Remote Sensing Scene Classification Using Sparse Representation-Based Framework With Deep Feature Fusion

Shaohui Mei, *Senior Member, IEEE*, Keli Yan, Mingyang Ma, *Student Member, IEEE*, Xiaoning Chen, Shun Zhang, *Member, IEEE*, and Qian Du, *Fellow, IEEE*

***Abstract*—Scene classification of high-resolution remote sensing (RS) images has attracted increasing attentions due to its vital role in a wide range of applications. Convolutional neural networks (CNNs) have recently been applied on many computer vision tasks and have significantly boosted the performance including imagery scene classification, object detection, and so on. However, the classification performance heavily relies on the features that can accurately represent the scene of images, thus, how to fully explore the feature learning ability of CNNs is of crucial importance for scene classification. Another problem in CNNs is that it requires a large number of labeled samples, which is impractical in RS image processing. To address these problems, a novel sparse representation-based framework for small-sample-size RS scene classification with deep feature fusion is proposed. Specially, multilevel features are first extracted from different layers of CNNs to fully exploit the feature learning ability of CNNs. Note that the existing well-trained CNNs, e.g., AlexNet, VGGNet, and ResNet50, are used for feature extraction, in which no labeled samples is required. Then, sparse representation-based classification is designed to fuse the multilevel features, which is especially effective when only a small number of training samples are available. Experimental results over two benchmark datasets, e.g., UC-Merced and WHU-RS19, demonstrated that the proposed method can effectively fuse different levels of features learned in CNNs, and clearly outperform several state-of-the-art methods especially with limited training samples.**

***Index Terms*—Deep feature learning, remote sensing (RS), scene classification, small training size, sparse representation.**

I. Introduction

SCENE classification of remote sensing (RS) images has received increasing attentions. In recent years, with the rapid development of satellite RS technology and a series of earth observation programs, RS images have greatly promoted the development of techniques to scene classification, object detection, and so on. The aim of scene classification is to automatically assign semantic labels to given images based on *a priori* knowledge. Scene classification has been used in many practical applications, such as land-use/land-cover investigation, environmental monitoring, traffic supervision, and urban planning [1]–[4]. Although great efforts have been made, scene classification still is challenging in the field of RS image processing areas, because it is requisite to interpret RS images with more intelligent approaches [5]–[8].

In the past decades, many researchers assumed that RS images from the same category owned similar statistically holistic attributes, and lots of attentions had been paid to constructing various effective features. Color and texture histograms are representative low-level features that were early used for such a purpose [9], [10]. Afterward, scale invariant feature transform (SIFT) and histogram of oriented gradients (HOG), which can extract local features of scene images [11], [12], have improved the performance of scene classification [13]–[15]. However, these low-level features may not adequately represent semantic information of complex RS images. To overcome this limitation, many mid-level features are utilized for scene classification, in which the bag-of-words (BOW) model [16]–[18] is one of the most effective methods. For example, a scene classifier with local-global BOW features was proposed in [19], which can combine local and global features at the histogram level. In addition, other models based on mid-level features such as the latent Dirichlet allocation (LDA) and spatial class LDA model [20]–[23] were proposed for scene classification. However, the aforementioned features generally require *a priori* knowledge and domain expert experience, and lack robustness and flexibility.

Recently, deep learning (DL) has demonstrated its advantages in the field of computer vision. In particular, convolutional neural network (CNN) based methods have greatly improved the performance of image classification and object detection, such as classical AlexNet [24], VGGNet [25], Inception Net [26], and ResNet [27]. These CNN-based frameworks can automatically learn to extract high-level discriminative features, which have been widely used. Meanwhile, CNN-based methods are also used in RS, and achieve promising results. In 2016, the deep CNN was first used for scene classification in RS and greatly enhanced the performance [28]. Zhou *et al.* [29] investigated the extraction of deep feature representations based on pretrained

Manuscript received April 21, 2021; revised May 9, 2021; accepted May 21, 2021. Date of publication May 27, 2021; date of current version June 16, 2021. This work was supported in part by the National Natural Science Foundation of China under Grant 61671383 and the Youth Program of Shaanxi Natural Science Foundation under Grant 2020JQ-205. *(Corresponding author: Shaohui Mei.)*

Shaohui Mei, Keli Yan, Mingyang Ma, Xiaoning Chen, and Shun Zhang are with the School of Electronics and Information, Northwestern Polytechnical University, Xi'an 710129, China (e-mail: meish@nwpu.edu.cn; yankeli@mail.nwpu.edu.cn; mamingyang@mail.nwpu.edu.cn; chenxn@dgpt.edu.cn; szhang@nwpu.edu.cn).

Qian Du is with the Department of Electrical and Computer Engineering, Mississippi State University, Starkville, MS 39762 USA (e-mail: du@ece.msstate.edu).

Digital Object Identifier 10.1109/JSTARS.2021.3084441

CNN architectures for scene classification tasks. Some work attempted feature fusion for scene classification. For example, a fusion strategy for integrating multilayer features of a pretrained CNN for scene classification and achieved the competitive performance compared with fully trained CNN models, fine-tuning CNN models, and other related models [30]. The multilayer stacked covariance pooling (MSCP) was proposed to combine multilayer feature maps obtained by pretrained CNN models [31]. In addition, metric learning was also introduced to learn much more discriminative features to boost the performance of scene classification [32]. Recently, extensive CNN-based methods have been proposed to address the limitations in RS. For example, a scale-free CNN was introduced to address the problem that fine-tuning process often needs fixed-size input images [33]. The marginal center loss with an adaptive margin model was presented to overcome the limitation of images with large intraclass variations [34]. Recently, a deep few-shot learning method for the scene classification has been developed [35]. Generally, DL-based classification algorithms, such as CNN-based ones, require a large amount of labeled data for the system to learn, which amplifies the computational complexity and also the risk of underfitting. SRs own powerful ability to represent higher dimensional data using few measured values, which is especially effective for classification task with small number of training samples. Thus, in order to conduct scene classification under small-training-sample situation, we adopt SRC as the final classifier.

The aforementioned CNN-based methods implement scene classification through designing more effective features. Therefore, extracting those features that adequately represent the scene in images plays a crucial role in RS scene classification. However, existing methods focus on extracting high-level top-layer features, but ignoring the intermediate-layer features for convolutional layers. Actually, different types of features, including top layer features and intermediate convolutional features, own different strengths and limitations in a specific task. The top layer features can well represent the overall content of RS images, while intermediate features for convolutional layers may describe sufficient partial details and object information. Besides the performance of SRC is limited by the features and algorithms. Thus, we propose the multilayer feature extraction and fusion strategy. In this way, the similarity of a test sample to the training samples will be measured in the space formed by multilevel features extracted from CNNs, and such compromise between top-layer features and intermediate convolutional features will be benifical for classification.

In brief, we analyze the advantages and disadvantages of SRC and CNNs under the condition of small samples, and combine these two methods to solve the problem of scene classification of small sample RS images. To make fully use of the advantages of multilevel features learned from CNNs, a novel sparse representation framework is constructed to fuse and balance the contribution of these two types of features in this article. Different from the previous methods based on CNNs, which need to be trained with large-scale scene images, our proposed method extracts different levels of features from well-trained CNNs, avoiding the limitation of training CNN with large-scale RS imagery samples. In addition, this approach collaborates multilevel features by sparse representation and achieves much more competitive performance.

The major contributions of this work are as follows.

1) We propose the strategy to fuse multilevel features including those from intermediate convolutional layers and the top layer for scene classification, which is greatly different from the existing work using single feature from the top layer.
2) We present a novel sparse representation classification (SRC) framework, which builds the fusion classifier corresponding to multilevel features, and fuses their contributions for scene classification of RS images.
3) The proposed method addresses the few-shot classification problem of RS images, since multilevel features are extracted from the well-trained CNNs. As a result, competitive results are obtained through the SRC framework based on multilayer framework.

The remainder of this article is organized as follows. In Section II, the recent CNN-based scene classification methods and the progress of feature extraction are introduced. In Section III, the details of our proposed sparse representation framework and feature fusion strategy are described. In Section IV, experiments are conducted to validate the proposed method. Finally, Section V concludes this article.

II. Related Work

A. Features for Scene Classification

Features used for representing scene images for classification can be divided into the following two categories: handcrafted features and DL features.

1) Handcrafted Features: Most early methods in scene classification of RS images are based on handcrafted features. For example, Zohrevand *et al.* [11] applied the local SIFT features to extract key points and the corresponding descriptors of scene images. Sun *et al.* [36] presented a popular method called boosted HOG features to detect pedestrians and vehicles in static images. Gan *et al.* [37] proposed a measure of continuous interval rotating detection sliding window of HOG feature in RS images for ship detection. As the development of scene classification technology, researchers have proposed methods using the combination of multiple different features. For example, Chu and Zhao [38] proposed a feature fusion scheme for scene classification by integrating the global GIST and local SIFT with weights, and improved the classification performance. Local region characteristics and overall structure of scene images are used for scene classification by combining different local and global descriptors [39]. Zhao *et al.* [18] proposed a concentric circle-structured multiscale BOW method using multiple features for land-use scene classification. Nevertheless, the representation ability of handcrafted features grows weaker with the increasing complexity of scene classification tasks.

2) DL Features: CNNs have been widely applied as the feature extractor in computer vision tasks due to their surpassing performance. Cheng *et al.* [28] investigated the use of deep CNNs for scene classification. Fang *et al.* [40] adopted

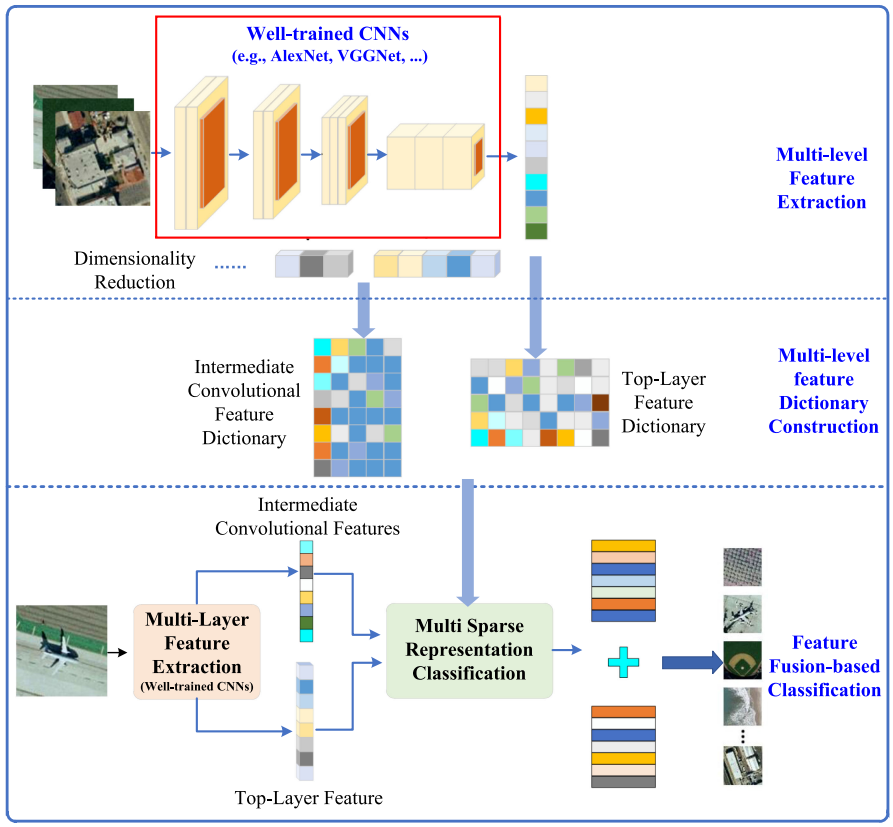

Fig. 1. Overall flowchart of our proposed SRC framework to fuse multilevel features extracted from both the intermediate convolutional layers and the top layers in a CNN for RS imagery scene classification.

the pretrained CaffeNet model with fine-tuning, the proposed method was robust and efficient. Liu *et al.* [41] presented a random-scale stretched CNN to solve scale variations of the same object in different scenes. Li *et al.* [42] employed a deep feature fusion model, which extracted features from pretrained and fine-tuned ResNet50 and VGG16. Tian *et al.* [43] proposed a CapsNet-based network structure called Res-CapsNet for RS scene classification, and achieved improved performance. Generally, these methods require a large number of training samples whether by using handcraft features or deep features. When the number of training sample is limited, their performance may degrade a lot.

B. Sparse Representation Classification

Sparse representation is a signal representation in a small vector space comprising of few nonzero entries. In the recent image classification applications, sparse representation has become a vital method because of its ability to represent higher dimensional data. Ali *et al.* [44] proposed a mathematical approach to map the sparse representation vector to Euclidian distances and achieved a better performance. Hsu *et al.* [45] proposed to integrate spectral and spatial information into a joint sparse representation simultaneously in order to increase performance of hyperspectral image classification. Rong *et al.* [46] presented a spectral-spatial classification framework based on joint superpixel-constrained and weighted sparse representation for HSI classification. Sumarsono *et al.* [47] improved the performance of various classifiers using the traditional linear discriminant analysis followed by maximum likelihood classifier with low-rank subspace representation. Sheng *et al.* [48] presented a cluster structured sparse coding method by unifying sparse coding and structural clustering. In other vision fields, Jiang *et al.* [49] proposed a face recognition algorithm based on sparse representation and feature fusion to improve the accuracy of face recognition. Lan *et al.* [50] proposed a new joint sparse representation model to properly select appropriate features for robust feature-level fusion to address different types of variations such as illumination, occlusion, and pose.

Although these algorithms based on sparse representation or feature fusion have achieved great classification performance, they use the low-level features such as Gabor and HOG instead of deep features learned in CNN. These handcrafted features limit the classification ability of sparse representation framework. In order to combine the advantages of CNNs and SRC, we proposed a novel sparse representation-based framework with deep feature fusion strategy.

III. Proposed Method

A. Overview of the Proposed Classification Scheme

In order for RS scene classification under small training samples, as shown in Fig. 1, we proposed a novel sparse representation-based feature fusion framework to explore the

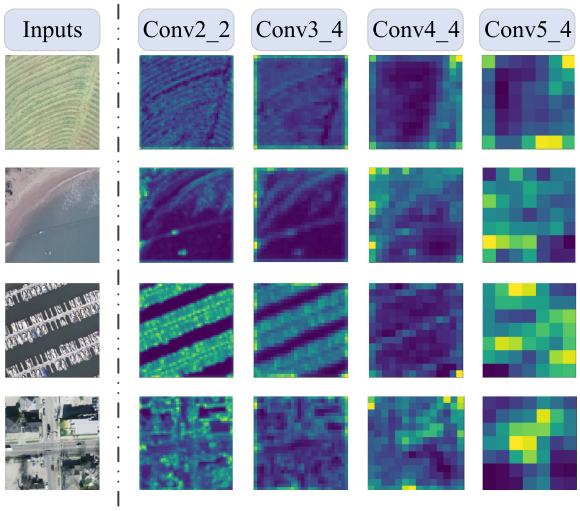

Fig. 2. Visualizations of convolutional feature maps. The feature maps are extracted from different convolutional layers of VGG19-Net.

feature learning advantage of well-trained CNNs and small-training-sample classification superiority of SR classifier. The proposed framework mainly consists of the following two modules.

1) Feature Extraction and Dictionary Construction: Aiming at addressing the few-shot classification, the well-trained CNNs, e.g., AlexNet [24], VGGNet [25], ResNet [27], etc., is used for feature representation, by which the large amount of labeled samples to train CNNs are avoided. Moreover, in order to well explore the feature representation ability of existing well-trained CNNs, multilevel features are used for feature representation of RS scenes.

2) Feature Fusion and SRC: A sparse representation model is used to fuse these multilevel features for RS scene classification, which balance the contribution of features from different layers of CNNs. Moreover, the proposed SR-based fusion also does not require large amount of training samples for classification.

B. Feature Extraction and Dictionary Construction

Generally, features from last or second fully connected layers of a well-trained CNN are used as features to represent the scene. However, the output of many intermediate layers also imply many useful features. Fig. 2 visualizes some feature maps derived from different convolutional layers of VGG19-Net. It is observed that these features generated by intermediate convolutional layers have high-level semantic representation. In order to fully explore the feature representation ability of these well-trained CNNs, features from not only fully connected layers, but also intermediate convolutional layers are used to represent RS scene.

Generally, the feature output of intermediate convolutional layers is highly redundant, which significantly increases the computation complexity and even weakens the performance of subsequent scene classification. Therefore, in order to reduce redundant information and computational complexity, a simple but effective strategy called global average pooling (GAP) is adapted to generate a new processed feature. Assume that the feature extracted from the jth feature map of the ith used intermediate convolutional layer is denoted as $\mathbf{f}^{\text{Convi}}_{1\le j\le \text{ch}_i} \in \mathbb{R}^{w_i * h_i}$, in which w_i, h_i, and ch_i is the width, height, and the channel number of the ith used intermediate convolutional layer. When the GAP strategy is used for dimensionality reduction, the feature from the ith used intermediate convolutional layer $\mathbf{f}^{\text{Convi}}$ can be obtained as

$$\mathbf{f}^{\text{Convi}} = \left[\text{GAP}\left(\mathbf{f}^{\text{Convi}}_1\right), \ldots, \text{GAP}\left(\mathbf{f}^{\text{Convi}}_{\text{ch}_i}\right)\right] \in \mathbb{R}^{\text{ch}_i}. \quad (1)$$

As a result, the feature extracted from all the selected convolutional layers can be denoted as

$$\mathbf{f}^{\text{Conv}} = \left[\mathbf{f}^{\text{Conv1}}, \ldots, \mathbf{f}^{\text{Convn}}\right] \in \mathbb{R}^{d_1} \quad (2)$$

in which n is number of the selected convolutional layers, and $d_1 = \sum_1^n \text{ch}_i$ is the dimension of the multiconvolutional feature $\mathbf{f}^{\text{Conv}}$. Similarly, the feature extracted from the top layer is denoted as $\mathbf{f}^{\text{Top}} \in \mathbb{R}^{d_2}$.

The second step of this module is separately constructing dictionaries for SRC with multilevel features. Let the convolutional feature dictionary consisting of c classes be denoted as

$$\mathbf{D}^{\text{Conv}} = \left[\mathbf{D}^{\text{Conv}}_1, \mathbf{D}^{\text{Conv}}_2, \ldots, \mathbf{D}^{\text{Conv}}_c\right] \in \mathbb{R}^{d_1 \times N} \quad (3)$$

with N being the total number of training samples for c classes. $\mathbf{D}^{\text{Conv}}_i$ is the multiconvolutional features of n_i training samples ($\sum_{i=1}^{c} n_i = N$) from the ith class, denoted as

$$\mathbf{D}^{\text{Conv}}_i = \left[\mathbf{f}^{\text{Conv}}_{i_1}, \mathbf{f}^{\text{Conv}}_{i_2}, \ldots, \mathbf{f}^{\text{Conv}}_{i_{n_i}}\right] \in \mathbb{R}^{d_1 \times n_i}. \quad (4)$$

In addition to the feature dictionary from convolutional layers, another feature dictionary from fully connected layers is also considered, which is represented as

$$\mathbf{D}^{\text{Top}} = \left[\mathbf{D}^{\text{Top}}_1, \mathbf{D}^{\text{Top}}_2, \ldots, \mathbf{D}^{\text{Top}}_c\right] \in \mathbb{R}^{d_2 \times N} \quad (5)$$

where $\mathbf{D}^{\text{Top}}_i, i = 1, 2, \ldots, c$ represents features of fully connected layers for the ith training sample. Generally, the outputs of fully connected layers from a well-trained CNNs are used for such $\mathbf{D}^{\text{Top}}_i, i = 1, 2, \ldots, c$.

C. Feature Fusion and SRC

The SRC framework is first introduced for face recognition and proved to be an effective tool for classification. In SRC, it is assumed that a testing sample can be well approximated by a linear combination of a few atoms from an overcomplete dictionary in which the number of atoms is far more than the dimensions. Under the ideal conditions, the coefficients of the atoms that have no relationship with the class of the testing sample tend to be zeros, which leads the coefficient vector to be sparse. In other words, the testing sample can be represented by training samples of the same class but with different weights. Thus, its class label can be predicted by finding a set of training samples that produce the best approximation. Mathematically, to find these training samples, we need to solve the following

optimization problem

$$\min_{\boldsymbol{\alpha}} \|\boldsymbol{\alpha}\|_0 \quad \text{s.t.} \quad \mathbf{y} = \mathbf{D}\boldsymbol{\alpha} \tag{6}$$

where $\mathbf{y}$ represents a testing sample, $\mathbf{D}$ is the constructed dictionary, and $\boldsymbol{\alpha}$ is the coefficient of sparse representation, besides $\|\boldsymbol{\alpha}\|_0$ counts the number of nonzero elements in $\boldsymbol{\alpha}$.

After obtaining the sparse coefficient $\hat{\boldsymbol{\alpha}}$, the class-specific residuals of $\mathbf{y}$ can be computed as

$$r_i(\mathbf{y}) = \|\mathbf{y} - \mathbf{D}_i\hat{\boldsymbol{\alpha}}_i\|_2, \quad i = 1, 2, \ldots, c. \tag{7}$$

Finally, a predicted classification label is determined as

$$\text{class}(\mathbf{y}) = \arg\min\{r_i\}, \quad i = 1, 2, \ldots, c. \tag{8}$$

In the feature representation step, two kinds of feature dictionary are constructed to represent RS scenes: convolutional feature dictionary $\mathbf{D}^{\text{Conv}}$ and fully connected feature dictionary $\mathbf{D}^{\text{Top}}$. When these two kinds of features are used, each of them can be modeled by the sparse representation model defined in (6). Consequently, in order to fuse these multilevel features of CNNs for RS scene classification, an efficient mathematical model based on SRC is proposed as

$$\begin{aligned} &\min \theta_1 \left\|\boldsymbol{\alpha}^{\text{Top}}\right\|_0 + \theta_2 \left\|\boldsymbol{\alpha}^{\text{Conv}}\right\|_0 \\ &\text{s.t.} \quad \mathbf{y}^{\text{Top}} = \mathbf{D}^{\text{Top}}\boldsymbol{\alpha}^{\text{Top}} \\ &\qquad \mathbf{y}^{\text{Conv}} = \mathbf{D}^{\text{Conv}}\boldsymbol{\alpha}^{\text{Conv}} \end{aligned} \tag{9}$$

in which $\boldsymbol{\alpha}^{\text{Top}}$ and $\boldsymbol{\alpha}^{\text{Conv}}$ represent the coefficient of sparse representation using features from fully connected layers and convolutional layers, respectively, θ_1 and θ_2 are the parameters to balance the reconstruction from different kinds of features by SRC, and $\theta_1 + \theta_2 = 1$.

For a given testing sample $\mathbf{y}$, we first extract the features $\mathbf{y}^{\text{Top}}$ and $\mathbf{y}^{\text{Conv}}$ through the CNN-based feature extractor as mentioned in Section III-B. Then, we adopt the orthogonal matching pursuit (OMP) to estimate the two sparse representation coefficient $\hat{\boldsymbol{\alpha}}^{\text{Top}}$ and $\hat{\boldsymbol{\alpha}}^{\text{Conv}}$. Next, the corresponding residual $r_i^{\text{Top}}(\mathbf{y}^{\text{Top}})$ and $r_i^{\text{Conv}}(\mathbf{y}^{\text{Conv}})$ are obtained by (7). Finally, the residuals are fused with weighting hyperparameters θ_1 and θ_2. The fusion model is formulated as

$$\begin{aligned} r_i^F(\mathbf{y}) &= \theta_1 r_i^{\text{Top}}(\mathbf{y}^{\text{Top}}) + \theta_2 r_i^{\text{Conv}}(\mathbf{y}^{\text{Conv}}) \\ &= \theta_1 \left\|\mathbf{y}^{\text{Top}} - \mathbf{D}_i^{\text{Top}}\hat{\boldsymbol{\alpha}}_i^{\text{Top}}\right\|_2 + \theta_2 \left\|\mathbf{y}^{\text{Conv}} - \mathbf{D}_i^{\text{Conv}}\hat{\boldsymbol{\alpha}}^{\text{Conv}}\right\|_2 \end{aligned} \tag{10}$$

where $\theta_1 + \theta_2 = 1$, and $r_i^F(\mathbf{y})$ is the final feature residual after fusion. According to $r_i^F(\mathbf{y})$ of (10), a label of $\mathbf{y}$ can be found by using (8).

The overall procedure is summarized in Algorithm 1. Obviously, the proposed sparse representation framework and feature fusion strategy can correctly classify testing samples that are misclassified into fault categories if using single-level feature alone.

Algorithm 1: The Overall Procedure of the Proposed Method.

Input: $\mathbf{Y}_{train}$, $\mathbf{Y}_{test}$, parameters θ_1 and θ_2
Output: $class(\mathbf{y})$.

```
Fine-tune the pretrained CNN or train it from scratch.
Construct the dictionaries D^{Conv} and D^{Top} according
    to (3) and (5) on the training samples Y_{train}.
for y in Y_{test} do
    Find the α̂^{Top} and α̂^{Conv} by solving (6) using OMP.
    Compute r_i^{Top} and r_i^{Conv} according to (7).
    Fuse the residuals r_i^{Top} and r_i^{Conv} using (10).
    Attach a label to the testing sample y using (8).
end for
```

D. Computational Complexity

In this section, we analyze the computational complexity of the proposed methods according to the steps of Algorithm 1. Since the complexity of OMP with the dictionary size being $m \times n$ is about $2Kmn + 3\ K^2\ m$[51], the total complexity of getting $\hat{\boldsymbol{\alpha}}^{\text{Top}}$ and $\hat{\boldsymbol{\alpha}}^{\text{Conv}}$ is $2K(d_1 + d_2)N + 3\ K^2(d_1 + d_2)$. Then, that of computing and fusing residuals is about $(d_1 + d_2)N$, which can be ignored compared to the abovementioned calculation. Thus, the total complexity of the proposed algorithm is about $2K(d_1 + d_2)N + 3\ K^2(d_1 + d_2)$, which is greatly influenced by K, but little by d_1, d_2, and N. Since this article mainly focuses on small samples, the size of the dictionary and sparsity is not large. As a result, the proposed method is efficient.

IV. Experiments

In this section, we adapt the proposed SRC-based framework for scene classification of RS images. To demonstrate the effectiveness and superiority of the proposed method, we conduct different experiments on two challenging datasets including UC-Merced21 [52], and WHU-RS19 [19], [53], [54].

A. Experimental Setup

Feature extraction: We adapt three classical CNNs, including AlexNet, VGG19-Net, and ResNet50 to extract the proposed multilevel features. Specifically, AlexNet is trained from scratch, VGG19-Net and ResNet50 is fine-tuned with a small size training samples (up to 10% of all training samples). Note that in AlexNet and VGG19-Net, the top-layer feature is from the last fully connected layer. As for the ResNet50, the last convolutional layer in ResNet50 is selected as the top-layer feature.

Hyperparameters setting: When fusing the representation residuals, two hyperparameters θ_1 and θ_2 balancing the impact of different features on scene classification need to be preset. We tune the value of θ_1 from the range of $\{0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1.0\}$ and θ_2 subjects to $\theta_2 = 1 - \theta_1$.

Compared implementations: To completely demonstrate the effectiveness of our proposed method, we implement the following variants for our method.

1) The SRC methods with the features extracted from only intermediate convolutional layers of three CNNs, are denoted as "AlexNet+LF+SRC," "VGG19+LF+SRC," and "ResNet50+LF+SRC."

Fig. 3. Sample images of UC-Merced dataset. (1) Agriculture, (2) Airplane, (3) Baseballdiamond, (4) Beach, (5) Buildings, (6) Chaparral, (7) Denseresidential, (8) Forest, (9) Freeway, (10) Golfcourse, (11) Harbor, (12) Intersection, (13) Mediumresidential, (14) Mobilehomepark, (15) Overpass, (16) Parkinglot, (17) River, (18) Runway, (19) Sparseresidential, (20) Storagetanks, and (21) Tenniscourt.

2) The SRC methods with the features extracted from only top layers of the VGG19-Net and AlexNet, are denoted as "AlexNet+GF+SRC," "VGG19+GF+SRC." While for ResNet50 that with the feature extracted from the last convolutional layer is denoted as "ResNet50+GF+SRC".
3) The SRC methods of our complete implementation, which fuses the proposed multilevel features, are denoted as "AlexNet+GLF+SRC," "VGG19+GLF+SRC," and "ResNet50+GLF+SRC."

In addition, the following two CNN-based implementations are also considered.

4) The AlexNet trained from scratch, the fine-tuned VGG19-Net and the fine-tuned ResNet50.
5) The end-to-end CNNs, which also fuse multilevel features used in our method, are denoted as "AlextNet+GLF" and "VGG19+GLF."

B. Experiments on UC-Merced Dataset

The first dataset is the UC-merced land use dataset consisting of 21 classes, including agricultural, airplane, baseball diamond, beach, buildings, chaparral, dense residential, forest, freeway, golf course, harbor, intersection, medium density residential, mobile home park, overpass, parking lot, river, runway, sparse residential, storage tanks, and tennis courts. There are 100 images for each class, in which each image measures 256×256 pixels. The images were manually extracted from large images from the USGS National Map Urban Area Imagery collection for various urban areas around the country. The pixel resolution of this public domain imagery is 1 ft. Sample images of each land-use class are illustrated in Fig. 3.

TABLE I
NUMBER OF SAMPLES CORRESPONDING TO DIFFERENT TRAINING RATIOS ON UCM DATASET

Ratios	2%	4%	6%	8%	10%
Samples	2	4	6	8	10

The UC-Merced dataset adopts 2% to 10% of the samples as the training set, and the rest are used for testing. The training ratio increases from 2% to 10%, with the increasing of 2%. Therefore, the number of samples corresponding to these training ratios for each class is 2, 4, 6, 8, or 10 shots. And the details are shown in Table I. The classification results are listed in Table II, in which the best result for each backbone network is highlighted in bold. The following conclusions can be obtained from Table II.

1) With the increase of training sample ratio from 2% to 10%, the performance of almost all methods is improved. For example, the classification accuracy of AlexNet is 25.656% when the training ratio is 2%, and is increased to 48.360% under the training ratio of 10%. The accuracy of "fine-tuned VGG19" is increased from 38.630% to 66.720%, when the training ratio is increased from 2% to 10%. The fine-tuned ResNet50 also has the similar increased accuracy.

2) The proposed multilevel features are effective for classification even in the end-to-end networks, especially when the training sample ratio is very small. When the training ratio is 2%, "AlexNet+GLF" and "VGG19+GLF" improve the classification accuracy by 1.649% and 4.664%, compared with "AlexNet" and "fine-tuned VGG19," respectively.

3) The SRC has better a classification performance than the corresponding end-to-end CNN under the condition of same features, especially in limited training samples. When only the top-layer feature is used for SRC, "AlexNet+GF+SRC" has boosted the classification accuracy by about 1% under all training ratios, compared with the original "AlexNet." More effectively, the method based on VGG19-Net backbone "VGG19+GF+SRC" has enhanced the accuracy by more than 10%. In addition, the accuracy increasement of "ResNet50+GF+SRC" is about 20% under the training ratio of 2%, while it becomes about 2% under the training ratio of 10%. When only the intermediate-layer feature is used for SRC, the same conclusion as that obtained when only the top-layer feature is used, can be drawn.

4) The features extracted from the intermediate convolutional layers also play a vital role for classification as the top-layer features do, even are more important. By comparing "AlexNet (or VGG19 or ResNet50)+LF+SRC" and "AlexNet (or VGG19 or ResNet50)+GF+SRC," the classification accuracy is at the similar level. This is also the reason that the features fusion strategy is proposed to enhance the classification performance.

5) The proposed SRC framework, which fuses the multilevel features has clear superiority in scene classification of RS images. The accuracy of the proposed method is obviously better than that of the corresponding CNN, including the original network structure and the improved structure with multilevel features. For example, compared to AlexNet, "AlexNet+GLF+SRC" boosts the overall classification accuracy

TABLE II
OVERALL CLASSIFICATION ACCURACY COMPARISON ON UCM DATASET

Methods	Accuracy(%)				
	2%	4%	6%	8%	10%
AlexNet	25.656	33.235	40.121	47.153	48.360
AlexNet+GLF	27.305	34.176	43.212	50.880	48.942
AlexNet+GF+SRC	26.531	34.176	41.084	48.758	48.889
AlexNet+LF+SRC	28.717	37.748	45.744	50.880	51.217
AlexNet+GLF+SRC	**28.863**	**38.889**	**46.707**	**52.950**	**54.074**
fine-tuned VGG19	38.630	49.355	54.255	60.611	66.720
VGG19+GLF	43.294	54.970	57.900	60.973	64.021
VGG19+GF+SRC	55.928	67.212	72.087	73.602	77.301
VGG19+LF+SRC	57.434	68.551	69.656	73.291	75.608
VGG19+GLF+SRC	**57.823**	**69.296**	**72.391**	**74.327**	**77.407**
fine-tuned ResNet50	40.180	56.845	69.149	73.290	84.402
ResNet50+GF+SRC	60.884	72.470	76.899	82.557	86.190
ResNet50+LF+SRC	60.009	71.528	75.836	80.538	85.291
ResNet50+GLF+SRC	**61.618**	**73.611**	**77.710**	**83.592**	**87.672**

TABLE III
NUMBER OF SAMPLES CORRESPONDING TO DIFFERENT TRAINING RATIOS ON WHU-RS19 DATASET

Ratios	2%	4%	6%	8%	10%
Samples	1	2	3	4	5

by 3.2% to 6% under different training ratios. Compared to fine-tuned VGG19-Net, "VGG19+GLF+SRC" enhances the overall accuracy by 19% under 2% training ratio, the improvement is reduced when the training ratio is 10%, but also exceeds 10%. Besides, "ResNet50+GLF+SRC" also improves classification performance to varying degrees. The accuracy is also better than that of the SRC methods using only top-layer features or convolutional-layer features.

C. Experiments on WHU-RS19 Dataset

The second RS dataset is a 19-class Google image dataset of WHU-RS19 designed by Wuhan University. The dataset is acquired from Google Earth and mainly covers urban areas, and there are 50 images for each of the following classes: airport, beach, bridge, commercial area, desert, farmland, football field, forest, industrial area, meadow, mountain, park, parking, pond, port, railway station, residential area, river, and viaduct. Each image measures 600 × 600 pixels, with a 0.5 m–8 m spatial resolution. Fig. 4 shows representative images of each class.

The WHU-RS19 dataset still randomly chooses 2% to 10% samples of each class for training and the rest for testing. The relationship between training ratio and sample size is shown in Table III; the classification results are listed in Table IV, in which the best result for each backbone network is highlighted in bold. On a whole, the same conclusions drawn from the experiments on UC-Merced dataset can be obtained. Therefore, on this dataset, we give some special cases and analysis as follows.

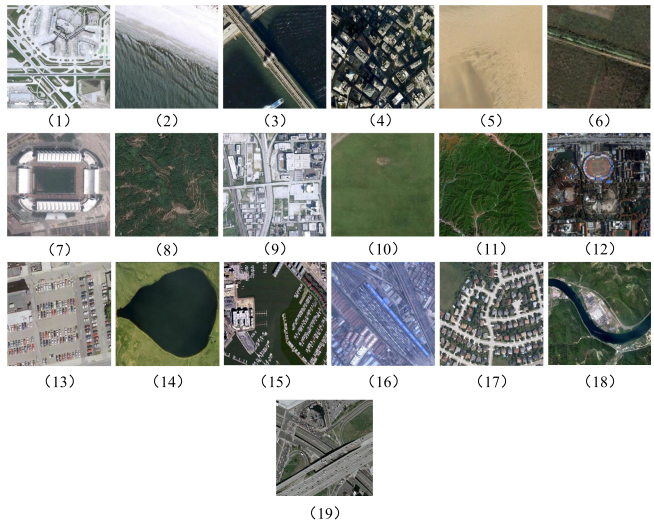

Fig. 4. Sample images of WHU-RS19 dataset. (1) Airport, (2) Beach, (3) Bridge, (4) Commercial, (5) Desert, (6) Farmland, (7) Footballfield, (8) Forest, (9) Industrial, (10) Meadow, (11) Mountain, (12) Park, (13) Parking, (14) Pond, (15) Port, (16) Railwaystation, (17) Residential, (18) River, and (19) Viaduct.

1) The fine-tuned ResNet50 performs the best except under the training ratio of 2%, and the fine-tuned VGG19 performs better than AlexNet. The accuracy of fine-tuned "ResNet50" is only 14.286% under 2% training ratio, which is far lower than AlexNet and VGG19-Net. However, it achieves higher accuracies than the two under the training ratios of 4%, 6%, 8%, and 10%. The reason may be that very few samples cause underfitting to a deeper CNN.

2) The SRC-based method generally offers a better performance than the corresponding end-to-end CNN under the same features. However, "AlexNet+GF+SRC" weakens the classification performance under the 2% and 10% training ratios. One possible reason is the large size of convolution kernel

TABLE IV
OVERALL CLASSIFICATION ACCURACY COMPARISON ON WHU-RS19 DATASET

Methods	Accuracy(%)				
	2%	4%	6%	8%	10%
AlexNet	31.686	38.816	42.105	46.110	53.567
AlexNet+GLF	33.830	40.24	42.777	48.741	53.918
AlexNet+GF+SRC	30.397	38.816	43.113	47.597	51.579
AlexNet+LF+SRC	35.338	40.789	44.793	48.512	54.503
AlexNet+GLF+SRC	**35.553**	**41.886**	**45.465**	**49.428**	**56.491**
fine-tuned VGG19	28.679	37.719	43.001	51.030	62.105
VGG19+GLF	33.730	42.760	47.140	49.656	54.153
VGG19+GF+SRC	47.798	65.351	69.317	77.002	77.193
VGG19+LF+SRC	49.409	69.079	70.436	76.545	77.427
VGG19+GLF+SRC	**50.698**	**69.847**	**70.997**	**77.459**	**78.596**
fine-tuned ResNet50	14.286	43.531	54.983	69.908	76.490
ResNet50+GF+SRC	65.736	79.715	84.786	86.155	88.421
ResNet50+LF+SRC	67.776	79.824	84.450	85.126	87.602
ResNet50+GLF+SRC	**69.387**	**80.482**	**85.682**	**86.842**	**88.655**

and the shallow network structure limit the ability to extract features of AlexNet. In addition, "AlexNet (VGG19 or ResNet50)+LF+SRC" performs obviously better than AlexNet (VGG19 or ResNet50) under all training ratios.

3) The proposed method can effectively improve the scene classification accuracy, especially under the case of few training samples, even only 1 training sample for each class, such as the training ratio is 2% of 50 samples. It is worth mentioning that although the accuracy of "fine-tuned ResNet50" is only 14.286% under the training ratio of 2%, the accuracy of the proposed "ResNet50+GLF+SRC" is 69.387%, with an improvement more than 55%.

In conclusion, compared to classic CNN classifiers, our proposed SRC framework with feature fusion strategy has effectively boosted the scene classification performance for RS images.

D. Comparison With State-of-the-Art Methods

To further demonstrate the superiority of our method, we conduct a comprehensive comparison with state-of-the-arts that have been evaluated on the UC-Merced and WHU-RS19 datasets. The comparison methods include attention recurrent convolutional network (ARCNet) [55], the method based on the improved cross-entropy loss (ICEL) [56], and that based on MSCP [31].

The comparison results of accuracy on two datasets are show in Tables V and VI, respectively, in which the best results for different training ratio are highlighted in bold. The accuracy of the proposed method on both datasets is obviously higher than that of the other comparison methods under all the ratios of training samples. It is observed from Table V that when the training ratio is 10%, the proposed method improves the accuracy by at least 2.9% on UCM dataset. As the training ratio decreases from 10% to 2%, the advantage of our method becomes more prominent, which indicates that it effectively improves the scene classification performance especially for few-shot classification.

TABLE V
COMPARISON BETWEEN OURS AND SOME STATE-OF-THE-ART METHODS ON UCM DATASET

Methods	Accuracy(%)				
	2%	4%	6%	8%	10%
ARCNet [55]	47.667	60.119	73.455	75.763	84.021
ICEL [56]	49.854	61.310	70.061	78.157	82.12
MSCP [31]	14.14	59.23	75.87	82.65	84.75
Ours	**61.618**	**73.611**	**77.710**	**83.592**	**87.672**

TABLE VI
COMPARISON BETWEEN OURS AND SOME STATE-OF-THE-ART METHODS ON WHU-RS19 DATASET

Methods	Accuracy(%)				
	2%	4%	6%	8%	10%
ARCNet [55]	41.890	63.158	68.192	74.020	81.871
ICEL [56]	24.812	63.158	66.517	75.515	75.906
MSCP [31]	/	/	48.98	79.75	85.68
Ours	**69.387**	**80.482**	**85.682**	**86.842**	**88.655**

Table VI presents similar results as Table V. It is worth mentioning that WHU-RS19 dataset has less RS images than UCM dataset. Therefore, our method achieves greater gain on WHU-RS19 in Table VI than on UCM in Table V under the training ratio of 2%–6%. All these results demonstrate the effectiveness and superiority of our few-shot RS scene classification.

E. Explorations on Hyperparameters

Since the contribution for SRC of each type of feature may affect the final classification accuracy, it should be explored for better fusion results and empirical settings for similar fusion work. The hyperparameter settings have been introduced in Section IV-A. The classification accuracies based on AlexNet

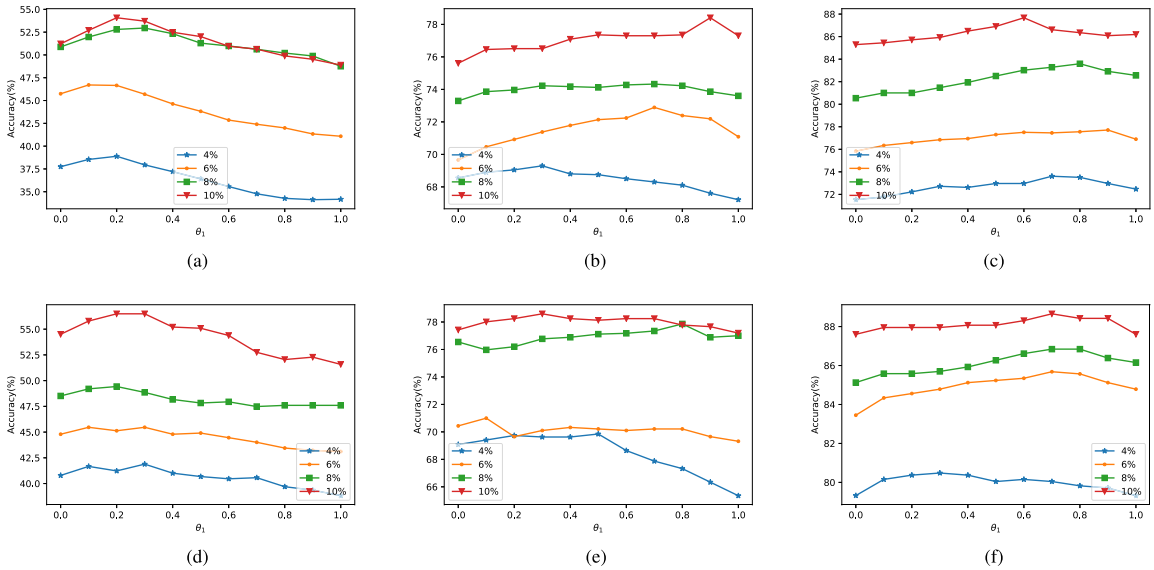

Fig. 5. Classification accuracy with different θ_1 values based on AlexNet, VGG19-Net, and ResNet50 on UC-Merced and WHU-RS19 datasets. (a) AlexNet on UC-Merced21. (b) VGG19 on UC-Merced21. (c) ResNet50 on UC-Merced21. (d) AlexNet on WHU-RS19. (e) VGG19 on WHU-RS19. (f) ResNet50 on WHU-RS19.

with different values of θ_1 on UC-Merced and WHU-RS19 are shown in Fig. 5(a) and (d), respectively. In addition, those based on VGG19-Net and ResNet50 are shown in Fig. 5(b), (c), (e), and (f).

When θ_1 is 0, the top features do not contribute anything in feature fusion, while intermediate features contribute all. On the contrary, the top features contribute all when $\theta_1 = 1$.

A suitable range of θ_1 can always be found, which makes the classification accuracy of multilevel feature fusion higher than that of single feature. This proves the advantage of multilayer feature fusion for classification. For AlexNet, when θ_1 is small, the performance is better, indicating that the features of the intermediate layer play a more important role in fused feature for classification in AlexNet. For deeper CNNs including VGG19 and ResNet50, the large hyperparameter θ_1 leads to the best classification accuracy in most cases.

F. Explorations on Intermediate Features Fusion

In terms of classical CNNs with the top layers, such as AlexNet and VGG19-Net, we propose the feature fusion strategy to complement the top-layer feature and the intermediate convolutional-layer feature each other. However, an issue that cannot be ignored is how to choose the intermediate features from different convolutional stages, or which stages should be selected. To make fully use of the intermediate features, we conduct the experiments on WHU-RS19 dataset to solve these problems.

Inspired by the architectures of AlexNet and VGG19-Net, intermediate features from the last three convolutional stages, denoted as F_{conv1}, F_{conv2}, and F_{conv3} numbered from the distance to the output layer, are selected. Then, the new feature is formed by concatenating the last one, two, and three intermediate features, that is, F_{conv1}, $F_{\text{conv1}} + F_{\text{conv2}}$, $F_{\text{conv1}} + F_{\text{conv2}} + F_{\text{conv3}}$.

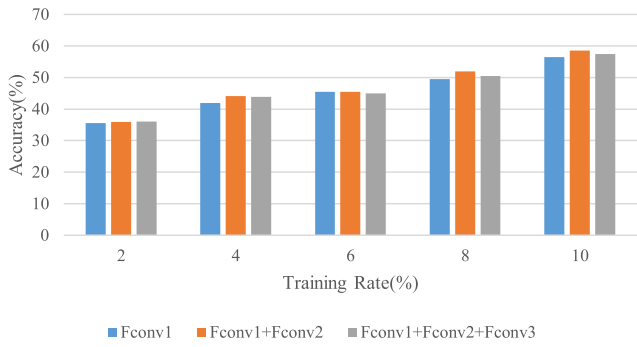

Fig. 6. Classification result of different intermediate features based on AlexNet.

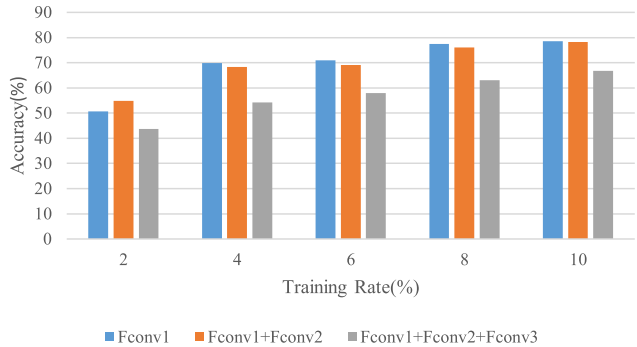

Fig. 7. Classification result of different intermediate features based on VGG19-Net.

Note that no fully connected feature in top layer is considered in this experiment. Figs. 6 and 7 show the classification result of concatenated features based on AlexNet and VGG19-Net backbones, respectively.

TABLE VII
COMPARISON BETWEEN OURS AND DATA AUGMENTATION METHOD ON WHU-RS19 DATASET

Methods	Accuracy				
	2%	4%	6%	8%	10%
VGG19	30.29	42.544	54.871	67.62	72.865
f-VGG19	34.479	44.627	56.439	70.59	68.187
f-VGG19+DataAug	45.435	**65.57**	68.757	76.201	78.012
f-VGG19+GLF+SRC	51.88	62.281	**75.924**	79.748	80.935
VGG19+GLF+SRC	**52.202**	64.583	73.348	**81.579**	**82.222**

It is observed from Fig. 6 that the new feature concatenated by F_{conv1} and F_{conv2} of AlexNet achieves the best classification performance among the three new features. However, for VGG19-Net, Fig. 7 demonstrates that F_{conv1} can provide almost the best result except under the training ratio of 2%. These experimental results indicate that the CNN with weaker feature extraction ability, which is caused by shallow architecture, large-size convolutional kernels or very few samples for training, needs more intermediate convolutional features to improve the classification performance, but the redundance brought by more intermediate features is also an issue to be considered. While for the network with strong feature extraction ability, more intermediate features may degrade the classification performance.

G. Explorations on Fine-Tuned and Data Augmentation

In order to demonstrate that the proposed method has a significant improvement under the condition of small-sample-size for RS scene classification, the following variants are implemented over WHU-RS19 dataset.

1) "VGG19": The original VGG19 is directly used for scene classification by fine-tuning the weights of the last fully connected layer.
2) "f-VGG19": Compared to the original VGG19, all layers involved in VGG19 are fine-tuned to achieve better performance.
3) "VGG19+GLF+SRC": Our proposed SRC-based method fuses the multilevel features of the original VGG19.
4) "f-VGG19+GLF+SRC": Our proposed SRC-based method fuses the multilevel features of the "f-VGG19."
5) "f-VGG19+DataAug": The VGG19 is retrained using data augmentation strategy. For the data augmentation, the rotation, horizontal and vertical shift, and horizontal flip are all considered.

The results of all these algorithms over WHU-RS19 dataset are shown in Table VII, and the best results for different training ratio are highlighted in bold. The following can be observed.

1) The f-VGG19 can achieve better results than VGG19 in almost all training ratios except for the training set of 10%. This is because fine-tuning all the layers is better than just fine-tuning the last classification layer.

2) When combined with multilevel features and SRC, the results of f-VGG19+GLF+SRC are a little worse than VGG10+GLF+SRC, indicating that fine-tuning with a small number of samples may degrade the feature representation ability of VGG19 for RS images.

3) Compared to fine-tuned VGG19 with data augmentation, our proposed method achieves much better improvements in most instances. Only at 4% training ratio, the results of the two are roughly the same. These experiments demonstrate the effectiveness of our proposed sparse representation-based framework, which fuses multilevel features for scene classification of small samples RS images.

V. CONCLUSION

In this article, we propose a novel few-shot classification framework using sparse representation to fuse the multilevel features extracted from CNNs to boost the performance of RS imagery scene classification. The proposed method aims to solve the following problems. First, the existing CNN-based methods extract multilevel features from different layers of CNNs, but feed only single level to the classifier and neglect the other features with important information. Second, the training of CNNs requires many training samples, which is generally unavailable in the application of RS. Thus, in order to address these problems, the proposed framework includes the two main modules. The one is multilayer feature extraction, which can extract different levels of features from both the top layer and the intermediate layers to obtain more rich representation for scene classification. The other one is feature fusion based SRC, in which a simple but highly effective strategy is devised to fuse multilevel features for classification. Experimental results on two public benchmark datasets demonstrate that the proposed few-shot classification framework using sparse representation embedded with multilevel deep feature fusion certainly boosts the classification performance compared to single-feature-based methods. Moreover, the proposed method achieves the state-of-the-art results under the case of limited training samples, even only 1 or 2 training samples per class.

REFERENCES

[1] S. Chen and Y. Tian, "Pyramid of spatial relatons for scene-level land use classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 53, no. 4, pp. 1947–1957, Apr. 2015.
[2] Q. Weng, Z. Mao, J. Lin, and W. Guo, "Land-use classification via extreme learning classifier based on deep convolutional features," *IEEE Geosci. Remote Sens. Lett.*, vol. 14, no. 5, pp. 704–708, May 2017.
[3] X. Li *et al.*, "Spatial-temporal super-resolution land cover mapping with a local spatial-temporal dependence model," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 7, pp. 4951–4966, May 2019.

[4] D. Ienco, R. Gaetano, C. Dupaquier, and P. Maurel, "Land cover classification via multitemporal spatial data by deep recurrent neural networks," *IEEE Geosci. Remote Sens. Lett.*, vol. 14, no. 10, pp. 1685–1689, Oct. 2017.

[5] Y. Zhong, Q. Zhu, and L. Zhang, "Scene classification based on the multifeature fusion probabilistic topic model for high spatial resolution remote sensing imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 53, no. 11, pp. 6207–6222, Nov. 2015.

[6] M. A. Dede, E. Aptoula, and Y. Genc, "Deep network ensembles for aerial scene classification," *IEEE Geosci. Remote Sens. Lett.*, vol. 16, no. 5, pp. 732–735, May 2019.

[7] S. Song, H. Yu, Z. Miao, Q. Zhang, Y. Lin, and S. Wang, "Domain adaptation for convolutional neural networks-based remote sensing scene classification," *IEEE Geosci. Remote Sens. Lett.*, vol. 16, no. 8, pp. 1324–1328, Aug. 2019.

[8] F. Zhang, B. Du, and L. Zhang, "Scene classification via a gradient boosting random convolutional network framework," *IEEE Trans. Geosci. Remote Sens.*, vol. 54, no. 3, pp. 1793–1802, Mar. 2016.

[9] F. Naccari, S. Battiato, A. Bruna, A. Capra, and A. Castorina, "Natural scenes classification for color enhancement," *IEEE Trans. Consum. Electron.*, vol. 51, no. 1, pp. 234–239, Feb. 2005.

[10] J. G. Aviña-Cervantes, S. Ledezma-Orozco, M. Torres-Cisneros, D. Hernández-Fusilier, J. González-Barbosa, and A. Salazar-Garibay, "Color texture histograms for natural images interpretation," in *Proc. Mex. Int. Conf. Artif. Intell., Special Session*, 2007, pp. 131–140.

[11] A. Zohrevand, A. Ahmadyfard, A. Pouyan, and Z. Imani, "A sift based object recognition using contextual information," in *Proc. Iranian Conf. Intell. Syst.*, 2014, pp. 1–4.

[12] M. Villamizar, J. Scandaliaris, A. Sanfeliu, and J. Andrade-Cetto, "Combining color-based invariant gradient detector with hog descriptors for robust image detection in scenes under cast shadows," in *Proc. IEEE Int. Conf. Robot. Autom.*, 2009, pp. 1997–2002.

[13] M. Wei and P. Xiwei, "Wlib-sift: A distinctive local image feature descriptor," in *Proc. Int. Conf. Inf. Commun. Signal Process.*, 2019, pp. 379–383.

[14] Z. Ni, "B-sift: A binary sift based local image feature descriptor," in *Proc. Int. Conf. Digit. Home*, 2012, pp. 117–121.

[15] S. Banerji, A. Sinha, and C. Liu, "Scene image classification: Some novel descriptors," in *Proc. IEEE Int. Conf. Syst., Man, Cybern.*, 2012, pp. 2294–2299.

[16] Q. Zhu, Y. Zhong, B. Zhao, G. Xia, and L. Zhang, "The bag-of-visual-words scene classifier combining local and global features for high spatial resolution imagery," in *Proc. Int. Conf. Fuzzy Syst. Knowl. Discov.*, 2015, pp. 717–721.

[17] T. Chen, K. Yap, and L. Chau, "From universal bag-of-words to adaptive bag-of-phrases for mobile scene recognition," in *Proc. IEEE Int. Conf. Image Process.*, 2011, pp. 825–828.

[18] L. Zhao, P. Tang, and L. Huo, "Land-use scene classification using a concentric circle-structured multiscale bag-of-visual-words model," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 7, no. 12, pp. 4620–4631, Dec. 2014.

[19] Q. Zhu, Y. Zhong, B. Zhao, G. Xia, and L. Zhang, "Bag-of-visual-words scene classifier with local and global features for high spatial resolution remote sensing imagery," *IEEE Geosci. Remote Sens. Lett.*, vol. 13, no. 6, pp. 747–751, Jun. 2016.

[20] B. Zhao, Y. Zhong, and L. Zhang, "Hybrid generative/discriminative scene classification strategy based on latent Dirichlet allocation for high spatial resolution remote sensing imagery," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, 2013, pp. 196–199.

[21] R. Bahmanyar, D. Espinoza-Molina, and M. Datcu, "Multisensor earth observation image classification based on a multimodal latent Dirichlet allocation model," *IEEE Geosci. Remote Sens. Lett.*, vol. 15, no. 3, pp. 459–463, Mar. 2018.

[22] H. Tang *et al.*, "A multiscale latent Dirichlet allocation model for object-oriented clustering of VHR panchromatic satellite images," *IEEE Trans. Geosci. Remote Sens.*, vol. 51, no. 3, pp. 1680–1692, Mar. 2013.

[23] M. Lienou, H. Maitre, and M. Datcu, "Semantic annotation of satellite images using latent Dirichlet allocation," *IEEE Geosci. Remote Sens. Lett.*, vol. 7, no. 1, pp. 28–32, Jan. 2010.

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

[25] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*.

[26] C. Szegedy *et al.*, "Going deeper with convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 1–9.

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

[28] G. Cheng, C. Ma, P. Zhou, X. Yao, and J. Han, "Scene classification of high resolution remote sensing images using convolutional neural networks," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, 2016, pp. 767–770.

[29] W. Zhou, Z. Shao, and Q. Cheng, "Deep feature representations for high-resolution remote sensing scene classification," in *Proc. Int. Workshop Earth Observation Remote Sens. Appl.*, 2016, pp. 338–342.

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

[31] N. He, L. Fang, S. Li, A. Plaza, and J. Plaza, "Remote sensing scene classification using multilayer stacked covariance pooling," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 12, pp. 6899–6910, Dec. 2018.

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

[33] J. Xie, N. He, L. Fang, and A. Plaza, "Scale-free convolutional neural network for remote sensing scene classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 9, pp. 6916–6928, Sep. 2019.

[34] T. Wei, J. Wang, W. Liu, H. Chen, and H. Shi, "Marginal center loss for deep remote sensing image scene classification," *IEEE Geosci. Remote Sens. Lett.*, vol. 17, no. 6, pp. 968–972, Jun. 2020.

[35] D. Alajaji, H. S. Alhichri, N. Ammour, and N. Alajlan, "Few-shot learning for remote sensing scene classification," in *Proc. Mediterranean Middle-East Geosci. Remote Sens. Symp.*, 2020, pp. 81–84.

[36] D. Sun and J. Watada, "Detecting pedestrians and vehicles in traffic scene based on boosted hog features and SVM," in *Proc. Int. Symp. Intell. Signal Process.*, 2015, pp. 1–4.

[37] L. Gan, P. Liu, and L. Wang, "Rotation sliding window of the hog feature in remote sensing images for ship detection," in *Proc. Int. Symp. Comput. Intell. Des.*, 2015, pp. 401–404.

[38] J. Chu and G. Zhao, "Scene classification based on sift combined with gist," in *Int. Conf. Inf. Sci., Electron. Elect. Eng.*, 2014, pp. 331–336.

[39] B. F. Cura and E. Surer, "Scene classification: A comprehensive study combining local and global descriptors," in *Proc. Signal Process. Commun. Appl. Conf.*, 2019, pp. 1–4.

[40] Z. Fang, W. Li, J. Zou, and Q. Du, "Using CNN-based high-level features for remote sensing scene classification," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, 2016, pp. 2610–2613.

[41] Y. Liu, Y. Zhong, F. Fei, and L. Zhang, "Scene semantic classification based on random-scale stretched convolutional neural network for high-spatial resolution remote sensing imagery," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, 2016, pp. 763–766.

[42] Y. Li, Q. Wang, X. Liang, and L. Jiao, "A novel deep feature fusion network for remote sensing scene classification," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, 2019, pp. 5484–5487.

[43] T. Tian, X. Liu, and L. Wang, "Remote sensing scene classification based on res-capsnet," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, 2019, pp. 525–528.

[44] A. Julazadeh, M. Marsousi, and J. Alirezaie, "Classification based on sparse representation and euclidian distance," in *Proc. Vis. Commun. Image Process.*, 2012, pp. 1–5.

[45] P. Hsu and Y. Cheng, "Hyperspectral image classification via joint sparse representation," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, 2019, pp. 2997–3000.

[46] J. Rong, H. Fu, A. Zhang, G. Sun, H. Huang, and Y. Hao, "Hyperspectral image classification based on joint superpixel-constrained and weighted sparse representation," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, 2019, pp. 3313–3316.

[47] A. Sumarsono and Q. Du, "Hyperspectral image classification with low-rank subspace and sparse representation," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, 2015, pp. 2864–2867.

[48] G. Sheng, W. Yang, L. Yu, and H. Sun, "Cluster structured sparse representation for high resolution satellite image classification," in *Proc. Int. Conf. Signal Process.*, 2012, pp. 693–696.

[49] C. Jiang, M. Wang, X. Tang, and R. Mao, "Face recognition method based on sparse representation and feature fusion," in *Proc. Chin. Automat. Congr.*, 2019, pp. 396–400.

[50] X. Lan, A. J. Ma, P. C. Yuen, and R. Chellappa, "Joint sparse representation and robust feature-level fusion for multi-cue visual tracking," *IEEE Trans. Image Process.*, vol. 24, no. 12, pp. 5826–5841, Dec. 2015.

[51] J. Wang, S. Kwon, and B. Shim, "Generalized orthogonal matching pursuit," *IEEE Trans. Signal Process.*, vol. 60, no. 12, pp. 6202–6216, Dec. 2012.
[52] Y. Yang and S. Newsam, "Bag-of-visual-words and spatial extensions for land-use classification," in *Proc. 18th SIGSPATIAL Int. Conf. Adv. Geographic Inf. Syst.*, 2010, pp. 270–279.
[53] B. Zhao, Y. Zhong, G. Xia, and L. Zhang, "Dirichlet-derived multiple topic scene classification model for high spatial resolution remote sensing imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 54, no. 4, pp. 2108–2123, Apr. 2016.
[54] B. Zhao, Z. Yanfei, Z. Liangpei, and H. Bo, "The fisher kernel coding framework for high spatial resolution scene classification," *Remote Sens.*, vol. 8, no. 2, p. 157, 2016.
[55] Q. Wang, S. Liu, J. Chanussot, and X. Li, "Scene classification with recurrent attention of VHR remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 2, pp. 1155–1167, Feb. 2019.
[56] A. Bahri, S. Ghofrani Majelan, S. Mohammadi, M. Noori, and K. Mohammadi, "Remote sensing image classification via improved cross-entropy loss and transfer learning strategy based on deep convolutional neural networks," *IEEE Geosci. Remote Sens. Lett.*, vol. 17, no. 6, pp. 1087–1091, Jun. 2020.

Shaohui Mei (Senior Member, IEEE) received the B.S. degree in electronics and information engineering and the Ph.D. degree in signal and information processing from Northwestern Polytechnical University, Xi'an, China, in 2005 and 2011, respectively.

He is currently an Associated Professor with the School of Electronics and Information, Northwestern Polytechnical University. From October 2007 to October 2008, he was a Visiting Student with the University of Sydney, Camperdown, NSW, Australia. His research interests include hyperspectral remote sensing image processing and applications, intelligent signal and information acquisition and processing, video processing, and pattern recognition.

Dr. Mei is currently an Associated Editor for the IEEE JOURNAL OF SELECTED TOPICS IN APPLIED EARTH OBSERVATIONS AND REMOTE SENSING and a Reviewer of more than 20 international famous academic journals. He was the recipient of Excellent Doctoral Dissertation Award of Shaanxi Province in 2014, Best Paper Award of IEEE ISPACS 2017, and Best Reviewer of IEEE JSTARS in 2019. He was also the Registration Chair of IEEE China Summit and International Conference on Signal and Information Processing2014.

Keli Yan received the M.S. degree in signal and information processing from Northwestern Polytechnical University, Xi'an, China, in 2018 and 2021, respectively.

His research interests include remote sensing image processing and scene classification.

Mingyang Ma (Student Member, IEEE) received the B.S. degree in 2015 in communication engineering from Northwestern Polytechnical University, Xi'an, China, where he is currently working toward the Ph.D. degree in information and communication engineering.

His main research interests include image processing and video summarization.

Xiaoning Chen received the B.S. degree in electronic information science and technology in 2004 and the M.S. degree in communication and information system in 2011 from Northwest University, Xian, China where she is currently working toward the Ph.D. degree in information and communication engineering.

Her research interests include computer vision and remote sensing image analysis.

Shun Zhang (Member, IEEE) received the B.S. and Ph.D. degrees in electronic engineering from Xi'an Jiaotong University, Xi'an, China, in 2009 and 2016, respectively.

He is currently an Associate Professor with the School of Electronic and Information, Northwestern Polytechnical University, Xi'an. His research interests include machine learning, computer vision and human–computer interaction, with a focus on object detection, visual tracking, person reidentification, image classification, feature extraction, and sparse representation.

Qian Du (Fellow, IEEE) received the Ph.D. degree in electrical engineering from the University of Maryland at Baltimore County, Baltimore, MD, USA, in 2000.

She is currently the Bobby Shackouls Professor with the Department of Electrical and Computer Engineering, Mississippi State University, Starkville, MS, USA. Her research interests include hyperspectral remote sensing image analysis and applications, pattern classification, data compression, and neural networks.

Dr. Du is a Fellow of IEEE and the SPIE-International Society for Optics and Photonics. She was a Co-Chair for the Data Fusion Technical Committee of the IEEE Geoscience and Remote Sensing Society from 2009 to 2013, and the Chair for Remote Sensing and Mapping Technical Committee of the International Association for Pattern Recognition from 2010 to 2014. She was an Associate Editor for the IEEE JOURNAL OF SELECTED TOPICS IN APPLIED EARTH OBSERVATIONS AND REMOTE SENSING, the *Journal of Applied Remote Sensing*, and the IEEE SIGNAL PROCESSING LETTERS. Since 2016, she has been the Editor-in-Chief of the IEEE JOURNAL OF SELECTED TOPICS IN APPLIED EARTH OBSERVATIONS AND REMOTE SENSING. She is the General Chair for the 4th IEEE GRSS Workshop on Hyperspectral Image and Signal Processing: Evolution in Remote Sensing in Shanghai, China, in 2012. She was the recipient of the 2010 Best Reviewer Award from the IEEE Geoscience and Remote Sensing Society.