Mathematical Biosciences and Engineering

MBE, 20(7): 12889–12907.
DOI: 10.3934/mbe.2023575
Received: 24 March 2023
Revised: 18 May 2023
Accepted: 25 May 2023
Published: 05 June 2023

http://www.aimspress.com/journal/MBE

Research article

Few-shot remote sensing scene classification based on multi subband deep feature fusion

Song Yang[1,2], Huibin Wang[1,*], Hongmin Gao[1] and Lili Zhang[1]

[1] College of Computer and Information, Hohai University, Nanjing 211100, China
[2] Faculty of Electronic Information Engineering, Huaiyin Institute of Technology, Huaian 223001, China

* **Correspondence:** Email: hbwang@hhu.edu.cn.

Abstract: Recently, convolutional neural networks (CNNs) have performed well in object classification and object recognition. However, due to the particularity of geographic data, the labeled samples are seriously insufficient, which limits the practical application of CNN methods in remote sensing (RS) image processing. To address the problem of small sample RS image classification, a discrete wavelet-based multi-level deep feature fusion method is proposed. First, the deep features are extracted from the RS images using pre-trained deep CNNs and discrete wavelet transform (DWT) methods. Next, a modified discriminant correlation analysis (DCA) approach is proposed to distinguish easily confused categories effectively, which is based on the distance coefficient of between-class. The proposed approach can effectively integrate the deep feature information of various frequency bands. Thereby, the proposed method obtains the low-dimensional features with good discrimination, which is demonstrated through experiments on four benchmark datasets. Compared with several state-of-the-art methods, the proposed method achieves outstanding performance under limited training samples, especially one or two training samples per class.

Keywords: remote sensing scene classification; deep feature fusion; discriminant correlation analysis; discrete wavelet transform

1. Introduction

Remote sensing (RS) images play a significant role in urban planning, land cover and land use (LCLU), agriculture management, etc. [1−4], not only due to the high spatial resolution, but also abundant structural patterns. To use these RS images sufficiently, remote sensing scene classification (RSSC) is imperative. The appropriate feature representation method plays a key role in RSSC. Due to the limited labeled RS images, it is still a challenging and complex issue to represent and classify RS scenes by using more intelligent and convenient methods.

Various efforts have been devoted to developing various methods for feature representation. Traditional methods are bag-of-the-visual words (BoVW) and many improvements or extensions of BoVW [1,5,6]. Since AlexNet [7] got the best score in the Large-Scale Visual Recognition Challenge in 2012, plenty of deep learning methods have sprung up [8−11]. Generally, CNN-based methods can be classified into three categories: training CNNs from scratch, fine-tuning pre-trained CNNs, and using pre-trained CNNs as feature extractors. Full-training-based methods mainly focused on the building of deep networks to enhance accuracy. Such methods usually improve currently available advanced models or rebuild the CNN structure to obtain astonishing scene classification results [12−16]. Wu et al. [16] took convolutional neural networks (CNNs) as a backbone to construct a deep-learning-based framework for multimodal RS data classification. Fine-tuning methods usually involve adjusting pre-trained CNNs or optimizing their loss functions to improve classification accuracy [17−21]. However, they generally require a significant number of labeled training samples, high-performance computer equipment, and take a very long time to fine-tune pretrained CNNs or train a new network.

The CNN-based methods mentioned above utilize practical features to classify remote sensing scenes. However, it is nontrivial to obtain the features that can adequately represent the scene in the case of few training samples. The lack of available data will make the neural network overfitting, which will lead to performance degradation. To tackle this problem, many few-shot based methods have been developed [22–25]. Wu et al. [26] proposed a "U-Net in U-Net" framework to detect small objects in infrared images. Mei et al. [27] presented a sparse representation-based framework and obtained a satisfactory result. However, in the case of very few samples, this method is not enough to describe the key semantic features, and there is still a lack of discrimination for remote sensing images of the same category with the diversity of direction scales. Zeng et al. [28] proposed a prototype calibration to enhance the representation of feature in few-shot RSSC task. Yang et al. [29] emphasized the importance of the underlying features in the classification of small samples, which improves the ability to characterize the feature of small samples, but the computational complexity is large. The feature-wise transformation can be employed for RSSC and land-cover mapping tasks [30]. Tseng et al. [31] used feature-wise transformation layers for addressing the problem of few-shot classification under domain shifts for metric-based methods. Chen et al. [32] proposed a feature-wise transformation module address the difficulty of cross-domain RSSC tasks with few training samples, and pointed out that transfer-based methods may outperform sophisticated few-shot learners. Chowdhury et al. [33] proposed a library of pre-trained feature extractors combined with a feed-forward network to solve few-shot image classification task. Recently, few-shot learning is presented to address a series of few-shot tasks. Discriminative learning of adaptive match network (DLA-MatchNet), an end-to-end network, was proposed for boosting a few-shot RSSC [34]. Deep nearest neighbor neural network (DN4) is proposed to exploit deep local descriptors and the image-to-class measure for classification, which is one of the most advanced networks for few-shot scene classification of remote sensing

images [35]. Huang et al. [25] proposed a meta-learning-based task-adaptive embedding network to enhance the generalization performance of the model for few-shot settings. These few-shot classification methods almost focused on the *C*-way *K*-shot problems. In addition, many studies focused on deep learning-based fusion strategy to generate a more comprehensive feature representation [36–39]. Hong et al. [39] proposed a cross fusion strategy to solve the multi-modality learning issue.

Although the above methods have acquired high accuracy, the features extracted from the deep learning approaches are usually high-dimensional and redundant. To further improve the classification performance of RSSC, there is still a thorny road to go between improving feature utilization and reducing computational complexity. Chaib et al. [40] adopted discriminant correlation analysis (DCA) to combine the deep features extracted from different fully connected layers CNNs, which provide an efficient and low-cost feature fusion strategy. However, the categories with a small distance between classes are become closer in the mapping space, which leads to overlap in the mapping space. Motivated by this, an improved DCA strategy is proposed in this paper. The key difference from the related method lies in that it reconstructed the between-class scatter matrix by introducing the distance coefficient, which helps adjusting the distance between classes and avoiding cross overlap in mapping space. In our approach, the features from different CNNs and different frequency bands are integrated by the improved DCA, thus enriches the expression of feature semantics and overcomes the limitation of the number of categories, especially for small number of categories in the remote sensing data set. Specifically, a discrete wavelet-based multilevel feature fusion (DWMLFF) strategy is proposed to fuse multi-sub-band features extracted from different CNNs for few-shot RSSC. The discrete wavelet transform (DWT) is employed as a decomposer to extend the multi-sub-band information of limited samples to surmount the problem of insufficient features in few-shot RSSC. The transfer learning-based CNN model is used as a feature extractor to generated the deep feature of original image and multi-sub-bands. Furthermore, an improved DCA method is proposed to integrate all the obtained deep feature. In the improved DCA method, we reconstruct the between-class scatter matrix by introducing distance coefficient, which addresses the overlap of categories in the mapping space to distinguish easily confused categories. The proposed method gives full play to the advantages of different wavelet sub-bands, and utilizes an improved DCA strategy to deeply integrate different frequency components to obtain low-dimensional and high-discriminative features for few-shot RSSC.

2. Materials and methods

The proposed method is comprised of the following parts: discrete wavelet transform, feature extraction from the pretrained deep CNNs, and feature fusion. Figure 1 is the framework of our approach. The discrete wavelet transform is employed to decompose the original image into various components at different frequency intervals. Then, the original image, and the generated low-low (LL) subbands in different level are fed into pre-trained CNNs to obtain the deep features, separately. Next, all the obtained feature are integrated by the improved DCA method. Finally, The LIBSVM is employed for replacing the softmax layer of CNN and monitoring classification on well-known datasets.

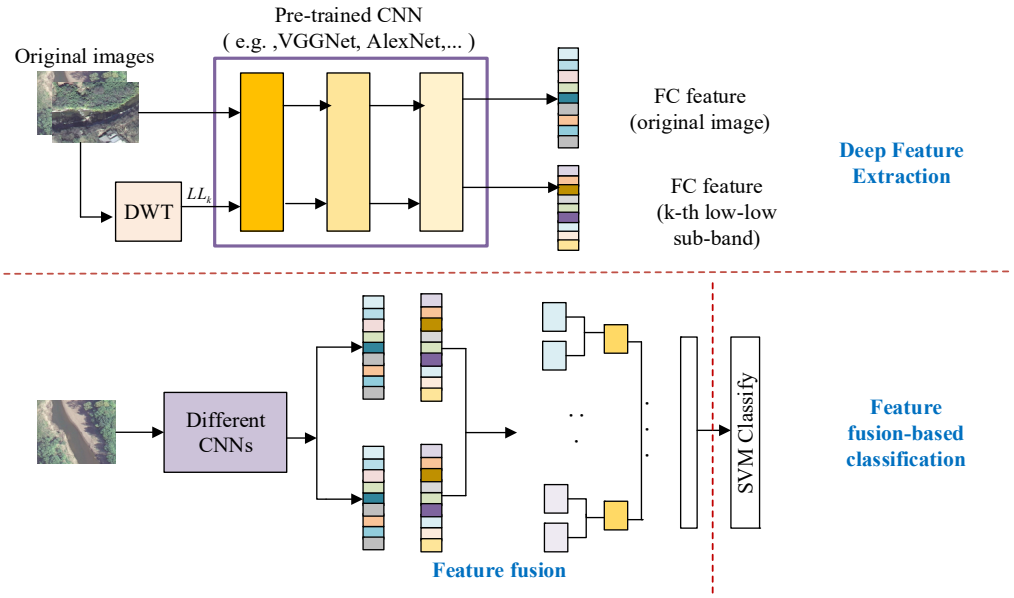
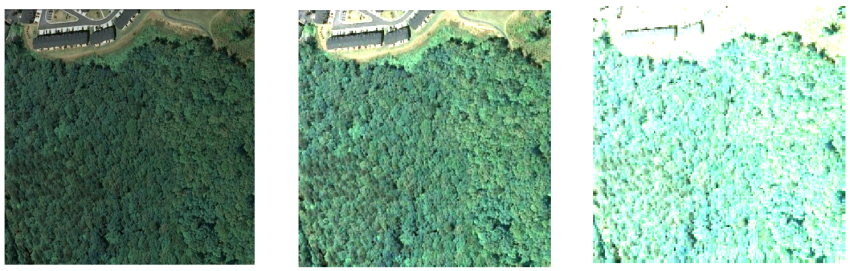


Figure 1. The framework of the proposed method.

2.1. Discrete wavelet transform

In image processing, the discrete wavelet transform (DWT) is proposed to decompose an image into various components at different frequency intervals. Figure 2 shows an example of low frequency components of the RS image. Wang et al. [41] validates that the low-frequency component is much more generalizable than the high-frequency component. The low-frequency components of the images obtained by discrete wavelet transform, which are used for deep feature extraction, can take full advantage of the image feature information. The low-frequency components of the images obtained by DWT, which are used for deep feature extraction, can take full advantage of the image feature information. Inspired by this, CNN features of low-frequency components at different levels can be used to construct feature pyramids.



(a) (b) (c)

Figure 2. An example of low-frequency components of different subbands. The low-frequency components are obtained by the Haar wavelet function. (a) the original image, (b) the first low-low component, (c) the second low-low component.

For the input image X, the basic wavelet function f_w is used for DWT to calculate LL_k coefficients as follows.

$$\left[LL_1, LH_1, HL_1, HH_1\right] = DWT(X, f_w) \tag{1}$$

$$\left[LL_k, LH_k, HL_k, HH_k\right] = DWT(LL_{k-1}, f_w), k = 2, 3, \ldots, \tag{2}$$

where LL_k, LH_k, HL_k and HH_k are the low-low, low–high, high-low and high-high filter coefficients of the $k-th$ level. Low-low subbands are used for the subsequent feature extraction.

2.2. Feature extraction

In the past decade, several typical CNN models have been developed, such as AlexNet [7], VGG-Net [9], GoogleNet [10], Resnet [11], etc. These models have different structures and different representational abilities. The CNNs pre-trained on ImageNet already can obtain powerful and rich features. Our approach focuses on the fusion of multiple subband deep features extracted from distinct off-the-shelf CNN models. In order to reduce computational complexity and improve recognition accuracy, AlexNet, VGG-Net, and ResNet 50 are introduced for feature extraction. There is a correlation between different features of convolutional networks. These features extracted from diverse CNN models are redundant and different, and their fusion can be applied to represent the RS images.

2.3. Improved DCA method

The basis for the combination of different features is their redundancies and differences. The DCA method [42] contributes to further disperse classes that are far away from each other in the mapping space, which provide an efficient and low-cost feature fusion strategy. However, the categories that are less distinct from each other are closer together. To address the superposition of categories in the mapping space, we reconstructed the between-class scatter matrix.

Assuming an image set *I*, *c* is the number of categories in *I*, *n* the number of trained features, we defined two feature matrices, *X* and *Y*, respectively.

The between-class scatter matrix in *X* is expressed as

$$S_{bx(p\times p)} = \sum_{i=1}^{c} n_i w_{disc} \left(\overline{x_i} - \overline{x}\right)\left(\overline{x_i} - \overline{x}\right)^T = W_{disc} \Phi_{bx} \Phi_{bx}^T . \tag{3}$$

$$\Phi_{bx(p\times c)} = \left[\sqrt{n_1}\left(\overline{x_1} - \overline{x}\right), \sqrt{n_2}\left(\overline{x_2} - \overline{x}\right), \ldots\ldots \sqrt{n_c}\left(\overline{x_c} - \overline{x}\right)\right], \tag{4}$$

$$W_{disc} = \frac{erf\left(dist(\overline{x_i}, \overline{x})\right)}{\sqrt{\arctan(dist(\overline{x_i}, \overline{x}))}} \tag{5}$$

where $dist(\overline{x_i}, \overline{x}) = (\overline{x_i} - \overline{x})^T (\overline{x_i} - \overline{x})$. $erf(.)$ is the error function. W_{disc} is the distance coefficient. $\overline{x_i}$ and $\overline{x}$ are the mean of the feature vector in the $i-th$ category and in the whole *X* set, respectively.

In order to separate the classes, *X* is projected into a new space. The projection X' is described as [42] by mapping matrix W_{bx}.

$$X'_{(r\times n)} = W^T_{bx(r\times p)} X_{(p\times n)} \tag{6}$$

in which $W_{bx} = \Phi_{bx} A\Lambda^{-1/2}$ unitizes S_{bx} and reduces the dimension of X from $p\times n$ to $r\times n$. r is the feature length of the transformed features [42]:

$$r \le \min\left(c-1, rank\left(X\right), rank\left(Y\right)\right). \tag{7}$$

The other feature set Y is processed in a similar way.

Supposing $S'_{xy} = X'Y'^T$ is the between-set covariance matrix of the transformed feature set. To maximize the pairwise correlation across X and Y, S'_{xy} needs to be diagonalized.

$$S'_{xy(r\times r)} = U\Sigma V^T \Rightarrow U^T S'_{xy} V = \Sigma. \tag{8}$$

Similar to the previous step, let $W_{cx} = U\Sigma^{-1/2}$, $W_{cy} = V\Sigma^{-1/2}$, then

$$\left(U\Sigma^{-1/2}\right)^T S'_{xy}\left(V\Sigma^{-1/2}\right) = I. \tag{9}$$

Next, the transformed feature set can be described as:

$$X^* = W^T_{cx} X' = W^T_{cx} W^T_{bx} X = W_x X \tag{10}$$

$$Y^* = W^T_{cy} Y' = W^T_{cy} W^T_{by} Y = W_y Y \tag{11}$$

where W_x, and W_y are the last transformation matrices for X and Y, respectively, thereby minimizing the correlation between-class.

The transformed features are fused to obtain the combination. There are two classic fusion approaches: parallel strategy and serial strategy. The parallel strategy is to add the feature vectors, and the serial strategy is to concatenate different features into one single feature. The final feature dimension is related to the number of classes. If the number of classes is small, the fusion features will not be rich enough, which will affect the subsequent classification performance. To enrich the information of fusion features, the proposed method is performed by concatenation after transforming.

2.4. Discrete wavelet-based multilevel feature fusion

Motivated by the idea of DWT in image processing, the discrete wavelet-based multilevel feature fusion (DWMLFF) method is proposed. This method can fuse the information extracted from different wavelet subbands. Figure 3 shows the details of the DWMLFF method. The method is explained as follows:

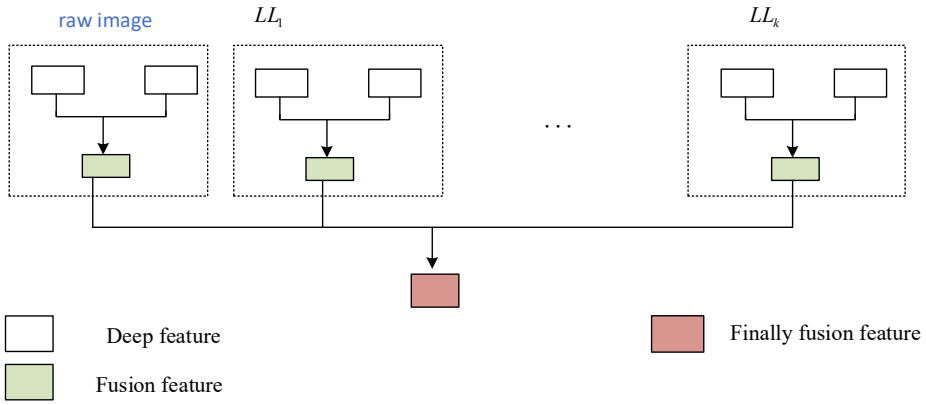


Figure 3. Details of the DWMLFF method

Step 1: Read each image from the remote sensing data set. For the input image X, the Haar wavelet function is used for DWT to calculate LL_k coefficient.

Step 2: Generate Semantic features from each image. Given a scene image, the original image I_{i}, and the generated $k-th$ low-low (LL) subbands are fed into pre-trained CNNs, separately. The features are generated from FC layers of pre-trained models. The output of $k-th$ on the FC layer extracted by the pre-trained model can be represented as f_i. The out of the $k-th$ approximate image on the FC layer extracted by the pre-trained model can be described as f_{LL_k}.

Step 3: The discrete wavelet-based multi-level feature fusion. For original image I_{i}, the deep features obtained from different CNNs, f_i^1 and f_i^2, are combined into a single feature by DCA. The same method is used to get the initial fusion features of the $k-th$ approximate image. Finally, the fusion features are concatenated to construct the final feature.

The maximum dimension of the initial fusion features is $2\times(c-1)$, and the full size of the final features is $2\times(c-1)\times(k+1)$. For small sample datasets, this value is much less than the dimension of the features directly extracted from the CNN model.

Finally, The LIBSVM library [43] is employed for replacing the softmax layer of CNN and monitoring classification on well-know datasets.

2.5. Data sets

To verify the feasibility of the DWMLFF method, four famous public datasets, the UC Merced dataset [1], the WHU-RS19 dataset [44], the AID dataset [4] and the NWPU-RESISC45 dataset [45], are employed in our experiments, respectively. The training ratios of these datasets are similar to [27].

The UC Merced dataset is extracted from a lot of optical images of the US Geological Survey National Map Urban Area Imagery. It includes 21 scene classes, and each category contains 100 RGB images with 256×256 pixels. The spatial resolution of these images is 1 foot per pixel. Figure 4 shows an example image for each category. It can be clearly seen that there are many similarities among 'forest', 'medium residential' and 'mobile home park'. Similarities can lead to severe difficulty in distinguishing them.

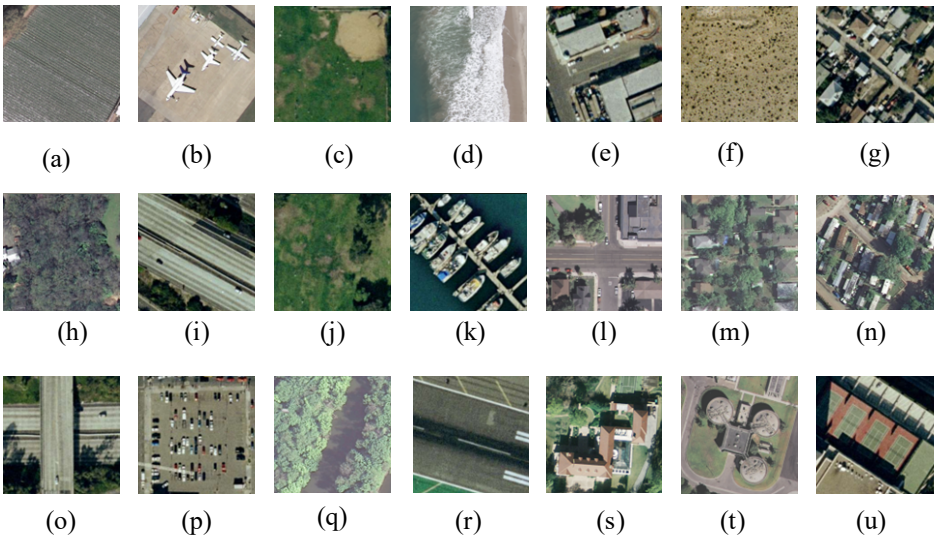


Figure 4. Some example images from UC Merced dataset. (a) agricultural, (b) airplane, (c) baseball diamond, (d) beach, (e) buildings, (f) chaparral, (g) dense residential, (h) forest, (i) freeway, (j) golf course, (k) harbor, (l) intersection, (m) medium residential, (n) mobile home park, (o) overpass, (p) parking lot, (q) river, (r) runway, (s) sparse residential, (t) storage tanks, (u) tennis court.

The WHU-RS19 dataset contains 19 challenging categories, which are exported from Google Earth. In this dataset, the size of the images is 600×600 pixels. The image samples in the same class are collected from different regions of satellite images with various resolutions. These samples contain different orientations, scales, and illumination. Some example images from this dataset are shown in Figure 5. The resolution of the images in this dataset is variable, which causes more challenges for RSSC in RS19 than that in the UC Merced dataset.

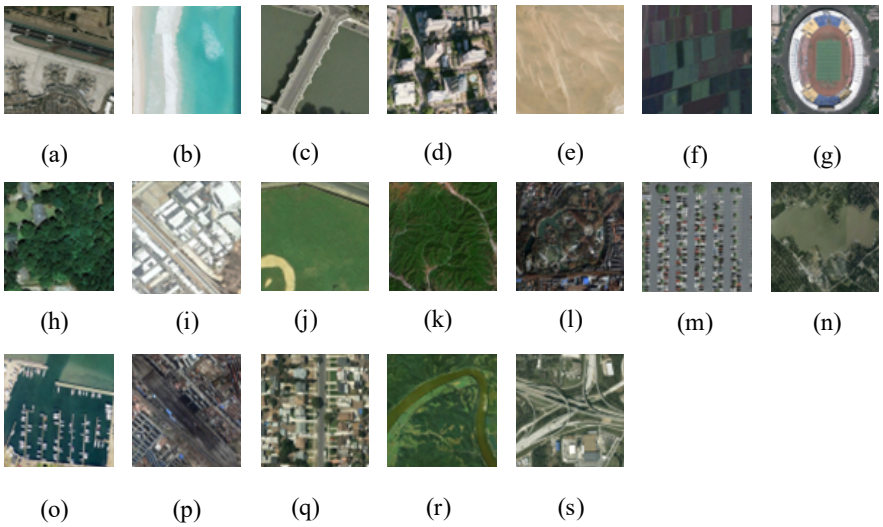


Figure 5. Samples of WHU-RS19 dataset. (a) airport, (b) beach, (c) bridge, (d) commercial, (e) desert, (f) farmland, (g) football field, (h) forest, (i) industrial, (j) meadow, (k) mountain, (l) park, (m) parking, (n) pond, (o) port, (p) railway station, (q) residential, (r) river, (s) viaduct.

The AID dataset, which are acquired from Google Earth, consists of 10,000 images within 30 aerial scene types. The pixel-resolution changes from about 0.5 to 8 m. The size of each image in AID30 is the same as that in RS19. The scene classes include: airport, bare land, baseball field, beach, bridge, center, church, commercial, dense residential, desert, farmland, forest, industrial, and so on. The images in AID dataset are extracted at different time and seasons under different conditions, which increases the intra-class diversities of the data.

The NWPU-RESISC45 dataset (NR45) was created by the researchers of Northwestern Polytechnical University. It contains 31,500 RGB images, covering 45 scene categories with 700 images in each class. The scene classes include: airplane, airport, baseball diamond, basketball court, beach, bridge, chaparral, church, circular farmland, cloud, commercial area, and so on. The spatial resolution varies from 0.2 to 30 m for most of the scene classes. The size of each image in NR45 is the same as that in the UC Merced dataset.

According to the above analysis, the images of these datasets show many low-level features of ordinary optical images. The features extracted by the CNNs pre-trained on ImageNet can be used for scene classification of these datasets.

3. Results and discussion

3.1. Experimental results of the UC Merced dataset

The recognition accuracy of different methods based on the UC Merced data set is shown in Table 1, in which the best results are highlighted in bold. The training ratio is from 2 to 10%, and the results show that the ratio of train samples affects the classification accuracy. The lower proportion of training samples leads to more complex scene recognition and a lower recognition ratio. Due to the limited number of classes in the dataset, the feature dimensions of the DCA method is small, which makes the DCA method cannot obtain sufficient discriminant information. However, the DWMLFF method by introducing different frequency band features from multiple sub-bands, is not only better than single feature without fused, but also better than the DCA method. The recognition accuracy of the DWMLFF method is more than 10% higher than that of the single CNN method.

Table 1. Comparison of different methods on the UC Merced dataset.

Models	Method	2%	4%	6%	8%	10%
AlexNet	Pretrained	60.617	71.181	78.227	82.702	84.698
VGG-19	Pretrained	54.995	66.486	77.902	80.342	84.392
ResNet50	Pretrained	54.982	66.445	69.048	72.913	74.544
AlexNet & VGG-19	DCA	64.609	70.357	76.813	81.191	83.951
	DWMLFF	66.102	71.233	78.292	82.613	84.523
ResNet 50 & VGG-19	DCA	65.892	74.453	80.168	82.081	85.343
	DWMLFF	66.629	77.434	80.678	83.022	85.439
AlexNet & ResNet 50	DCA	66.803	74.211	80.152	82.392	85.195
	DWMLFF	**67.185**	**77.996**	**81.521**	**83.536**	**86.067**

The per-class recognition accuracies using the single feature, DCA, and DWMLFF methods are shown in Figure 6. In Figure 6, the training rate is 4%, the single feature is generated from AlexNet,

the DCA method and DWMLFF method fuse the features extracted from AlexNet and ResNet50.

It can be seen that the classification accuracies of most classes have been enhanced by using DWMLFF. However, the performance of DCA on the categories of 'runway', 'intersection' and 'storage tanks' is better than that of the DWMLFF method. It can be explained that the multi-subband low-frequency components of these images cannot provide enough additional effective features but increase redundancy, leading to the degradation of classification performance.

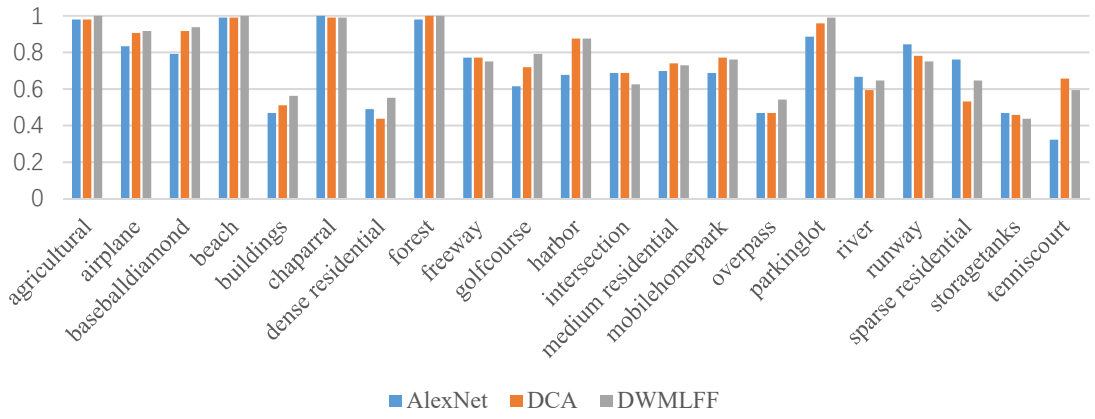


Figure 6. Per-class recognition accuracy of the UC Merced data set based on different methods.

3.2. Experimental results of the WHU-RS19 data set

The classification accuracy of different methods on the WHU-RS19 data set is described in Table 2, in which the training ratio is from 2 to 10%. The results of the experiment demonstrate that feature fusion between different networks can effectively improve classification performance. As shown in Table 2, the recognition performance of the DWMLFF method is obviously better than that of the DCA method and the single feature method. For example, compared to ResNet 50, 'AlexNet & ResNet 50' by DWMLFF improves the overall accuracy by more than 11% under different training ratios. Compared to DCA method and individual CNNs, the DWMLFF method has outstanding advantages in small sample RSSC. By comparing Tables 1 and 2, it is worth mentioning that our DWMLFF method achieves a greater signification gain on WHU-RS19 than on UC Merced. That can be explained by the fact that the WHU-RS19 dataset has fewer categories than the UC Merced dataset, and there are less images in each class of the WHU-RS19 dataset than in the first dataset.

Figure 7 shows the per-class recognition accuracy on the WHU-RS19 dataset of different methods. Obviously, for almost categories of the WHU-RS19 dataset, the recognition accuracy of the fused method is better than the single feature method. Furthermore, the DWMLFF method performs better than the DCA and single feature methods in all categories except the 'river' class. It is due to the fact that the number of classes in this dataset is such small that the dimensions of the feature which is fused by the DCA method is too small to provide sufficient discriminant information. The DWMLFF method of introducing different frequency band features from multiple subbands is superior to the DCA method.

Table 2. Comparison with other methods on the WHU-RS19 data set (Bold indicates the best results).

Models	Method	Accuracy (%)				
		2%	4%	6%	8%	10%
AlexNet	Pretrained	58.904	73.485	76.772	81.097	81.481
VGG-19	Pretrained	59.079	71.067	77.764	81.576	83.158
ResNet50	Pretrained	64.483	74.767	78.713	81.765	83.209
AlexNet & VGG-19	DCA	65.945	77.594	82.131	84.563	88.419
	DWMLFF	72.110	85.866	90.126	90.079	93.536
ResNet50 & VGG-19	DCA	67.809	80.331	84.894	87.648	90.231
	DWMLFF	72.738	85.274	90.549	92.467	94.239
AlexNet & ResNet50	DCA	68.357	81.489	84.958	88.741	89.835
	DWMLFF	**75.587**	**86.580**	**91.368**	**92.514**	**94.519**

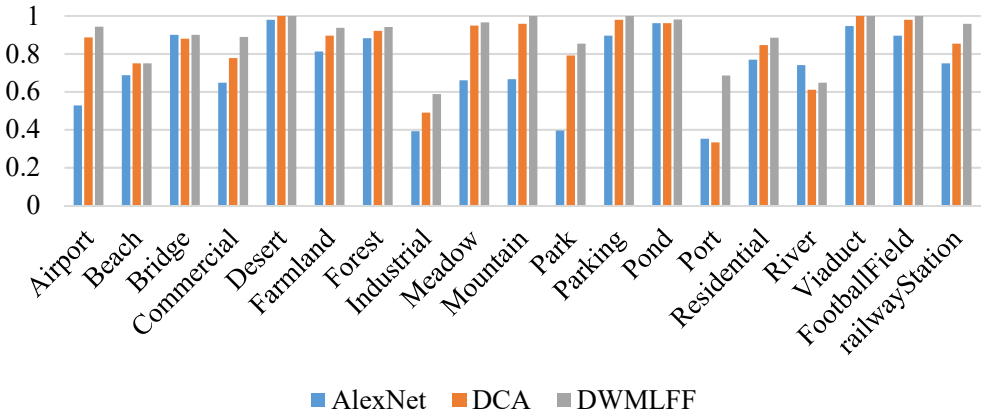


Figure 7. Per-class recognition performance on the WHU-RS19 data set based on single feature and fusion features.

In order to clearly observe the dispersion of different methods on the WHU-RS19, the image features are visualized. In Figure 8, features generated from different methods of the WHU-RS19 data set are visualized for comparison. Various colors on the graphs represent different categories in the data set, and the points represent the feature of images in the data set.

As shown in Figure 8, the single feature extracted from AlexNet or ResNet50 forms some overlaps that are in a confused order. On the contrary, the fusion feature generated by DCA and DWMLFF forms clusters that are clearly separated. Compared to the DCA method, the DWMLFF method using the multi-subband information of images can be utilized to achieve a better representation, obtain a higher convergence among the identical classes, and gain a greater distinguish among different classes.

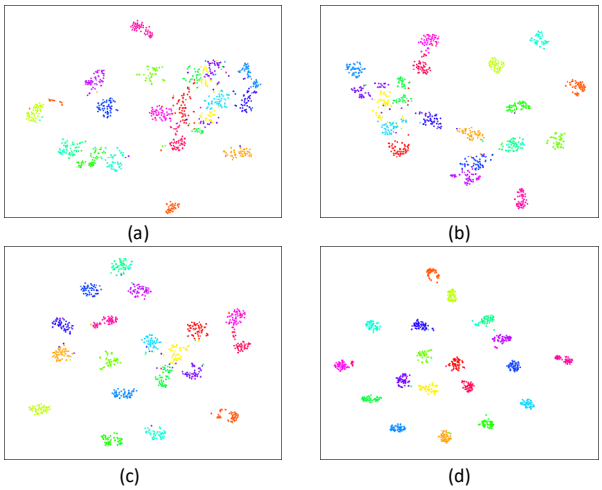


Figure 8. Visualization of features generated from different methods on the WHU-RS19 data set. (a) single feature of AlexNet, (b) single feature of ResNet50, (c) fusion feature by DCA, (d) fusion feature by DWMLFF.

3.3. Experimental results of the AID dataset and the NR45 dataset

The experimental results of different methods on AID and NR45 are shown in Tables 3 and 4, respectively. The training ratio is from 2 to 10%.

Table 3. Comparison with other methods on the AID data set (Bold indicates the best results).

Models	Method	Accuracy (%)				
		2%	4%	6%	8%	10%
AlexNet	Pretrained	68.90	76.48	79.97	81.87	82.95
VGG-19	Pretrained	68.90	75.49	78.25	79.92	81.96
ResNet50	Pretrained	64.83	75.07	78.71	78.77	82.21
AlexNet & VGG-19	DCA	68.63	76.61	80.19	82.83	84.58
	DWMLFF	71.39	78.63	82.43	84.72	85.91
ResNet50 & VGG-19	DCA	69.41	78.314	81.92	83.81	83.24
	DWMLFF	71.42	79.265	83.13	85.12	86.03
AlexNet & ResNet50	DCA	72.203	78.97	82.89	83.96	84.22
	DWMLFF	**73.895**	**80.253**	**83.58**	**85.99**	**86.17**

Similar to the above experiments, the classification performance of the DWMLFF method on the AID dataset and the NR45 dataset is significantly superior to that of the DCA method and the single feature method. By comparing the classification results in Tables 1–4, it can be found that the DWMLFF method achieves a greater signification gain on WHU-RS19 than on other datasets. It indicates that in the case of small samples and fewer categories, our method has more obvious advantages.

Table 4. Comparison with other methods on the NR45 data set (Bold indicates the best results).

Models	Method	Accuracy (%)				
		2%	4%	6%	8%	10%
AlexNet	Pretrained	63.33	70.48	72.97	76.87	78.01
VGG-19	Pretrained	64.91	71.49	73.67	77.29	77.68
ResNet50	Pretrained	65.813	72.07	73.71	76.77	79.363
AlexNet & VGG-19	DCA	66.31	72.18	74.22	76.56	79.11
	DWMLFF	68.20	72.63	74.57	77.93	80.09
ResNet50 & VGG-19	DCA	69.521	73.51	74.24	77.63	81.17
	DWMLFF	**71.735**	74.97	75.49	78.28	83.35
AlexNet & ResNet50	DCA	67.58	73.06	75.33	78.21	81.76
	DWMLFF	70.74	**75.46**	**76.83**	**79.76**	**83.48**

3.4. Comparison with advanced methods

To effectively analyze the performance of our method, we conducted a comparison between the DWMLFF method and the state-of-the-art methods. The results of the accuracy comparison on UC Merced and WHU-RS19 are shown in Tables 5 and 6, respectively, in which the best results are highlighted in bold. In Tables 4 and 5, the methods with '*' indicate that the experiments are performed around the 5-way *K*-shot, which indicates that *K* labeled samples are used to recognize samples from 5 scene classes. For UC Merced, training ratios of 1 and 5% are added in experiments to obtain a more comprehensive comparison.

Table 5. Comparison with the advanced methods on the UC Merced data set. The methods with '*' indicate that the experiments are performed around the 5-way K-shot.

Method	Training ratios of labelled samples for each class						
	1%	2%	4%	5%	6%	8%	10%
MSCP [20]	/	13.15	59.23	/	75.76	82.28	84.63
ARCNet [22]	/	47.667	60.119	/	73.45	75.77	84.01
ICEL [21]	/	45.86	61.31	/	70.06	78.14	82.12
GLF+SRC [27]	/	61.618	73.611	/	77.710	**83.592**	**87.672**
SIFT & ResNet [29]	/	66.07	73.4	/	77.68	81.01	84.85
RS_MetaNet [46]	55.29	/	/	71.42	/	/	75.16
*DLA [28]	53.76	/	/	63.01	/	/	/
*SPNet [47]	57.64	/	/	73.52	/	/	/
*DN4 [23]	57.25	/	/	**79.74**	/	/	/
*TAE-Net [25]	**60.21**	/	/	77.44	/	/	/
DWMLFF (Ours)	56.336	**67.185**	**77.996**	79.680	**81.521**	83.536	86.067

As shown in Tables 5 and 6, the accuracy performance of the DWMLFF method is superior to that of most methods, the less the number of training samples, the more advantageous DWMLFF method is. Although the experiment results of the 'GLF + SRC' method and '5-way k-shot' methods are better than that of our DWMLFF method in some cases, the DWMLFF method is simpler to

implement, and generates smaller feature dimensions.

Table 6. Classification performance of different methods on the WHU-RS19 dataset. The methods with '*' indicate that the experiments are performed around the 5-way K-shot.

Method	number of labelled samples of each class				
	1	2	3	4	5
MSCP [20]	/	/	48.952	79.75	85.68
ARCNet [22]	41.885	63.159	68.192	74.022	81.871
ICEL [21]	24.801	63.158	66.523	75.315	75.866
GLF+SRC [27]	69.387	80.482	85.682	86.842	88.655
*DLA [28]	68.27	/	/	/	79.89
*SPNet [47]	**81.06**	/	/	/	88.04
*TAE-Net [25]	73.67	/	/	/	88.95
DWMLFF (ours)	75.587	**86.580**	**91.368**	**92.514**	**94.519**

3.5. Impact of multi-subband number

The number of subbands k can affect the final classification accuracy. In order to analyze the effect of k on classification performance, and obtain better fusion results, experiments were conducted. The experimental configuration is the same as before, except for k.

Figure 9 shows that the number of subbands of DWT affects the classification accuracy. Figure 9(a),(b) shows the effect of different k values on the overall accuracy on the two datasets, respectively. For different data sets, the impact of k value on the classification accuracy is different. Due to the larger size and higher resolution of WHU-RS19, the low-frequency components from multiple subbands contain richer information, resulting in a more significant gain on WHU-RS19 than on UC Merced.

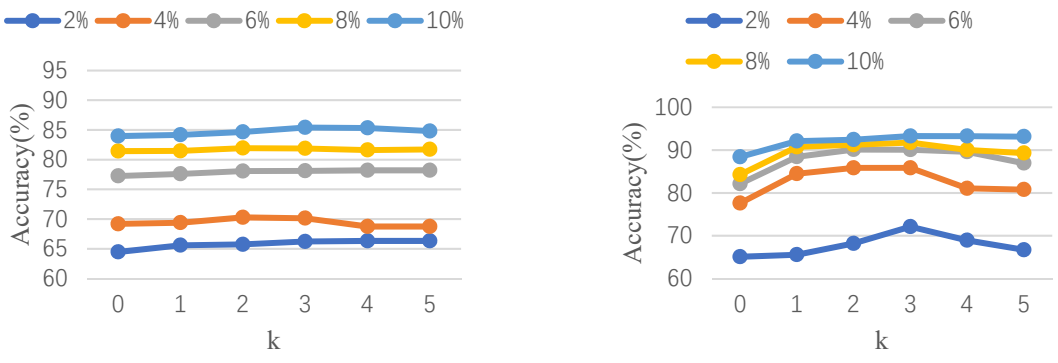


Figure 9. Recognition rate of different subbands on two datasets. (a) UC Merced; (b) WHU-RS19.

When the number of subbands is small, the dimension of the feature is tiny, which cannot provide satisfactory semantic information, resulting in low classification accuracy. As the value of k increases, the feature dimension also increases, and the classification accuracy is improved. However, when the k is too large, some redundant information will be introduced, resulting in no further significant improvement in accuracy. It is worth mentioning that if k is greater than 3, the recognition accuracy will decrease to different degrees under the training ratio of 2–4%. When the training rate is set to 2%, there are only two images in each category in the UC Merced dataset as training samples. There are even fewer images in the WHU-RS19 dataset as training samples, with only one shot in each category. In this case, increasing different frequency components of the training samples to obtain more complex feature information will lead to overfitting, thus reducing the recognition accuracy.

The dimension of fusion features obtained by our method is closely related to the number of wavelet subbands. According to Eq (7), the maximum dimension of the final feature fused by our strategy is $2\times(k+1)\times(c-1)=36(k+1)$. The dimension of the features of various ways on the WHU-RS19 data set under the 4% training ratio is depicted in Table 7.

Table 7. Feature dimension of various methods on the WHU-RS19.

Method	k	Accuracy (%)	Dimension of feature
AlexNet		72.4	4096
VGG-19		70.6	4096
DCA		76.2	36
DWMLFF (AlexNet & VGG-19)	1	81.7	72
	2	84.2	108
	3	86.1	144
	4	82.6	180
	5	81.4	216

Compared with the single feature extracted from pretrained CNNs, the recognition accuracy of the DWMLFF method dramatically increases, while the dimension of feature decreases significantly. Compared with the DCA method, the feature dimension of the DWMLFF method increases slightly, but the classification accuracy is greatly improved. As can be seen from Figure 9 and Table 7, we can find that the appropriate value of k is helpful in improving the classification performance, which is the advantage of making full use of different frequency components from wavelet subbands.

4. Conclusions

In this paper, a multi-subband feature fusion method, namely DWMLFF, is proposed for few-shot RSSC. To surmount the problem of insufficient features in few-shot RSSC, the DWT is employed as a decomposer to obtain the multi-subband information of limited samples. The original image and the LL subbands of different level are fed into pretrained CNN models, which improves the feature generation capability of pretrained CNN models. In order to maximize the difference between categories, the improved DCA strategy is proposed. In the improved DCA strategy, the distance coefficient is introduced to reconstructed the between-class scatter matrix, which helps adjusting the distance between classes and avoiding cross overlap in the mapping space. Finally, the features extracted from different CNNs and different frequency components are fused by the improved DCA.

The proposed method gives full play to the advantages of different wavelet subbands, and utilizes an improved DCA strategy to obtain low-dimensional and high-discriminative features for RSSC. The experimental results on four well-known datasets indicate that the proposed method achieves outstanding performance in RSSC with few training data, especially with one or two training samples per category. In the future, we would focus on the relationship between different modalities to achieve automatic and accurate classification, which we would apply for post-disaster identification with limited training samples.

Acknowledgments

This work was supported by Jiangsu Engineering Research Center of Lake Environment Remote Sensing Technologies, Huaiyin Institute of Technology.

Conflict of interest

The authors declare there is no conflict of interest.

References

1. Y. Yang, S. Newsam, Bag-of-visual-words and spatial extensions for land-use classification, in *Proceedings of Sigspatial International Conference on Advances in Geographic Information Sys*tems, (2010), 270–279. https://doi.org/10.1145/1869790.1869829
2. Z. Yang, X. Mu, F. Zhao, Scene classification of remote sensing image based on deep network and multi-scale features fusion, *Optik*, **171** (2018), 287–293. https://doi.org/10.1016/j.ijleo.2018.06.024
3. T. Hieu, W. Adrian, Remote sensing of coastal hydro-environment with portable unmanned aerial vehicles (pUAVs) a state-of-the-art review, *J. Hydro-environ. Res.*, **37** (2021), 32–45. https://doi.org/10.1016/j.jher.2021.04.003
4. G. S. Xia, J. Hu, F. Hu, B. Shi, X. Bai, Y. Zhong, et al., AID: A benchmark data set for performance evaluation of aerial scene classification, *IEEE Trans. Geosci. Remote Sens.*, **55** (2017), 3965–3981, https://doi.org/10.1109/TGRS.2017.2685945
5. R. Cinbis, J. Verbeek, C. Schmid, Approximate fisher kernels of non-iid image models for image categorization, *IEEE Trans. Pattern Anal. Mach. Intell.*, **38** (2016), 1084–1098. https://doi.org/10.1109/TPAMI.2015.2484342
6. F. Hu, G. S. Xia, Z. Wang, X. Huang, L. Zhang; H. Sun, Unsupervised feature learning via spectral clustering of multidimensional patches for remotely sensed scene classification, *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, **8** (2015), 2015–2030. https://doi.org/10.1109/JSTARS.2015.2444405
7. A. Krizhevsky, I. Sutskever, G. E. Hinton, Imagenet classification with deep convolutional neural networks, *Commun. ACM*, **60** (2017), 84–90. https://doi.org/10.1145/3065386
8. D. Hong, L. Gao, J. Yao, B. Zhang, A. Plaza, J. Chanussot, Graph convolutional networks for hyperspectral image classification, *IEEE Trans. Geosci. Remote Sens.*, **59** (2021). 5966–5978. https:// doi.org/ 10.1109/TGRS.2020.3015157

9. K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, preprint, arXiv:1409.1556. https://doi.org/10.48550/arXiv.1409.1556
10. C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, et al., Going deeper with convolutions, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, (2015), 1–9.
11. K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, (2016), 770–778.
12. Y. Guo, J. Ji, X. Lu, H. Huo, T. Fang, D. Li, Global-local attention network for aerial scene classification, *IEEE Access*, **7** (2019), 67200–67212. https://doi.org/10.1109/ACCES2019.2918732
13. J. Xie, N. He, L. Fang, A. Plaza, Scale-free convolutional neural network for remote sensing scene classification, *IEEE Trans. Geosci. Remote Sens.*, **57** (2019), 6916–6928. https://doi.org/10.1109/TGRS.2019.2909695
14. H. Xie, Y. Chen, P. Ghamisi, Remote sensing image scene classification via label augmentation and intra-class constraint, *Remote Sens.*, **13** (2021), 2566–2586. https://doi.org/10.3390/rs13132566
15. N. He, L. Fang, S. Li, J. Plaza, A. Plaza, Skip-connected covariance network for remote sensing scene classification, *IEEE Trans. Neural Networks Learn. Syst.*, **31** (2020), 1461–1474. https://doi.org/10.1109/TNNLS.2019.2920374.
16. X. Wu, D. Hong, J. Chanussot, Convolutional neural networks for multimodal remote sensing data classification, *IEEE Trans. Geosci. Remote Sens.*, **60** (2022). https://doi.org/10.1109/TGRS.2021.3124913.
17. Y. Liu, C. Y. Suen, Y. Liu, L. Ding, Scene classification using hierarchical Wasserstein CNN, *IEEE Trans. Geosci. Remote Sens.*, **57** (2019), 2494–2509. https://doi.org/10.1109/TGRS.2018.2873966
18. J. Fang, Y. Yuan, X. Lu, Y. Feng, Robust space-frequency joint representation for remote sensing image scene classification, *IEEE Trans. Geosci. Remote Sens.*, **57** (2019), 7492–7502. https://doi.org/10.1109/TGRS.2019.2913816
19. H. Sun, S. Li, X. Zheng, X. Lu, Remote sensing scene classification by gated bidirectional network, *IEEE Trans. Geosci. Remote Sens.*, **58** (2019), 82–96. https://doi.org/10.1109/TGRS.2019.2931801
20. N. He, L. Fang, S. Li, A. Plaza, J. Plaza, Remote sensing scene classification using multilayer stacked covariance pooling, *IEEE Trans. Geosci. Remote Sens.*, **56** (2018), 6899–6910. https://doi.org/10.1109/TGRS.2018.2845668
21. A. Bahri, S. G. Majelan, S. Mohammadi, M. Noori, K. Mohammadi, Remote sensing image classification via improved crossentropy loss and transfer learning strategy based on deep convolutional neural networks, *IEEE Geosci. Remote Sens. Lett.*, **17** (2020), 1087–1091. https://doi.org/10.1109/LGRS.2019.2937872
22. Q. Wang, S. Liu, J. Chanussot, X. Li, Scene classification with recurrent attention of VHR remote sensing images, *IEEE Trans. Geosci. Remote Sens.*, **57** (2019), 1155–1167. https://doi.org/10/1109/TGRS.2018.2864987
23. Y. Chen, Y. Li, H. Mao, X. Chai, L. Jiao, A novel deep nearest neighbor neural network for few-shot remote sensing image scene classification, *Remote Sens.*, **15** (2023), 666–684. https://doi.org/10.3390/rs15030666

24. N. Jiang, H. Shi, J. Geng, Multi-scale graph-based feature fusion for few-shot remote sensing image scene classification, *Remote Sens.*, **14** (2022), 5550–5568. https://doi.org/10.3390/rs14215550
25. W. Huang, Z. Yuan, A. Yang, C. Tang, X. Luo, TAE-Net: Task-adaptive embedding network for few-shot remote sensing scene classification, *Remote Sens.*, **14** (2022), 111–119. https://doi.org/10.3390/rs14010111
26. X. Wu, D. Hong, J. Chanussot, UIU-Net: U-Net in U-Net for infrared small object detection, *IEEE Trans. Image Process.*, **32** (2023), 364–376. https://doi.org/10.1109/TIP.2022.3228497
27. S. Mei, K. Yan, M. Ma, X. Chen, Q. Du, Remote sensing scene classification using sparse representation-based framework with deep feature fusion, *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.*, **14** (2021), 5867–5878. https://doi.org/10.1109/JSTARS.2021.3084441
28. Q. Zeng, J. Geng, K. Huang, W. Jiang, J. Guo, Prototype calibration with feature generation for few-shot remote sensing image scene classification, *Remote Sens.*, **13** (2021), 2728–2747. https://doi.org/10.3390/rs13142728
29. S. Yang, H. Wang, H. Gao, L. Zhang, Feature fusion method based on discriminant correlation analysis for land use classification with few-shot, in *proceedings of the International Conference on Computer Engineering and Artificial Intellige*nce, (2022), 671–675. https://doi.org/10.1109/ICCEAI55464.2022.00143
30. Q. Chen, Z. Chen, W. Luo, Feature transformation for cross-domain few-shot remote sensing scene classification, preprint, arXiv:2203.02270. https://doi.org/10.48550/arXiv.2203.02270
31. H. Y. Tseng, H. Y. Lee, J. B. Huang, M. Yang, Cross-domain few-shot classification via learned feature-wise transformation, preprint, arXiv:2001.08735. https://doi.org/10.48550/arXiv.2001.08735
32. W. Chen, Y. Liu, Z. Kira, Y. Wang, J. Huang, A closer look at few-shot classification, preprint, arXiv:1904.04232. https://doi.org/10.48550/arXiv.1904.04232
33. A. Chowdhury, M. Jiang, C. Jermaine, Few-shot image classification: Just use a library of pre-trained feature extractors and a simple classifier, in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, (2021), 9425–9434.
34. L. Li, J. Han, X. Yao, G. Cheng, L. Guo, DLA-MatchNet for few-shot remote sensing image scene classification, *IEEE Trans. Geosci. Remote Sens.*, **59** (2021), 7844–7853. https://doi.org/10.1109/TGRS.2020.3033336
35. W. Li, L. Wang, J. Xu, J. Huo, Y. Gao, J. Luo, Revisiting local descriptor based image-to-class measure for few-shot learning, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, (2019), 7253–7260.
36. R. Cao, L. Fang, T. Lu, N. He, Self-attention-based deep feature fusion for remote sensing scene classification, *IEEE Geosci. Remote Sens. Lett*., **18** (2020), 43–47. https://doi.org/10.1109/LGRS.2020.2968550
37. J. Li, K. Zheng, J. Yao, L.Gao, D. Hong, Deep unsupervised blind hyperspectral and multispectral data fusion, *IEEE Geosci. Remote Sens. Lett.*, **19** (2022). https://doi.org/10.1109/LGRS.2022.3151779
38. J. Li, D. Hong, L. Gao, J. Yao, K. Zheng, B. Zhang, et al., Deep learning in multimodal remote sensing data fusion: A comprehensive review, *Int. J. Appl. Earth Obs. Geoinf.*, **112** (2022), 102926. https://doi.org/10.1016/j.jag.2022.102926

39. D. Hong, L. Gao, N. Yokoya, J. Yao, J. Chanussot, Q. Du, et al., More diverse means better: multimodal deep learning meets remote-sensing imagery classification, *IEEE Trans. Geosci. Remote Sens.*, **59** (2021). 4340–4354. https://doi.org/ 10.1109/TGRS.2020.3016820
40. S. Chaib, H. Liu, Y. Gu, H. Yao, Deep feature fusion for VHR remote sensing scene classification, *IEEE Trans. Geosci. Remote Sens.*, **55** (2017), 4775–4784. https://doi.org/10.1109/TGRS.2017.2700322
41. H. Wang, X. Wu, Z. Huang, E. P. Xing, High-frequency component helps explain the generalization of convolutional neural networks, in *IEEE Conference on Computer Vision and Pattern Recognition*, (2020), 8681–8691.
42. M. Haghighat, M. Abdel-Mottaleb, W. Alhalabi, Discriminant correlation analysis: real-time feature level fusion for multimodal biometric recognition, *IEEE Trans. Inf. Forensics Secur.*, **11** (2016), 1984–1996. https://doi.org/10.1109/TIFS.2016.2569061
43. C. Chang, C. Lin, LIBSVM: A library for support vector machines, *ACM Trans. Intell. Syst. Technol.*, **2** (2011), 1–27. https://doi.org/10.1145/1961189.1961199
44. G. Xia, W. Yang, J. Delon, Y. Gousseau, H. Sun, H. Maître, Structural high-resolution satellite image indexing, in *ISPRS TC VII Symposium-100 Years ISPRS*, (2010), 298–303.
45. G. Cheng, J. Han, X. Lu, Remote sensing image scene classification: benchmark and state of the art, in *Proceedings of the IEEE*, **105** (2017), 1865–1883. https://doi.org/10.1109/JPROC.2017.2675998
46. H. Li, Z. Cui, Z. Zhu, L. Chen, J. Zhu, H. Huang, et al., RS-MetaNet: Deep metametric learning for few-shot remote sensing scene classification, *IEEE Trans. Geosci. Remote Sens.*, **59** (2020), 6983–6994. https://doi.org/10.1109/TGRS.2020.3027387
47. G. Cheng, L. Cai, C. Lang, X, Yao, J, Chen, L. Guo, et al., SPNet: Siamese-prototype network for few-shot remote sensing image scene classification, *IEEE Trans. Geosci. Remote Sens.*, **60** (2022), 5608011. https://doi.org/10.1109/TGRS.2021.3099033

AIMS Press