variations problem of RS scene images; (b) the proposed multiscale feature learning method has a positive effect on constructing the few-shot scene images classifiers, showing broad applicability.

5) Effectiveness of Our Classifier: Except for the feature Extractor method, another key step of the FSRSSC task is selecting a suitable metric benchmark to build a high-quality classifier. Because of irrelevant complex background in the RS scene images, the classifier for FSRSSC task needs to decrease the impact of irrelevant background as effectively as possible. Hence, for the Conv5 and the proposed multiscale feature extraction method, we continue to explore the impact of proposed classifiers on RS image classification performance.

From Table X, regardless of using Conv5 or the proposed multiscale feature learning method, it can be observed that compared to the prototype classifier, the subspace classifier can significantly enhance the classification performance of the remote sensing datasets. Specifically, for the Conv5, compared to the prototype classifier, the subspace classifier improves the classification accuracy by 2.51% to 5.74% in 1-shot scenarios and by 2.59% to 10.57% in 5-shot scenarios. For our multiscale feature learning method, compared to the prototype classifier, the subspace classifier improves the classification accuracy by 3.26% to 7.43% in 1-shot scenarios and by 0.81% to 14.23% in 5-shot scenarios.

In addition to the above experimental data, we make a more intuitive comparison of the classification ability of the prototype and subspace classifiers. We ensure the feature extraction network to be consistent and use the proposed multiscale feature extraction network. Then, we classify the remote sensing scene images with the prototype classifier and subspace classifier, respectively. Fig. 8 presents the scene images that can be correctly classified by the subspace classifier but misclassified by the prototype classifier. We find that these scene images both have complex backgrounds. For example, the scene images with the category "Basketball Court" all contain roads, trees, houses,

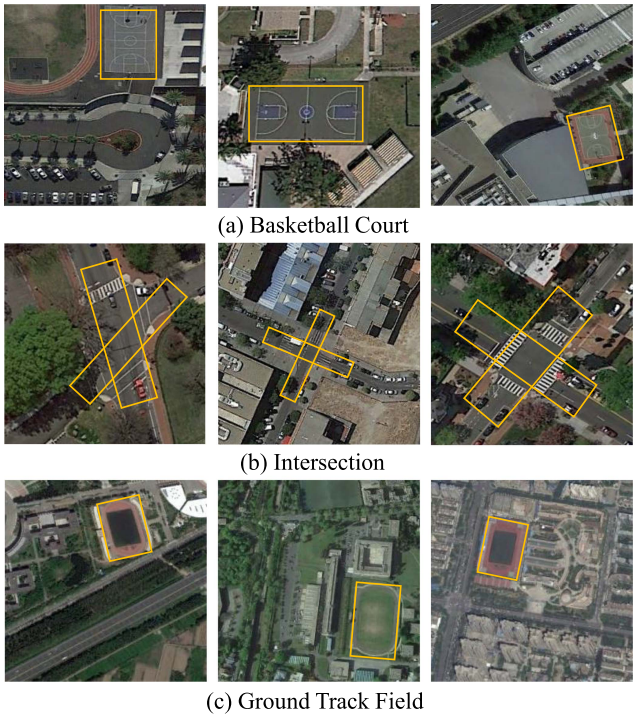


Fig. 8. Images correctly classified by our subspace classifier but wrongly classified by prototype classifier.

and cars; the scene images with the category "Intersection" all appear trees, houses, and cars; and the scene images with the category "Ground Track Field" all have roads, trees, and houses. These objects, unrelated to the category semantics, lead to the complex background of remote sensing scene images and pose greater challenges to the FSRSSC task. The widely used prototype classifier has no ability to classify these scene images with complex backgrounds, whereas our subspace classifier can correctly classify them, thus improving the overall classification accuracy. It indicates that our subspace classifier can acquire more accurate commonality features of each class in few-shot scenarios, to reduce the adverse impact of irrelevant complex backgrounds in remote sensing images.

V. Conclusion

This work proposes a multiscale feature learning-based FSRSSC approach via a subspace classifier. For the RS scene images, our method initially tackles the feature extraction issue under the few-shot framework by focusing on two aspects: addressing the challenge of scale variations in the scene images and then obtaining more discriminative representations of the scene images. Specifically, for the large-scale variations in the scene images, we design four different scale feature extraction branches based on self-attention mechanisms to learn the scene image features at different scales. Corresponding global information in each scale will be incorporated into the respective scale features to emphasize the important regions of the scene images. In addition, a channel attention-based multiscale feature fusion module is used to effectively integrate important information from different scales. By emphasizing the semantic information of key channels, more accurate scene image feature representations can be obtained. On the basis of performance gains brought by the above multiscale feature learning method, the subspace classifier is further used to learn commonality features of each class under the few-shot scenarios, aiming to reduce the adverse impact of irrelevant complex background in RS images. For three public RS scene datasets, our proposed approach demonstrates relatively competitive performance in few-shot classification, and the proposed multiscale feature learning method can effectively enhance the accuracy of scene image feature representation.

References

[1] C. Qiu et al., "Few-shot remote sensing image scene classification: Recent advances, new baselines, and future trends," *ISPRS J. Photogrammetry Remote Sens.*, vol. 209, pp. 368–382, 2024.
[2] D. Yu and C. Fang, "Urban remote sensing with spatial Big Data: A review and renewed perspective of urban studies in recent decades," *Remote Sens.*, vol. 15, no. 5, 2023, Art. no. 1307.
[3] 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.
[4] Q. Wang, Y. Liu, Z. Xiong, and Y. Yuan, "Hybrid feature aligned network for salient object detection in optical remote sensing imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–15, 2022.
[5] J. Yao, B. Zhang, C. Li, D. Hong, and J. Chanussot, "Extended vision transformer (ExViT) for land use and land cover classification: A multimodal deep learning framework," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, pp. 1–15, 2023.
[6] Y. Lin, L. He, D. Zhong, Y. Song, L. Wei, and L. Xin, "A high spatial resolution aerial image dataset and an efficient scene classification model," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, pp. 1–13, 2024.
[7] L. Zhang and L. Zhang, "Artificial intelligence for remote sensing data analysis: A review of challenges and opportunities," *IEEE Geosci. Remote Sens. Mag.*, vol. 10, no. 2, pp. 270–294, Jun. 2022.
[8] L. Fang, Y. Jiang, Y. Yan, J. Yue, and Y. Deng, "Hyperspectral image instance segmentation using spectral–spatial feature pyramid network," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, pp. 1–13, 2023.
[9] A. Qin et al., "Distance constraints-based generative adversarial networks for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, pp. 1–16, 2023.
[10] J. Hu, H. Yang, X. Du, and H. Wang, "Application of few-shot learning in high resolution remote sensing image classification and recognition," *J. Chongqing Univ. Posts Telecommun. (Natural Sci. Ed.)*, vol. 34, no. 3, pp. 410–422, 2022.
[11] L. Lei, B. Huang, M. Ye, F. Yao, and Y. Qian, "Cross-domain residual deep NMF for transfer learning between different hyperspectral image scenes," *Int. J. Wavelets, Multiresolution Inf. Process.*, vol. 21, no. 02, 2023, Art. no. 2250046.
[12] O. A. Penatti, K. Nogueira, and J. A. D. 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.
[13] A. Ma, Y. Wan, Y. Zhong, J. Wang, and L. Zhang, "SceneNet: Remote sensing scene classification deep learning network using multi-objective neural evolution architecture search," *ISPRS J. Photogrammetry Remote Sens.*, vol. 172, pp. 171–188, 2021.
[14] W. Wang, Y. Chen, and P. Ghamisi, "Transferring CNN with adaptive learning for remote sensing scene classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–18, 2022.
[15] J. Wang, W. Li, M. Zhang, R. Tao, and J. Chanussot, "Remote sensing scene classification via multi-stage self-guided separation network," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, pp. 1–12, 2023.
[16] J. Lu, P. Gong, J. Ye, J. Zhang, and C. Zhang, "A survey on machine learning from few samples," *Pattern Recognit.*, vol. 139, 2023, Art. no. 109480.
[17] G. Cheng, X. Sun, K. Li, L. Guo, and J. Han, "Perturbation-seeking generative adversarial networks: A defense framework for remote sensing image scene classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–11, 2022.
[18] H. Li, W. Deng, Q. Zhu, Q. Guan, and J.-C. Luo, "Local-global context-aware generative dual-region adversarial networks for remote sensing scene image super-resolution," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, pp. 1–14, 2024.

[19] Y. Wang, Q. Yao, J. T. Kwok, and L. M. Ni, "Generalizing from a few examples: A survey on few-shot learning," *ACM Comput. Surv.*, vol. 53, no. 3, pp. 1–34, 2020.

[20] A. Qin et al., "Deep updated subspace networks for few-shot remote sensing scene classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, pp. 1–14, 2024.

[21] F. Hao, F. He, L. Liu, F. Wu, D. Tao, and J. Cheng, "Class-aware patch embedding adaptation for few-shot image classification," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2023, pp. 18859–18869.

[22] Y. Jia, J. Gao, W. Huang, Y. Yuan, and Q. Wang, "Exploring hard samples in multi-view for few-shot remote sensing scene classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, pp. 1–14, 2023.

[23] J. Wu, C. Lang, G. Cheng, X. Xie, and J. Han, "Retentive compensation and personality filtering for few-shot remote sensing object detection," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 34, no. 7, pp. 5805–5817, Jul. 2024.

[24] C. Lang, G. Cheng, B. Tu, and J. Han, "Global rectification and decoupled registration for few-shot segmentation in remote sensing imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, pp. 1–11, 2023.

[25] C. Lang, J. Wang, G. Cheng, B. Tu, and J. Han, "Progressive parsing and commonality distillation for few-shot remote sensing segmentation," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, pp. 1–10, 2023.

[26] C. Lang, G. Cheng, B. Tu, C. Li, and J. Han, "Base and meta: A new perspective on few-shot segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 9, pp. 10669–10686, Sep. 2023.

[27] Y. Jia, J. Gao, W. Huang, Y. Yuan, and Q. Wang, "Holistic mutual representation enhancement for few-shot remote sensing segmentation," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, pp. 1–13, 2023.

[28] Y. Song, T. Wang, P. Cai, S. K. Mondal, and J. P. Sahoo, "A comprehensive survey of few-shot learning: Evolution, applications, challenges, and opportunities," *ACM Comput. Surv.*, vol. 55, no. 13s, pp. 1–40, 2023.

[29] X. Luo, Z. Tian, T. Zhang, B. Yu, Y. Y. Tang, and J. Jia, "PFENet++: Boosting few-shot semantic segmentation with the noise-filtered context-aware prior mask," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 46, no. 2, pp. 1273–1289, Feb. 2024.

[30] G. Li and Y. Yu, "Visual saliency detection based on multiscale deep CNN features," *IEEE Trans. Image Process.*, vol. 25, no. 11, pp. 5012–5024, Nov. 2016.

[31] 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, pp. 1–15, 2022.

[32] 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.

[33] Y. Yang et al., "An explainable spatial-frequency multi-scale transformer for remote sensing scene classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, pp. 1–15, 2023.

[34] J. Snell, K. Swersky, and R. Zemel, "Prototypical networks for few-shot learning," in *Proc. Annu. Conf. Neural Inf. Process. Syst.*, 2017, vol. 30, pp. 4080–4090.

[35] L. Li, J. Han, X. Yao, G. Cheng, and L. Guo, "DLA-MatchNet for few-shot remote sensing image scene classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 9, pp. 7844–7853, Sep. 2021.

[36] G. Cheng et al., "SPNet: Siamese-prototype network for few-shot remote sensing image scene classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–11, 2022.

[37] X. Li, D. Shi, X. Diao, and H. Xu, "SCL-MLNet: Boosting few-shot remote sensing scene classification via self-supervised contrastive learning," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–12, 2022.

[38] J. Li, M. Gong, H. Liu, Y. Zhang, M. Zhang, and Y. Wu, "Multiform ensemble self-supervised learning for few-shot remote sensing scene classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, pp. 1–16, 2023.

[39] H. Abdi and L. J. Williams, "Principal component analysis," *Wiley Interdiscipl. Rev., Comput. Statist.*, vol. 2, no. 4, pp. 433–459, 2010.

[40] C. Simon, P. Koniusz, R. Nock, and M. Harandi, "Adaptive subspaces for few-shot learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 4135–4144.

[41] A. Afrasiyabi, J.-F. Lalonde, and C. Gagné, "Mixture-based feature space learning for few-shot image classification," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 9021–9031.

[42] 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.

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

[44] Y. Liu, Y. Zhong, and Q. Qin, "Scene classification based on multiscale convolutional neural network," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 12, pp. 7109–7121, Dec. 2018.

[45] 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.

[46] H. Sun, S. Li, X. Zheng, and X. Lu, "Remote sensing scene classification by gated bidirectional network," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 1, pp. 82–96, Jan. 2020.

[47] 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.

[48] F. Sung, Y. Yang, L. Zhang, T. Xiang, P. H. Torr, and T. M. Hospedales, "Learning to compare: Relation network for few-shot learning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 1199–1208.

[49] O. Vinyals, C. Blundell, T. Lillicrap, K. Kavukcuoglu, and D. Wierstra, "Matching networks for one shot learning," in *Proc. Annu. Conf. Neural Inf. Process. Syst.*, vol. 29, 2016, pp. 3630–3638.

[50] B. Zhang et al., "SGMNet: Scene graph matching network for few-shot remote sensing scene classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–15, 2022.

[51] H. Zhang, I. Goodfellow, D. Metaxas, and A. Odena, "Self-attention generative adversarial networks," in *Proc. Int. Conf. Mach. Learn.*, 2019, pp. 7354–7363.

[52] 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.

[53] 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.

[54] A. Paszke et al., "Automatic Differentiation in PyTorch," in *Proc. Annu. Conf. Neural Inf. Process. Syst.*, 2017, vol. 30, pp. 1–4.

[55] D. P. Kingma and J. Ba, "ADAM: A method for stochastic optimization," in *Proc. Int. Conf. Learn. Representations*, 2015, pp. 1–15.

[56] C. Finn, P. Abbeel, and S. Levine, "Model-agnostic meta-learning for fast adaptation of deep networks," in *Proc. Int. Conf. Mach. Learn. PMLR*, 2017, pp. 1126–1135.

[57] Z. Li, F. Zhou, F. Chen, and H. Li, "Meta-SGD: Learning to learn quickly for few-shot learning," 2017, *arXiv:1707.09835*. [Online]. Available: http://arxiv.org/abs/1707.09835

[58] C. Zhang, Y. Cai, G. Lin, and C. Shen, "DeepEMD: Few-shot image classification with differentiable earth mover's distance and structured classifiers," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 12200–12210.

[59] D. Wertheimer, L. Tang, and B. Hariharan, "Few-shot classification with feature map reconstruction networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 8008–8017.

[60] Y. Liu, W. Zhang, C. Xiang, T. Zheng, D. Cai, and X. He, "Learning to affiliate: Mutual centralized learning for few-shot classification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 14391–14400.

[61] F. Hao, F. He, J. Cheng, and D. Tao, "Global-local interplay in semantic alignment for few-shot learning," *IEEE Trans Circuits Syst Video Technol*, vol. 32, no. 7, pp. 4351–4363, Jul. 2022.

[62] W. Chen, C. Si, Z. Zhang, L. Wang, Z. Wang, and T. Tan, "Semantic prompt for few-shot image recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 23581–23591.

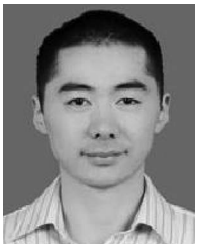
[63] M. Zhai, H. Liu, and F. Sun, "Lifelong learning for scene recognition in remote sensing images," *IEEE Geosci. Remote. Sens. Lett.*, vol. 16, no. 9, pp. 1472–1476, Sep. 2019.

[64] H. Li et al., "RS-MetaNet: Deep metametric learning for few-shot remote sensing scene classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 8, pp. 6983–6994, Aug. 2021.

[65] J. Ma, W. Lin, X. Tang, X. Zhang, F. Liu, and L. Jiao, "Multi-pretext-task prototypes guided dynamic contrastive learning network for few-shot remote sensing scene classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, pp. 1–16, 2023.

[66] M. Hamzaoui, L. Chapel, M.-T. Pham, and S. Lefèvre, "Hyperbolic prototypical network for few shot remote sensing scene classification," *Pattern Recognit. Lett.*, vol. 177, pp. 151–156, 2024.

[67] J. Deng, Q. Wang, and N. Liu, "Masked second-order pooling for few-shot remote sensing scene classification," *IEEE Geosci. Remote Sens. Lett.*, vol. 21, pp. 1–5, 2024.



Anyong Qin received the B.S. degree in information and computational science and the Ph.D. degree in computer science from Chongqing University, Chongqing, China, in 2012, and 2019, respectively.

He is currently a Lecturer with the School of Communications and Information Engineering, Chongqing University of Posts and Telecommunications, Chongqing. His research interests include pattern recognition, image processing, machine learning, and remote sensing images.

Fuyang Chen received the bachelor's degree in communication engineering from Henan Normal University, Henan, China, in 2021. She is currently working toward the master's degree in communication engineering with the Chongqing University of Posts and Telecommunications, Chongqing, China.

Her research interests include image processing, machine learning, few-shot learning and remote sensing images.

Qiang Li received the master's degree in communication engineering from the Chongqing University of Posts and Telecommunications, Chongqing, China, in 2002.

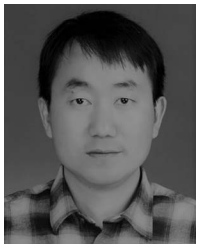
He is currently a Professor with the School of Communications and Information Engineering, Chongqing University of Posts and Telecommunications. His research interests include audio and video signal processing.

Tiecheng Song received the Ph.D. degree in signal and information processing from the University of Electronic Science and Technology of China, Chengdu, China, in 2015.

From 2015 to 2016, he was with the Multimedia Laboratory, Nanyang Technological University, Singapore, as a Visiting Student. He is currently an Associate Professor with the School of Communications and Information Engineering, Chongqing University of Posts and Telecommunications, Chongqing, China. His research interests include image processing and computer vision.

Yu Zhao received the Ph.D. degree in information and communication engineering from Harbin Engineering University, Harbin, China, in 2018.

He is currently a Lecturer with the School of Communications and Information Engineering, Chongqing University of Posts and Telecommunications, Chongqing, China. His research interests include signal processing, audio processing and machine learning.



Chenqiang Gao received the B.S. degree in computer science from the China University of Geosciences, Wuhan, China, in 2004, and the Ph.D. degree in control science and engineering from the Huazhong University of Science and Technology, Huazhong, Wuhan, in 2009.

He joined the School of Communications and Information Engineering, Chongqing University of Posts and Telecommunications, Chongqing, China, in 2009. He joined the Informedia Group, School of Computer Science, Carnegie Mellon University, Pittsburgh, PA, USA, in 2012, where he was a Visiting Scholar on multimedia event detection and surveillance event detection. In 2013, he became a Postdoctoral Fellow and continued work on MED and SED until 2014, when he returned to CQUPT. He joined the School of Intelligent Systems Engineering, Sun Yat-sen University, Shenzhen, China, in 2023. His research interests include image processing, infrared target detection, action recognition, and event detection.