IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received December 18, 2019, accepted January 5, 2020, date of publication January 22, 2020, date of current version February 17, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.2968771

Remote Sensing Scene Classification Based on Multi-Structure Deep Features Fusion

WEI XUE, XIANGYANG DAI, AND LI LIU
School of Automation, China University of Geosciences, Wuhan 430074, China
Hubei Key Laboratory of Advanced Control and Intelligent Automation for Complex Systems, Wuhan 430074, China

Corresponding author: Wei Xue (xuew@cug.edu.cn)

This work was supported in part by the National Natural Science Foundation of China under Grant 61805215.

ABSTRACT Convolutional neural networks (CNNs) have been widely used in remote sensing scene classification due to their excellent performance in natural image classification. However, the complementarity of features extracted by different CNNs is seldom exploited, which limits the further improvement of classification accuracy. To solve this problem, we propose a classification method based on multi-structure deep features fusion (MSDFF). First, a data augmentation method based on random-scale cropping is adopted to achieve the expansion of limited data. Then, three popular CNNs are respectively used as feature extractors to capture deep features from the image. Finally, a deep feature fusion network is adopted to fuse these features and implement the classification. The effectiveness of the proposed method is verified on UC Merced, AID, and NWPU-RESISC45 datasets. The proposed method can achieve better performance than state-of-the-art scene classification methods.

INDEX TERMS Convolutional neural network, scene classification, feature extraction, multi-structure deep features fusion.

I. INTRODUCTION

With the development of satellite remote sensing technology, remote sensing scene classification has played a significant role in many practical applications such as land management, urban planning, and environment monitoring [1]–[4]. Robust feature extraction is a key technique for scene classification. Feature extraction methods can be divided into three categories: low-level, middle-level, and deep feature methods. Due to the semantic gap between low-level features and high-level semantic features, it is difficult to obtain satisfactory results using low-level features such as spectra, colors, textures and shapes.

Many mid-level feature methods have been proposed to overcome the semantic gap. One popular mid-level approach is the bag-of-visual-words (BoVW) model [5], which is derived from document classification in text analysis. By considering the lack of spatial information in the BoVW representation, many BoVW variants have been proposed [6]–[8]. Yang and Newsam [6] developed the spatial pyramid co-occurrence kernel to capture both the absolute and relative spatial arrangement of words. Zhao *et al.* [7] proposed a concentric circle-based spatial-rotation-invariant representation to encode the spatial information. Chen and Tian [8] proposed the pyramid-of-spatial-relation model to effectively incorporate spatial information into the BoVW model for the land use classification. Two probabilistic topic models inspired by BoVW are latent Dirichlet allocation (LDA) [9] and probabilistic latent semantic analysis (pLSA) [10]. Zhao *et al.* [11] proposed P-LDA and F-LDA by using LDA as part of the classifier and the topic feature extractor, respectively. To combine different features, Kusumaningrum *et al.* [12] used the edge orientation histogram, CIELab color moments, and gray-level co-occurrence matrix to extract spectral, texture, and structure information in the LDA model. Zhong *et al.* [13] employed the LDA model and the pLSA model to capture semantic information with a multi-feature fusion approach and achieved enhanced performance compared with those where a single feature is employed. However, these mid-level methods all require prior knowledge of manual feature extraction and lack the flexibility for different scenes.

Recently, many deep learning methods have been proposed to capture deep features of image adaptively [14], [15]. Compared with low-level and mid-level methods, deep learning methods can learn more abstract and discriminative semantic features, and achieve excellent image scene classification performance [16]–[24]. Convolutional neural

The associate editor coordinating the review of this manuscript and approving it for publication was Donato Impedovo.

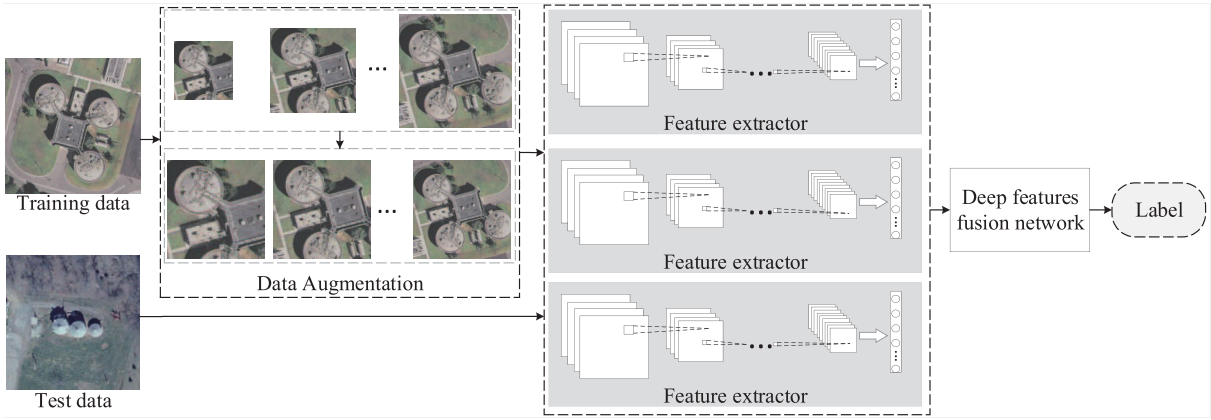

FIGURE 1. Framework of the proposed method.

networks (CNNs) are the most popular deep learning-based networks, which can directly extract features from raw image pixels without prior knowledge. Compared with the hand-crafted feature-based methods, CNNs employ an end-to-end feature learning approach and exhibit better performance for high spatial resolution (HSR) image scene classification [16], [17]. In practice, CNNs need a large amount of data for training, which not only incurs additional complexity but also causes overfitting especially in case of insufficient data. Transfer learning-based scene classification methods were proposed to address this problem for HSR imagery scene classification [25]–[27] and can speed up the training progress and guarantee the performance of deep learning models with a moderate amount of data. In addition, feature fusion strategy is also introduced into CNNs to further improve the performance of scene classification [28]–[31]. Liu *et al.* [28] combined the activations from intermediate and FC layers to generate a converted CNN with a directed acyclic graph topology. Yu and Liu [29] proposed a new two-stream deep architecture to fuse two different types of deep convolutional features extracted by the original red-green-blue (RGB) stream and the saliency stream. Ye *et al.* [30] proposed a parallel multi-stage structure model formed by three different level sub-models to obtain low-, middle-, and high-level comprehensive semantic representation. Cheng *et al.* [31] proposed discriminative CNNs (D-CNNs) trained by optimizing a new discriminative objective function and used a metric learning regularization to make the D-CNN models more discriminative.

Although these methods based on CNNs can improve the classification performance, only one CNN structure is used to extract the features, and the complementarity of the features for different structures of CNNs is seldom considered. Thus, we propose a remote sensing image scene classification method based on multi-structure deep features fusion (MSDFF). First, a data augmentation method based on random-scale cropping is employed to expand the limited data. Then, three popular CNNs are used as feature extractors to capture deep features from the image respectively. Finally, a deep feature fusion network is adopted to fuse these features and implement the classification. Experimental results on the UC Merced, AID, and NWPU-RESISC45 datasets demonstrate that the proposed MSDFF produces excellent performance.

The main contributions of this study are as follows: (1) three types of CNNs, CaffeNet, VGG-VD16, and GoogLeNet, are introduced to extract the complementary features; (2) an improved network including a concatenation layer and four fully connected layers with a softmax is proposed to integrate different features adaptively.

The rest of this paper is organized as follows. In Section II, the details of the proposed method are described. In Section III, the testing dataset, experimental results, and discussion are presented. In Section IV, the conclusions are summarized.

II. PROPOSED METHOD

In this section, the proposed remote sensing scene classification method based on MSDFF is described. The framework of the proposed method comprising data augmentation, feature extraction, and features fusion and classification is shown in Fig. 1.

A. DATA AUGMENTATION

As a supervised learning method, a neural network requires a large amount of data to train the model. Insufficient data will decrease the scene classification accuracy [32]. Therefore, in practice, data augmentation becomes the first step in deep model training. Effective data augmentation increases not only the number of training samples but also the diversity of training samples.

In remote sensing scene image classification, object scale variation is an important factor that affects classification performance. Here, a data augmentation method based on random-scale cropping is used to tackle the scale variation problem. In this paper, the image is represented as a tensor with dimensions $H \times W \times C$, where H, W, and C represent the height, width, and channel number of the image, respectively.

First, mean normalization is performed on the training image by subtracting the average of the pixel values of the training set for each pixel value, which aims to remove the common portion of the image and highlight the

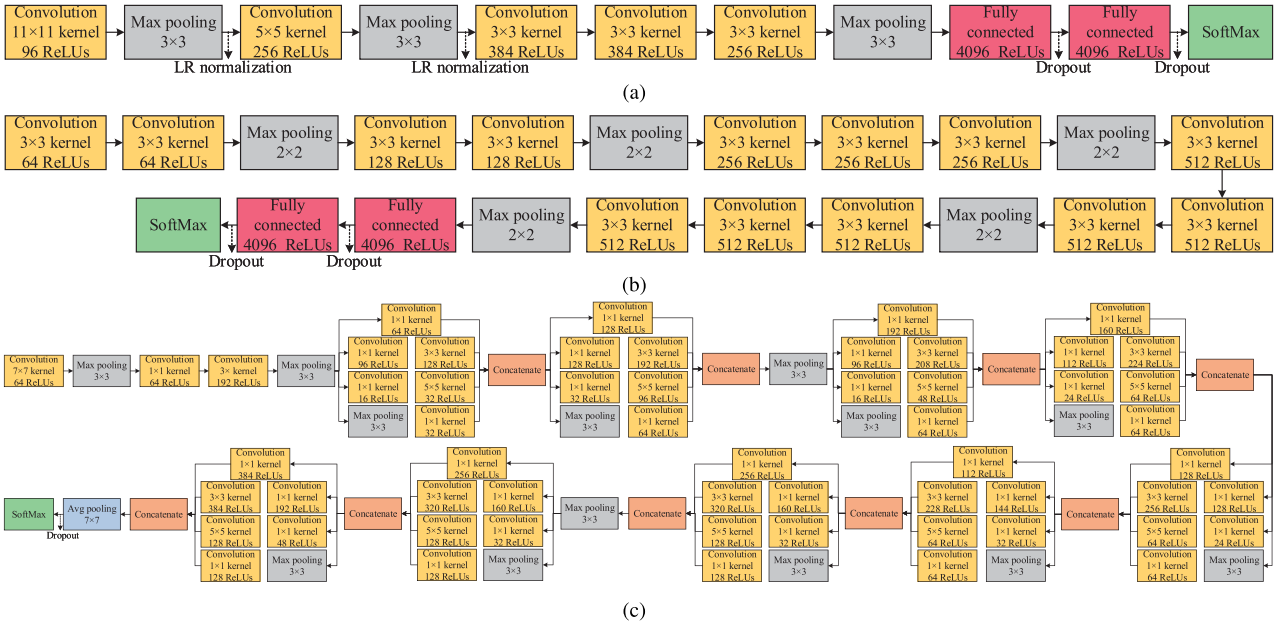

FIGURE 2. Frameworks of the three CNNs. (a) CaffeNet. (b) VGG-VD16. (c) GoogLeNet.

individual differences. The mean normalization is written as

$$I_{norm} = I - \frac{1}{N_{Train}} \sum_{i=1}^{H} \sum_{j=1}^{W} p\,(i, j). \tag{1}$$

where I_{norm} is the normalized image, N_{Train} is the number of training images, and $p\,(i, j)$ indicates the pixel value of a location of the image.

Second, patches with a random scale are cropped from the image and stretched to the specified scale as the input, which allows the CNNs to extract features that are robust to the scale variation. The ratio of the patch scale to the original image scale is defined as the crop ratio, which is given by

$$crop\ \ ratio = \frac{h \times w}{H \times W}. \tag{2}$$

where h and w are the height and width of the patches.

B. FEATURE EXTRACTION

The features extracted by CNNs are distinctive and representative, but most researchers have focused on features from the last convolution layer or the fusion features from different layers in one CNN, and the complementarity of features from different CNNs is overlooked. However, the fusion features from different CNNs can capture more potential information for improving classification performance. Here, three popular CNNs, CaffeNet, VGG-VD16, and GoogLeNet, are selected as feature extractors in the proposed architecture. On the one hand, compared with other CNNs, the three networks have relatively simple structures, and they can process images faster as feature extractors. On the other hand, the three networks have different depths and widths, which allow them to extract complementary features from the same image. The characteristics of each CNN model are described next, and the source of the features for each model is also specified.

1) CAFFENET

CaffeNet [33] uses 1000 different types of labeled images (a total of 1.2 million) provided by ImageNet Large-Scale Visual Recognition Challenge (ILSVRC) for parameter training. The framework of the model is shown in Fig. 2(a). CaffeNet consists of only five convolutional layers, three max-pooling layers and three fully connected layers (including a softmax), which enable the extraction of features of high depth from images. Here, CaffeNet is used as a feature extractor to extract features from the second fully connected layer, which generates a feature vector of 4096 dimensions.

2) VGG-VD16

VGG-Net [34] won the localization and classification tracks in ILSVRC-2014. Due to the use of very small convolution filters in all layers, the depth of the network can be increased easily by adding more convolutional layers. Here, the VGG-VD16 is used as the feature extractor. Fig. 2(b) shows the framework of VGG-VD16. VGG-VD16 contains 13 convolutional layers, five pooling layers, and three fully connected layers (including a softmax). Due to the similarity to CaffeNet for fully connected layers, the output of the second fully connected layer of VGG-VD16 is used as the feature vector of 4096 dimensions.

3) GOOGLENET

The framework of GoogLeNet [35] is shown in Fig. 2(c). Its main novelty is the inception modules, which use multiple filters with different sizes at the same layer to reduce dimension and improve the network representations. Here, GoogLeNet is used as a feature extractor by extracting features from the last pooling layer, which generates a feature vector of 1024 dimensions.

TABLE 1. Details of the three datasets.

Datasets	Resolution	Size	Classes	Total Number	Obtained	Geolocations
UC Merced	0.3m	256 × 256	21	2100	Aerial imagery	USA
AID	0.5-8m	600 × 600	30	10000	Google Earth	Around the world
NWPU-RESISC45	0.2-30m	256 × 256	45	31500	Google Earth	More than 100 countries

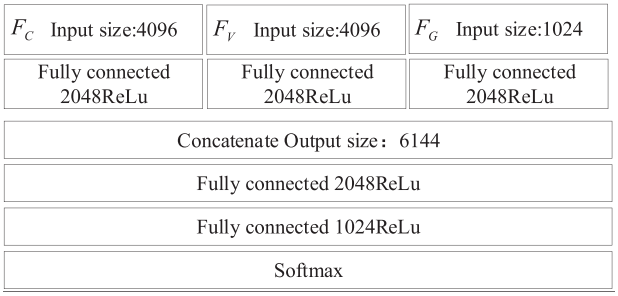

FIGURE 3. Framework of the proposed feature fusion network.

FIGURE 4. Scene classes in the UC Merced dataset. (1) agricultural. (2) airplane. (3) baseball diamond. (4) beach. (5) buildings. (6) chaparral. (7) high-density residential. (8) forest. (9) freeway. (10) golf course. (11) harbor. (12) intersection. (13) medium-density residential. (14) mobile home park. (15) overpass. (16) parking lot. (17) river. (18) runway. (19) low-density residential. (20) storage tanks. (21) tennis courts.

FIGURE 5. Scene classes in the AID dataset. (1) airport. (2) bare land. (3) baseball field. (4) beach. (5) bridge. (6) center. (7) church. (8) commercial. (9) high-density residential. (10) desert. (11) farmland. (12) forest. (13) industrial. (14) meadow. (15) medium-density residential. (16) mountain. (17) park. (18) parking. (19) playground. (20) pond. (21) port. (22) railway station. (23) resort. (24) river. (25) school. (26) low-density residential. (27) square. (28) stadium. (29) storage tanks. (30) viaduct.

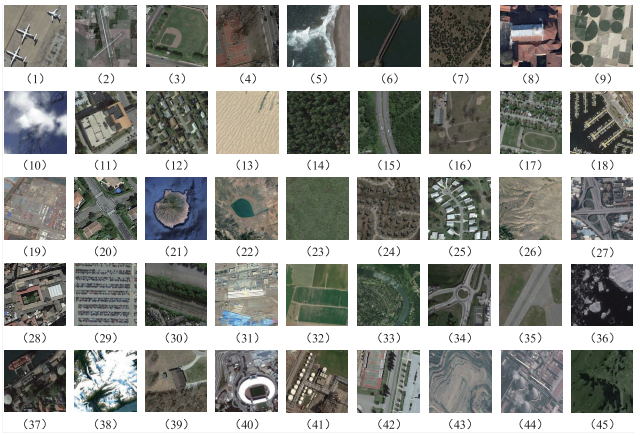

FIGURE 6. Scene classes in the NWPU-RESISC45 dataset. (1) airplane. (2) airport. (3) baseball diamond. (4) basketball court. (5) beach. (6) bridge. (7) chaparral. (8) church. (9) circular farmland. (10) cloud. (11) commercial area. (12) high-density residential. (13) desert. (14) forest. (15) freeway. (16) golf course. (17) ground track field. (18) harbor. (19) industrial area. (20) intersection. (21) island. (22) lake. (23) meadow. (24) medium-density residential. (25) mobile home park. (26) mountain. (27) overpass. (28) palace. (29) parking lot. (30) railway. (31) railway station. (32) rectangular farmland. (33) river. (34) roundabout. (35) runway. (36) sea ice. (37) ship. (38) snow berg. (39) low-density residential. (40) stadium. (41) storage tank. (42) tennis court. (43) terrace. (44) thermal power station. (45) wetland.

C. FEATURES FUSION AND CLASSIFICATION

Feature fusion is a robust and efficient strategy for image scene classification, and the fused features contain rich information that describes the image scene in detail. The aim of using feature fusion is to combine multiple relevant features into a single feature vector with more discriminative information than the original input feature vectors. Serial strategy and parallel strategy are two traditional strategies for feature fusion.

Serial feature fusion simply concatenates two features into one single feature where F_1 and F_2 are two features extracted from an input image with m, n vector dimension, respectively, and then the fused feature is F_f with the size of $(m + n)$.

Parallel feature fusion combines two features into a complex vector $F_f = F_1 + iF_2$ where i is the imaginary unit.

However, the existing feature fusion methods do not make full use of the original input feature and lack the flexibility to integrate multiple features. Here, a deep feature fusion network is proposed by integrating feature vectors from different CNNs. The framework of proposed deep feature fusion network is shown in Fig. 3. F_C, F_V, and F_G are normalized feature vectors extracted from CaffeNet, VGG-VD16, and GoogLeNet, respectively. The normalization is written as

$$X_{norm} = \frac{X - X_{\min}}{X_{\max} - X_{\min}}. \tag{3}$$

where X_{norm} is the normalized data, X is the original data, and X_{max} and X_{min} are the maximum and minimum of X.

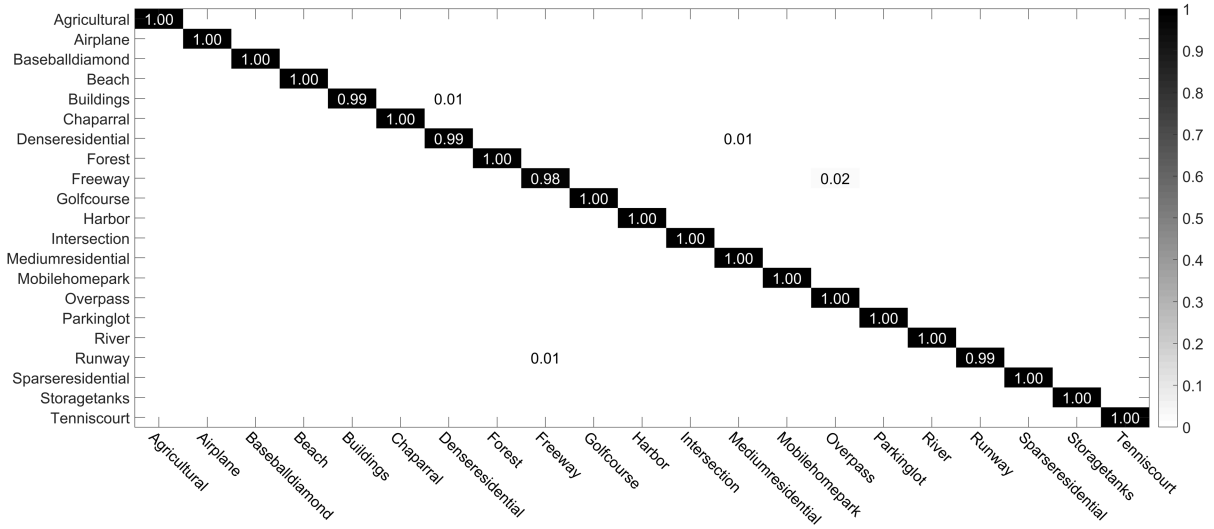

FIGURE 7. Confusion matrix for the proposed method on the UC Merced dataset for the proposed method with a training ratio of 80%.

TABLE 2. Classification results of various methods on the UC Merced dataset.

Methods	50% Training set	80% Training set
SPCK [12]	-	73.14
ResNet-101 [33]	84.38	-
GoogLeNet [5]	92.70	94.30
VGG-VD-16 [5]	94.14	95.20
Fine-tuned GoogLeNet [29]	-	97.10
Fine-tuned ResNet-50 [36]	-	97.97
Triplet networks [27]	-	97.99
Two-Stream Fusion [35]	96.97	98.02
Deep CNN Transfer [32]	-	98.49
PMS [36]	-	98.81
Proposed method	98.85	99.76

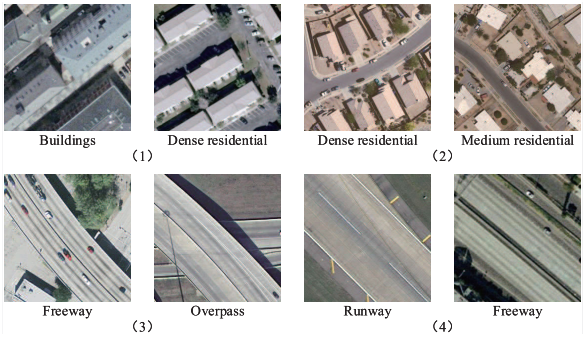

FIGURE 8. Confused scene classes in the UC Merced dataset. (1) buildings and high-density residential. (2) high-density residential and medium-density residential. (3) freeway and overpass. (4) runway and freeway.

The network consists of one concatenation layer and four fully connected layers with a softmax, which can integrate different features adaptively to improve classification performance. Moreover, dropout is introduced into each fully connected layer to speed up training and prevent overfitting. The strategy reduces the coadaptation of neurons by deactivating hidden neuron outputs with a probability of p during training. The probability of dropped neurons changes randomly at each epoch, which changes the architecture and reduces overfitting compared with training without dropout.

In the concatenation layer, the linear concatenation is used to fuse the output vector, which generates a vector of 6144 dimensions. The output vector of ith sample $f(i)$ can be defined as

$$f(i) = \bigcup_{n=1}^{3} f^n(i). \tag{4}$$

where $\cup$ indicates the concatenation operation, $f^n(i)$ is the nth feature vector, and n is the index of the model.

Inspired by the idea of shortcut connections in ResNet [36], [37], three outputs of the first fully connected layer are added to form the second fully connected layer, which allows feature reuse and enhances feature propagation. The shortcut connections mean the direct addition from front layer to the layer considered. The output of the lth layer in the building block of ResNet is defined as

$$y_l = F\left(F\left(y_f, \{W_i\}\right) + y_f\right). \tag{5}$$

where y_f and y_l are the input and output vectors of the building block, $F(\cdot)$ denotes activate function, W_i represents the connection weight, and the biases are omitted for simplifying notations.

The proposed deep feature fusion approach can be regarded as a collection of nonlinear high-dimensional projections due to the fully connected layers. Meanwhile, shortcut connections could realize feature reuse and strengthen feature propagation. Such a feature fusion approach helps to extract more representative features from the fusion of different deep features, and it can lead to enhanced classification performance.

III. RESULTS AND DISCUSSION

A. DATASET DESCRIPTION

The proposed method is evaluated on three publicly available land-use datasets: UC Merced [5], AID [5], and NWPU-RESISC45 datasets [17]. The details of each dataset are listed in Table 1.

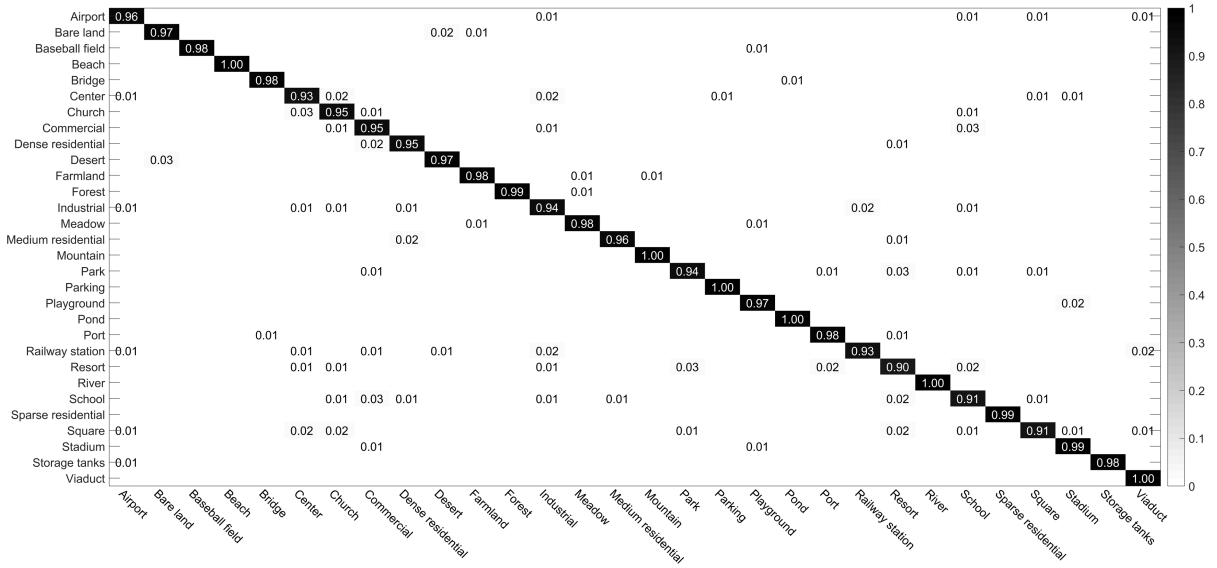

FIGURE 9. Confusion matrix for the proposed method on the AID dataset with a training ratio of 50%.

TABLE 3. Classification results of various methods on the AID dataset.

Methods	20% Training set	50% Training set
IFK(SIFT) [5]	70.60	77.33
GoogLeNet [5]	83.44	86.39
ResNet-101 [33]	-	88.88
VGG-VD-16 [5]	86.59	89.64
CaffeNet [5]	86.86	89.53
Two-Stream Fusion [35]	92.32	94.58
Fine-tuned ResNet-50 [36]	-	94.64
PMS [36]	-	95.56
Proposed method	93.47	96.74

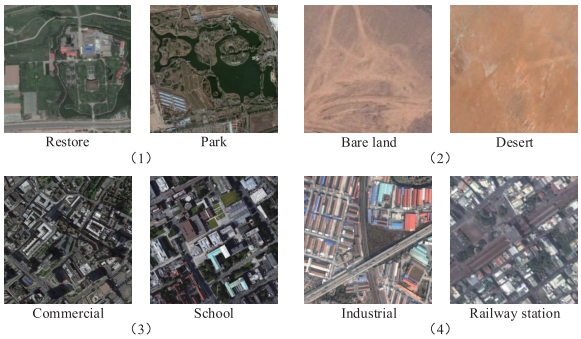

FIGURE 10. Confused scene classes in the AID dataset. (1) restore and park. (2) bare land and desert. (3) commercial and school. (4) industrial and railway station.

UC Merced dataset consists of 21 distinctive scene classes, as shown in Fig. 4. Each class contains 100 images with a size of 256 × 256 pixels and RGB channels.

AID dataset consists of 30 different land-use and land cover classes, as shown in Fig. 5. It contains 10000 images with a size of 600 × 600 pixels and RGB channels.

NWPU-RESISC45 dataset consists of 45 distinctive scene classes, as shown in Fig. 6. Each class contains 700 images with a size of 256 × 256 pixels and RGB channels.

B. EXPERIMENTAL SETUP

1) IMPLEMENTATION DETAILS

The Tensorflow framework is employed to implement the proposed method. For the data augmentation, the range of the crop ratio is set to [0.6, 0.8], and the number of patches cropped from each original image is set to 20. To evaluate the proposed method accurately, we adopt two commonly used training ratios for each dataset. For the UC Merced dataset, the ratios of the training sample are 50% and 80%. For the AID dataset, the ratios are 20% and 50%. For the NWPU-RESISC45 dataset, the ratios are 10% and 20%. To obtain reliable results on all three datasets, we repeat all results 10 times for each ratio to report the average classification accuracy. The stochastic gradient descent optimization algorithm is used to train the deep feature fusion network, and the learning rate of the network is set to 0.0001.

2) EVALUATION MEASURES

Two commonly used measures, overall accuracy and confusion matrix, are used to evaluate the classification performance quantitatively. The overall accuracy is defined as the number of correctly classified samples divided by the total number of samples. It is a direct measure to reveal the classification performance on the whole dataset and can be defined as

$$OA=\frac{1}{N}\sum_{i=1}^{n} C_{ii} \quad i=1,2,\cdots,n. \tag{6}$$

where C_{ii} is the number correctly classified samples for class i, n is the number of classes, and N is the number of the total samples.

The confusion matrix is a specific table layout that allows direct visualization of the performance in each class. It reports the errors and confusions among different classes by calculating the correct and incorrect classification of the test images for each class and accumulating the results in the table.

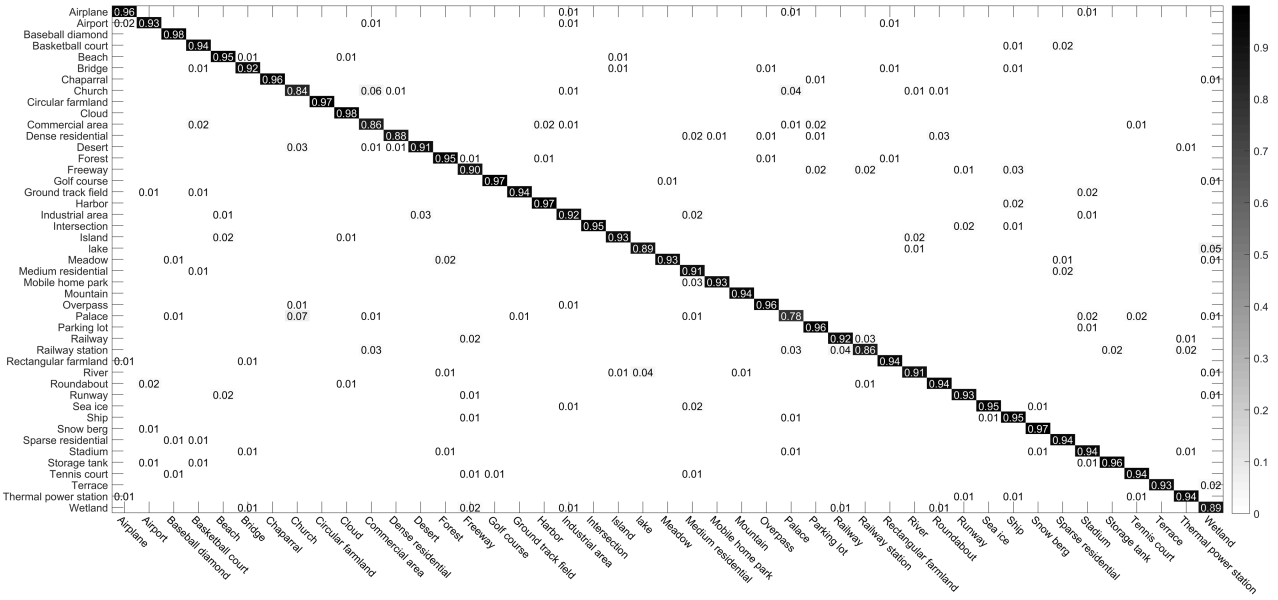

FIGURE 11. Confusion matrix for the proposed method on the NWPU-RESISC45 dataset with a training ratio of 20%.

TABLE 4. Classification results of various methods on the NWPU-RESISC45 dataset.

Methods	10% Training set	20% Training set
LLC [12]	38.83	40.03
Fine-tuned AlexNet [22]	81.22	85.16
Fine-tuned GoogLeNet [22]	82.57	86.02
ResNet-101 [33]	-	87.03
Fine-tuned VGGNet-VD16 [22]	87.15	90.36
D-CNN with VGGNet-16 [37]	89.22	91.89
Proposed method	91.56	93.55

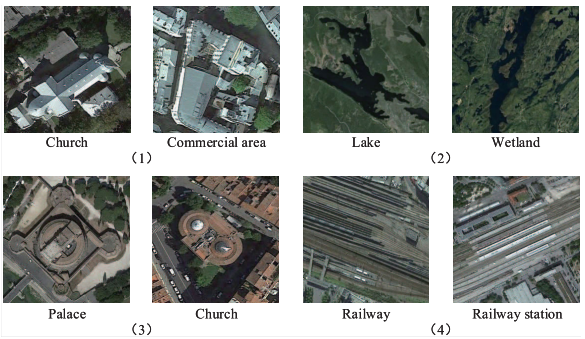

FIGURE 12. Confused scene classes in the NWPU-RESISC45 dataset. (1) church and commercial area. (2) lake and wetland. (3) palace and church. (4) railway station and railway station.

C. EXPERIMENTAL RESULTS

1) CLASSIFICATION OF THE UC MERCED DATASET

In this section, the proposed method is compared with the other nine methods on the UC Merced dataset. The classification results of various methods are shown in Table 2. As shown in Table 2, the proposed method, by fusing three deep features, obtains the highest overall accuracies of 98.85% and 99.76% for 50% and 80% training ratios, respectively.

Fig. 7 shows the confusion matrix for the proposed method on the UC Merced dataset with a training ratio of 80%. As shown in Fig. 7, in addition to buildings, high-density residential, freeway, and runway, other scene classes obtain a classification accuracy of 1.

The confusion occurs between buildings and high-density residential, between high-density residential and medium-density residential, between freeway and overpass, and between runway and freeway. As shown in Fig. 8, the confused scene classes contain similar structures.

2) CLASSIFICATION OF THE AID DATASET

The proposed method is compared with the other seven methods on the AID dataset. The classification results of various methods are shown in Table 3. The proposed method obtains the highest classification accuracies of 93.47% and 96.74% for 20% and 50% training ratios, respectively. The accuracy of the proposed method is 1.18% higher than that of PMS [30] with a 50% training ratio. The improvement of classification accuracy is mainly proved by the fusion of multi-structure deep features.

Fig. 9 shows the confusion matrix for the proposed method on the aid dataset with the training ratio of 50%. Most classes achieve a classification accuracy of over 95%, and some classes achieve the classification accuracy of 1. The classes with small inter-class distinction, such as high-, medium-, and low-density residential areas, could also be classified accurately. However, the confusion also occurs between some classes, as shown in Fig. 10.

3) CLASSIFICATION OF THE NWPU-RESISC45 DATASET

The proposed method is compared with the other five methods on the NWPU-RESISC45 dataset, and their classification results are shown in Table 4. The proposed method significantly improves the classification performance. For 10% and 20% training ratios, the overall accuracy of the proposed method is 2.34% and 1.66% higher than that of the suboptimal method, respectively.

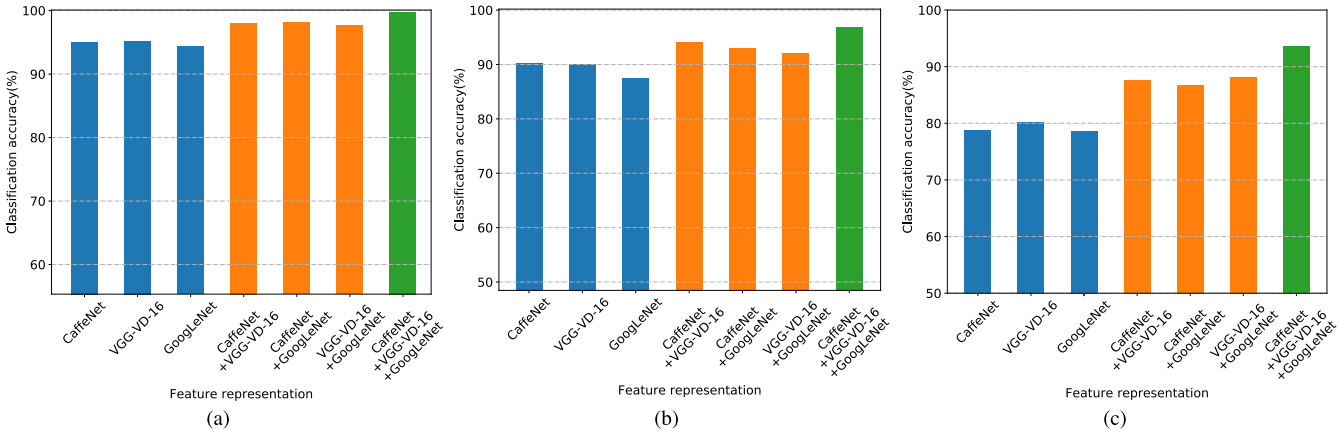

FIGURE 13. Classification accuracy of feature fusion solutions. (a) UC Merced dataset. (b) AID dataset. (c) NWPU-RESISC45 dataset.

TABLE 5. Average time (seconds) for raining and testing on the three datasets.

Datasets	Training	Testing
UC Merced	0.579	0.150
AID	0.727	0.181
NWPU-RESISC45	0.705	0.163

Fig. 11 shows the confusion matrix for the proposed method on the NWPU-RESISC45 dataset with a training ratio of 20%. Fourteen scene classes achieve a classification accuracy of over 95%. The most confused scene classes are church, commercial area, lake, wetland, palace, church, railway, and railway station, as shown in Fig. 12.

D. RUNNING PERFORMANCE

All the programs are performed on the Windows 10 operating system installed a GPU with a 6 GB graphic memory and a 3.20 GHz CPU with an 8 GB RAM. Table 5 lists the average training and testing time on the three datasets for the proposed method.

E. DISCUSSION

In this section, the influence of the number of fused features and crop ratio on the classification accuracy of the three datasets is analyzed. In the following experiments, samples with 80%, 50%, and 20% training ratios are randomly selected from each class of the UC Merced, AID, and NWPU-RESISC45 datasets, respectively.

First, the effect of the number of fused features on the classification accuracy for the three datasets is analyzed. Different numbers of features are fused by setting input interfaces of the deep feature fusion network. Fig. 13 shows the performance of these methods on the three datasets.

As shown in Fig. 13, the fused features can effectively improve the classification performance due to the complementarity among features extracted by different networks. Moreover, as the number of fused features increases, the classification accuracy increases for the relatively complex NWPU-RESISC45 dataset, indicating that the feature fusion solution is more effective for the complex dataset.

Second, the effect of the crop ratio used in data augmentation on the classification accuracy for the three

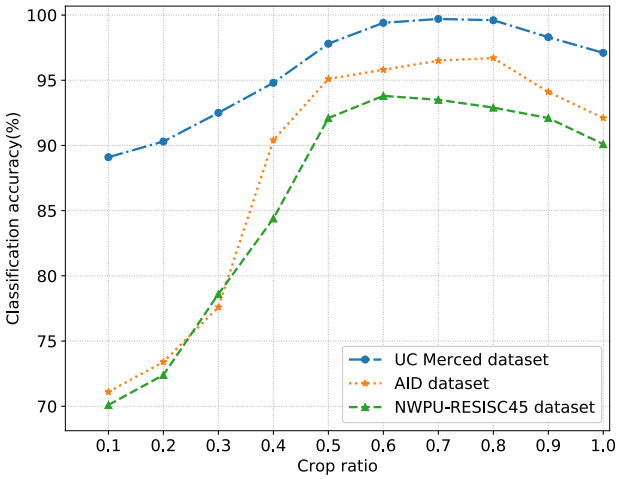

FIGURE 14. Classification accuracies with different crop ratios.

datasets is analyzed. Crop ratio is the ratio of the patch scale to the original image scale. Different crop ratios can generate different scales of patches from an image, and patches with different scales may contain different information. The classification accuracies with different crop ratios on the three datasets are shown in Fig. 14.

As shown in Fig. 14, the proposed method obtains the best classification performance for the three datasets when the crop ratio is between 0.6 and 0.8, which verifies that the crop ratio set in the previous experiment is optimal. When the crop ratio is too small, the classification accuracy is lower than that obtained directly from the original image (crop ratio = 1) because a small cropping area cannot contain a sufficient amount of distinguishable information.

IV. CONCLUSION

In this paper, a MSDFF framework has been proposed for remote sensing scene classification. In the proposed method, three types of CNNs, CaffeNet, VGG-VD16, and GoogLeNet, are used to extract the complementary features for the classification, and one fusion network including a concatenation layer and four fully connected layers with a softmax is used to combine different features adaptively.

Experiments were conducted on three public datasets with various land cover events. Our results demonstrate that the proposed method achieves higher overall accuracy than several existing methods. In addition, the effects of the number of fused features and crop ratio on the classification accuracy were studied. It turns out that the classification accuracy increases with the number of fused features increasing. The crop ratio in data augmentation also affects the classification performance greatly, and the optimal range of the crop ratio is between 0.6 and 0.8. Future work will investigate more efficient structures of feature fusion network to further improve the classification accuracy.

REFERENCES

[1] Q. Hu, W. Wu, T. Xia, Q. Yu, P. Yang, Z. Li, and Q. Song, "Exploring the use of Google earth imagery and object-based methods in land use/cover mapping," *Remote Sens.*, vol. 5, no. 11, pp. 6026–6042, Nov. 2013.
[2] C. Lu, X. Yang, Z. Wang, and Z. Li, "Using multi-level fusion of local features for land-use scene classification with high spatial resolution images in urban coastal zones," *Int. J. Appl. Earth Observ. Geoinf.*, vol. 70, pp. 1–12, Aug. 2018.
[3] W. Li, "Deep learning-based classification methods for remote sensing images in urban built-up areas," *IEEE Access*, vol. 7, pp. 36274–36284, Mar. 2019.
[4] L. Zhang, L. Zhang, D. Tao, and X. Huang, "On combining multiple features for hyperspectral remote sensing image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 50, no. 3, pp. 879–893, Mar. 2012.
[5] Y. Yang and S. Newsam, "Bag-of-visual-words and spatial extensions for land-use classification," in *Proc. ACM 18th SIGSPATIAL Int. Conf. Adv. Geogr. Inf. Syst.*, San Jose, CA, USA, 2010, pp. 270–279.
[6] Y. Yang and S. Newsam, "Spatial pyramid co-occurrence for image classification," in *Proc. Int. Conf. Comput. Vis.*, Barcelona, Spain, Nov. 2011, pp. 1465–1472.
[7] L.-J. Zhao, P. Tang, and L.-Z. 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.
[8] 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.
[9] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent Dirichlet allocation," *J. Mach. Learn. Res.*, vol. 3, pp. 993–1022, Mar. 2003.
[10] A. Bosch, A. Zisserman, and X. Muñoz, "Scene classification via pLSA," in *Proc. Eur. Conf. Comput. Vis.*, vol. 4, May 2006, pp. 517–530.
[11] B. Zhao, Y. Zhong, and L. Zhang, "Scene classification via latent Dirichlet allocation using a hybrid generative/discriminative strategy for high spatial resolution remote sensing imagery," *Remote Sens. Lett.*, vol. 4, no. 12, pp. 1204–1213, Dec. 2013.
[12] R. Kusumaningrum, H. Wei, R. Manurung, and A. Murni, "Integrated visual vocabulary in latent Dirichlet allocation–based scene classification for IKONOS image," *J. Appl. Remote Sens*, vol. 8, no. 1, Jan. 2014, Art. no. 083690.
[13] 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.
[14] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," *Commun. ACM*, vol. 60, no. 6, pp. 84–90, May 2017.
[15] J. Han, D. Zhang, G. Cheng, L. Guo, and J. Ren, "Object detection in optical remote sensing images based on weakly supervised learning and high-level feature learning," *IEEE Trans. Geosci. Remote Sens.*, vol. 53, no. 6, pp. 3325–3337, Jun. 2015.
[16] G.-S. Xia, J. Hu, F. Hu, B. Shi, X. Bai, Y. Zhong, L. Zhang, and X. Lu, "AID: A benchmark data set for performance evaluation of aerial scene classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 7, pp. 3965–3981, Jul. 2017.
[17] G. Cheng, J. Han, and X. Lu, "Remote sensing image scene classification: Benchmark and state of the art," *Proc. IEEE*, vol. 105, no. 10, pp. 1865–1883, Oct. 2017.
[18] N. Liu, L. Wan, Y. Zhang, T. Zhou, H. Huo, and T. Fang, "Exploiting convolutional neural networks with deeply local description for remote sensing image classification," *IEEE Access*, vol. 6, pp. 11215–11228, Jan. 2018.
[19] F. P. S. Luus, B. P. Salmon, F. Van Den Bergh, and B. T. J. Maharaj, "Multiview deep learning for land-use classification," *IEEE Geosci. Remote Sens. Lett.*, vol. 12, no. 12, pp. 2448–2452, Dec. 2015.
[20] 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.
[21] Q. Zou, L. Ni, T. Zhang, and Q. Wang, "Deep learning based feature selection for remote sensing scene classification," *IEEE Geosci. Remote Sens. Lett.*, vol. 12, no. 11, pp. 2321–2325, Nov. 2015.
[22] Y. Liu and C. Huang, "Scene classification via triplet networks," *IEEE J. Sel. Top. Appl. Earth Observat. Remote Sens.*, vol. 11, no. 1, pp. 220–237, Jan. 2018.
[23] 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.
[24] M. Castelluccio, G. Poggi, C. Sansone, and L. Verdoliva, "Land use classification in remote sensing images by convolutional neural networks," 2015, *arXiv:1508.00092*. [Online]. Available: http://arxiv.org/abs/1508.00092
[25] M. Oquab, L. Bottou, I. Laptev, and S. Josef, "Learning and transferring mid-level image representations using convolutional neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 1717–1724.
[26] J. Yosinski, J. Clune, Y. Bengio, and H. Lipson, "How transferable are features in deep neural networks," in *Proc. Adv. Neural Inf. Proc. Syst.*, 2014, pp. 3320–3328.
[27] K. Nogueira, O. A. Penatti, and J. A. Dos Santos, "Towards better exploiting convolutional neural networks for remote sensing scene classification," *Pattern Recognit.*, vol. 61, pp. 539–556, Jan. 2017.
[28] Y. Liu, Y. Liu, and L. Ding, "Scene classification based on two-stage deep feature fusion," *IEEE Geosci. Remote Sens. Lett.*, vol. 15, no. 2, pp. 183–186, Feb. 2018.
[29] Y. Yu and F. Liu, "A two-stream deep fusion framework for high-resolution aerial scene classification," *Comput. Intell. Neurosci.*, vol. 2018, pp. 1–13, Jan. 2018.
[30] L. Ye, L. Wang, Y. Sun, L. Zhao, and Y. Wei, "Parallel multi-stage features fusion of deep convolutional neural networks for aerial scene classification," *Remote Sens. Lett.*, vol. 9, no. 3, pp. 294–303, Mar. 2018.
[31] 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.
[32] Y. Zhou, X. Liu, J. Zhao, D. Ma, R. Yao, B. Liu, and Y. Zheng, "Remote sensing scene classification based on rotation-invariant feature learning and joint decision making," *EURASIP J. Image Video Process.*, vol. 2019, p. 3, Dec. 2019.
[33] Y. Jia, "Caffe: Convolutional architecture for fast feature embedding," in *Proc. 22nd ACM Int. Conf. Multimedia*, Nov. 2014, pp. 675–678.
[34] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*. [Online]. Available: http://arxiv.org/abs/1409.1556
[35] C. Szegedy, "Going deeper with convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 1–9.
[36] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Las Vegas, NE, USA, Jun. 2016, pp. 770–778.
[37] B. Li and Y. He, "An improved ResNet based on the adjustable shortcut connections," *IEEE Access*, vol. 6, pp. 18967–18974, 2018.

WEI XUE received the Ph.D. degree from the Shanghai Institute of Microsystem and Information Technology, Chinese Academy of Science, Shanghai, China, in 2008. He is currently an Associate Professor with the School of Automation, China University of Geosciences, Wuhan, China. His research interests include radar signal processing and remote sensing image classification.

XIANGYANG DAI is currently pursuing the M.S. degree with the School of Automation, China University of Geosciences, Wuhan, China. His research interests include deep learning and image processing.

LI LIU received the Ph.D. degree from the Huazhong University of Science and Technology, Wuhan, China, in 2016. He is currently an Associate Professor with the School of Automation, China University of Geosciences, Wuhan. His research interests include laser signal processing and deep learning.