Preprints.org

Article

Not peer-reviewed version

Few-Shot Remote Sensing Scene Classification Based on Diffusion Augmentation and Multimodal Feature Fusion

Zhou Yang , Siming Han * , Ming Wu

Posted Date: 5 March 2026

doi: 10.20944/preprints202603.0447.v1

Keywords: few-shot remote sensing scene classification (FSRSSC); diffusion augmentation; multiscale feature fusion

Preprints.org is a free multidisciplinary platform providing preprint service that is dedicated to making early versions of research outputs permanently available and citable. Preprints posted at Preprints.org appear in Web of Science, Crossref, Google Scholar, Scilit, Europe PMC.

Copyright: This open access article is published under a Creative Commons CC BY 4.0 license, which permit the free download, distribution, and reuse, provided that the author and preprint are cited in any reuse.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 5 March 2026 doi:10.20944/preprints202603.0447.v1

Disclaimer/Publisher's Note: The statements, opinions, and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions, or products referred to in the content.

Article

Few-Shot Remote Sensing Scene Classification Based on Diffusion Augmentation and Multimodal Feature Fusion

Zhou Yang, Siming Han * and Ming Wu

Rocket Force University of Engineering, China; yzmailbox2015@163.com; wming029@163.com

* Correspondence: hsming@163.com; Tel.: +86-15319725545

Highlights

In this article, we propose a novel framework named as MMFF-Net. It can be utilized for the few-shot remote sensing scene classification (FSRSSC).

What are the main findings?

1. Diffusion augmentation.
2. Multimodal feature fusion.

What are the implications of the main findings?

3. The diffusion augmentation could be employed to augment the samples with high-quality in FSRSSC, which can address the challenge of labeled data scarcity.
4. The multimodal feature fusion combines the visual and semantic features, which can obtain more representative image features.

Abstract

Few-shot remote sensing scene classification (FSRSSC) entails identifying images scene classes from limited labeled samples, facing the challenges of labeled data scarcity, as well as the intricacy and variety of remote sensing images with high intraclass variance and interclass similarity. To address these challenges, we propose a novel framework named as MMFF-Net in this article, which consists of four key components: diffusion augmentation (DA), multiscale feature fusion (MSFF), dual attention fusion module (DAFM), and information interaction mutual attention (IIMA). The DA is utilized to augment support set samples with high-quality. In addition, the MSFF focuses on obtaining the local spatial details, and the DAFM is utilized to fuse the local feature and the global feature. What is more, the IIMA module is employed to interact between the query set and support set information. What is more, we use word2vec to obtain the semantic features for reducing the disparity between them and the visual features with LSE Loss. The comparative experimental results with multiple models on three benchmark remote sensing scene (RSS) datasets validate the effectiveness of the proposed MMFF-Net, showcasing the superiority and feasibility of our approach in most FSRSSC cases.

Keywords: few-shot remote sensing scene classification (FSRSSC); diffusion augmentation; multiscale feature fusion

1. Introduction

Remote sensing scene classification has been widely used in the fields of land use [5] urban planning [6], environmental monitoring [7], disaster detection [8], and so on. In recent years, the methods of remote sensing scene classification based on deep learning have made great progress. However, these methods typically require a large number of labeled samples [9] for model training,

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 5 March 2026 doi:10.20944/preprints202603.0447.v1

and the scarcity of labeled data limits their generalization ability heavily, especially in real-world scenarios where labeled data are lacking or unavailable.

In order to deal with the limitation of labeled data scarcity, few-shot remote sensing scene classification (FSRSSC) has gradually become a research hotspot. Different from the general classification tasks, FSRSSC aims to learn transferable prior knowledge from a large number of labeled data base class for classification of the remote sensing scene with limited samples. In accordance with the training approach, FSRSSC methods can be categorized as follows：1) Transfer learning-based methods [6–10]; 2) Attention mechanisms-based methods [7,11–16]; 3) Novel network architectures-based methods [11,12,17–19].

The current FSRSSC methods mainly extract remote sensing image features through convolutional neural networks and convert them into feature vectors through the pooling layer, thereby achieving classification. However, since remote sensing images are captured from a longer distance and a wider range of viewpoints through sensing devices such as satellites, aircraft or other unmanned aerial vehicles, a single image often contains complex and diverse ground scenes as well as spatially scattered objects. It results in the large intraclass and small interclass variances in remote sensing images, as is shown in Figure 1. If only the global features are paid attention to, it may lead to the loss of the useful local details in the images, thereby affecting the classification accuracy. Therefore, while extracting the global features of remote sensing images, the local features with more details are combined to enhance the representational ability in this paper. It should be noted that it is crucial that the extracted local features must be related to the classified image closely, otherwise, the irrelevant local features will reduce the classification accuracy. Meanwhile, semantic textual information is leveraged to refine the prototypes of the support set in few-shot scenarios, bringing them closer to the true class prototypes, thereby enhancing accuracy [24].

(**a**) airport

(**b**) baseball_diamond(**c**) dense_residential(**d**) basketball_court(**e**) mobile_home_park

Figure 1. The samples of large intraclass and small interclass variances in remote sensing images. (**a**) Description of the large intraclass variances of the airport category; (**b**) Description of the small interclass variances among baseball_diamond, dense_residential, basketball_court and mobile_home_park.

In summary, to solve the scarcity problem of labeled data and the separability problem of few-shot classifier for the FSRSSC task, an effective approach is proposed in this paper which is based on data augmentation and multimodal feature fusion. The primary contributions of this work can be summarized as follows:

1) We use the diffusion mode to implement data enhancement, which is applied to solve the challenges of data scarcity in FSRSSC, and in the case of meeting the actual data distribution, it reduces the similarity and performance gap between synthetic and real data.

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 5 March 2026 doi:10.20944/preprints202603.0447.v1

2) Proposal frames are divided for marking local landscape according to the type and spatial distribution of ground objects, and the local feature is extracted from the above frames to improve the feature representation.

3) A multimodal feature fusion module (MFFM) is constructed in this paper. This module obtains the image characteristics with high quality by combining the global channel features, the local space and the semantic ones, to develop classification accuracy in the FSRSSC task.

2. Related Work

2.1. Few-Shot Remote Sensing Scene Classification

FSRSSC is an important branch of remote sensing image classification, which aims to learn a classifier to recognize remote sensing image classes during training with limited labeled examples. Xing et al. [6] proposed to exploit two pre-trained models to classify the remote-sensing scene, respectively. The models comprehensively consider the contribution of various decisions and further improve the discrimination of features. Gong et al. [7] introduced a novel two-path aggregation attention network with quad-patch data augmentation. The network makes it easier to focus on the key clues in a targeted manner. Ji et al. [8] introduced a baseline model using a standard cross-entropy loss leveraging two auxiliary objectives to capture intrinsical characteristics across the semantic classes. Li et al. [9] proposed a multiform ensemble self-supervised learning framework for FSRSSC. The framework can achieve an effective compromise between expensive computational cost and classification accuracy. Ji et al. [10] proposed a two-stage framework that first learns a general-purpose representation and then propagates knowledge in a transductive paradigm. This framework finds an expected prototype having the minimal distance to all samples within the same class. Although transfer learning has shown good results, it requires additional training time and a large amount of labeled data to adapt to the features of the target domain, which may lead to a decline in its performance with few shots. This paper addresses the issue of data scarcity by using the diffusion mode.

Li et al. [11] introduced an attention metric network to improve the performance of one-shot scene classification. The network is composed of a self-attention embedding network and a cross-attention metric network. Li et al. [12] proposed a discriminative learning of adaptive match network. The network can incorporate the channel attention and spatial attention modules, achieving "discriminative learning." Li et al. [13] introduced an end-to-end framework called self-supervised contrastive learning-based metric learning network. The framework learns representative image features from few annotated samples through multi-task learning, and fuses multi-scale spatial features through a novel attention module. Zeng et al. [14] proposed a task-specific contrastive learning model. This model learns feature correlations and reduces the background interference through a self-attention and mutual-attention module. Xu et al. [15] introduced an end-to-end metric learning framework named attention-based contrastive learning network. The framework employs the attention-based feature optimization module to align and enhance target image features. Zhang et al. [16] proposed a few-shot multi-class ship detection algorithm with attention feature map and multi-relation detector. The algorithm enhances the features of the target and optimizes the detection head of YOLO. Although attention mechanisms can improve the classification performance, but they ignore the semantic information of the images. This paper enables the multimodal representations to more closely approximate the real prototype with semantic embeddings.

Cui et al. [17] proposed a method called meta-kernel networks via integrating a parametric linear classifier into the meta-learning framework. The PLC learns prior knowledge and the meta-kernel strategy remaps low-dimensional indistinguishable features to a high-dimensional space. Li et al. [18] proposed a RS-MetaNet. It raises the learning level by organizing training in a meta way, and learns a metric space that can well classify remote sensing scenes from a series of tasks. Qin et al. [19] proposed the deep updated subspace network. The network uses class subspace as a metric benchmark to represent the commonality of a category and can effectively mitigate the negative

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 5 March 2026 doi:10.20944/preprints202603.0447.v1

impact of irrelevant objects on the classifier. Although novel network architectures have acquired good performance, but they have a large number of parameters to be tuned with experience. This paper uses the general four convolution blocks (Conv4) as base network, and the proposal frame and self-attention module are applied to obtain the local features. This network architecture has reduced the total number of parameters while improving the classification performance.

2.2. Data Augmentation

In both general and few-shot image classification, data augmentation expands the number of available images per class and generates novel classes and tasks [25]. To overcome the data scarcity problem for FS-RSSC, several methods have been proposed to augment the training data in different ways. For example, the vast amount of visual data was harvested simply through 2d rotations [22–24] and crops [29]. And some methods used mixup [30] and cutmix [31]. In recent years, some research has employed more refined strategies. Li et al. [32] used GAN networks to emulate the target data distribution. Gong et al. [11] proposed the quad patch method to generate synthetic samples by cutting and reassembling patches from existing images. Ni et al. [33] found that it was more effective by increasing the number of query samples and tasks during meta-testing than increasing the number of support samples during meta-training. Yang et al. [34] proposed to enrich the feature space by simulating the distribution of neighboring classes with spatial vector enhancement. Chen et al. [35] performed data augmentation on images using the SimSiam method. Hou et al. [36] proposed a random augmentation sampling strategy for augmented samples. However, the samples produced through the methods above are always either low quality or may encounter pattern collapse or distortion. In this paper, we embed iterative minor changes into the initial data to generate new samples. It can reduce the performance gap between synthetic and real data, as well as capturing the initial data distribution and ensuring data coverage [37].

3. Materials and Methods

In this section, taking into account the constraints of FSRSSC data scarcity, coupled with the challenges posed by the high intraclass and low interclass variance, we introduce a multimodal feature fusion network (MMFF-Net), as shown in Figure 2. It consists of diffusion augmentation (DA), multiscale feature fusion (MSFF), dual attention fusion module (DAFM), and information interaction mutual attention (IIMA) module. The DA uses the diffusion model for data enhancement through increasing and reducing gaussian noise to the original support data. And the MSFF focuses on obtaining the local spatial details. In addition, the DAFM is utilized to extract critical remote sensing image features with parallel spatial and channels attention. What is more, the IIMA is employed to facilitate interactions between the target features of the support set and the query set, generating the cross-attention map.

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 5 March 2026 doi:10.20944/preprints202603.0447.v1

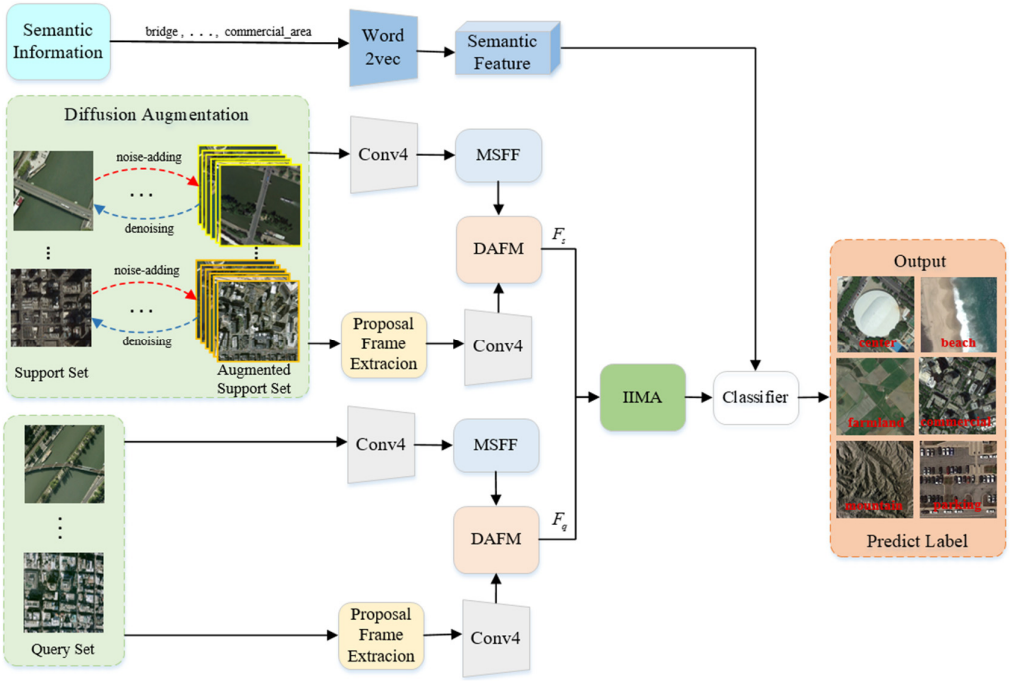

Figure 2. Architecture of multimodal feature fusion network (MMFF-Net).

3.1. Problem Definition

In this study, each RS scene dataset D is divided into three sets: training set D_{train}, validation set D_{val}, and test set D_{test}. The scene categories in D_{train}, D_{val} and D_{test} are distinct and mutually exclusive, that is, $D_{train} \cap D_{val} = \emptyset$, $D_{train} \cap D_{test} = \emptyset$, $D_{val} \cap D_{test} = \emptyset$, $D_{train} \cup D_{val} \cup D_{test} = D$. There are three corresponding tasks which are training tasks, validation tasks, and test tasks, respectively. In each meta-task, we utilize the classical N-way K-shot setup, that is, N distinct categories are randomly selected with K + L scene images in every category. The K labeled samples are designated support set $D_S = (x_i^s, y_i^s)_{i=1}^{N \times K}$ and the remaining L scene images are for query set $D_Q = (x_j^q, y_j^q)_{j=1}^{N \times L}$, where x_i^s denotes the ith support image and y_i^s represents the label of x_i^s. It is similar for x_j^q and y_j^q. The network is trained on the D_S and its performance is evaluated on the D_Q. It is important to note that for each task, the remote sensing scene images in the D_S will not appear in the D_Q ($D_S \cap D_Q = \emptyset$). In this paper, we set N to 5 and K to 1 or 5 by the general experience, corresponding to the five-way one-shot and five-way five-shot settings.

3.2. Method Overview

The architecture of the MMFF-Net proposed in this paper is depicted in Figure 2. We utilize DA for data enhancement in the D_S. Then the enhancement data is put into Conv4 followed by MSFF for extracting local features F_L based on spatial attention. Meanwhile the representative global features F_G based on channel attention are extracted from the proposal frames of the enhancement data through Conv4. Then the F_L and F_G are fused to F_s by DAFM. In order to interact between the D_S and D_Q information, the IIMA module is utilized for the fusion features F_{fusion}, which is more distinct than F_s and F_q. And semantic features are extracted through the Word2Vec [38] model trained on the Wikipedia corpus during training. Finally, the WM Loss function based on the semantic correlation matrices is employed to maintain clear boundaries between different scene categories, ensuring distinctiveness among them.

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 5 March 2026 doi:10.20944/preprints202603.0447.v1

3.3. Diffusion Augmentation

The DA utilizes the diffusion model to enhance the data, where adds and reduces Gaussian noise to the Ds image features to generate labeled samples for augmentation. As is shown in Figure 3, the DA module consists of two processes: forward noise-adding and reverse denoising. In the process of the forward process, the diffusion module gradually adds Gaussian distribution noise to the Ds. For a sample x in the Ds, it is transformed from x_0^s to x_1^s, x_2^s,…, x_T^s based on the time step t. The transitioning probability from x_t^s to x_{t-1}^s at step t is derived as following:

$$P\left(x_t^s \mid x_{t-1}^s\right) = G\left(\sqrt{1-\alpha_t}\, x_{t-1}^s, \alpha_t N\right) \tag{1}$$

where x_t^s and x_{t-1}^s denote the remote sensing images with noise-adding at time steps t and t-1. $G(.)$ represents the Gaussian distribution, $\{\alpha_t\}_{t=0}^T \in (0,1)$ is the diffusion rate at time step t, and N denotes the standard normal distribution. Thus $\sqrt{1-\alpha_t}\, x_{t-1}^s$ and $\alpha_t N$ represent the mean and variance of $P\left(x_t^s \mid x_{t-1}^s\right)$ respectively.

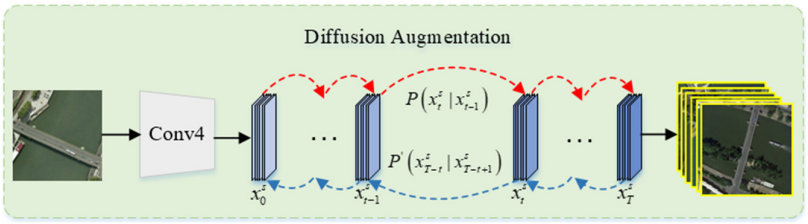

Figure 3. Structure of DA module.

During the reverse denoising process, the diffusion module gradually denoises the remote sensing images with noise-added generated in the forward process. At the tth step, the remote sensing image instance x_{T-t}^s is obtained from the instance x_{T+1-t}^s with Gaussian noise through the denoising process, and the conditional probability is computed according to the following formula:

$$P'\left(x_{T-t}^s \mid x_{T-t+1}^s\right) = G\left(\mu_\theta(x_{T-t+1}^s, t), \sigma_{T-t+1}^2 N\right) \tag{2}$$

where $\mu_\theta(x_{T-t+1}^s, t)$ and σ_{T-t+1}^2 represent respectively the mean and variance of $P'\left(x_{T-t}^s \mid x_{T-t+1}^s\right)$. Thus the augmented image dataset is named D_{aug}, which comprises the Ds and the augmentation images generated through the diffusion module. It will be utilized for training the proposed MMFF-Net and optimizing it parameters.

3.4. Dual Attention Fusion Modul (DAFM)

The proposed DAFM is shown as Figure 4. The conv4 is utilized to extract image features as the basic network. It consists of four convolution blocks, each of which includes 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 [12]. The input RSS image is denoted as $x_i \in \Re^{C\times H\times W}$, where its channels number, height and weight are C, H and W respectively. Then a feature map F can be obtained through the Conv4, which is represented as the following equation:

$$F = g_\varphi(x_i; \varphi) \tag{3}$$

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 5 March 2026 doi:10.20944/preprints202603.0447.v1

where $F \in \Re^{C' \times H' \times W'}$, g_{φ} with the parameters φ represents the Conv4. Based on the feature map F, we can acquire the precise representations of the RSS image through DAFM, which effectively captures and fuses the abundant feature information from both spatial attention and channel attention. The spatial attention is employed to focus on spatial details, such as object positions and local structures, through a local branch in the image, whereas the channel attention is used to integrate global semantic information through a global branch.

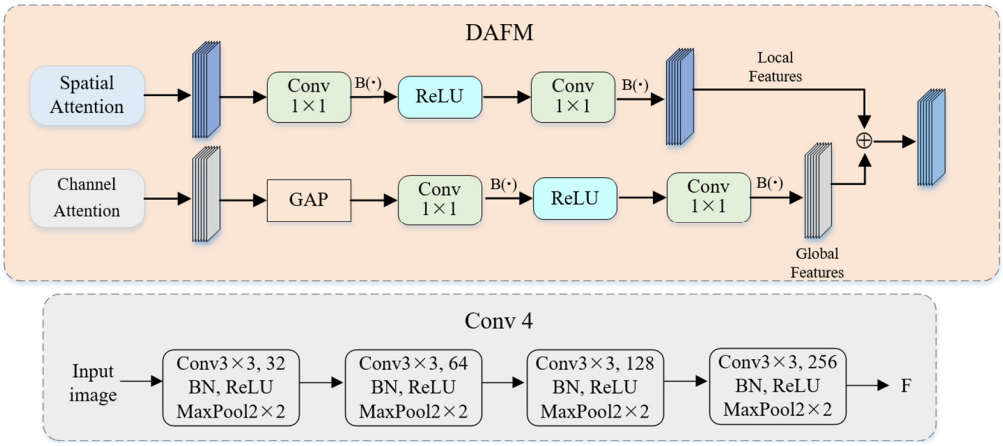

Figure 4. Structure of DAFM module and Conv4.

1) Local Feature Based on Spatial Attention: We utilize the multiscale feature fusion (MSFF) module to obtain the local spatial details. As is shown in Figure 5, we employ four different scale feature extraction branches based on the output of the Con4. Each branch comprises a convolutional, a normalized, and a ReLU layer, and the convolution kernels of the four branches are set to the typical sizes (1×1, 3×3, 5×5 and 7×7). Then the four branch outputs are separately input to the self-attention module (SAM) for catching the relationships among various regions in the RSS images, and the feature map F_c is respectively converted into three different feature spaces *U*, *V* and *W* through a 1×1 convolution.

In order to reduce the channels number, the input channels C_{in} is set *C*′/4 and output channels is set C_{in}- /8 in *U* and *V*, while the output channels C_{out} is set *C*′/4 in *W* [39]. Then *U*, *V* and *W* are reshaped to $\tilde{U}$, $\tilde{V}$ and $\tilde{W}$ along the spatial dimensions. Then the relationships between positions in the spaces $\tilde{U}$ and $\tilde{V}$ can be obtained

$$R = \tilde{U}^T \tilde{V} \tag{3}$$

where $R \in \Re^{H'W' \times H'W'}$, $\tilde{U}^T \in \Re^{H'W' \times (C_{in}/8)}$, $\tilde{V}^T \in \Re^{(C_{in}/8) \times H'W'}$, R_{ij} denotes correlation between the *i*th pixel position and the *j*th pixel position in the feature space. The self-attention feature map between the spaces $\tilde{U}$ and $\tilde{V}$ can be computed according to the following equations:

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

where Softmax(·) represents rowwise normalization, A_{ij} denotets the attention weight of the *j*th column on the *i*th row pixel position, and $\sum_{j=1}^{H'W'} A_{ij} = 1$. And the self-attention feature map F_a of the three spaces can be calculated as follows:

$$reshape(R^{C_{in} \times H'W'}) = R^{C_{in} \times H' \times W'} \tag{5}$$

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 5 March 2026 doi:10.20944/preprints202603.0447.v1

$$F_a = reshape(\tilde{W}A^T) \tag{6}$$

where the *i*th column of $\tilde{W} \in \Re^{C_{in} \times H'W'}$ represents values of all channels at the *i*th pixel position in the space $\tilde{W}$. Then, the final feature map $F_c^{'}$ through the SAM is denoted as follows:

$$F_c^{'} = \gamma F_a + F_c \tag{7}$$

where $F_c^{'} \in \Re^{C_{in} \times H' \times W'}$, and γ denotes the learnable scale parameter.

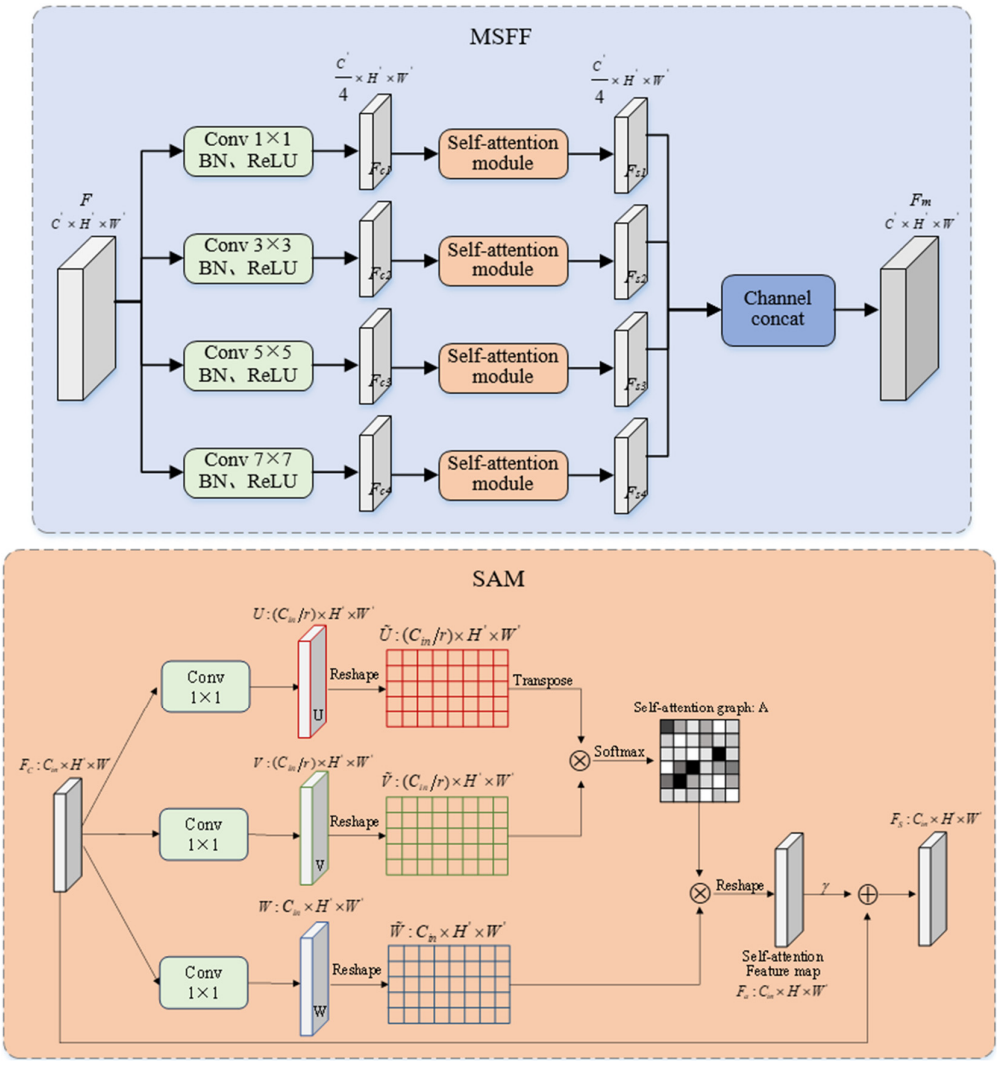

Figure 5. Structure of MSFF module and SAM.

Then we can obtain four different scale feature maps with self-attention in the feature learning branches. The feature maps are presented as $F_{c1}^{'}$, $F_{c2}^{'}$, $F_{c3}^{'}$, and $F_{c4}^{'}$ with the same size $\Re^{C_{in} \times H'W'}$, and they are concatenated to the multiscale feature map F_m with the size $\Re^{C' \times H' \times W'}$. Then the local feature F_L is obtained based on spatial attention.

2) Global Feature Based on Channel Attention: In order to obtain the representative local landscapes in the RSS images, we employ a proposal frame extraction strategy [40] which is designed to emphasize on the spatial distribution and types of land cover within the scene. It bases on the case that the more diverse and denser the distribution of land features, the more likely a proposal frame

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 5 March 2026 doi:10.20944/preprints202603.0447.v1

represents a local landscape. As is shown in Figure 6, we choose (including, but not limited to) three typical representative land feature types from the diverse types and manifestations of local landscapes: buildings, vegetation and impervious surfaces, and extract the corresponding semantic pixel set by using MBI index [41], NDI index [42] and BCI index [43]. Then the set of proposal frames can be extracted with a clustering index DS_{loc} through combining the richness of feature types and their distribution density.

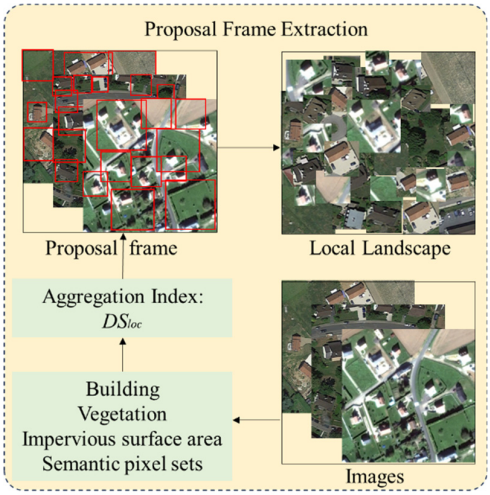

Figure 6. Structure of proposal frame extraction module.

The DS_{loc} is computed according to the following formula:

$$DS_{loc} = \sum B_m (1 - \frac{d_i}{H}) \Big/ \sum (1 - \frac{d_i}{H}) \tag{8}$$

where H denotes the diagonal length of the current proposal frame F_{rame}, d_i signifies the distance from the particular pixel i to the center pixel of F_{rame}, and B_m (m = 1, 2, 3) denotes respectively the values of MBI, ND and BCI for the given pixel i. The specific steps in detail are as follows:

Step 1: For a particular pixel i in the image, it is used as the center and performed proposal frame growth with 8-CONNECTIVITY [44]. Proposal frame growing stops and the current proposal frame is recorded as $Frame_i$ when the DS_{loc} mentioned above increases continuously for three iterations and then decreases continuously for three iterations. Otherwise, if it reaches the image boundary during proposal frame growth, there is no corresponding proposal frame for pixel i.

Step 2: Repeat Step 1 by traversing all pixels in the image to obtain all the proposal frame collection, denoted as $Frame_{all}$.

Step 3: To avoid falling into local optima, the total number of image pixels is represented as $pixel_{total}$ and divided into ten equal intervals. On this basis, the number of pixels is calculated in each proposal frame of $Frame_{all}$ and they are clustered based on the respective intervals. Finally, the optimal proposal frame collection $Frame_{opt}$ is obtained by choosing the proposal frames with the highest DS_{loc} value (or tied highest values) from each interval for the image x_i.

Then the $Frame_{opt}$ is input into the Conv4 followed by channel attention for extracting the representative global feature F_G.

3) Feature Fusion Based on Dual Attention: After generating the F_G and F_L, the DAFM is utilized for their fusion. As is shown in Figure 4, To minimize the number of parameters, the parameter matrices for Conv$_1$1×1 and Conv$_2$1×1 are set respectively as $M_1 \in \mathbb{R}^{c \times c/r}$ and $M_2 \in \mathbb{R}^{c/r \times c}$, and r is a constant. As is mentioned above in the paper, F_G and F_L concentrate on global semantics and local detail of the input RSS image respectively. They are fused according to the following formula:

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 5 March 2026 doi:10.20944/preprints202603.0447.v1

$$F_s = F_L \oplus F_G \tag{9}$$

where $\oplus$ denotes element-wise addition, thus the F_s is more detailed and richer in the image information.

3.5. Information Interaction Mutual Attention (IIMA)

To interact the information between the query and support sets, the IIMA module is utilized to generate a cross-attention map, represented as A_s (A_q), which serves to emphasize the regions corresponding to the target objects. It can assist the network in learning more distinct image features. As is shown in Figure 7, the IIMA module comprises two primary modules:

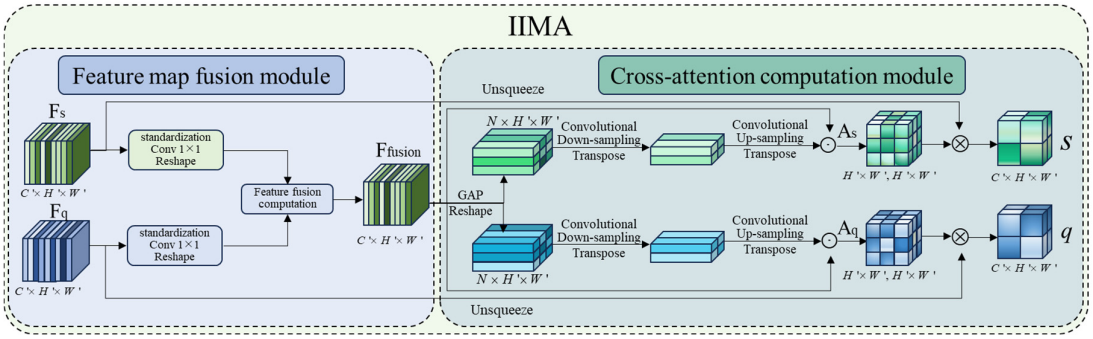

Figure 7. Structure of IIMA module.

1) Feature map fusion module: The support set feature map $F_s \in \Re^{C' \times H' \times W'}$ and the query set feature map $F_q \in \Re^{C' \times H' \times W'}$ are standardized by a 1 × 1 convolution operation across the channel dimension. Then the fusion feature $F_{fusion} \in \Re^{H' \times W' \times H' \times W'}$ is calculated as follows:

$$F_{fusion}(x_s, x_q) = sim(F_s, F_q) \tag{9}$$

where x_s and x_q represent the corresponding spatial positions within the feature maps F_s and F_q, and sim denotes the cosine similarity function.

2) Cross-attention computation module: To pay more attention to the information of the target regions, this cross-attention module is utilized for identifying these regions through the similarities shared by the support and query sets. In the support set branch, the H' and W' dimensions of the feature map F_{fusion} are eliminated by a GAP operation, then the resulting tensor is downsampled and upsampled respectively through the convolution module. The weights ω of the target regions are computed as follows:

$$\omega = conv(Relu(GAP(F_{fusion}))) \tag{10}$$

where *conv* represents the convolution operation, while *Relu* denotes a nonlinear activation function, and *GAP* signifies the global average pooling. Then the final attention map $A_s \in \Re^{H' \times W' \times H' \times W'}$ is calculated as follows:

$$A_s(x_s) = \sum_{x_q} \frac{\exp(C(\omega^T x_s, \omega^T x_q)/\tau)}{\sum_{x_s}^{H' \times W'} \exp(C(\omega^T x_s, \omega^T x_q)/\tau)} \tag{11}$$

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 5 March 2026 doi:10.20944/preprints202603.0447.v1

where x_s and x_q represent the spatial positions within the final feature map, and $C(\omega^T x_s, \omega^T x_q)$ signifies the matching values of x_s and x_q according to their similarity, while τ denotes the temperature factor.

In the support set D_S, the final embedding feature map is signified by $s \in \Re^{C_s \times H \times W}$, which is calculated as the following equation:

$$s = \sum_{x_s} A_s(x_s) \otimes F_s(x_s) \tag{12}$$

where $\otimes$ denotes element-wise multiplication operation. Similarly, the embedding feature map $q \in \Re^{C_q \times H \times W}$ in the query set D_Q is computed as follows:

$$q = \sum_{x_q} A_q(x_q) \otimes F_q(x_q) \tag{13}$$

In the IIMA module, the cross-attention map is the critical portion, which focuses on the information of the target regions within both the query set and the support set. Thus the module can emphasize the relevance of features and the final feature map is more distinct.

3.6. Loss Function

In order to compare the final embedding feature map F (s or q) with the semantic information of the scene, we utilize the Word2Vec model trained on the Wikipedia corpus to transform each scene class name into a semantic vector, as is shown in Figure 2, and obtain the semantic vector sets $u = \{u_1, u_2, ..., u_v\}$, where v is the total number of scene categories. For a scene class u_i, the loss function based on the popular LSE Loss (least square embedding loss) [45] is as the following equation:

$$L(u) = \sum \|F - u_i\|^2 + \lambda \|Z\| \tag{14}$$

where F represents s or q, λ is a regularization parameter, and Z denotes a randomly initialized encoding matrix for the semantic information branch to align the visual feature and semantic feature.

4. Results

4.1. Dataset Introduction

To validate the efficacy of our proposed approach, experiments were conducted on three public benchmark RS datasets: UC Merced [46], AID [47], and NWPU-RESISC45 [48]. In order to make a fair comparison, the datasets were divided into training, validation, and test sets with the scheme used in [16,49,50]. Table I shows the details of three RS scene datasets used for the FSRSSC task.

The UC Merced (UCM) dataset was created by the UC Merced Computer Vision Laboratory. The dataset consists of 21 scene categories with 100 images in each class, accumulating to a total of 2100 images. Every image is 256 × 256 pixels with a spatial resolution of 0.3 m per pixel. As is shown in Table I, the training set includes 10 classes, the validation set includes 5 ones, and the testing set includes the remaining 6 ones [16,49].

The AID dataset, created by Wuhan University and Huazhong University of Science and Technology, consists of 30 scene categories with a total of 1000 RGB images. Each class comprises 220~420 images with 600 × 600 pixels, and the spatial resolution varies from 0.5 to 8 m. As shown in Table 1, we adopt 16, 7 and 7 classes for training, validation, and testing, respectively [50].

The NWPU-RESISC45 dataset, provided by Northwestern Polytechnical University (NWPU), contains 45 scene categories with a total of 31500 images. Each class has 700 images with 256 × 256

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 5 March 2026 doi:10.20944/preprints202603.0447.v1

pixels. The spatial resolution of the most images varies from 0.2 to 30 m per pixel. As shown in Table I, the number of classes for the training, validation, and testing are 25, 10, and 10, respectively [16,49].

Table 1. Three remote sensing scene datasets used for the FSRSSC task.

Datasets	Training classes	Validation classes	Testing classes
UCM	airplane, baseball diamond, chaparral, denser residential, freeway, golf course, harbor, mobile home park, overpass, parking lot	agricultural, tanks, forest, runway, sparse residential	beach, river, buildings, medium residential, intersection, storage, tennis court
AID	airport, park, bare land, center, desert, farmland, industrial, medium residential, parking, playground, pond, viaduct, railway station, resort, school, stadium	baseball field, bridge, church, commercial area, port, meadow, river, storage tanks	beach, commercial, dense residential, mountain, forest, sparse residential, square
NWPU-RESISC45	airport, baseball diamond, bridge, chaparral, church, cloud, desert, forest, freeway, golf course, harbor, island, lake, meadow, mountain, overpass, palace, rectangular farmland, railway, roundabout, sea ice, sparse residential, stadium, wetland, thermal	beach, terrace, power station, industrial area, mobile home park, railway station, river, snowberg, storage tank, tennis court	airplane, basketball court, circular farmland, dense residential, ground track field, intersection, medium residential, parking lot, runway, ship,

4.2. Implementation Details

All models were implemented within the PyTorch framework, and run on hardware with an NVIDIA A40 GPU. Prior to model input, images underwent preprocessing to a fixed spatial resolution of 256×256 pixels. And each task was set as 5-way 1-shot and 5-way 5-shot as usual, with each meta-task comprising 15 query samples chosen for each class. During the meta-training phase, we trained for 100 epochs, with each epoch consisting of 1000 training meta-tasks and 600 validation ones. The classification accuracy of our method was the average of 5 runs randomly sampling 2000 meta-tasks in each run, with a 95% confidence interval. The final chosen hyperparameter settings were outlined as follows: the learnable scale parameter γ in the SAM was set to 0.6, the constant r used to minimize the number of parameters in the DAFM was set to 8, the temperature coefficient τ for the IIMA module was set to 5, and the regularization parameter λ was set to 0.05. Furthermore, we utilized SGD to refine the model parameters with an initial learning rate of 0.0001 and a weight decay factor of 0.0005. At 1-shot and 5-shot, the learning rate decayed every 100 epochs and 50 ones with a decay factor of 0.5 and 0.1 respectively.

4.3. Results Comparison of Multiple Models

To clearly validate the advantages of our MMFF-Net, we conduct a direct comparison with several FSL methods, including MatchingNet [51], MAML [52], ProtoNet [53], RelationNet [54], LLSR[55], DeepEMD [47], DLA-MatchNet [16], FRN [56], MCL-Katz [57], GLIML [58], SPNet [49], SCL-MLNet [17], SPFSR [59], MPCLNet [60], HProtoNet [61], MSoP-Net2024 [62], and FS-RSSCvS [39]. The experimental comparison results on the three RS datasets are shown in Table 2. The optimal results are emphasized in bold, while the sub-optimal results are marked with an underline, respectively.

Table 2. Classification accuracy of various methods.

Method	Year	5-way 1-shot			5-way 5-shot		
		UCM	AID	NWPU-RESISC45	UCM	AID	NWPU-RESISC45
MatchingNet [47]	2016	34.70	33.87	37.61	52.71	50.40	47.10
MAML [48]	2017	48.86±0.74	43.20±0.77	48.40±0.82	60.78±0.62	60.37±0.75	62.90±0.69
ProtoNet [49]	2017	52.27±0.20	54.32±0.86	40.33±0.18	69.86±0.15	67.80±0.64	63.82±0.56
RelationNet [50]	2018	48.08±1.67	54.62±0.80	66.43±0.73	61.88±0.50	68.80±0.66	78.35±0.51

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 5 March 2026 doi:10.20944/preprints202603.0447.v1

LLSR [51]	2019	39.47	45.18	51.43	57.40	61.76	72.90
DeepEMD [43]	2020	58.47±0.76	61.04±0.77	64.39±0.84	70.42±0.58	74.51±0.55	78.01±0.56
DLA-MatchNet [12]	2021	53.76±0.62	61.99±0.94	68.80±0.70	63.01±0.51	75.03±0.67	81.63±0.46
FRN [52]	2021	50.89±0.37	**62.29±0.37**	64.98±0.42	68.34±0.30	79.33±0.24	81.65±0.25
MCL-Katz [53]	2022	50.73	55.28	63.30	68.95	75.66	80.78
GLIML [54]	2022	56.41±0.62	61.28±0.61	66.86±0.68	70.40±0.41	79.56±0.41	78.91±0.45
SPNet [45]	2022	57.64±0.73	-	67.84±0.87	73.52±0.51	-	83.94±0.50
SCL-MLNet [13]	2022	51.37±0.79	59.46±0.96	62.21±1.12	68.09±0.92	76.31±0.68	80.86±0.76
SPFSR [55]	2023	55.40±1.11	60.01±1.09	65.97±1.22	71.38±0.77	75.40±0.76	80.72±0.79
MPCLNet [56]	2023	56.46±0.21	60.61±0.43	55.94±0.04	76.57±0.07	76.78±0.08	76.24±0.12
HProtoNet [57]	2024	57.51±0.95	59.78±0.58	66.41±0.87	75.08±0.29	75.87±0.35	82.71±0.41
MSoP-Net [58]	2024	54.27±0.60	-	67.05±0.80	69.77±0.38	-	82.02±0.46
FS-RSSCvS [35]	2024	59.05±0.84	61.62±0.75	67.05±0.80	76.34±0.51	81.32±0.45	84.00±0.46
Ours		**60.23±0.75**	61.78±0.92	**69.58±0.36**	**77.65±0.48**	**82.76±0.64**	**84.82±0.51**

The comparison results of the classification accuracy on the UCM,AID and NWPU-RESISC45 datasets are shown in Table II. Our method attains better classification results than the other approaches listed above, except that its performance is slightly worse than that of DLA-MatchNet on AID within the 5-way 1-shot task. It maybe relative to the image sizes of the datasets, which are 600×600 pixels in the AID dataset, while we uniformly resized all the images to 256×256 pixels as inputs in the experiments of our method. It caused more feature loss to the images in the AID dataset while resizing. Thus the feature extraction ability of the proposed multiscale feature extraction network is negatively affected. At the same time, there is only one labeled sample per class within the 5-way 1-shot task. The inaccuracy of the extracted features affects the representation for the scene images and leads to the lower accuracy.

In addition, Table 3 records the algorithmic efficiency results of our method and the ones [39]. As is shown, parameter count (Params), floating-point operations (FLOPs), and inference time (Time) are compared. Although the efficiency of our method is not the best, the difference between it and the FRN or the MCL-Katz is not very obvious. It is superior to other methods for the FSRSSC according to the classification accuracy and efficiency.

Table 3. Algorithmic efficiency of various methods on the NWPU-RESISC45 dataset.

Method	Params	5-way 1-shot			5-way 5-shot		
	(M)	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	-	-	-	-	-	-
MSoP-Net	2.10	-	-	-	-	-	-
FS-RSSCvS	1.85	47	26.91	5.95	41	33.63	6.46
Ours	1.87	45	27.10	6.22	53	33.86	6.78

4. Discussion

To verify the effectiveness of the individual components within our MMFF-Net, ablation experiments were also conducted in both 5-way 1-shot and 5-way 5-shot on three datasets: UCM, AID, and NWPU-RESISC45, as is presented in Table 4.

There are mainly four parts in our MMFF-Net, they are DA, MSFF, DAFM, and IIMA. First, every part was separately conducted, it can be seen from the results that DA was more effective than the other three parts in both 5-way 1-shot and 5-way 5-shot, and MSFF was more effective in 5-way 1-shot, while DAFM was more effective in 5-way 5-shot. It might because that MSFF could extract

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 5 March 2026 doi:10.20944/preprints202603.0447.v1

more different scale features when the sample was only one in 5-way 1-shot, while DAFM could improve the feature representation capability through the dual attention fusion when the number of samples added to 5. In the second type of experiments, the three ones of four parts were conducted in combination. It is found that IIMA was the least effective in the four parts as same as the single part experiments. And the best results were obtained by combining the four parts which can complement each other, leading to improve the feature representation of the scene images.

Table 4. Results of ablation studies on the three datasets.

DA	MSFF	DAFM	IIMA	5-way					
				UCM		AID		NWPU-RESISC45	
				1-shot	5-shot	1-shot	5-shot	1-shot	5-shot
✓	×	×	×	43.73±0.35	58.26±0.45	43.84±0.65	62.46±0.54	49.24±0.57	65.15±0.78
×	✓	×	×	42.85±0.67	56.83±0.71	43.12±0.64	60.39±0.84	47.57±0.64	63.57±0.38
×	×	✓	×	42.79±0.58	57.63±0.47	42.57±0.46	61.78±0.36	48.95±0.46	64.13±0.56
×	×	×	✓	41.64±0.76	55.61±0.54	41.85±0.65	59.17±0.68	46.38±0.72	62.64±0.41
✓	✓	✓	×	59.36±0.54	76.95±0.73	60.32±0.57	80.41±0.65	68.85±0.43	83.83±0.76
✓	✓	×	✓	57.67±0.68	74.18±0.34	58.01±0.83	81.53±0.47	67.73±0.56	82.51±0.63
✓	×	✓	✓	58.21±0.83	75.36±0.47	59.18±0.65	79.67±0.49	66.57±0.82	81.34±0.58
✓	✓	✓	✓	**60.23±0.75**	**77.65±0.48**	**61.78±0.92**	**82.76±0.64**	**69.58±0.36**	**84.82±0.51**

5. Conclusions

In this article, an innovative framework is presented for few-shot remote sensing image classification, which is named as MMFF-Net. This framework comprises four core components: DA, MSFF, DAFM, and IIMA. To overcome the samples scarcity, we utilize the DA to augment support set samples. To combine the local and global features, the DAFM is applied for feature fusion, especially obtaining the local spatial details with the MSFF based on spatial attention and extracting the representative global feature in the proposal frames based on channel attention. Besides, we employ the IIMA module to interact between the query and support set information, it can enhance the distinctiveness of the extracted features. What is more, we use word2vec to obtain the semantic features for reducing the disparity between them and the visual features with LSE Loss. The comparative experimental results with multiple models on three benchmark remote sensing scene datasets validate the effectiveness of the proposed MMFF-Net, showcasing the superiority and feasibility of our approach in most FSRSSC cases.

In future work, we plan to focus on optimizing the approach efficiency with strategies such as lightweight architecture design, model compression, adaptive computation techniques, etc. The aim is to achieve an effective balance between classification accuracy and efficiency, thereby enhancing the applicability and feasibility of the model further in the real-world scenarios.

Author Contributions: Conceptualization, Z.Y. and S.H.; methodology, Z.Y. and S.H.; software, Z.Y., S.H. and M.W.; validation, Z.Y., S.H. and M.W.; investigation, Z.Y., S.H. and M.W.; resources, M.W.; writing—original draft preparation, Z.Y., S.H. and M.W.; writing—review and editing, Z.Y., S.H. and M.W.; supervision, Z.Y. and S.H.; funding acquisition, M.W. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Natural Science Basic Research Plan in Shaanxi Province of China, grant number 2025JC-YBMS-730.

Data Availability Statement: The remote sensing datasets employed in this study: UCM, AID, and NWPU-RESISC45, are publicly available.

Acknowledgments: This work was supported by the Natural Science Basic Research Plan in Shaanxi Province of China under Grant 2025JC-YBMS-730.

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

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 5 March 2026 doi:10.20944/preprints202603.0447.v1

Abbreviations

The following abbreviations are used in this manuscript:

FSRSSC	Few-shot remote sensing scene classification
MMFF-Net	multimodal feature fusion network
DA	Diffusion augmentation
MSFF	Multiscale feature fusion
DAFM	Dual attention fusion module
IIMA	Information interaction mutual attention
RSS	Remote sensing scene
SAM	Self-attention module
UCM	UC Merced
NWPU	Northwestern Polytechnical University
FLOPs	Floating-point operations

References

5. 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.
6. 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.
7. Q. Zhao, S. Lyu, Y. Li, Y. Ma, and L. Chen, "MGML: Multigranularity multilevel feature ensemble network for remote sensing sceneclassification," IEEE Trans. Neural Netw. Learn. Syst., vol. 34, no. 5, pp. 2308–2322, May 2023.
8. Z. Ji, L. Hou, X. Wang, G. Wang, and Y. Pang, "Dual contrastive networkfor few-shot remote sensing image scene classification," IEEE Trans. Geosci. Remote Sens., vol. 61, 2023, Art. no. 5605312.
9. J. Li, K. Zheng, J. Yao, L. Gao, and D. Hong, "Deep unsupervised blind hyperspectral and multispectral data fusion," IEEE Geosci. Remote Sens. Lett., vol. 19, 2022, Art. no. 6007305.
10. L. Xing, S. Shao, Y. Ma, Y. Wang, W. Liu, and B. Liu, "Learning to cooperate: Decision fusion method for few-shot remote-sensing scene classification," IEEE Geosci. Remote Sens. Lett., vol. 19, pp. 1–5, 2022.
11. M. Gong, J. Li, Y. Zhang, Y. Wu, and M. Zhang, "Two-path aggregation attention network with quad-patch data augmentation for few-shot scene classification," IEEE Trans. Geosci. Remote Sens., vol. 60, 2022, Art. no. 4511616.
12. H. Ji, Z. Gao, Y. Zhang, Y. Wan, C. Li, and T. Mei, "Few-shot scene classification of optical remote sensing images leveraging calibrated pretext tasks," IEEE Trans. Geosci. Remote Sens., vol. 60, 2022, Art. no. 562551.
13. J. Li, M. Gong, H. Liu, Y. Zhang, M. Zhang, and Y. Wu, "Multi-form ensemble self-supervised learning for few-shot remote sensing scene classification," IEEE Trans. Geosci. Remote Sens., vol. 61, 2023, Art. no. 4500416.
14. H. Ji, H. Yang, Z. Gao, C. Li, Y. Wan, and J. Cui, "Few-shot scene classification using auxiliary objectives and transductive inference," IEEE Geosci. Remote Sens. Lett., vol. 19, pp. 1–5, 2022.
15. X. Li, F. Pu, R. Yang, R. Gui, and X. Xu, "AMN: Attention metric network for one-shot remote sensing image scene classification," Remote. Sens., vol. 12, no. 24, Dec. 2020, Art. no. 4046.
16. 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.
17. 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, 2022, Art. no. 5801112.
18. Q. Zeng and J. Geng, "Task-specific contrastive learning for few-shot remote sensing image scene classification," ISPRS J. Photogramm., vol. 191, pp. 143–154, Sep. 2022.
19. Y. Xu et al., "Attention-based contrastive learning for few-shot remote sensing image classification," IEEE Trans. Geosci. Remote Sens., vol. 62, Apr. 2024, Art. no. 5620317.
20. H. Zhang, X. Zhang, G. Meng, C. Guo, and Z. Jiang, "Few-shot multi-class ship detection in remote sensing images using attention feature map and multi-relation detector," Remote Sens., vol. 14, no. 12, 2022, Art. no. 2790.）.

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 5 March 2026 doi:10.20944/preprints202603.0447.v1

21. Z. Cui, W. Yang, L. Chen, and H. Li, "MKN: Metakernel networks for few-shot remote sensing scene classification," IEEE Trans. Geosci. Remote Sens., vol. 60, 2022, Art. no. 4705611.
22. 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.
23. A. Qin et al., "Deep updated subspace networks for few-shot remote sensing scene classification," IEEE Trans. Geosci. Remote Sens., vol. 62, Jan. 2024, Art. no. 5606714.
24. Teng L, Gao S. Multimodal Feature Calibration Network for Few-Shot Remote Sensing Image Scene Classification[C]//2025 IEEE 6th International Seminar on Artificial Intelligence, Networking and Information Technology (AINIT).0[2026-01-13]. DOI:10.1109/AINIT65432.2025.11035263.
25. C. Shorten and T. M. Khoshgoftaar, "A survey on image data augmentation for deep learning," J. Big Data, vol. 6, no. 1, pp. 1–48, 2019.
26. S. Gidaris, P. Singh, and N. Komodakis, "Unsupervised representation learning by predicting image rotations," 2018, arXiv:1803.07728.
27. Ji H, Gao Z, Lu Y, et al. Semi-supervised few-shot classification with multitask learning and iterative label correction[J]. IEEE Transactions on Geoscience and Remote Sensing, 2024, 62: 1-15.
28. Liu Y, Li J, Gong M, et al. Collaborative self-supervised evolution for few-shot remote sensing scene classification[J]. IEEE Transactions on Geoscience and Remote Sensing, 2024.
29. 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/CVF Conf. Comput. Vis. Pattern Recognit.,Jun. 2020, pp. 12203–12213.
30. J.-H. Kim, W. Choo, and H. O. Song, "Puzzle mix: Exploiting saliency and local statistics for optimal mixup," in Proc. Int. Conf. Mach. Learn., 2020, pp. 5275–5285.
31. S. Yun, D. Han, S. Chun, S. J. Oh, Y. Yoo, and J. Choe, "CutMix: Regularization strategy to train strong classifiers with localizable features," in Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV), Oct. 2019, pp. 6023–6032.
32. K. Li, Y. Zhang, K. Li, and Y. Fu, "Adversarial feature hallucination networks for few-shot learning," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2020, pp. 13470–13479.
33. R. Ni, M. Goldblum, A. Sharaf, K. Kong, and T. Goldstein, "Data augmentation for meta-learning," in Proc. 38th Int. Conf. Mach. Learn., 2021, pp. 8152–8161.
34. S. Yang, L. Liu, and M. Xu, "Free lunch for few-shot learning: Distribution calibration," 2021, arXiv:2101.06395.
35. Chen X, He K. Exploring simple siamese representation learning[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2021: 15750-15758.
36. Hou L, Ji Z, Wang X, et al. Diversity-infused network for unsupervised few-shot remote sensing scene classification[J]. IEEE Geoscience and Remote Sensing Letters, 2024, 21: 1-5.
37. Zhu Y, Han J, Pan B, et al. DiffPR-Net: Few-shot remote sensing scene classification based on generative diffusion and prototype rectified model[J]. IEEE Transactions on Geoscience and Remote Sensing, 2025.
38. T. Mikolov, K. Chen, G. S. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," in Proc. Int. Conf. Learn. Representations, 2013, pp. 1–12.
39. Qin A, Chen F, Li Q, et al. Few-shot remote sensing scene classification via subspace based on multiscale feature learning[J]. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 2024.
40. Wang C, Li J, Tanvir A, et al. Zero-shot remote sensing scene classification method based on local-global feature fusion and weight mapping loss[J]. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 2023, 17: 2763-2776.
41. X. Huang and L. Zhang, "Multidirectional and multiscale morphological index for automatic building extraction from multispectral GeoEye-1 imagery," Photogrammetric Eng. Remote Sens., vol. 77, pp. 721–732, 2011.
42. Y. Fu et al., "Winter wheat nitrogen status estimation using UAV-based RGB imagery and Gaussian processes regression," Remote Sens., vol. 12, no. 22, 2020, Art. no. 3778.

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 5 March 2026 doi:10.20944/preprints202603.0447.v1

43. C. Deng and C. Wu, "BCI: A biophysical composition index for remote sensing of urban environments," Remote Sens. Environ., vol. 127, pp. 247–259, 2012.
44. R. Tao and J. Qiao, "Fast component tree computation for images of limited levels" IEEE Trans. Pattern Anal. Mach. Intell., vol. 45, no. 3, pp. 3059–3071, Mar. 2023.
45. Türkşen, I. Burhan. "Fuzzy functions with LSE." Applied Soft Computing 8.3 (2008): 1178-1188.
46. Y. Yang and S. Newsam, "Bag-of-visual-words and spatial extensions for land-use classification," in Proc. 18th SIGSPATIAL Int. Conf. Adv.Geograph. Inf. Syst., Nov. 2010, pp. 270–279.
47. 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.
48. 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.
49. 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.
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. 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.
52. 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.
53. 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.
54. 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.
55. 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.
56. 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.
57. 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.
58. 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.
59. 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.
60. 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.
61. 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.
62. 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.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.