Attention-Aware Deep Feature Embedding for Remote Sensing Image Scene Classification

Xiaoning Chen, Zonghao Han, Yong Li, Mingyang Ma, *Student Member, IEEE*,
Shaohui Mei, *Senior Member, IEEE*, and Wei Cheng, *Member, IEEE*

***Abstract*—Due to the wide application of remote sensing (RS) image scene classification, more and more scholars activate great attention to it. With the development of the convolutional neural network (CNN), the CNN-based methods of the RS image scene classification have made impressive progress. In the existing works, most of the architectures just considered the global information of the RS images. However, the global information contains a large number of redundant areas that diminish the classification performance and ignore the local information that reflects more fine spatial details of local objects. Furthermore, most CNN-based methods assign the same weights to each feature vector causing the mode to fail to discriminate the crucial features. In this article, a novel method by Two-branch Deep Feature Embedding (TDFE) with a dual attention-aware (DAA) module for RS image scene classification is proposed. In order to mine more complementary information, we extract global semantic-based features of high level and local object-based features of low level by the TDFE module. Then, to focus selectively on the key global-semantics feature maps as well as the key local regions, we propose a DAA module to attain those key information. We conduct extensive experiments to verify the superiority of our proposed method, and the experimental results obtained on two widely used RS scene classification benchmarks demonstrate the effectiveness of the proposed method.**

***Index Terms*—Attention mechanism, convolutional neural network (CNN), dual attention-aware (DAA), remote sensing (RS), scene classification.**

I. Introduction

WITH the development of satellite imaging technology, the number of remote sensing (RS) images increases rapidly, especially the acquisition of high-resolution RS images. Analyzing and understanding these RS images, such as identification or classification, bring new opportunities for more accurate surface monitoring and management and have received extensive attention [1], [2]. Especially, the RS image scene classification, which attempts to allocate a label to the RS image based on a variety of semantic categories, has been widely used in the urban planning [3], environment monitoring [4], [5], agriculture development [6], geography exploration [7], disaster monitoring [8], and military [9].

In recent decades, a number of methods have been proposed for RS image scene classification. In the early, low-level feature-based methods are adopted for RS image scene classification [10], [11], which concentrates on designing various human-engineering features, such as histogram of oriented gradient [12], local binary pattern [13], and scale-invariant feature transform [14]. Although these methods performed well on images with simple objects, they failed to classify the complex and challenging RS images. To improve the feature representation, mid-level feature-based methods emerged, which obtain a global feature representation by encoding the local descriptors. A popular mid-level method named Bag-of-Visual-Word (BoVW) utilizes visual word occurrences histogram to describe an image [15]. Due to the simplicity and effectiveness of the BoVW, it has been extensively used in the RS image scene classification [16], [17], [18]. In order to attain spatial information, spatial pyramid matching (SPM) divides the image into several subregions to code the spatial pyramid [19]. Although the mid-level features methods have made a great performance, they are not still satisfied with the increasingly complex and challenging RS images due to the insufficient representation capacity [3], [20], [21]. Both low-level feature-based methods and mid-level feature-based methods are handcrafted feature-based and largely rely on professional knowledge of image processing. Therefore, these methods show weak performance when processing complicated and challenging RS images.

The deep-learning-based methods have recently achieved excellent performance in many fields including object recognition [22], image classification, [23], [24], semantic segmentation [25], and other fields [26]. Subsequently, the deep-learning-based methods have been utilized in RS image scene classification [27], [28] and also achieved excellent performance due to the powerful feature representation learning ability of deep convolutional neural networks (CNNs) [29], [30], such as AlexNet [31], VGGNet [32], GoogLeNet [33], and ResNet [34]. Those architectures move the burden from hand-engineering knowledge to the framework of deep CNNs and become the most commonly used backbones for RS image scene classification. Then, by virtue of the strong feature extraction capability of CNNs, a series of RS image classification methods have emerged based on CNNs. Han et al. [35] propose a pretrained model

Manuscript received 12 September 2022; revised 15 October 2022 and 4 December 2022; accepted 9 December 2022. Date of publication 16 December 2022; date of current version 11 January 2023. This work was supported by the Fundamental Research Funds for the Central Universities under Grant 3102019ZX015. *(Corresponding author: Yong Li.)*

Xiaoning Chen, Zonghao Han, Yong Li, Shaohui Mei, and Wei Cheng are with the School of Electronics and Information, Northwestern Polytechnical University, Xi'an 710129, China (e-mail: chenxiaoning2018@mail.nwpu.edu.cn; hanzonghao@mail.nwpu.edu.cn; ruikel@nwpu.edu.cn; meish@nwpu.edu.cn; pupil_119@nwpu.edu.cn).

Mingyang Ma is with the School of Information and Communications Engineering, Xi'an Jiaotong University, Xi'an 710049, China (e-mail: mamingyang@mail.nwpu.edu.cn).

Digital Object Identifier 10.1109/JSTARS.2022.3229729

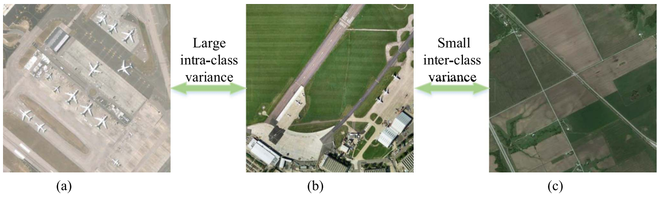

Fig. 1. (a)–(c) Categories of the airport, airport, and farmland from the AID dataset, respectively. The complex spatial distributions in RS scene images bring larger intraclass variance and smaller interclass variance. Panels (a) and (b) have a big difference in their scene but come from the same category of Airport, which shows that even if the same category may have a different structure of the scene. Panels (b) and (c) are from the different categories of Airport and Farmland, respectively, but they have very similar scene distribution, which reveals that even if the similar arrangement also may exist in different categories.

based on AlexNet to solve the nonconvergence caused by the insufficient training samples for RS scene classification. In order to attain more effective feature representation, Zeng et al. [36] propose an end-to-end CNN-based architecture by merging the local features and the global features. Liu et al. [37] propose a weighted SPM method to boost the performance of the RS image scene classification. In [38], a multisource compensation network is introduced to address the distribution inconsistency and category insufficiency. Although excellent results have been achieved in RS scene classification [39], [40], [41], there are still great challenges. Due to the special acquisition method, RS images present the characteristics of multiscale, multitarget, and complex structures [42]. Therefore, in order to build a distinguishing feature representation, Xu et al. [43] take advantage of the feature fusion strategy by multilayers for the RS scene classification. Tian et al. [44] propose multiscale dense networks to extract more effective features by automatically transforming between small networks.

However, the current CNN-based methods mainly tend to learn global semantic-level feature representation from raw images for the RS image scene classification. The high-level semantics features are helpful for explicitly understanding scenes via strong activations to semantics. Nevertheless, the global semantic-level feature ignores the spatial information between local objects, which is also crucial for classification [45], [46], [47]. To optimize the training model, Zhang et al. [48] take into account globe consistency and local particularity in the loss function. Therefore, only utilizing the global semantic-level features for scene classification to those images, which have significant intraclass variances and small interclass dissimilarity may lead to allocating wrong labels [49]. For example, Fig. 1(a) and (b) has an extremely diverse distribution of objects and would be assigned to two different categories, but in fact, they are the same category. Conversely, Fig. 1(b) and (c) has absolutely similar global scene semantic distribution but they are different categories actually. The low-level features reflect the fine details of the local objects and are able to capture the clear boundaries of small objects, which are beneficial to complement the loss of spatial information of the high-level features [29], [41]. In this case, the low-level local features are indispensable for RS image classification. Feature pyramid network (FPN) [50] is proposed for passing high-level semantic information to low-level local information. FPN is originally proposed to be applied to object detection, which usually involves assigning a specific category to a single object. However, RS image usually contains multiobjects and complex background. Therefore, people often need to assign a label by local objects as well as a global scene to the RS image. Inspired by FPN, we propose a Two-branches Deep Feature Embedding (TDFE) module, which contains two-level feature aggregation for global semantic-based features and local object-based features, respectively.

Moreover, the global information of the RS images contains a great number of redundant areas, which diminish the performance of classification. Similarly, the local information also has various local objects and some of them interfere with the results of classification. Therefore, how to selectively focus on the key parts of the image is crucial for scene classification. To deal with this problem, the attention mechanism [51], [52], [53] that suppresses irrelevant features and focuses on the important features has been widely applied in RS image scene classification. Wang et al. [54] introduce an attention recurrent convolutional network (ARCNet) according to the human visual system (HVS), which can highlight the crucial areas. In order to exploit the local semantic representation, Bi et al. [55] propose an APDC-Net that utilizes the spatial attention mechanism. Ding et al. [56] propose a local attention network (LANet) to improve the capability of feature representation with patch-level local attention. Considering the influence of the similarities between images and the spatial rotation, Tang et al. [30] proposed an attention consistent network (ACNet) to enhance the performance. In order to capture class-specific features, Li et al. [57] propose an augmentation operation, which can capture discriminative regions. Fan et al. [58] combine attention mechanisms and residual units to allocate large weight to the important areas and ignore redundant parts adaptively. Woo et al. [52] utilize channel attention and spatial attention modules along two separate dimensions to boost the capability of feature extraction and propose the convolutional block attention module (CBAM). The channel attention mechanism is capable of discovering the key global semantic information of each feature map along the channel axis while the spatial attention mechanism aims to focus on the important local objective information by the spatial axis. It is well known that the CBAM is composed of the concatenation of channel attention and spatial attention modules, which is generally embedded in each convolution block in a deep convolutional network to refine features. However, TDFF module contains two branches with different roles, which are the high-level global semantic-based features extraction branch and the low-level local object-based features extraction branch, respectively. In order to enhance the roles of these two branches, we need to design attention modules that can explore the discriminating high-level semantic features and the important low-level local object features, respectively. Therefore, according to the analysis, we propose a dual attention-aware (DAA) module to perform channel attention on the high-level global features and spatial attention on the low-level local features, so as to obtain key global features and local features, respectively.

Therefore, according to the above analysis, we proposed a promising method by introducing the TDFE module consisting

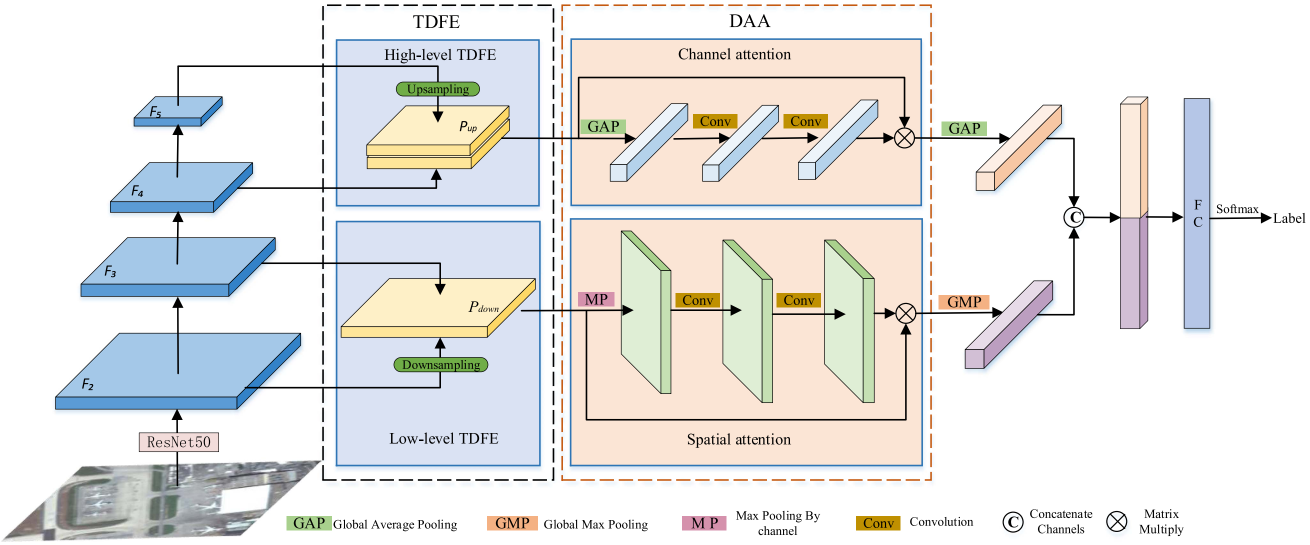

Fig. 2. Overall architecture of the proposed TDFE-DAA for the RS image scene classification.

of the high-level semantic-based feature embedding and the low-level object-based feature embedding, and the DAA module into the vanilla framework to obtain the crucial information of the RS images, so as to the performance of the classification.

Overall, our contributions are summarized as follows.

1) A two-branches deep feature embedding (TDFE) module is proposed, through which more robust global semantic-based features and local object-based features are both extracted, respectively.
2) A DAA module is proposed. With the DAA module, the key high-level semantic-based features are extracted from the channel dimension, and the key local object-based features are captured along the spatial dimensions.

We organize the rest of this article as follows. The details of the proposed method are introduced in Section II. In Section III, the experiments and analysis are presented. Section IV discusses the scalability of our method and the effectiveness of each module in our proposed method. Finally, we make conclusions in Section V.

II. Proposed Method

In this section, our proposed TDEF-DAA for RS image scene classification will be explained in detail. The overall architecture of our proposed method is shown in Fig. 2, which contains four modules. The first module is the backbone network of TDEF-DAA, which attempts to extract features from RS images. The second module is the TDFE, which consists of two branches: 1) the high-level branch aiming to integrate the global semantic information and 2) the low-level branch incorporating the local object spatial information. The next DAA module is composed of channel attention and spatial attention for obtaining the crucial representation of global features and local features, respectively. The last module concatenates the feature vectors from the DAA module in the channel dimension to get the final discriminative feature representation for classification.

A. Two-Branches Deep Feature Embedding

1) Motivation: Currently, CNN-based methods mainly tend to learn global semantic-level feature representation from raw images for RS image scene classification [36]. The global semantic information is the strong activations of the last layers of CNNs, which have global receptive fields [59]. Consequently, the high-level semantics features are helpful for explicitly understanding scenes via strong activations to semantics. However, the global semantic-level feature ignores the spatial information between local objects, which is also crucial for classification [41]. More seriously, only using the global semantic-level feature for classification may reduce the accuracy when the separability of interclass is small and the variance of intraclass is large. In order to make up for the global high-level semantic information, the low-level local object features that reflect more fine spatial details of local objects are utilized as complementation. Therefore, according to the above analysis, we propose a novel module named Two-branches Deep Feature Embedding (TDFE) to generate the discriminative representation of global semantic-based features and local object-based features simultaneously. The specific composition is shown in Fig. 3.

2) Two-Branches Deep Feature Embedding: Fig. 3 shows the overall structure of our proposed TDFE module. This module consists of two branches of feature fusion modules, which are the aggregation of the higher-level semantic features and the low-level local features, respectively. The detailed description of TDFE is as follows.

In our method, we take ResNet50, which has five hierarchies [25] as our basic backbone to extract features from RS images. We select four hierarchies, which are the output of the last convolutional layer of the last residual block in each stage from up to bottom as the feature candidate. Formally, we utilize [F2, F3, F4, F5] to denote the output feature maps of conv2_3, conv3_4, conv4_6, conv5_3 of ResNet50. Furthermore, inspired by FPN [50], we concatenated the top-two adjacent feature

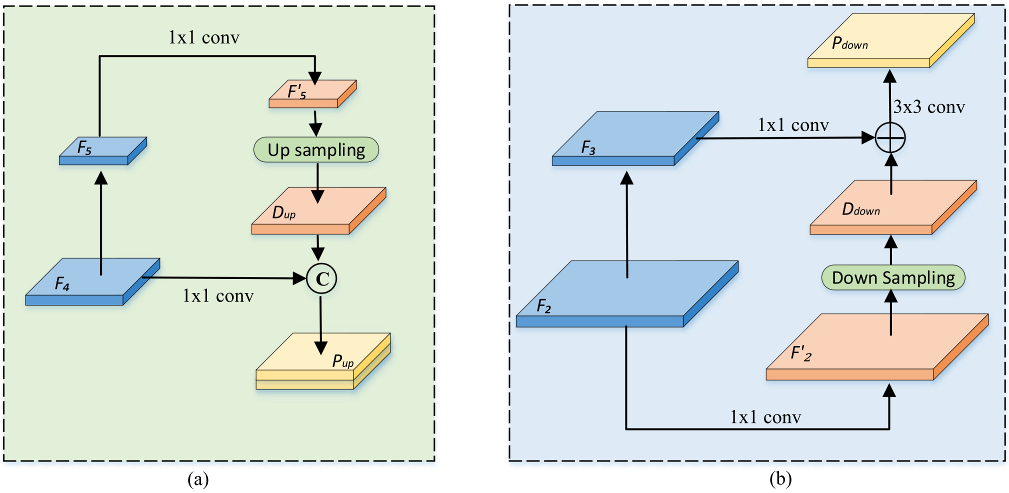

Fig. 3. Detailed description of TDFE. (a) High-level branch. (b) Low-level branch.

maps to attain the high-level semantic-based features, and add the down-two adjacent feature maps to obtain the low-level object-based features. We do not use the output of the block of conv1 due to its too small receptive field and too much memory footprint [50].

High-level Branch: The detailed structure of the high-level branch is shown in the Fig. 3(a). Given the ith feature $\mathbf{F}_i \in \mathbb{R}^{C_i \times H_i \times W_i}$, $i = 1, 2, 3, 4, 5$, where i denotes the ith stage of ResNet50, C_i represents channel dimension and H_i, W_i are spatial sizes. The higher-level feature maps $\mathbf{F}_4$ and $\mathbf{F}_5$ are adopted in this branch. First, we get the feature maps $\mathbf{F}'_i$ by the convolutional layer of 1×1 to each bottom-up feature maps to reduce the channel dimension, as follows:

$$\mathbf{F}'_i = \mathcal{F}_1\left(\mathbf{F}_i, \omega_1\right), i = 1, 2, 3, 4, 5 \tag{1}$$

where $\mathcal{F}_\mathbf{1}(\cdot, \omega_\mathbf{1})$ represents a 1×1 convolution with parameters ω_1. In this branch, a deconvolution process is adopted to upsample the feature map $\mathbf{F}'_5$ by a factor of 2 due to that the higher level feature map the coarser spatial information. And following the deconvolution, a batch normalization (BN) and a rectified linear unit (ReLU) are used. The process of the deconvolution is as follows:

$$\mathbf{D}_{\text{up}} = \mathcal{D}\left(\mathbf{F}'_5, \psi\right) \tag{2}$$

where $\mathcal{D}(\mathbf{F}'_5, \psi)$ denotes a deconvolutional layer with a kernel size of 3×3 and parameters ψ. Therefore, a feature map named $\mathbf{D}_{\text{up}}$ with the same spatial resolution as the $\mathbf{F}'_4$ is obtained. After that, we concatenate the channels of $\mathbf{D}_{\text{up}}$ and $\mathbf{F}'_4$ to obtain more global-semantic information. Finally, the final feature maps of a high-level branch named $\mathbf{P}_{\text{up}}$ are generated that the channel dimension is the sum of $\mathbf{D}_{\text{up}}$ and $\mathbf{F}'_4$

$$\mathbf{P}_{\text{up}} = \mathcal{F}_1\left(\mathbf{F}_4, \omega_1\right) \textcircled{c} \mathbf{D}_{\text{up}} \tag{3}$$

where denotes the concatenation operation along channelwise.

Low-level Branch: The low-level feature map reflects the fine details of the local objects and is able to capture the clear boundaries of small objects. Therefore, this branch merges the two lower-level features to aggregate complementary information of adjacent features to get more discriminative feature maps. The detailed structure of the low-level branch is shown in Fig. 3(b). The lower-level feature maps $\mathbf{F}_2$ and $\mathbf{F}_3$ are adopted in this branch. Different from the high-level branch, the low-level branch uses the downsampling to generate the same spatial dimension features as the adjacent higher-level features. Equation (1) is used to get the feature map $\mathbf{F}'_2$, and then, the convolution function with a kernel size of 3×3 and factor 2 of stride is applied to downsample the feature $\mathbf{F}'_2$ to obtain the feature map $\mathbf{D}_{\text{down}}$. The process of the downsampling is as follows:

$$\mathbf{D}_{\text{down}} = \mathcal{F}_2\left(\mathcal{F}_1\left(\mathbf{F}_2, \omega_1\right), \omega_2\right) \tag{4}$$

where $\mathcal{F}_2(\cdot, \omega_2)$ denotes a convolution of 3×3 kernel and 2 stride with parameters ω_2. Unlike the high-level branch that aggregates features by concatenation, in order to capture more spatial information of low-level features, the outputs of the downsample are aggregated by elementwise addition. Besides, a 3×3 convolution is applied to alleviate the aliasing effect. The process of the final output of the low-level branch is represented by $\mathbf{P}_{\text{down}}$. The formulation is as follows:

$$\mathbf{P}_{\text{down}} = \mathcal{F}_3\left(\left(\mathcal{F}_1\left(\mathbf{F}_3, \omega_1\right) \oplus \mathbf{D}_{\text{down}}\right), \omega_3\right) \tag{5}$$

where $\oplus$ denotes the addition operation of elementwise and $\mathcal{F}_\mathbf{3}(\cdot, \omega_\mathbf{3})$ represents a 3×3 convolution with parameters ω_3.

B. DAA Module

1) Motivation: Intuitively, global semantic information contains much of redundant areas that may mitigate the performance of feature representation. In the same way, the local semantic information has various dispensable objects, which may reduce the results of classification. Therefore, it is crucial to extract key features and remove redundant information for classification. The attention mechanism aims to restrain the irrelevant features and focus on the important features. Especially, the channel attention mechanism can concentrate on the key global semantic information of each feature map along the channel

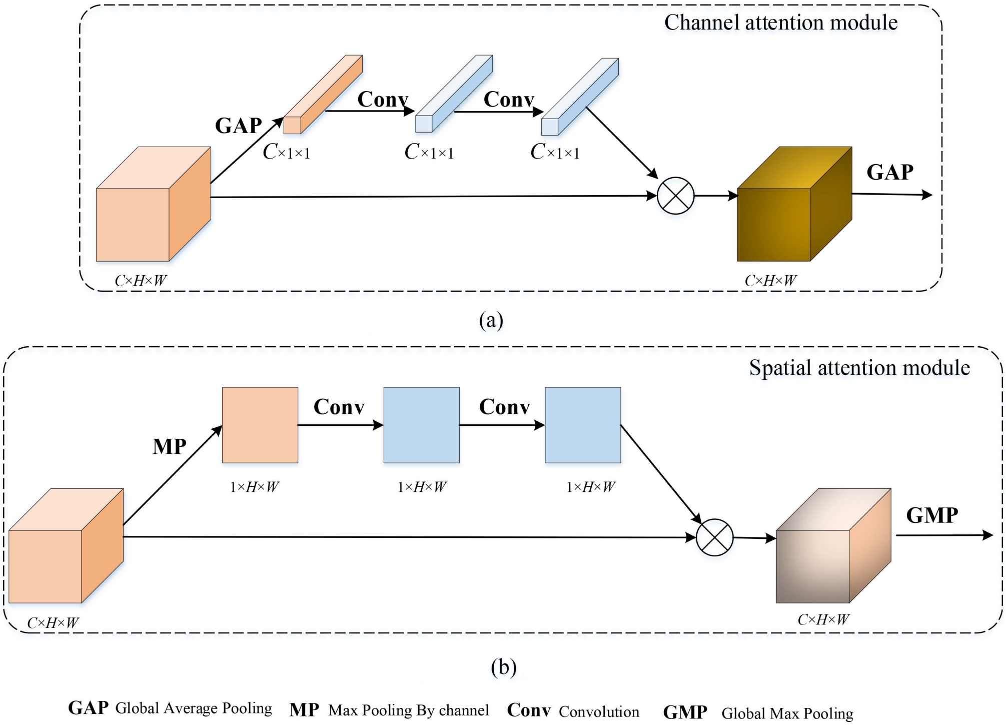

Fig. 4. Detailed description of DAA. (a) Channel attention module. (b) Spatial attention module.

axis. The spatial attention mechanism can focus on the crucial local objective information of each feature map by the spatial axis and boost the attention network to the key local objective and the key small object of the image. Therefore, we propose a DAA module, which introduces the channel attention module to the higher-lever features to capture the key global semantics and the spatial attention module to the lower-level features to focus on the important spatial object information.

2) DAA Module: Fig. 4 shows the detailed structure of the DAA module. These attention modules consist of a channel attention module and spatial attention module to boost the capability of feature extraction for higher-level semantic features and lower-level local features, respectively. The detailed description of the DAA module is as follows.

Channel attention module: In order to capture the important semantic information, we embed the channel attention module to the higher-level feature maps to map the interchannel relationships so as to obtain the more discriminative features. The channel attention module receives the higher-level feature maps $\mathbf{P}_{\text{up}}$, which are generated by TDFE. As shown in Fig. 4(a), this module first executes global average pooling (GAP) by the spatial dimension to extract the global semantic features $\mathbf{P}_{\text{avg}} \in \mathbb{R}^{C\times 1\times 1}$. Then, two 1×1 convolution operations are employed to generate the channel-attention weight map. The formulation is as follows:

$$\mathbf{p}_{\text{avg}}^{l} = \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} f_l(i,j), l = 1, 2, \ldots, C \tag{6}$$

$$\mathbf{A}_{\text{cha}} = \mathcal{F}_5\left(\mathcal{F}_4\left(\mathbf{P}_{\text{avg}}, \omega_4\right), \omega_5\right) \tag{7}$$

in (6), $f_l(i,j)$ express the elements at position (i,j) in the lth channel, $\mathbf{p}_{\text{avg}}^{l}$ is the lth element in $\mathbf{P}_{\text{avg}}$, and C is the total number of feature channels.

In (7), $\mathcal{F}_4(\cdot, \omega_4)$ and $\mathcal{F}_5(\cdot, \omega_5)$ denote 1×1 convolution with parameters ω_4 and ω_5, respectively. Following each convolutional layer, a BN layer and a ReLU layer are applied. Table I shows the details of those two convolutions. After attaining the weights of channels, we weight the original feature map $\mathbf{P}_{\text{up}}$ by those weights to highlight the significant channels and diminish the insignificant channels, as follows:

$$\mathbf{P}_{\text{up}}' = \mathbf{P}_{\text{up}} \otimes \mathbf{A}_{\text{cha}} \tag{8}$$

where "$\otimes$" represents the multiplication operation of elementwise executed by expanding along the spatial dimension.

Spatial attention module: Similar to the channel attention module, the spatial attention module is embedded in the low-level feature maps to focus on the more important parts and restrain irrelevant parts of an image. The spatial attention module receives the lower-level feature maps $\mathbf{P}_{\text{down}}$, which are generated by TDFE. Fig. 4(b) shows the architecture of spatial attention. First, we execute max-pooling (MP) operation along the channel dimension to highlight the informative regions [60] to generate a more efficient feature descriptor $\mathbf{P}_{\max} \in \mathbb{R}^{1\times H\times W}$. Subsequently, two 3×3 convolutional operations are applied to obtain a spatial-attention weight map $\mathbf{A}_{\text{spa}} \in \mathbb{R}^{1\times H\times W}$. For each convolutional layer, a BN layer and a ReLU layer are applied, which can be formulated as

$$\mathbf{A}_{\text{spa}} = \mathcal{F}_7\left(\mathcal{F}_6\left(\mathbf{P}_{\max}, \omega_6\right), \omega_7\right) \tag{9}$$

$\mathcal{F}_{\mathbf{6}}(\cdot, \omega_{\mathbf{6}})$ and $\mathcal{F}_{\mathbf{7}}(\cdot, \omega_{\mathbf{7}})$ denote 3×3 convolution with parameters ω_6 and ω_7, respectively. Table I shows the details of those

TABLE I
DETAILS OF THE ATTENTION BLOCKS IN THE DAA MODULE (BACKBONE IS RESNET50)

DAA module	layer	Kernel Size	Padding	Input Size	Output Size
Channel attention	GAP	14×14	0	1024×14×14	1024×1×1
	Conv1	1×1	0	1024×1×1	1024×1×1
	Conv2	1×1	0	1024×1×1	1024×1×1
Spatial attention	MP	1×1	0	256×28×28	1×28×28
	Conv1	3×3	1	1×28×28	1×28×28
	Conv2	3×3	1	1×28×28	1×28×28

two convolutions. Performing the same operation as channel attention, we weight the original feature map $\mathbf{P}_{\text{down}}$ to highlight the informative regions and weaken the insignificant regions, as follows:

$$\mathbf{P}'_{\text{down}} = \mathbf{P}_{\text{down}} \otimes \mathbf{A}_{\text{spa}} \tag{10}$$

where "$\otimes$" represents the multiplication operation of element-wise executed by expanding along the channel dimension.

C. Scene Classification

First, GAP [61] is utilized for the feature map of the output of the channel attention module by (6) to obtain the feature map $\mathbf{T}_{\text{avg}}$, which can strengthen the correspondence between scene semantic information and the categories. Simultaneously, the global max pooling (GMP) is applied to the output of the spatial attention module, which can reinforce the ability to discriminate the salience regions and generate the feature map $\mathbf{T}_{\max}$. And the expression of GMP is as follows:

$$\mathbf{T}_{\max} = \left[\mathbf{t}^{1}_{\max}, \mathbf{t}^{2}_{\max}, \ldots, \mathbf{t}^{C}_{\max}\right] \tag{11}$$

$$\mathbf{t}^{l}_{\max} = \max(f^{l}) \tag{12}$$

where $\mathbf{t}^{l}_{\max}$ is the lth element in $\mathbf{T}_{\max}$, C is the total number of feature channels, and f^{l} express the feature map of the lth channel.

Following, we concatenate the channels of $\mathbf{T}_{\text{avg}}$ and $\mathbf{T}_{\max}$ to further enhance the feature representation. After attaining the comprehensive feature vectors including the enhanced channel information and the strengthened spatial information, a fully connected (FC) layer and a Softmax layer are applied to predict the category label of the input image.

Given $z_i, (i = 1, 2, \ldots, C)$ is the output of the FC layer, where C is the total number of the category labels. The formulation of the Softmax function is as follows:

$$\alpha_i = \frac{\exp(z_i)}{\sum_{j}^{C} \exp(z_j)} \tag{13}$$

$$\text{label} = \arg\max_{i}(\alpha_i) \tag{14}$$

where α_i stands for the probability of the input image belonging to the ith category. And by (14), the final label is determined.

To optimize our proposed model, the cross-entropy loss function is applied for classification [62]. The cross-entropy loss function is given as follows:

$$\text{Loss} = -\frac{1}{N}\sum_{n=1}^{N}\sum_{c=1}^{C} y^{n}_{c} \log(\hat{y}^{n}_{c}) \tag{15}$$

where y denotes the real scene label, $\hat{y}$ denotes the predicted scene label, N is the number of samples in a minibatch, and C is the number of categories.

III. EXPERIMENTS

A. Dataset

To demonstrate the effectiveness of our proposed TDFE-DAA, two publicly available RS image benchmarks are employed in experiments. One is the well-known UC Merced Land-Use dataset (UCM) [6], and another one is the aerial image dataset (referred to as AID) [63].

UCM: The UCM dataset contains 2100 RS scene images of 21 classes in total collected by the United States Geological Survey (USGS) National Map. Each class consists of 100 images, and the resolution is 256 × 256 with a spatial resolution of 30 cm per pixel. A detailed description of the UCM dataset is given in Table II, and Fig. 5 gives one example image of each category in the UCM dataset.

AID: The AID dataset consists of 30 classes including 10 000 RS scene images in total, which are collected by Wuhan University from the Google Earth platform. Each class contains 220 to 420 images, and the spatial resolution varies from 1 to 8 m with the image resolution fixed to 600 × 600. A detailed description of the AID dataset is given in Table II, and Fig. 6 gives some example images including all categories in the AID dataset.

B. Experimental Setup

Data setting: In the experiment, we randomly select the training ratios of 80% and 50% for the UCM dataset, 50% and 20% for the AID dataset, and then the remaining are for testing. The input images are all resized to 224 × 224.

Implementation Details: All experiments are done based on the Pytorch library with an NVIDIA RTX 3080 GPU for acceleration. The model of the backbone is pretrained on ImageNet. The Adam algorithm [64] is employed to optimize the model weights, the initial learning rate is set to 1e-5, and the weight decay penalty is 0.1 for 20 epochs.

TABLE II
DESCRIPTION OF EACH DATASET

Dataset	Size	Spatial resolution	Number of classes	Total number
UCM	256 × 256	30cm	21	2100
AID	600 × 600	1m–8m	30	10000

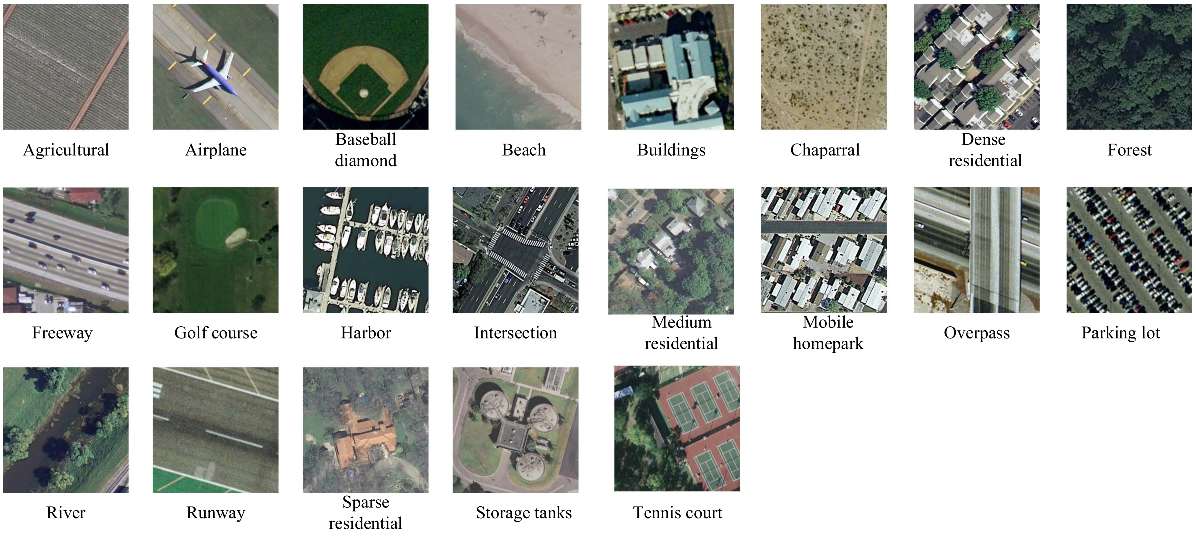

Fig. 5. Example images of the UCM dataset.

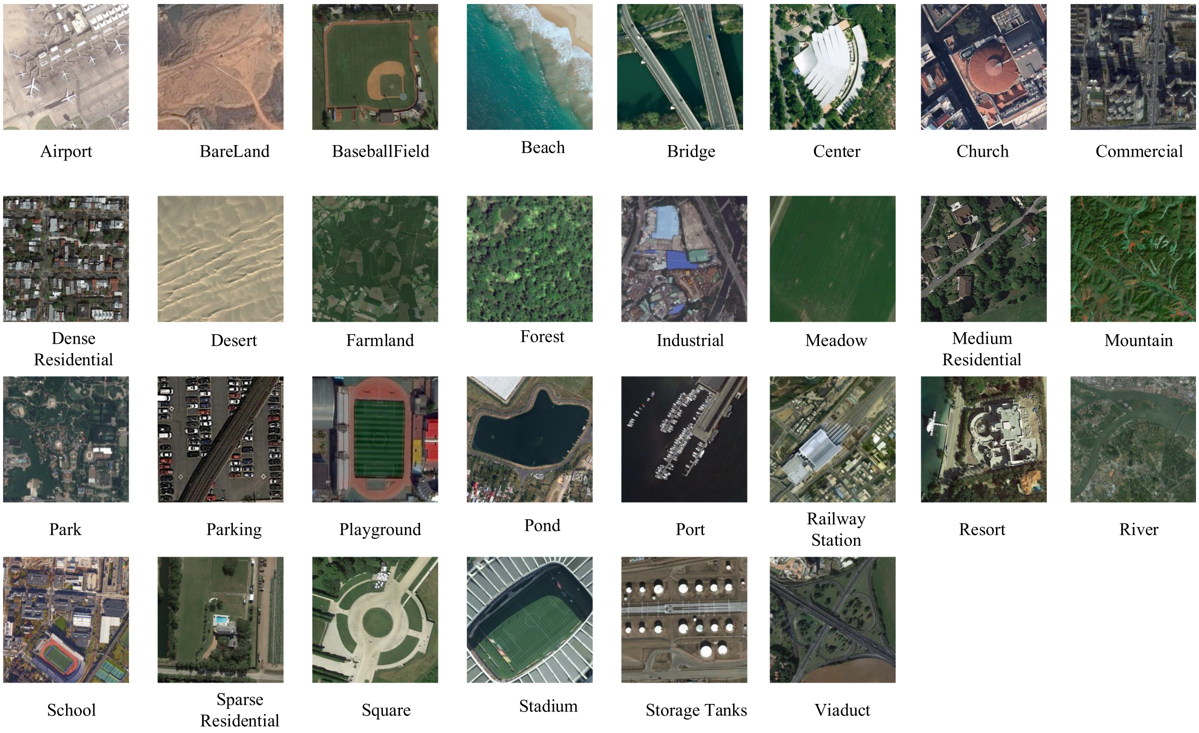

Fig. 6. Example images of the AID dataset.

TABLE III
COMPARISON OF THE OA (%) ON THE UCM DATASET

Methods	Overall Accuracy (%)	
	80%	**50%**
BoVW [6]	74.12±3.30	71.90±0.79
GoogLeNet [6]	94.31±0.89	92.70±0.60
CaffeNet [6]	95.02±0.81	93.98±0.67
VGG-VD-16 [63]	95.21±1.20	94.14±0.60
TEXNet [65]	96.79±0.49	95.89±0.37
VGG-VD-16-CapsNet [66]	98.81±0.22	95.33±0.18
VGG-VD-16-SAFF [47]	97.02±0.78	/
VGG-VD-16-DCF [67]	97.10±0.85	95.42±0.71
ResNet-LGFFE [68]	98.62±0.88	/
EFPN-DSE [29]	98.12±0.57	96.65±0.36
TDFE-DAA(Ours)	**99.05±0.72**	**97.32±0.56**

The boldface values represents the best value in the experiment.

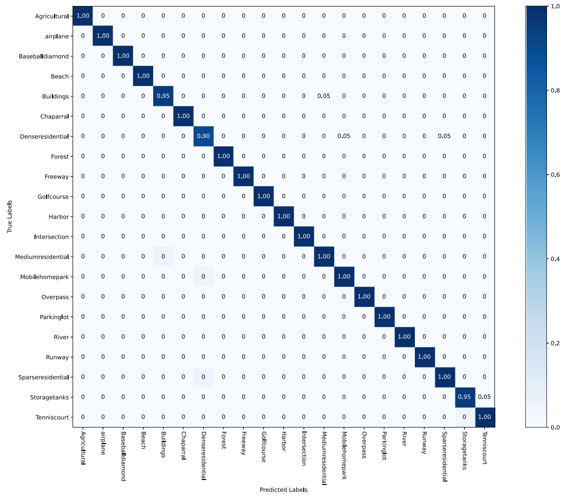

Fig. 7. Confusion matrix of TDFE-ADD under the 80% training proportion on the UCM dataset.

Evaluation Metrics: To quantitatively estimate the performance of our proposed method, overall accuracy (OA) and confusion matrix is selected. The OA is the ratio of the correct predictions with the overall predictions. Furthermore, to analyze the detailed accuracy of each category, a confusion matrix is used, which denotes the probability that the category of each row predicts the categories in the corresponding column.

C. Comparison With State of the Arts

To fully verify the advance of our proposed method, we compared it with some state of the arts, including BoVW [6], GoogleNet [6], CaffeNet [6], VGG-VD-16 [63], TEXNet [65], VGG16-CapsNet [66], VGG-VD-16-SAFF [47], VGG-VD-16-DCF [67], ResNet-LGFFE [68], and EFPN-DSE-TDFF [29]. For those methods whose codes can be obtained, we trained and tested the models with the default settings, and for those models of which the corresponding code cannot be got, we used the original results in their works. In order to have a fair comparison, we use the same ratios in the same dataset with data augmentation operation of random horizontal flipping for all models and repeat ten times for the experimental results with the corresponding means and standard deviations of the OA.

1) Experimental Results on UCM Dataset: We selected 80% and 50% images randomly as the training set, respectively, and the rest of the images are categorized as the testing set. We compared some state of the arts with our proposed method on the UCM dataset. Table III shows the experimental results. As can be found in Table III, the BoVW got the lowest results of OA, which also tells the truth that the deep-level-based approaches have better performance than the method based on mid-level features. We can also find from Table III that our proposed method has got the best performance. Compared with other methods, the OA values attained by our proposed TDFE-DAA are increased by 3.84% (VGG-VD-16), 2.03% (VGG-VD-16-SAFF), 0.93% (EFPN-DSE), respectively, on the UCM dataset with the ratio of 80%. And with the ratio of 50%, the OA values of our method are 3.18% higher than VGG-VD-16, 1.90% higher than VGG-VD-16-DCF, and 0.67% higher than EFPN-DSE, respectively. And there are twofold reasons for achieving the best performance of our proposed method. First, the Two-branch Deep Feature Embedding (TDFE) architecture, not only captures the global-semantic features but also takes into account the local-object features. Second, with the DAA mechanisms, the key channels and the important regions in RS images can be fully explored.

Furthermore, we conduct the confusion matrix to verify the advantage of our proposed method. As can be seen from Fig. 7, under the 80% training ratio, we can find that most results are equal to 1 except for the "Buildings," "Dense Residential," and "Storage Tank." The "Dense Residential," get the worst result of classification, with 5% being mistakenly classified as the "Mobile Home Park" and 5% into the "Sparse Residential." As is known, those categories have seriously similar objects in their scenes. In addition, 5% of images from the category of "Buildings" are misclassified into the "Medium Residential" while 5% images from "Storage Tank" are mistakenly classified as "Tennis Count." As is known, those categories have common properties including diverse objects, complex distribution, and similar scenes, which makes scene classification difficult.

Fig. 8 presents the results of the confusion matrix under the 50% training ratio, and we can find that 11 of 21 categories attain 1, and 19 categories reach results of more than 90%. Only two categories are less than 90%, including "Dense Residential" (88%) and "Sparse Residential" (88%). As we know, those two categories contain similar objects, such as "building," "tree," and "road," which causes the difficulty to distinguish.

2) Experimental Results on AID Dataset: To further verify the advancement of our proposed method, we compared various state-of-the-art methods with ours on another broadly used RS scene classification dataset named AID dataset. We selected 50% and 20% images randomly as the training set, respectively, and the rest of the images are categorized as the test set. The comparison results are shown in Table IV. We can find the same

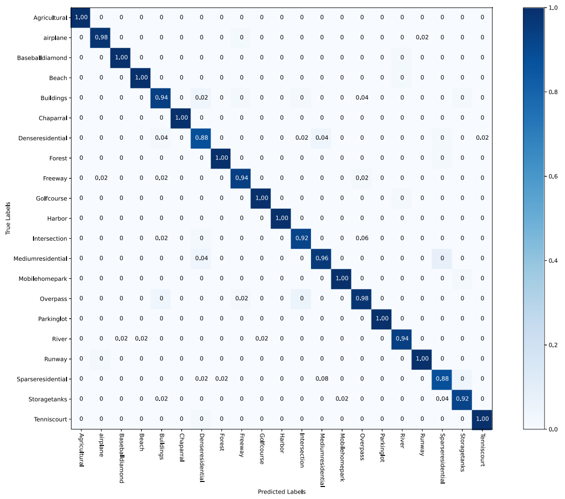

Fig. 8. Confusion matrix of TDFE-ADD under the 50% training proportion on the UCM dataset.

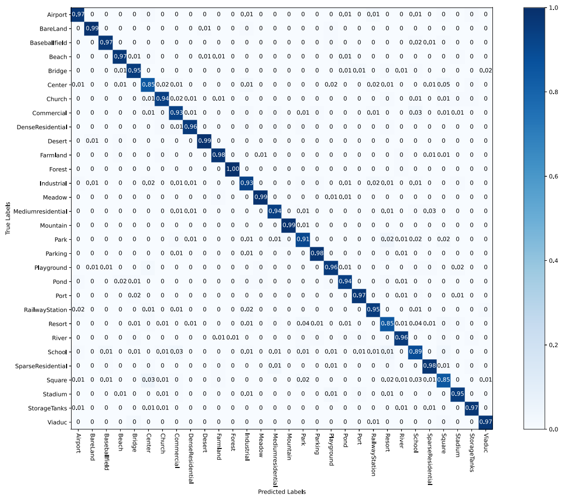

Fig. 9. Confusion matrix of TDFE-ADD under the 50% training proportion on the AID dataset.

TABLE IV
COMPARISON OF THE OA (%) ON THE AID DATASET

Methods	Overall Accuracy (%)	
	50%	**20%**
BoVW [6]	67.65±0.49	61.40±0.41
GoogLeNet [6]	86.39±0.55	83.44±0.40
CaffeNet [6]	89.53±0.31	86.86±0.47
VGG-VD-16 [63]	89.64±0.36	86.59±0.29
TEXNet [65]	92.96±0.18	90.87±0.11
VGG-VD-16-CapsNet [66]	94.74±0.17	91.63±0.19
VGG-VD-16-SAFF [47]	93.83±0.28	90.25±0.29
ResNet-LGFFE [68]	94.46±0.48	90.83±0.55
EFPN-DSE [29]	93.35±0.09	89.28±0.05
TDFE-DAA(Ours)	**95.36±0.15**	**92.15±0.75**

The boldface values represents the best value in the experiment.

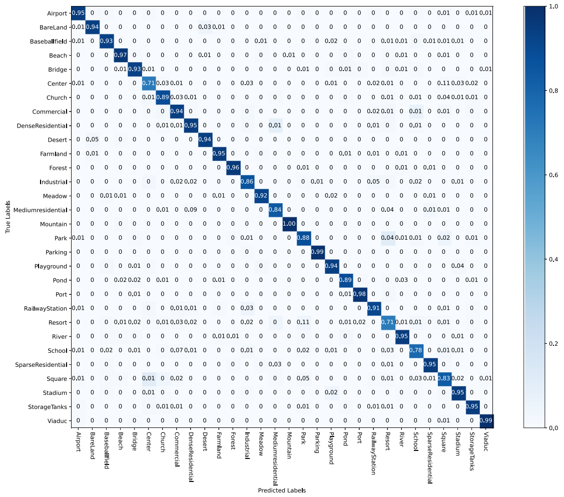

Fig. 10. Confusion matrix of TDFE-ADD under the 20% training proportion on the AID dataset.

conclusion as the UCM dataset that the deep-level-based approaches have better performance than the method based on mid-level features. In Table IV, we can also find that our proposed method has got the best performance. Compared with other methods, the OA values obtained by our proposed TDFE-DAA is increased by 5.72% (VGG-VD-16), 1.53% (VGG-VD-16-SAFF), and 1.83% (EFPN-DSE), respectively, on the AID dataset with the ratio of 50%. Furthermore, at the ratio of 20%, our method still achieves the best performance. Therefore, on the AID dataset, we obtain the same superior performance as the UCM dataset, which proves that our method has the property of robustness.

In order to further demonstrate the superiority of our proposed method, we show the confusion matrixes under 50% and 20% training ratio in Figs. 9 and 10, respectively. As can be seen from Fig. 9, 26 of 30 categories reach the results of accuracy of more than 90%. Only the categories of "Center" (85%), "Resort" (85%), "School" (89%), and "Square" (85%) obtain a bit of severe misclassification. This is caused by the similar geometrical structures or distributions of a ground object. There are still several categories that can be exactly classified, such as "BareLand" (99%), "Desert" (99%), "Forest" (99%), and "Meadow" (99%).

The confusion matrix under the 20% training ratio is shown in Fig. 10. We can find that there are 21 of 30 categories that achieve an accuracy of over 90%, and even 13 of 30 categories attain an accuracy of more than 95%. Although, the training ratio is reduced to 20%, there are several categories that still achieve an excellent effect of classification, such as "Beach" (97%), "Forest" (96%), "Mountain" (100%), "Parking" (99%), "Port" (98%), and "Viaduc" (99%). This further shows the effectiveness of our proposed method for RS scene classification.

From the above experiments, we can summarize that the proposed TDFE-DAA shows excellent performance on both UCM and AID datasets. It can find that our proposed method is

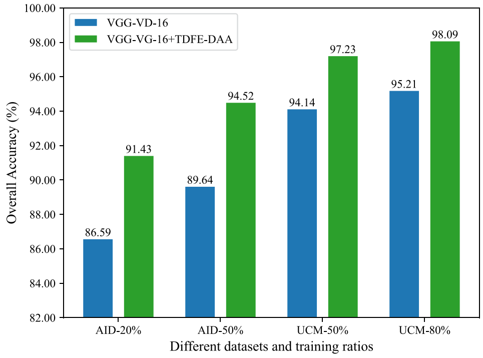

Fig. 11. Scalability analysis on the backbone of VGG-VD-16.

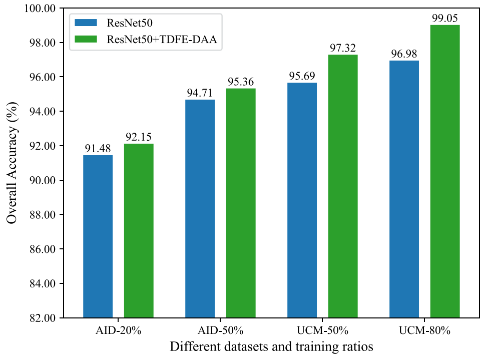

Fig. 12. Scalability analysis on the backbone of ResNet50.

TABLE V
RESULTS OF OA (%) FOR DIFFERENT METHODS WITH THE TRAINING RATIOS OF 80% AND 50% ON THE UCM DATASET

Architecture	Overall Accuracy (%)	
	80%	**50%**
VGG-VD-16	95.21	94.14
VGG-VD-16+TDFE	97.14	95.90
VGG-VD-16+DAA	97.61	95.75
VGG-VD-16+TDFE-DAA (ours)	**98.09**	**97.23**
ResNet50	96.98	95.69
ResNet50+TDFE	98.33	96.76
ResNet50+DAA	98.09	96.95
ResNet50+TDFE-DAA (ours)	**99.05**	**97.31**

The boldface values represents the best value in the experiment.

TABLE VI
RESULTS OF OA (%) FOR DIFFERENT METHODS WITH THE TRAINING RATIOS OF 50% AND 20% ON THE AID DATASET

Architecture	Overall Accuracy (%)	
	50%	**20%**
VGG-VD-16	89.64	86.59
VGG-VD-16+TDFE	94.16	90.66
VGG-VD-16+DAA	94.25	91.16
VGG-VD-16+TDFE-DAA (ours)	**94.52**	**91.43**
ResNet50	94.71	91.48
ResNet50+TDFE	95.07	91.69
ResNet50+DAA	95.08	91.63
ResNet50+TDFE-DAA (ours)	**95.36**	**92.15**

The boldface values represents the best value in the experiment.

far superior to those mid-level-based methods, as well as better than those CNN-based methods. Those results further demonstrate the effectiveness of our architecture, including the TDFE module that concatenation the high-level global features and low-level local features, and the DAA module that highlights the crucial global features and key local features. And those results also indicate the superiority and robustness of our proposed architecture for RS image scene classification.

IV. DISCUSSION

To comprehensively verify the advance of our proposed method, we conduct various experiments including scalability and the effect of each module.

A. Scalability

Our proposed method of TDFE-DAA is well suited to apply in the most popular CNNs, such as VGG-VD-16 and ResNet50, which are commonly used in scene classification. We conduct a number of experiments with two different backbones of CNNs on the UCM dataset and the AID dataset accompanied by the same training ratios as before to validate the scalability of our proposed method. We utilize the same setting as before, and for the VGG-VD-16, we use the same strategy as the ResNet-50, which is to choose the last output of the convolutional layer of each stage as the initial input of the TDFE module. The detailed results are presented in Figs. 11 and 12. We can find from Fig. 11 that the performance of our proposed TDFE-DAA-VGG-VD-16 is all better than the VGG-VD-16 on each training ratio. With the backbone of ResNet50, we can get the same conclusion as the VGG-VD-16 shown in Fig. 12. From those results, it can demonstrate the scalability of our method.

B. Effects of Different Modules

To certify how the TDFE module and the DAA module affect our architecture, we conduct various ablation experiments. All the ablation experiments are performed on the UCM dataset and the AID dataset, and the results are presented in Tables V and VI.

Effects of TDFE: In order to comprehensively certify our proposed TDFE (Backbone+TDFE), we perform experiments on two different backbones of VGG-VD-16-based and Resnet50-based, respectively. We can observe from Table V that the VGG-VD-16 combined with the TDFE module obtained higher results than only using the VGG-VD-16 and increase by 1.93% and 1.76% under 80% and 50% training ratios on the UCM dataset, respectively. Similarly, for the backbone of ResNet50,

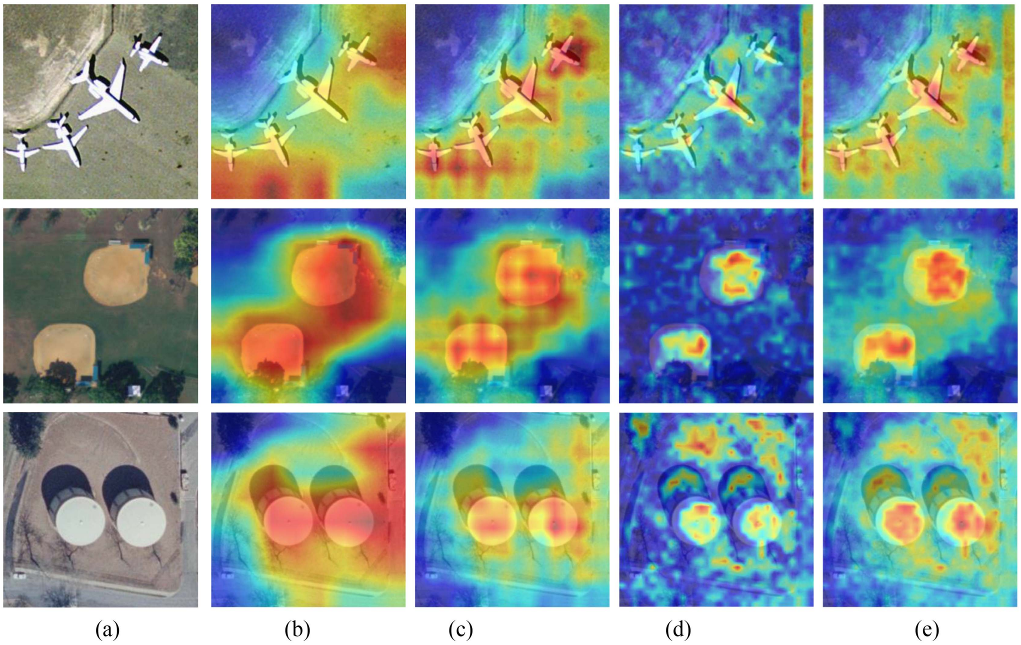

Fig. 13. Visualization results of Grad-CAM. (a) Input. (b) ResNet50. (c) Channel attention. (d) Spatial attention. (e) DAA.

we arrive at a consistent conclusion as the backbone of VGG-VD-16. On the 80% and 50% training ratios for the UCM dataset, the OA values obtained by ResNet50 combined with the TDFE are higher at 1.35% and 1.07% than just utilizing ResNet50, respectively. In addition, it can be found from Table VI that on the AID dataset, we can get a consistent conclusion as on the UCM dataset. Therefore, it can be seen from the above observations that the TDFE module is exactly effective for RS image scene classification.

Effects of DAA: To further verify the effects of the DAA module, we perform the same experiments as mentioned above. From Tables V and VI, we can obtain consistent results as the TDFE module. Compared with two backbones, by utilizing our DAA module, our proposed methods (backbone+DAA) achieve better performance. In detail, the VGG-VD-16+DAA increased the OA by 2.40% and 1.61%, as well as, the ResNet50+DAA increased the OA by 1.11% and 1.26% under 80% and 50% training ratio for the UCM dataset, respectively. For the AID dataset, we can get a consistent conclusion as on the UCM dataset. Those results strongly prove the superiority and effectiveness of our proposed module of DAA. To further illustrate the effectiveness of our proposed DAA module, the Gradient-weighted Class Activation Mapping (Grad-CAM) [69] is utilized to visualize how the attention module affects the performance of feature representation. From Fig. 13, we can find that the activated field by ResNet50 is relatively large, resulting in the inclusion of a lot of redundant information which may interfere with the classification accuracy. As can be seen from Fig. 13(c), the channel attention module covers the salient regions, which indicates that our proposed channel attention can more accurately response to important global-semantic features than the original ResNet50. Fig. 13(d) shows that the activation of local features by the spatial attention mechanism, we can find the lighter regions are some local details that are crucial for distinguishing other images. From Fig. 13(e), we can observe that the lighter regions are the addition of the activation of the channel attention with the activation of the spatial attention, which illustrates that our proposed DAA can obtain more discriminative features.

Effects of TDFE-DAA: We can find from Tables V and VI that the OA obtained by our proposed ResNet50+TDFE-DAA is higher than that obtained by both ResNet50+TDFE and ResNet50+DAA. Similarly, the VGG-VD-16 based to get the same conclusion. Especially compared to VGG-VD-16 and ResNet50, our proposed methods of VGG-VD-16+TDFE-DAA and ResNet50+TDFE-DAA obtained significant progress, which improves the OA by 2.88% and 2.07% for 80% training ratio and also 3.09% and 1.62% for 50% training ratio on the UCM dataset. For the AID dataset, we got the consistent conclusion. From those results, one can find that our introduced architecture achieves the best performance, and also proves the superiority and effectiveness of the TDFE module and the DAA module for RS image scene classification.

C. *Time Efficiency*

The time efficiency is important to assess a model; thus, we conduct the experiments on the UCM dataset and the AID dataset with 50% training ratio to show the time consumption. The time efficiency is measured by the training time of the corresponding model for an epoch, and the amount of weight parameters of different models is also counted, which are both shown in Table VII. It can be found from Table VII that the time taken by ResNet50-TDFE-DAA is very close to that of ResNet50, but the OA is improved by 1.63% and 0.65% on the UCM dataset and the AID dataset, respectively. In addition, we can find that VGG16+TDFE-DAA consumes less time than VGG-VD-16,

TABLE VII
COMPARE TIME COST WITH OTHER MODELS ON THE UCM DATASET AND THE AID DATASET WITH 50% TRAINING RATIOS

Architecture	Parameters(M)	UCM (50%)		AID (50%)	
		Time Cost (s/epoch)	Overall Accuracy (%)	Time Cost (s/epoch)	Overall Accuracy (%)
VGG-VD-16	134.4	14.78	94.14	49.21	89.64
ResNet50	23.6	6.95	95.69	26.65	94.71
EFPN-DSE	25.7	7.71	96.65	31.53	93.35
VGG-VD-16+TDFE-DAA (ours)	22.5	10.86	97.23	43.63	94.52
ResNet50+TDFE-DAA (ours)	30.8	7.76	97.32	31.68	95.36

which is due to that the original VGG-VD-16 uses extra two fully connected layers in the classification block to fuse features while our method replaces them with the proposed TDFE block consisted of convolutions with less parameters. It can be also found from Table VII that compared with EFPN-DSE, our proposed methods take almost the same time but achieve better performance whether on the UCM dataset or the AID dataset. In conclusion, our method can achieve a better result of OA with approximated even less time when compared with the original network and EFPN-DSE. Meanwhile, we can find that the parameter amount of ResNet50-TDFE-DAA is only increased by 7.2 M compared to ReNet50 while the OA is greatly increased. Obviously, the parameter amount of VGG16+TDFE-DAA is greatly reduced compared with VGG-VD-16, and the OA value is also improved by 3.09% and 4.88% on the UCM dataset and the AID dataset, respectively. Therefore, it can be concluded from the above observations that our proposed method strikes a balance between accuracy and time efficiency.

V. Conclusion

In this article, we introduce a competitive method named TDFE-DAA to further improve the performance of the RS image scene classification. The method consists of two modules, which are the TDFE module and the DAA module, respectively. The TDFE module contains two branches, which are a high-level branch and a low-level branch and is designed to aggregate the high-level but low-resolution features and the low-level but high-resolution features to enhance the image feature representation. The DAA module is introduced to further highlight the important features, which consist of the channel attention module and spatial attention module. The channel attention module is designed to capture the key channels and diminish the insignificant channels for the higher-level features and the spatial attention module focuses on the important spatial locations and suppresses irrelevant locations. We conducted a number of experiments to prove the advantages of our proposed method for the RS image scene classification.

References

[1] L. Zhang, Q. Zhang, B. Du, X. Huang, Y. Y. Tang, and D. Tao, “Simultaneous spectral-spatial feature selection and extraction for hyperspectral images,” *IEEE Trans. Cybern.*, vol. 48, no. 1, pp. 16–28, Jan. 2018.

[2] M. Voltersen, C. Berger, S. Hese, and C. Schmullius, “Expanding an urban structure type mapping approach from a subarea to the entire city of Berlin,” in *Proc. Joint Urban Remote Sens. Event*, 2015, pp. 1–4.

[3] X. Zheng, Y. Yuan, and X. Lu, “A deep scene representation for aerial scene classification,” *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 7, pp. 4799–4809, Jul. 2019.

[4] M. Weiss, F. Jacob, and G. Duveiller, “Remote sensing for agricultural applications: A meta-review,” *Remote Sens. Environ.*, vol. 236, 2020, Art. no. 111402.

[5] Q. Zhang, Q. Yuan, M. Song, F. Sun, and L. Zhang, “Cooperated spectral low-rankness prior and deep spatial prior for HSI unsupervised denoising,” *IEEE Trans. Image Process.*, vol. 31, pp. 6356–6368, Oct. 2022.

[6] 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.

[7] R. Avtar et al., “Exploring renewable energy resources using remote sensing and GIS—A review,” *Resources*, vol. 8, no. 3, 2019, Art. no. 149.

[8] Y. Yi, Z. Zhang, W. Zhang, H. Jia, and J. Zhang, “Landslide susceptibility mapping using multiscale sampling strategy and convolutional neural network: A case study in Jiuzhaigou region,” *Catena*, vol. 195, 2020, Art. no. 104851.

[9] K. Tiwari, M. K. Arora, and D. Singh, “An assessment of independent component analysis for detection of military targets from hyperspectral images,” *Int. J. Appl. Earth Observ. Geoinformation*, vol. 13, no. 5, pp. 730–740, 2011.

[10] G. Cheng, J. Han, L. Guo, Z. Liu, S. Bu, and J. Ren, “Effective and efficient midlevel visual elements-oriented land-use classification using VHR remote sensing images,” *IEEE Trans. Geosci. Remote Sens.*, vol. 53, no. 8, pp. 4238–4249, Aug. 2015.

[11] J. A. dos Santos, O. A. B. Penatti, and R. d. S. Torres, “Evaluating the potential of texture and color descriptors for remote sensing image retrieval and classification,” in *Proc. Int. Conf. Comput. Vis. Theory Appl.*, 2010, pp. 203–208.

[12] N. Dalal and B. Triggs, “Histograms of oriented gradients for human detection,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2005, pp. 886–893.

[13] J. Ren, X. Jiang, and J. Yuan, “Learning LBP structure by maximizing the conditional mutual information,” *Pattern Recognit.*, vol. 48, no. 10, pp. 3180–3190, 2015.

[14] D. G. Lowe, “Distinctive image features from scale-invariant keypoints,” *Int. J. Comput. Vis.*, vol. 60, no. 2, pp. 91–110, 2004.

[15] 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.

[16] Q. Zhu, Y. Zhong, B. Zhao, G.-S. 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.

[17] H. Sridharan and A. Cheriyadat, “Bag of lines (BoL) for improved aerial scene representation,” *IEEE Geosci. Remote Sens. Lett.*, vol. 12, no. 3, pp. 676–680, Mar. 2015.

[18] L. Zhou, Z. Zhou, and D. Hu, “Scene classification using a multi-resolution bag-of-features model,” *Pattern Recognit.*, vol. 46, no. 1, pp. 424–433, 2013.

[19] S. Lazebnik, C. Schmid, and J. Ponce, “Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories,” in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, 2006, vol. 2, pp. 2169–2178.

[20] L. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, “DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 4, pp. 834–848, Apr. 2018.

[21] F. P. 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.

[22] Y. Luo et al., “CE-FPN: Enhancing channel information for object detection,” *Multimedia Tools Appl.*, vol. 81, pp. 30685–30704, 2022.

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

[24] K. B. Obaid et al., "Deep learning models based on image classification: A review," *Int. J. Sci. Bus.*, vol. 4, no. 11, pp. 75–81, 2020.

[25] M. Hu, Y. Li, L. Fang, and S. Wang, "A2-FPN: Attention aggregation based feature pyramid network for instance segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 15343–15352.

[26] Q. Zhang, Q. Yuan, Z. Li, F. Sun, and L. Zhang, "Combined deep prior with low-rank tensor SVD for thick cloud removal in multitemporal images," *ISPRS J. Photogrammetry Remote Sens.*, vol. 177, pp. 161–173, 2021.

[27] R. Minetto, M. P. Segundo, and S. Sarkar, "Hydra: An ensemble of convolutional neural networks for geospatial land classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 9, pp. 6530–6541, Sep. 2019.

[28] S. Li, W. Song, L. Fang, Y. Chen, P. Ghamisi, and J. A. Benediktsson, "Deep learning for hyperspectral image classification: An overview," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 9, pp. 6690–6709, Sep. 2019.

[29] X. Wang, S. Wang, C. Ning, and H. Zhou, "Enhanced feature pyramid network with deep semantic embedding for remote sensing scene classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 9, pp. 7918–7932, Sep. 2021.

[30] X. Tang, Q. Ma, X. Zhang, F. Liu, J. Ma, and L. Jiao, "Attention consistent network for remote sensing scene classification," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 2030–2045, Jan. 2021.

[31] 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.

[32] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. Int. Conf. Learn. Representations*, 2015.

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

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

[35] X. Han, Y. Zhong, L. Cao, and L. Zhang, "Pre-trained AlexNet architecture with pyramid pooling and supervision for high spatial resolution remote sensing image scene classification," *Remote Sens.*, vol. 9, no. 8, 2017, Art. no. 848.

[36] D. Zeng, S. Chen, B. Chen, and S. Li, "Improving remote sensing scene classification by integrating global-context and local-object features," *Remote Sens.*, vol. 10, no. 5, 2018, Art. no. 734.

[37] B. Liu, J. Meng, W. Xie, S. Shao, Y. Li, and Y. Wang, "Weighted spatial pyramid matching collaborative representation for remote-sensing-image scene classification," *Remote Sens.*, vol. 11, no. 5, 2019, Art. no. 518.

[38] X. Lu, T. Gong, and X. Zheng, "Multisource compensation network for remote sensing cross-domain scene classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 4, pp. 2504–2515, Apr. 2020.

[39] 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.

[40] C. Xu, G. Zhu, and J. Shu, "A lightweight and robust lie group-convolutional neural networks joint representation for remote sensing scene classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–15, Jan. 2021.

[41] X. Chen, M. Ma, Y. Li, and W. Cheng, "Fusing deep features by kernel collaborative representation for remote sensing scene classification," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 12429–12439, Nov. 2021.

[42] D. Wang and J. Lan, "A deformable convolutional neural network with spatial-channel attention for remote sensing scene classification," *Remote Sens.*, vol. 13, no. 24, 2021, Art. no. 5076.

[43] K. Xu, H. Huang, Y. Li, and G. Shi, "Multilayer feature fusion network for scene classification in remote sensing," *IEEE Geosci. Remote Sens. Lett.*, vol. 17, no. 11, pp. 1894–1898, Nov. 2020.

[44] T. Tian, L. Li, W. Chen, and H. Zhou, "SEMSDNet: A multiscale dense network with attention for remote sensing scene classification," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 5501–5514, Apr. 2021.

[45] X. Lu, H. Sun, and X. Zheng, "A feature aggregation convolutional neural network for remote sensing scene classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 10, pp. 7894–7906, Oct. 2019.

[46] W. Li et al., "Classification of high-spatial-resolution remote sensing scenes method using transfer learning and deep convolutional neural network," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 13, pp. 1986–1995, May 2020.

[47] R. Cao, L. Fang, T. Lu, and N. He, "Self-attention-based deep feature fusion for remote sensing scene classification," *IEEE Geosci. Remote Sens. Lett.*, vol. 18, no. 1, pp. 43–47, Jan. 2021.

[48] Q. Zhang, Q. Yuan, J. Li, Z. Li, H. Shen, and L. Zhang, "Thick cloud and cloud shadow removal in multitemporal imagery using progressively spatio-temporal patch group deep learning," *ISPRS J. Photogrammetry Remote Sens.*, vol. 162, pp. 148–160, 2020.

[49] 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.

[50] T. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 2117–2125.

[51] A. Vaswani et al., "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017, pp. 5998–6008.

[52] S. Woo, J. Park, J. Lee, and I. S. Kweon, "CBAM: Convolutional block attention module," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 3–19.

[53] Y. Zhu, X. Guo, J. Liu, and Z. Jiang, "Multi-branch context-aware network for person re-identification," in *Proc. IEEE Int. Conf. Multimedia Expo*, 2019, pp. 712–717.

[54] 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.

[55] Q. Bi, K. Qin, H. Zhang, J. Xie, Z. Li, and K. Xu, "APDC-Net: Attention pooling-based convolutional network for aerial scene classification," *IEEE Geosci. Remote Sens. Lett.*, vol. 17, no. 9, pp. 1603–1607, Sep. 2020.

[56] L. Ding, H. Tang, and L. Bruzzone, "LANet: Local attention embedding to improve the semantic segmentation of remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 1, pp. 426–435, Jan. 2021.

[57] F. Li, R. Feng, W. Han, and L. Wang, "An augmentation attention mechanism for high-spatial-resolution remote sensing image scene classification," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 13, pp. 3862–3878, Sep. 2020.

[58] R. Fan, L. Wang, R. Feng, and Y. Zhu, "Attention based residual network for high-resolution remote sensing imagery scene classification," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, 2019, pp. 1346–1349.

[59] S. Liu, L. Qi, H. Qin, J. Shi, and J. Jia, "Path aggregation network for instance segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 8759–8768.

[60] N. Komodakis and S. Zagoruyko, "Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer," in *Proc. Int. Conf. Learn. Representations*, 2017.

[61] J. Ji, T. Zhang, L. Jiang, W. Zhong, and H. Xiong, "Combining multilevel features for remote sensing image scene classification with attention model," *IEEE Geosci. Remote Sens. Lett.*, vol. 17, no. 9, pp. 1647–1651, Sep. 2020.

[62] M. Zhu, L. Jiao, F. Liu, S. Yang, and J. Wang, "Residual spectral–spatial attention network for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 1, pp. 449–462, Jan. 2021.

[63] G. Xia et al., "AID: A benchmark data set for performance evaluation of aerial scene classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 7, pp. 3965–3981, Jul. 2017.

[64] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. Int. Conf. Learn. Representations*, 2015.

[65] R. M. Anwer, F. S. Khan, J. Van De Weijer, M. Molinier, and J. Laaksonen, "Binary patterns encoded convolutional neural networks for texture recognition and remote sensing scene classification," *ISPRS J. Photogrammetry Remote Sens.*, vol. 138, pp. 74–85, 2018.

[66] W. Zhang, P. Tang, and L. Zhao, "Remote sensing image scene classification using CNN-CapsNet," *Remote Sens.*, vol. 11, no. 5, 2019, Art. no. 494.

[67] N. Liu, X. Lu, L. Wan, H. Huo, and T. Fang, "Improving the separability of deep features with discriminative convolution filters for RSI classification," *ISPRS Int. J. Geo-Inf.*, vol. 7, no. 3, 2018, Art. no. 95.

[68] Y. Lv, X. Zhang, W. Xiong, Y. Cui, and M. Cai, "An end-to-end local-global-fusion feature extraction network for remote sensing image scene classification," *Remote Sens.*, vol. 11, no. 24, 2019, Art. no. 3006.

[69] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual explanations from deep networks via gradient-based localization," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 618–626.

Xiaoning Chen received the B.S. degree in electronic information science and technology and the M.S. degree in communication and information systems from Northwest University, Xi'an, China, in 2004 and 2011, respectively. She is currently working toward the Ph.D. degree in information and communication engineering with Northwestern Polytechnical University, Xi'an.

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

Zonghao Han received the B.S. degree in communication engineering from Ningxia University, Yinchuan, China, in 2020. He is currently working toward the Ph.D. degree in information and communication engineering with the School of Electronics and Information, Northwestern Polytechnical University, Xi'an, China.

His major research interests include hyperspectral remote sensing image processing and deep learning.

Yong Li received the B.S. degree in electronic engineering and the M.S. and Ph.D. degrees in circuits and systems from Northwestern Polytechnical University, Xi'an, China, in 1983, 1988, and 2005, respectively.

He joined the School of Electronic Information, Northwestern Polytechnical University, in 1983, where he has been a Professor, since 2002. His research interests include digital signal processing and radar signal processing.

Mingyang Ma (Student Member, IEEE) received the B.S. degree in communication engineering and the Ph.D. degree in communication and information systems from Northwestern Polytechnical University, Xi'an, China, in 2015 and 2021, respectively.

He is currently an Assistant Professor with the School of Information and Communications Engineering, Xi'an Jiaotong University, Xi'an. His main research interests include image processing and video summarization.

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 Associate Professor with the School of Electronics and Information. He was a Visiting Student with the University of Sydney, from October 2007 to October 2008. 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 was the recipient of the Excellent Doctoral Dissertation Award of Shaanxi Province in 2014, and the Best Paper Award of IEEE ISPACS 2017. He is currently a reviewer of more than 20 international famous academic journals and was awarded the Best Reviewer of the IEEE JOURNAL OF SELECTED TOPICS IN APPLIED EARTH OBSERVATIONS AND REMOTE SENSING in 2019. He was the Registration Chair of IEEE China Summit and International Conference on Signal and Information Processing (ChinaSIP) 2014.

Wei Cheng (Member, IEEE) received the B.S. degree in electronic information engineering and the M.S. and Ph.D. degrees in communication and information system from Northwestern Polytechnical University, Xi'an, China, in 2003, 2006, and 2011, respectively.

He is currently working as an Associate Professor with the School of Electronic Information, Northwestern Polytechnical University. His research interests include wireless sensor networks, ad hoc networks, machine learning, and radar signal processing.