This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JSTARS.2021.3103176, IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing

A Multiscale Dual-branch Feature Fusion and Attention Network for Hyperspectral Images Classification

Hongmin Gao, *Member, IEEE*, Yiyan Zhang, *Student Member, IEEE*, Zhonghao Chen, *Student Member, IEEE*, and Chenming Li

***Abstract*— Recently, hyperspectral image classification based on deep learning has achieved considerable attention. Many CNN classification methods have emerged and exhibited superior classification performance. However, most methods focus on extracting features by using fixed convolution kernels and layer-wise representation, resulting in feature extraction singleness. Additionally, the feature fusion process is rough and simple. Numerous methods get accustomed to fusing different levels of features by stacking modules hierarchically, which ignore the combination of shallow and deep spectral-spatial features. In order to overcome the preceding issues, a novel multiscale dual-branch feature fusion and attention network (MSDBFA) is proposed. Specifically, we design a multiscale feature extraction module (MSFE) to extract spatial-spectral features at a granular level and expand the range of receptive fields, thereby enhancing the multiscale feature extraction ability. Subsequently, we develop a dual-branch feature fusion interactive module (DBFM) that integrates the residual connection's feature reuse property and the dense connection's feature exploration capability, obtaining more discriminative features in both spatial and spectral branches. Additionally, we introduce a novel shuffle attention mechanism that allows for adaptive weighting of spatial and spectral features, further improving classification performance. Experimental results on three benchmark datasets demonstrate that our model outperforms other state-of-the-art methods while incurring the lower computational cost.**

***Index Terms*—Hyperspectral image classification, multiscale feature extraction module (MSFE), Dual-branch feature fusion (DBFM), Shuffle attention block, Convolutional neural network (CNN)**

I. Introduction

Hyperspectral images (HSIs) have recently gained increased attention in the field of remote sensing. Hyperspectral remote sensing is a multi-dimensional signal acquisition technology that combines imaging and spectroscopy technology, which not only detect two-dimensional space characters but also one-dimensional spectral information of targets. Hyperspectral images have the following advantages over conventional remote sensing images. To begin, the spectral resolution is high, making for the acquisition of continuous spectral curves for varieties of ground objects. Meanwhile, the spectral coverage range is expanded, allowing for more detection of ground object responses to electromagnetic waves. Additionally, HSIs incorporate both spatial and spectral features and contain a greater amount of detailed information. Exactly due to these characteristics, HSIs play a significant role in agricultural detection [1]-[2], medical diagnosis [3]-[4], atmospheric monitoring [5]-[6], hydrological detection [7] and other fields. The essence of the hyperspectral remote sensing image classification is assigning each pixel vector to a specific land cover class. How to fully exploit the abundant spatial and spectral features becomes a great challenge in HSIs classification.

The traditional classification methods of HSIs are all based on the handcrafted feature. Early-stage classification methods such as Support Vector Machine [8], Random Forest [9] and Multiple Logistic Regression [10], they are all aimed at utilizing one-dimensional spectral features to complete the classification. Although the large number of spectral bands usually implies more potential information, the classification accuracy rises at first and then decreases owning to the high-dimensional data characteristics of HSIs that contribute to the Hughes phenomenon [11]. As a result, more and more studies focus on the dimensionality reduction of the spectral dimension [12]. Currently, the widely used methods include PCA [13] and LDA [14]. While these methods compress the spectral dimension and reduce the spectral redundancy, noise typically exists, which is caused by lighting and imaging equipment. Due to the spatial resolution limitation and the complexity of the imaging process, the phenomena of the same land-cover may exhibit spectral dissimilarity, while the spectral properties of different materials may be indistinguishable [15].

In recent years, deep learning occupies a dominant position in computer vision due to its robust feature representation ability. Deep learning eliminates the tedious process of feature engineering. Through an end-to-end structure, the network can automatically extract abstract features hierarchically. Deep learning (DL) has achieved great success in the fields of image classification [16], target recognition [17] and semantic segmentation [18]. For the first time, Chen *et al.* [19] apply DL to HSIs classification. Till then, more and more deep learning methods [20]-[25] have been existed in HSIs classification. For instance, [26] perform the feature extraction and classification with Deep Belief Network simultaneously. Xu *et al.* [27] generate HSIs classification map by combination of stacking

This work was supported in part by the National Natural Science Foundation of China under Grant 62071168. in part by the National Key Research and Development Program of China under Grant 2018YFC1508106, in part by the Fundamental Research Funds for the Central Universities of China under Grant B200202183 and in part by the China Postdoctoral Science Foundation under Grant 2021M690885. (*Corresponding author: Chenming Li.*)

Hongmin Gao, Yiyan Zhang, Zhonghao Chen and Chenming Li are with th e College of Computer and Information, Hohai University, Nanjing 211100, C hina (e-mail: gaohongmin@hhu.edu.cn; zhangyiyan@hhu.edu.cn; chenzhongh ao@hhu.edu.cn; lcm@hhu.edu.cn).

encoders. While both of the preceding methods have demonstrated considerable success, whereas they rely on the spectral vector of pixels to complete the classification and miss the spatial distribution of image pixels. The spatial context information of the original data is destroyed, resulting in the loss of useful spatial information. As a result, the research on HSIs classifications need to be further carried on. Researchers begin to place great emphasis on the spatial structure information of HSIs. A lot of methods based on 2-D CNN have been proposed to apply in the HSIs classification [28]-[31]. For instance, Makantasis *et al.* [28] developed a neural network model based on 2-D CNN. The intermediate pixels are packed into fixed-size cubes by filling surrounding pixels and then sent into the neural network to extract spatial information. This data processing technique is quite novel and achieve an excellent classification performance. Li *et al.* [32] proposed a novel pixel-pair method to exploit the similarity between pixels and use a majority voting strategy to generate the final label. Pan *et al.* [33] designed a small-scale data-driven method, multi-grained network (MugNet) to deal with the limited samples in HSIs classifications. Cao *et al.* [34] developed a Bayesian HSIs classification method, which combines the CNN and a smooth Markov random field to exploit the spatial information. However, the most distinguishing features of HSIs are the spectral diversities. Studies frequently place a greater emphasis on spatial characteristics but appear to overlook spectral characteristics. Therefore, later researches begin to explore the combination of spatial and spectral features to complete the classification tasks. For the first time, [35] proposed a hyperspectral image classification algorithm based on the spectral-spatial features, in which spectral information was fused with the spatial information through the transformation of the network. The classification task was carried out on the fused features, and the results were excellent. Li *et al.* [36] proposed a double-branch spatial-spectral extraction and fusion method based on 2D convolutional network which further improved the discriminative feature extraction capacity. Liu *et al.* [37] introduced LSTM to HSIs classification in a novel way, including spectral and spatial LSTM blocks. The method passed each pixel's spectral-spatial features to the softmax layer, which generated two distinct types of results, and then used the decision fusion method to generate classification renderings.

CNN has strong feature extraction capabilities which can achieve the high-level abstract features by stacking modules and deepening the network layers. Nonetheless, On the one hand, a deeper network introduces additional parameters into the training process and lengthens the training time. On the other hand, gradient vanishing impairs back propagation and degrades the classification performance. For the former, the continued development of the high-performance graphics processing units (GPUs) [38] have resulted in a significant reduction in training time when confronted with a large training parameter network. For the latter, He *et al.* [39] propose a residual network that uses skip connections to ensure that gradients circulate smoothly in the deeper network, alleviating the problem of gradient vanishing. Soon after, residual networks gained popularity in the field of computer vision, and they were also applied to the classifications of HSIs. For instance, Zhong *et al.* [38] designed a spectral-spatial residual network (SSRN) with two consecutive residual blocks in order to learn the discriminative features in the HSIs, which can perform well with small training samples. Lee *et al.* [41] enhance the learning efficiency of traditional CNN models by introducing residual network and use multiscale convolution kernels to explore the spatial-spectral features in HSIs. Song *et al.* [42] develop a deep residual network with an attention mechanism to learn HSIs discriminative features and obtain further improvement in classification performance. Paoletti *et al.* [43] designed a deep pyramidal residual network for HSIs classification.

Recent works in attention mechanisms have shown it to be an extremely powerful tool to boost the classification performance According to the biological cognitive research, human being receive significant information by focusing on a few critical items and ignoring others [44]. Similarly, attention in neural networks has the same function, which has been successfully applied to various tasks in the computer vision [45]-[46]. In the HSIs classification tasks, many methods based on existing attention mechanism also emerged, demonstrating the effectiveness of improving the classification performance.

Meanwhile, multiscale feature extraction is a critical component of hyperspectral image classification, as it has a significant impact on the classification performance. Existing multiscale extractors [47] are limited to extracting features from fixed receptive fields, and thus cannot extract both global and local features simultaneously. To say the least, even if the multiscale features have been extracted from the front layers of the network, the fusion process of the features is rough, resulting in information loss in front layers.

Drawing intuition from the success of the above-mentioned methods, a novel 3D dual-branch feature extraction and fusion attention network is proposed for HSIs classification. The main contributions of this article are summarized as follows:

(1) Many existing classification methods based on CNN extract the multiscale spatial-spectral features with layer-wise representations and fixed kernel size. Different from them, we design a 3-D multiscale feature extraction module (MSFE) which refers to the multiple available receptive fields at a more granular level. The MSFE is capable of performing multiscale feature extraction in a lightweight and efficient manner.

(2) In order to fully excavate the potential of spatial and spectral feature representation of HSIs, a 3D dual-branch feature interactive module (DBFM) is proposed for classification. Different from the existing parallel processing network with stacked convolution modules, DBFM is a dual-branch structure that consists of multiple filters with additive links and concatenative links. To be brief, concatenative links focus on new effective feature exploration in HSIs, while additive links enhance the feature reuse in previous layers. We integrate the two types of links in DBFM for fusing the spatial and spectral features in different levels of the network and assimilate the particular features from previous layers.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JSTARS.2021.3103176, IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing

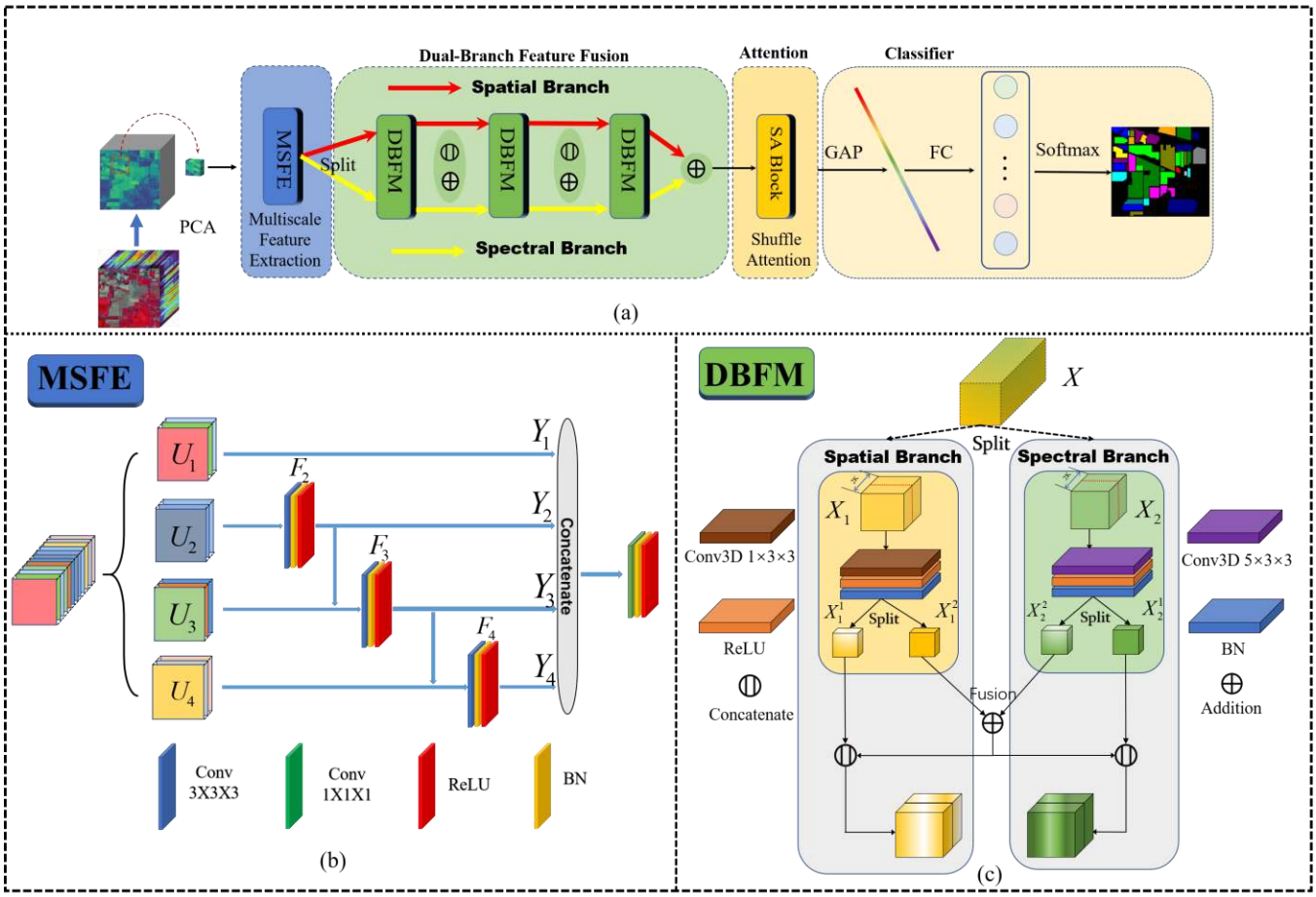


Fig. 1. (a) Overall flowchart of the proposed MSDBFA. (b) Structure of the MSFE. (c) Structure of the DBFM

(3) Given that the significant contribution of distinct spatial and channel features to classification results in HSIs, we introduce a 3-D spatial-channel attention block to boost the network's feature representation capability. Existing attention blocks focus on capturing the dependencies in spectral dimension, while our proposed 3-D attention block improves classification performance by creatively altering the conventional weight distribution method in both channel and spatial dimensions.

(4) Extensive experiments on four publicly available datasets are conducted. The results indicate that our model outperforms state-of-the-art methods.

The remainder of this article will be organized in the following manner. First, the proposed MSFE, DBFM, attention block and corresponding algorithms are described in Section II. Next, the section III details the associated experiments and analysis. Finally, Section IV concludes the article with observations and conclusions.

II. Proposed Method

This section begins with a brief overview of the proposed MSDBFA model. And next we will elaborate the multiscale feature extraction (MSFE), dual-branch feature interactive module (DBFM) and attention block.

A. Overview of proposed model

The main procedure of the proposed MSDBFA is shown in Fig. 1 (a). We take the Indian Pines dataset for example to illustrate the detail process of the algorithm. Firstly, principal component analysis (PCA) is applied to reduce the spectral dimension and suppress the band noise in original HSIs. Additionally, PCA effectively mitigates the Hughes effect and thereby improves classification performance. The HSIs are then segmented into 3-D image cubes centered on labeled pixels and sent to the multiscale feature extraction module (MSFE). The MSFE is intended to extract multi-scale spatial-spectral features at a granular level and thus expand the range of receptive fields. Following that, the MSFE-processed image is evenly divided into two feature subsets and fed into the 3-D dual-branch feature interactive module. To achieve deep feature fusion in both spatial and spectral dimensions, we use hierarchical layers comprised of three DBFM modules with different kernel filters. Each DBFM has a spatial and spectral branch corresponding to it. The two branches combine shallow and deep features via additive and concatenative links to produce discriminative spatial-spectral features. Additionally, a Shuffle attention block is inserted into the network to adaptively filter out critical features for classification, allowing

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JSTARS.2021.3103176, IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing

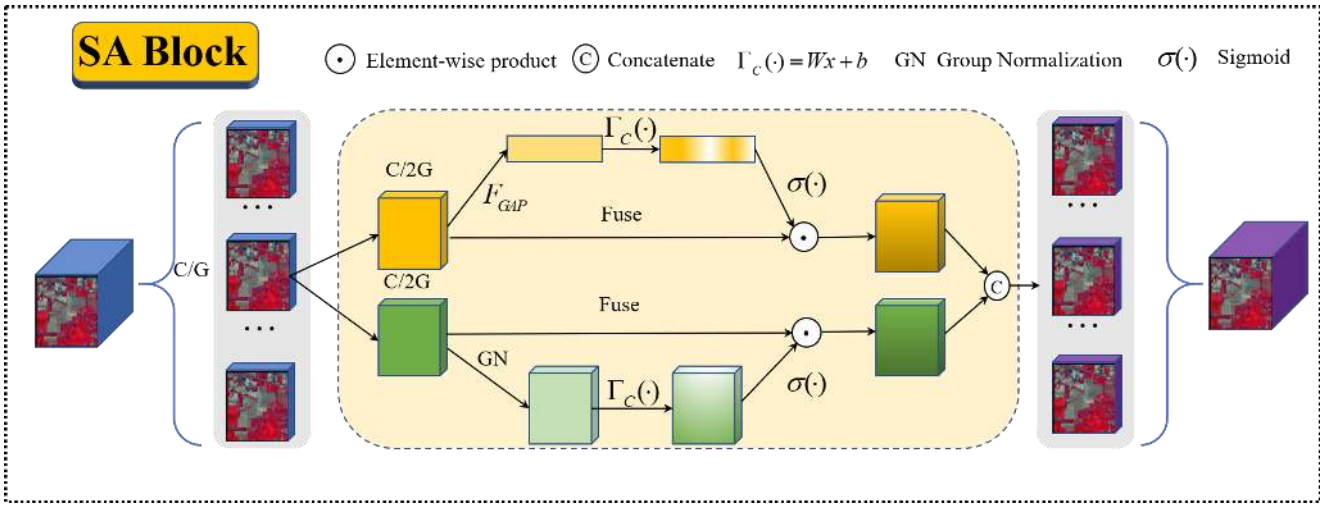


Fig. 2. Architecture of the SA Block

TABLE I
CONFIGURATION OF THE MSDBFA MODEL FOR THE INDIAN PINES DATASET (SPATIAL SIZE=15×15)

	Layer Name	Input Shape	Kernel size	padding	Stride	Filters	Output Shape
MSFE	Conv3D_pre	(1, 30, 15, 15)	(1, 1, 1)	(0, 0, 0)	(1, 1, 1)	32	(32, 30, 15, 15)
	Split	(32, 30, 15, 15)	——	——	——	——	#4 (8, 30, 15, 15)
	Conv3D_1	(8, 30, 15, 15)	(3, 3, 3)	(1, 1, 1)	(1, 1, 1)	8	(8, 30, 15, 15)
	Conv3D_2	(8, 30, 15, 15)	(3, 3, 3)	(1, 1, 1)	(1, 1, 1)	8	(8, 30, 15, 15)
	Conv3D_3	(8, 30, 15, 15)	(3, 3, 3)	(1, 1, 1)	(1, 1, 1)	8	(8, 30, 15, 15)
	Conv3D_4	(8, 30, 15, 15)	(3, 3, 3)	(1, 1, 1)	(1, 1, 1)	8	(8, 30, 15, 15)
	Concatenate	#4 (8, 30, 15, 15)	——	——	——	——	(32, 30, 15, 15)
——	Split	(32, 30, 15, 15)	(1, 1, 1)	(0, 0, 0)	(1, 1, 1)	16	#2 (16, 30, 15, 15)
DBFM1	Spa_Conv3D_1	(16, 30, 15, 15)	(1, 3, 3)	(0, 0, 0)	(2, 1, 1)	32	(32, 15, 13, 13)
	Spe_Conv3D_1	(16, 30, 15, 15)	(5, 3, 3)	(2, 0, 0)	(2, 1, 1)	32	(32, 15, 13, 13)
DBMF2	Spa_Conv3D_2	(32, 15, 13, 13)	(1, 3, 3)	(0, 0, 0)	(2, 1, 1)	64	(64, 8, 11, 11)
	Spe_Conv3D_2	(32, 15, 13, 13)	(5, 3, 3)	(2, 0, 0)	(2, 1, 1)	64	(64, 8, 11, 11)
DBFM3	Spa_Conv3D_1	(64, 8, 11, 11)	(5, 3, 3)	(2, 0, 0)	(2, 1, 1)	96	(96, 4, 9, 9)
	Spe_Conv3D_2	(64, 8, 11, 11)	(1, 3, 3)	(0, 0, 0)	(2, 1, 1)	96	(96, 4, 9, 9)
SA Block	Spa_attention	(96, 4, 9, 9)	——	——	——	——	(12, 4, 9, 9)
	Spe_attention	(96, 4, 9, 9)	——	——	——	——	(12, 4, 9 ,9)
	aggreation	#2 (12, 4, 9, 9)	——	——	——	——	(96, 4, 9, 9)
GLP	Global pooling	(96, 4, 9, 9)	——	——	——	——	(96, 1, 1, 1)
FC1	Dense1	(96,)	——	——	——	——	(128,)
	Dense2	(128,)	——	——	——	——	(16,)

the network to focus more on sensitive features while suppressing weaker ones. As a result, we will have discriminative feature maps for various classes. After completing above-mentioned operations, the feature maps are converted to vectors using an average pooling layer and then fed into the fully connected layers via softmax function to obtain the final classification maps.

B. Structure of MSFE

Multiscale feature representations are essential for various computer vision tasks. At the moment, the majority of methods rely on stacking multiple kernel filters in hierarchical layers to extract multiscale features. For instance, [46] makes use of spatial pyramid pooling to enhance the multiscale ability in each layer. [47] develops a feature pyramid that combines features at various scales.

However, these methods extract features in a layer-wise manner and with relatively fixed receptive fields. In contrast to these existing methods, we aim to improve the layer-wise multiscale representation capability and to achieve multiple available receptive fields at a more granular level. As a result, we developed the MSFE module for the purpose of extracting multiscale features from HSIs. As shown in Fig. 1(b), the input of original HSIs can be denoted as $U \in \mathbb{R}^{C\times D\times W\times H}$,where U represents the image patch. C, D, W and H denotes the channel, spectral dimension, width, and height of the image patch, respectively. we subdivided the original feature map into 4 subsets along the channel, denoted by U_i ,where $i \in \{1,2,3,4\}$. They all retain the same spatial sizes and spectral dimensions, but the channel count is reduced to 1/4 in comparison to the original feature map U . Subsequently,

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JSTARS.2021.3103176, IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing

each feature subset is sent to the convolutional sequential (Conv-BatchNorm-ReLU) with kernel size 3×3×3 to generate a new feature map. To avoid size inconsistency, we fill the data with the padding operation. The convolutional sequential operation is denoted by $F_i()$. To strengthen the feature reuse of previous layer and reduce the calculating parameters, we omit convolution for U_1 in the process of forward propagation, namely $U_1 = Y_1$. After adding with the output of $F_{i-1}()$,the remaining feature subsets $Y_{i,i\in\{2,3,4\}}$ are fed into $F_i()$. As a result, Y_i can be written as:

$$Y_i = \begin{cases} U_1 & i = 1 \\ F_i(U_1) & i = 2 \\ F_i(U_i + Y_{i-1}) & 2 < i \le 4 \end{cases} \tag{1}$$

According to the forward propagation, it can be found that the potential receptive field of each convolutional layer is a segmentation of $\{U_i, i \le 4\}$. Each time a convolutional operator is applied, the outputs will have a larger range of receptive field. Due to the effect of the combination, the final output of the module may contain multiple receptive fields with varying scales. In order to further improve the representative ability of the model, we concatenate all the feature subsets and pass them through a 1×1×1 convolution with ReLU activation to obtain more nonlinear characteristics.

C. Structure of DBFM

As is well known, ResNet [39] can be achieved by sequentially stacking residual blocks. The features are added element-wisely to the output ones through shortcut connections, which not only enhances the information propagation but also speeds up the network's training. While concatenative links in DenseNet [50] enable each layer to receive raw data generated by preceding layers, which is useful for exploring new feature. Fig. 3 shows the connection pattern differences between additive links and concatenative links. To sum up, we propose a 3D dual-branch feature interactive module for fusing these multiscale features in a novel way that includes multiple filters with additive and concatenative links in order to obtain discriminative spatial-spectral fused features. As is shown in the Fig. 1(c), the original input HSI cube X is indicated by $X \in \mathbb{R}^{C\times D\times H\times W}$, which has been evenly divided into two cubes along the channel C ,denoted by $X_i \in \mathbb{R}^{C/2\times D\times H\times W}, i \in \{1,2\}$. We pass the two feature subsets into spatial and spectral branch separately in the DBFM block. In the spatial branch, for the feature subset X_1 , we adopt 1×3×3 spatial kernels with subsampling strides of (2, 1, 1) to obtain feature maps with representative spatial features. Similarly for the spectral branch, we apply 5×3×3 kernels with strides of (2, 1, 1) to convolve with X_2 to achieve discriminative spectral features. Consequently, the outputs of two branches take advantage of the desired spatial and spectral features. To further reduce the computational parameters, we divide the feature

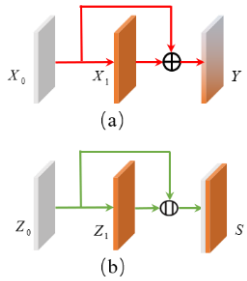


Fig. 3. Illustration of two type of links (a) additive links (b)concatenative links

maps X_1 , X_2 into two subsets respectively, denoted by $X_1^{\,j} \in \mathbb{R}^{C/2\times D\times H\times W} j \in \{1,2\}$, $X_2^{\,k} \in \mathbb{R}^{C/2\times D\times H\times W} k \in \{1,2\}$. Due to the X_1^1 and X_2^1 contain rich original features, we reserve them for subsequent concatenative link with previous fused feature map to explore new mied features. While for the X_1^2 and X_2^2, we combine them by additive links to strengthen the spatial-spectral feature fusion. Mathematically, it can be described as:

$$\begin{aligned} X_{fus} &= X_1^2 \oplus X_2^2 \\ X_{spatial} &= concat\left[X_1^1, X_{fus}\right] \\ X_{spectal} &= concat\left[X_2^1, X_{fus}\right] \end{aligned} \tag{2}$$

Where X_{fus} represents the fused features which contain discriminative spatial and spectral features. $cocnat[\cdot\cdot]$ is indicated that concatenative links between the original feature subset and fused features. It will enhance the feature fusion and explore some new features in other way.

D. Structure of Attention block

As we all know, various features contribute differently to the HSIs classification. Based on the fact, we introduce the attention mechanism here to allow the network to focus on useful features and neglect unsignificant ones. Existing attention block includes SENet [51], CBAM [52], GCNet [53], and so on. Among them, SENet is a representative channel attention architecture which apply the global average pooling and fully connected layers to recalibrate channel-wise feature response and remodel interdependencies between channels. GCNet is a lightweight and effective attention block which is used to construct global context feature. CBAM separates spatial and channel attention in order to capture representative features respectively, and then combines them to create a weighted feature map. Based on the fact that both spatial and channel attention are critical for HSI classification. Inspired by [54], we propose a novel lightweight spatial-channel attention block capable of effectively combining two distinct types of attention. As is shown in Fig. 2, for a given HSI cube $X \in \mathbb{R}^{C\times D\times W\times H}$, where C, D, W, H refers to the channel, dimension, width and height respectively. We divide X into G groups along the channel dimension, denoted by $X = [X_1, ..., X_G]$,

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JSTARS.2021.3103176, IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing

$X_k \in \mathbb{R}^{C/2G\times D\times H\times W}$. Each feature map will be segmented along the channel dimension into two branches, denoted by $X_{k1}, X_{k2} \in \mathbb{R}^{C/2G\times D\times H\times W}$. One branch is used to generate channel attention feature maps by acquiring the inter-relationship of channels, while the other branch generates spatial attention maps via the analysis of space location relationships.

For the channel attention branch, we first adjust the channel-wise dimension of the feature map X_{k1} via global average pooling operation, and then use sigmoid activation to recalibrate the channel-wise weights. The output of the channel attention can be obtained by:

$$X_{k1}^{'} = \sigma[w_1 * (\frac{1}{H\times W}\sum_{i=1}^{H}\sum_{j=1}^{W} X_{k1}(i,j)) + b_1] * X_{k1} \quad (3)$$

where $\frac{1}{H\times W}\sum_{i=1}^{H}\sum_{j=1}^{W} X_{k1}(i,j)$ represents the global average pooling (GAP) operation. $w_1 \in \mathbb{R}^{C/2G\times 1\times 1\times 1}$ and $b_1 \in \mathbb{R}^{C/2G\times 1\times 1\times 1}$ are the two dynamic parameters to scale the feature map. (i,j) refers to the specific spatial position in HSIs cube. σ is the sigmoid activation. For the spatial attention branch, we first use Group Norm (GN) to obtain spatial-wise statistics, and then, similarly to the channel attention branch, we introduce sigmoid activation to create a gating mechanism and generate the weighted feature map. As a result, the final output of spatial attention branch is followed by:

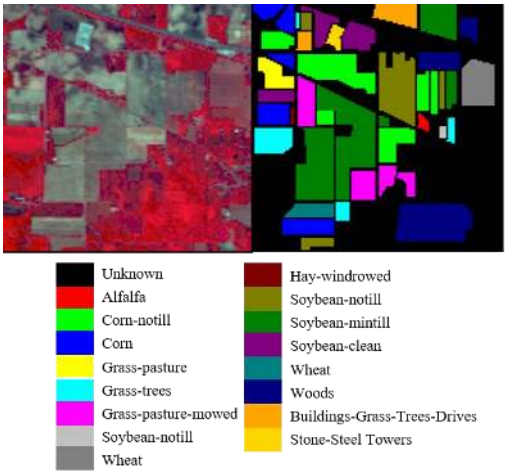
$$X_{k2} = \sigma(w_2 * GN(X_{k2}) + b_2) * X_{k2} \quad (4)$$

where $GN(\cdot)$ represents the Group Norm. Other parameters are consistent with the channel attention branch. To obtain the final weighted feature map, the representative features obtained by the spatial and channel attention branches require to be aggregated with concatenative links, denoted by $X_k = concat[X_{k1}, X_{k2}]$ .And then, we use the channel shuffle operation to allow cross-group information to flow along the channel dimension, which results in a more discriminative feature. The attention block is flexible and can be inserted anywhere in network. To optimize the classification performance, we place the attention block after the dual-branch feature interactive module. This allows the network to focus on and highlight the most critical components.

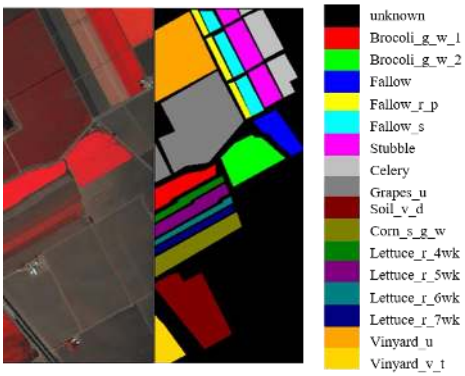
III. Experiments and Discussion

A. Data sets Descriptions

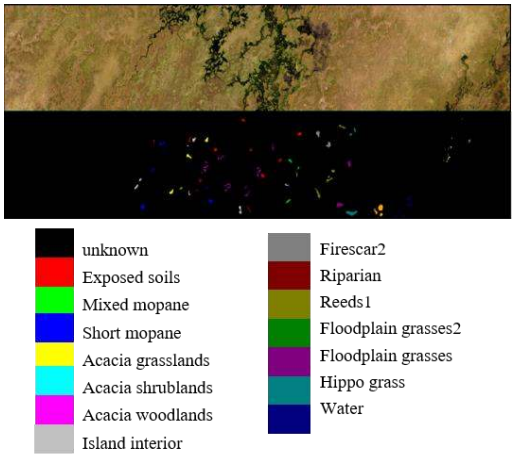
The IP dataset was collected in 1992 over the Indian Pines agriculture experimental area by the AVIRIS sensor. It has a spatial resolution of 145×145 pixels and 220 bands with a wavelength range of 0.4-2.5um. After eliminating the 20 bands contaminated by water vapor, there are 200 bands used for classification. The IP dataset contains 16 labeled material classes.



(a)



(b)



(c)

Fig. 4. False color image and the ground truth maps of the datasets. (a)Indian pines dataset, (b) Salinas dataset, (c)Botswana dataset

The Salinas dataset was captured by the AVIRIS sensor in the SA Valley in California. It covers 512×217 pixels with spatial size of 3.7m and 224 bands across the wavelength of 0.36-2.5um. In addition, 20 bands were abandoned due to the water absorption. In total, 16 labeled material classes are available in the SA dataset.

The Botswana dataset was collected by the NASA EO-1 Hyperion sensor over the Okavango Delta. The dataset is 1476×256 pixels in size with a spatial resolution of 30m per pixel, and 242 bands span the spectral wavelength range of

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JSTARS.2021.3103176, IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing

TABLE II
THE NUMBERS OF TRAINING AND TESTING SAMPLES FOR IP DATA SET

No.	Name	Training	Testing
1	Alfalfa	5	41
2	Corn-notill	143	1285
3	Corn-mintill	83	747
4	Corn	24	213
5	Grass-pasture	48	435
6	Grass-t	73	657
7	Grass-p-m	3	25
8	Hay-w	48	430
9	Oats	2	18
10	Soybean-notill	97	875
11	Soybean-mintill	246	2209
12	Soybean-clean	59	534
13	Wheat	20	185
14	Woods	126	1137
15	Buildings-g-t-d	39	347
16	Stone-s-t	9	84
	Total	1025	9224

TABLE III
THE NUMBERS OF TRAINING AND TESTING SAMPLES FOR SA DATA SET

No.	Name	Training	Testing
1	Brocoli_g_w_1	20	1989
2	Brocoli_g_w_2	37	3689
3	Fallow	20	1956
4	Fallow_r_p	14	1380
5	Fallow_s	27	2651
6	Stubble	39	3920
7	Celery	36	3543
8	Grapes_u	113	11158
9	Soil_v_d	62	6141
10	Corn_s_g_w	33	3245
11	Lettuce_r_4wk	11	1057
12	Lettuce_r_5wk	19	1908
13	Lettuce_r_6wk	9	907
14	Lettuce_r_7wk	11	1059
15	Vinyard_u	72	7196
16	Vinyard_v_t	18	1789
	Total	541	53047

TABLE IV
THE NUMBERS OF TRAINING AND TESTING SAMPLES FOR BT DATA SET

No.	Name	Training	Testing
1	Exposed soils	27	243
2	Mixed mopane	10	91
3	Short mopane	25	226
4	Acacia-g	21	194
5	Acacia-s	27	242
6	Acacia-w	27	242
7	Island-i	26	233
8	Firescars2	20	183
9	Riparian	31	283
10	Reeds1	25	223
11	Floodplain-g2	30	275
12	Floodplain-g	18	163
13	Hippo grass	27	241
14	water	9	86
	Total	323	2925

0.4-2.5um. After removing uncalibrated and noise bands, it contains 14 different categories in total. Fig. 4 (a)-(c) illustrates the false-color image and corresponding ground truth maps of three HSIs datasets.

B. Experimental Setup

All experiments are carried on with a desktop PC with NVIDIA RTX 2060 Super GPU and 64GB RAM. The proposed model DBMFA is implemented by using Pytorch with Python language. We take the overall accuracy (OA), average accuracy (AA), and kappa coefficient (kappa) to evaluate the classification performance quantitatively. Considering the unbalanced categories in three benchmarks, we use the different portions of training samples for each dataset to verify the effectiveness of our proposed model. Specifically, 1% labeled samples are randomly selected in SA dataset as the training set and the remaining 99% samples as the testing set. We randomly choose 10% samples per class for training and 90% for testing for the IP and Botswana dataset. In addition, the batch size and epochs are 16 and 200, respectively. We adopt the Adam to optimize the parameters. The initial learning rate is 0.0025 and decreases by 1% every 50 epochs. We repeat all the experiments five times and average the results in order to avoid errors.

C. Analysis of parameters

The classification performance is based on the proposed model structure and the selection of network parameters. PCA is first used to process the hyperspectral images in order to obtain the C principal components. And then, the input datasets are neighborhood blocks with C×d×s×s centered on the label pixels, where s×s refers to the spatial size of input data. We will elaborate on the analysis of the effects of these parameters.

1) Effect of principal component C

This section examines the effect of varying the number of principal components C on the proposed model's classification performance. We adopt PCA to decrease the dimension of the spectral. C is empirically set to 10, 20, 30, and 40. It can be observed in Fig. 5 (a) that the overall accuracies rise significantly from 10 to 30 and then plateau at 30. However, when the number of principal components exceeds 30, the overall accuracy begins to decline. The phenomenon demonstrates that in a certain degree, the greater the number of principal components, the more detailed the spectral information contained in HSIs. Simultaneously, the neural network can extract additional discriminative features from these components. As the number of principal components continues to increase, the classification performance degrades due to spectral redundancy. In addition, excessive principal components will inevitably generate computational complexity. Therefore, C is set to 30 for all three datasets.

2) Effect of spatial size s×s

In HSIs classification, the spatial size of the image cube means how many pixels are processed simultaneously by the neural network. We select image patches with different sizes to test the classification performance. Specifically, the spatial sizes are varied from 7×7 to 17×17 with the interval of 2. The Overall accuracies of our model classification performance on different spatial sizes are shown in Fig. 5 (b). From the observation of the figure, we can find the 7×7 spatial size has the worst performance, as it is too small to extract sufficient spatial-spectral information for classification. With the continuous expansion of spatial size, the patch contains more

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JSTARS.2021.3103176, IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing

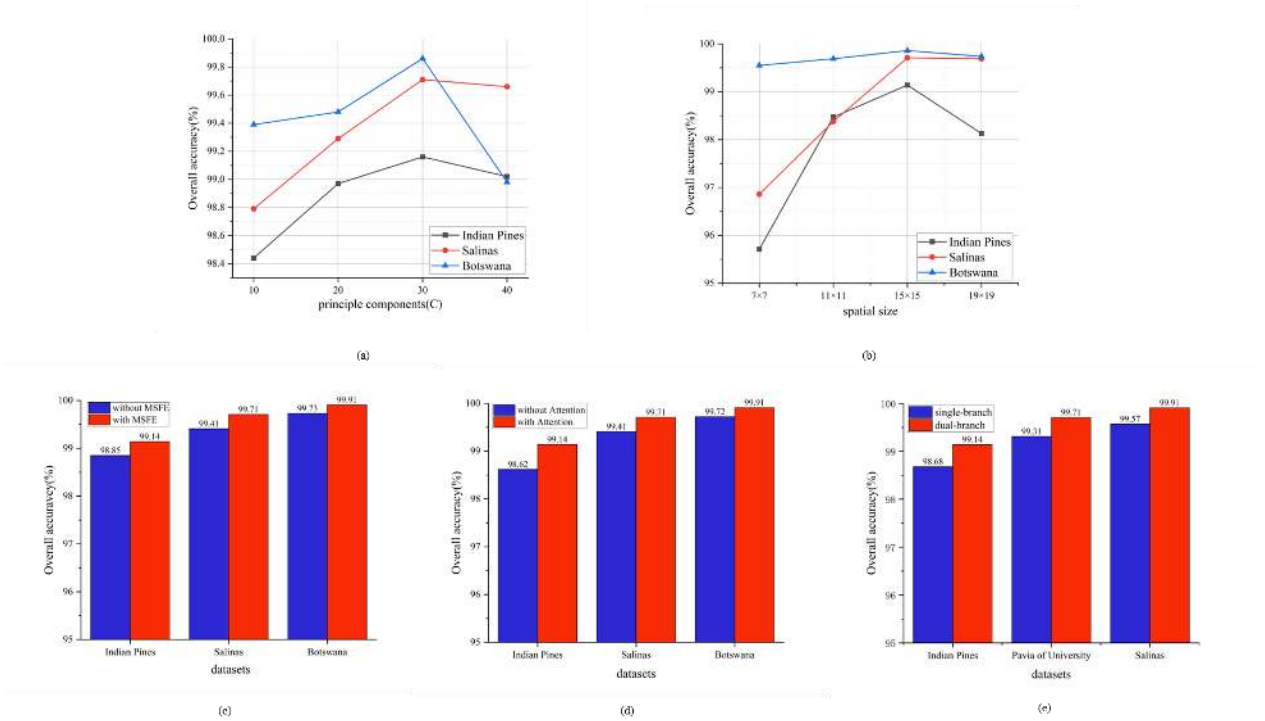


Fig. 5. The results of parameter analysis and ablation study(a)effect of C on overall accuracies on three HSIs datasets, (b) effect of spatial size on overall accuracies on three HSIs datasets, (c)effect of MSFE module on three HSIs datasets, (d)effect of attention block on three HSIs datasets, (e)effect of DBMA module on three HSIs datasets

TABLE V
CLASSIFICATION RESULTS OF THE PROPOSED MODEL WITH DIFFERENT TRAINING RATIOS

Training Ratio		1%	2%	5%	7%	9%	10%	15%
Indian Pines	OA	86.04	90.24	97	97.93	98.62	99.14	99.38
	AA	84.56	90.15	96.89	98.32	97.86	99.01	99.42
	Kappa	90.22	92.62	95.34	98.54	98.04	98.97	99.65
Training Ratio		0.1%	0.2%	0.5%	0.7%	0.9%	1%	1.5%
Salinas	OA	97.18	98.6	99.43	99.47	99.51	99.71	99.75
	AA	96.89	97.85	99.24	98.97	99.35	99.78	99.61
	Kappa	97.56	96.52	98.64	97.86	98.81	99.68	99.58
Training Ratio		1%	2%	5%	7%	9%	10%	15%
Botswana	OA	86.66	92.94	97.3	98.77	99.34	99.91	99.95
	AA	85.32	91.75	96.87	98.54	99.42	99.89	99.92
	Kappa	84.75	92.36	97.68	98.89	99.31	99.92	99.97

discriminative information, and classification performance improves steadily. The peak value for different datasets appears at 15×15. When the spatial size exceeds 15×15, the overall accuracies begin to decline due to the excessive redundant features. As a result, we conclude that either an excessively large or excessively small spatial size is detrimental to classification performance.

D. Impact of training ratio

It is a difficult and time-consuming task for HSIs to find sufficient samples. In this section, we will examine the model classification performance under different training ratios. We randomly select 1%, 2%, 5%, 7%, 9%, 10%, 15% of samples on IP and Botswana datasets. For the SA datasets, the training sets portion is set to 0.1%, 0.2%, 0.5%, 0.7%, 0.9%, 1% and 1.5% of each land-cover category. Table Ⅴ reports the overall accuracies of different ratios of training samples in three datasets. It can be observed that the overall accuracies rise steadily with the increase of the training samples. Simultaneously, our proposed model exhibits robust performance when training samples are insufficient.

E. Ablation Study

In order to demonstrate the effectiveness of the proposed MSFE module, attention block and DBFM module, we design three specific ablation experiments. The models used for comparison are consistent with the network of the proposed method except for the removal of the MSFE module and attention block from the original networks. While for the DBFM, we replace the module with single branch layer-wise 3D CNN. The principal components and the spatial size are set to 30 and 15×15 to guarantee the fairness of the experiments. The overall accuracies of three datasets of comparison models are displayed in Fig .5 (c)-(e). It can be observed that the MSFE module improve the value of overall accuracies by approximately 0.18%-0.41%.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JSTARS.2021.3103176, IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing

TABLE VI
CLASSIFICATION RESULTS OF DIFFERENT METHODS ON THE IP DATA SET

	Methods								
Class	SVM	MLR	RF	1D-CNN	2D-CNN	Hybird-SN	SSRN	A2S2KResNet	Proposed
1	75	46.15	83.33	50	**100**	83.72	**100**	95.45	95.35
2	79.58	73.57	69.78	72.58	83.79	94.12	98.07	**99.29**	98.91
3	74.97	65.38	73.85	81.39	83.33	**99.44**	97.43	98.94	98.93
4	55.67	43.96	64.1	58.93	89.33	97.96	96.37	**100**	98.57
5	81.12	89.73	88.31	91.4	96.73	99.06	98.94	97.23	**100**
6	90.22	93.37	79.14	90.00	98.78	97.68	96.84	97.5	**100**
7	**100**	83.33	**100**	**100**	**100**	66.67	**100**	93.33	**100**
8	94.83	86.56	85.06	91.87	93.56	98.62	97.47	96.25	**100**
9	80	**100**	**100**	**100**	**100**	**100**	**100**	75	**100**
10	72.04	71.79	74.41	76.99	91.21	97.17	95.31	**98.7**	98.61
11	78.64	68.57	63.62	75.5	88.17	97.63	98.5	**99.69**	98.79
12	86.95	56.05	55.64	76.43	90.05	97.62	97.89	96.13	99.25
13	93.37	93.30	89.19	93.85	**100**	97.87	**100**	**100**	99.46
14	91.57	86.86	90.12	91.6	94.48	97.67	97.96	98.92	**99.56**
15	80.51	81.20	60.09	74.19	93.81	99.13	97.05	99.33	**100**
16	97.79	**98.55**	98.53	98.63	98.51	90.41	94.8	94.8	96.43
OA(%)	81.78±0.10	75.49±0.25	73.09±0.19	80.6±0.55	90.27±0.49	97.17±0.52	97.69±0.35	98.63±0.02	**99.14±0.16**
AA(%)	83.26±1.07	77.39±0.43	79.69±0.46	82.71±0.53	93.85±1.69	94.67±1.53	97.91±0.34	96.28±0.13	**98.99±0.15**
Kappa×**100**	79.16±1.41	71.84±0.26	68.75±0.19	77.62±0.59	88.82±0.56	95.32±0.28	97.37±0.48	98.44±0.01	**99.02±0.18**

TABLE VII
CLASSIFICATION RESULTS OF DIFFERENT METHODS ON THE SA DATA SET

	Methods								
Class	SVM	MLR	RF	1D-CNN	2D-CNN	Hybird-SN	SSRN	A2S2KResNet	Proposed
1	99.95	**100**	99.9	**100**	**100**	99.35	**100**	**100**	**100**
2	96.78	98.42	99.54	98.64	**100**	**100**	99.78	**100**	**100**
3	95.5	96.29	83.69	92.48	**100**	97.55	**100**	99.23	**100**
4	99.63	93.28	96.62	97.37	89.59	99.64	95.95	96.88	**100**
5	95.67	92.46	98.65	94.71	99.51	**100**	99.68	99.14	**100**
6	99.85	99.97	99.92	**100**	99.80	**100**	99.36	99.95	99.97
7	98.68	99.69	97.9	99.02	**100**	99.63	**100**	**100**	**100**
8	79.25	76.14	71.8	80.45	87.90	95.42	90.55	92.94	98.93
9	98.9	98.18	94.24	98.21	99.98	99.9	99.75	**100**	**100**
10	93.94	91.33	87.22	97.27	95.86	98.57	99.28	95.82	**99.66**
11	95.36	98.39	87.09	89.96	**100**	80.69	**100**	98.94	99.81
12	97.97	95.31	96.76	96.61	95.88	99.47	99.56	98.95	**100**
13	95.61	94.78	97.03	96.24	95.36	98.04	**100**	99.75	99.45
14	95.36	92.02	84.03	97.28	99.70	98.97	93.78	90.05	**100**
15	73.15	64.13	62.48	74.64	95.43	96.51	96.94	93.83	**99.82**
16	97.78	98.78	96.99	99.94	**100**	96.53	**100**	**100**	99.72
OA(%)	90.25±0.46	89.79±1.12	86.18±0.11	90.81±0.38	95.58±0.58	97.68±0.74	97.09±0.18	97.01±0.67	**99.71±0.35**
AA(%)	94.58±0.38	93.07±0.50	9086±0.35	94.55±0.26	97.43±0.52	97.51±0.58	98.41±0.24	97.84±0.42	**99.83±0.26**
Kappa×100	90.02±0.44	88.62±1.24	84.48±0.13	89.75±0.42	95.07±0.64	97.42±0.63	96.76±0.11	96.67±0.37	**99.67±0.17**

TABLE VIII
CLASSIFICATION RESULTS OF DIFFERENT METHODS ON THE BOTSWANA DATA SET

	Methods								
Class	SVM	MLR	RF	1D-CNN	2D-CNN	Hybird-SN	SSRN	A2S2KResNet	Proposed
1	99.59	**100**	99.59	**100**	98.78	98.76	**100**	**100**	**100**
2	92.86	97.83	81.91	95.56	**100**	**100**	**100**	**100**	**100**
3	98.64	96.97	79.50	99.56	**100**	**100**	**100**	**100**	**100**
4	92.57	92.96	86.14	94.09	94.15	97.97	93.33	97.67	**99.74**
5	87.60	87.45	71.63	87.39	94.98	93.7	93.72	89.79	**99.79**
6	79.83	85.84	63.53	81.86	**100**	97.4	99.47	97.04	99.59
7	**100**	98.29	99.11	**100**	**100**	**100**	**100**	**100**	**100**
8	98.92	92.82	79.91	97.27	**100**	98.38	**100**	**100**	99.46
9	87.68	89.90	77.42	86.96	**100**	94.52	93.7	96.56	**100**
10	86.15	90.45	81.87	87.45	99.11	99.55	**100**	**100**	**100**
11	92.50	94.57	90.41	98.41	**100**	**100**	**100**	**100**	**100**
12	94.48	97.96	95.48	96.64	**100**	**100**	**100**	**100**	**100**
13	92.95	90.70	83.55	89.39	98.37	97.15	**100**	**100**	**100**
14	97.62	98.80	95.40	**100**	96.59	**100**	98.48	**100**	**100**
OA(%)	92.65±0.36	92.40±0.99	84.49±0.72	93.51±0.24	98.63±0.24	98.07±0.18	98.27±0.04	98.31±0.06	**99.91±0.05**
AA(%)	92.95±0.06	93.89±1.24	84.67±0.87	93.89±0.37	98.71±0.35	98.39±0.34	98.48±0.13	98.64±0.11	**99.90±0.04**
Kappa×100	92.05±0.39	91.77±1.07	83.21±0.79	90.12±0.80	98.51±0.46	97.91±0.26	98.19±0.24	98.17±0.13	**99.96±0.05**

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JSTARS.2021.3103176, IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing

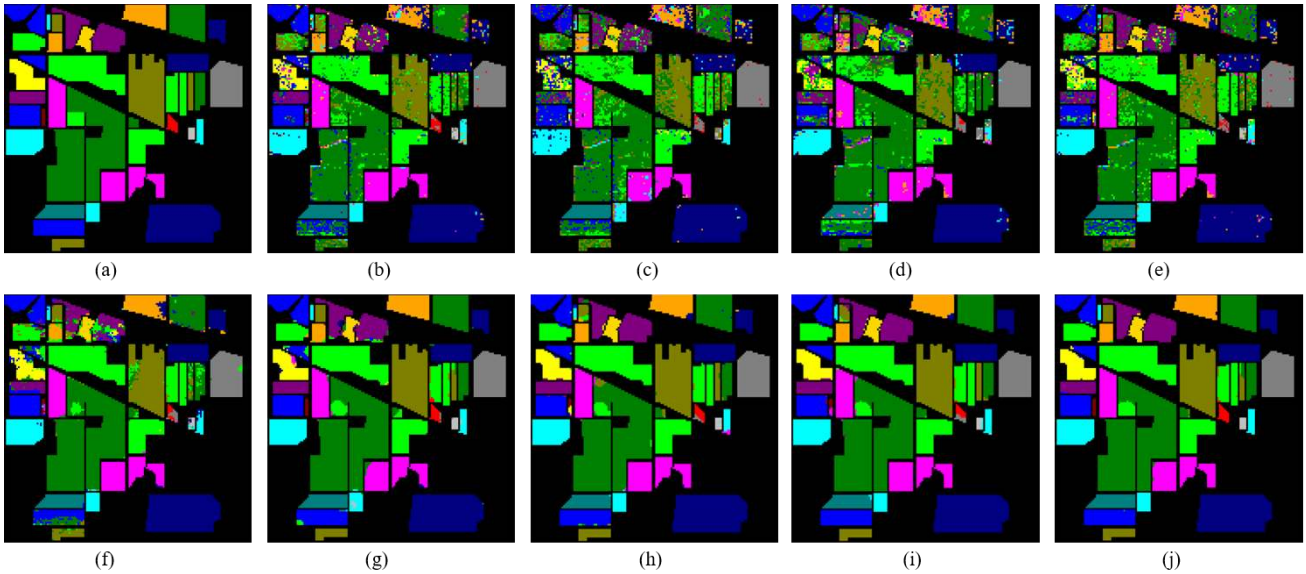


Fig. 6. Classification maps for IP. (a). Ground truth (b)-(j) Predicted classification maps for SVM (OA=81.78%), MLR (OA=75.49%), RF(OA=73.09%), 1D-CNN(OA=80.60%),2D-CNN (90.27%), Hybird-SN (OA=97.17%), SSRN (OA=97.69%), A2S2KResNet (OA=98.63%) and proposed HSMSN-HFF (99.14%).

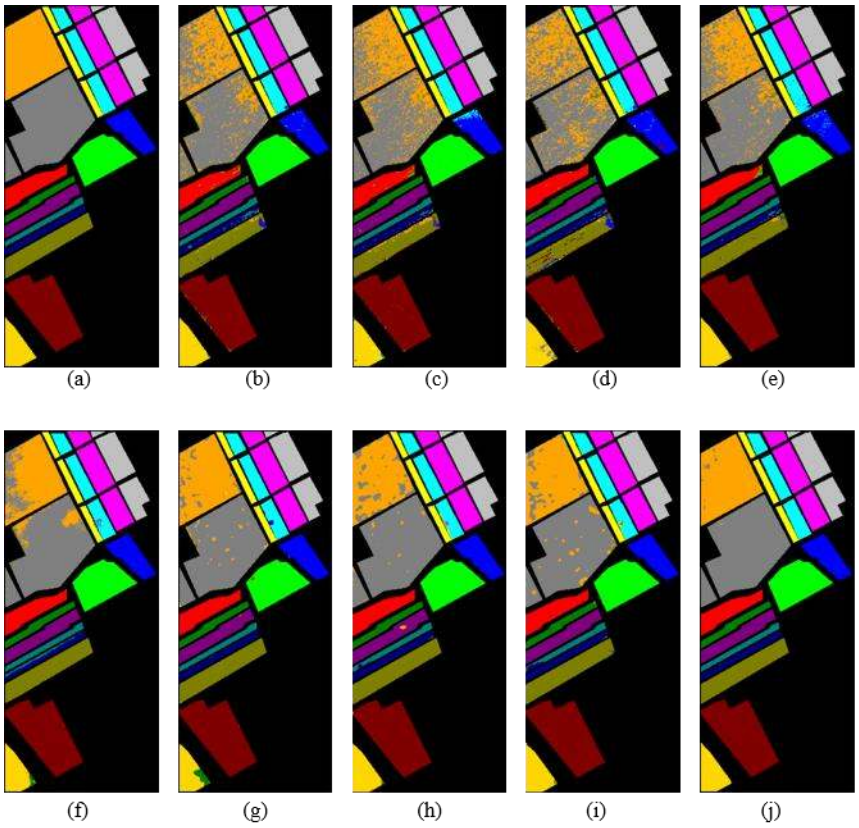


Fig. 7. Classification maps for SA. (a). Ground truth (b)-(j) Predicted classification maps for SVM (OA=90.25%), MLR (OA=89.79%), RF(OA=86.18%), 1D-CNN(OA=90.81%),2D-CNN (OA=95.58%), Hybird-SN (OA=97.68%), SSRN (OA=97.09%), A2S2KResNet (OA=97.01%) and proposed HSMSN-HFF (99.71%).

The reason for the phenomenon is that our MSFE module introduces multiple sizes of kernels to capture the rich spatial-spectral information and more effectively fuse information at different scales. Simultaneously, the model with the attention block achieves higher overall accuracies (approximately 0.19%-0.52%) than the model without the

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JSTARS.2021.3103176, IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing

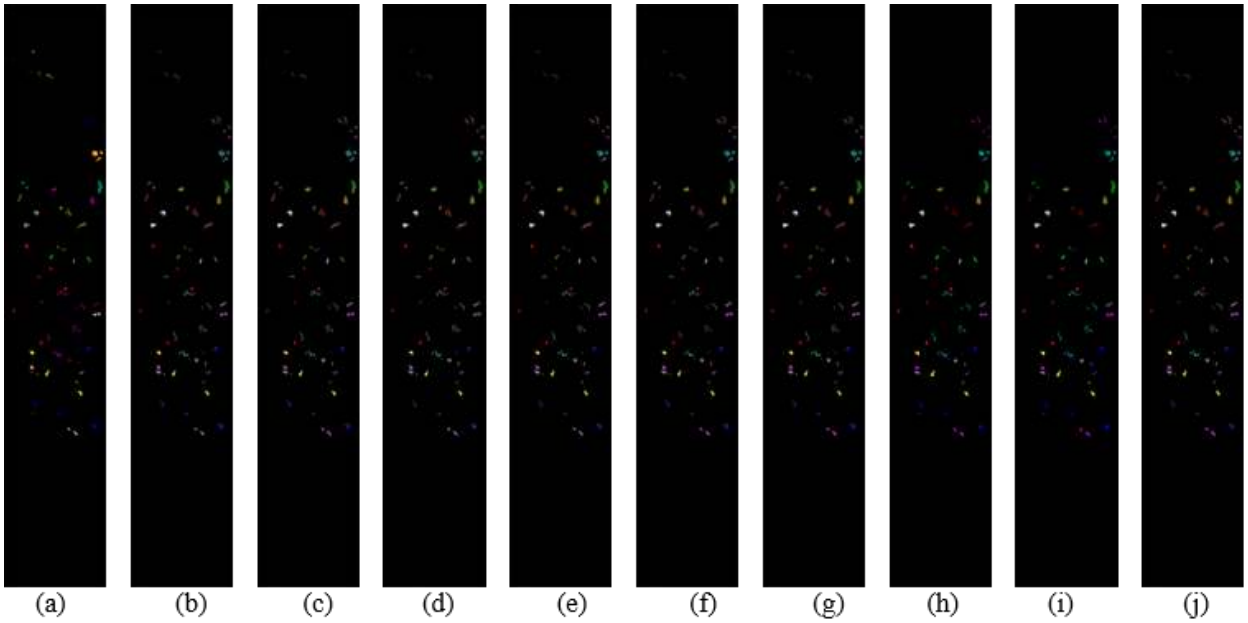


Fig. 8. Classification maps for Botswana. (a). Ground truth (b)-(j) Predicted classification maps for SVM (OA=92.65%), MLR (OA=92.40%), RF(OA=84.49%), 1D-CNN(OA=93.51%),2D-CNN (OA=98.63%), Hybird-SN (OA=98.07%), SSRN (OA=98.27%), A2S2KResNet (OA=98.31%) and proposed HSMSN-HFF (99.91%).

attention block, demonstrating that our proposed attention block can adaptively assign different weights to spatial-channel regions and selectively strengthen valuable features during the HSI feature extraction process. Simultaneously, our proposed DBMA module exhibits the superior performance compared with single-branch 3-D CNN network in three HSIs datasets. DBMA module is composed of multiple filters with additive links and concatenative links, which improves the spatial-spectral feature fusion with low resolution land-cover and captures discriminative features.

F. Compare with different methods

In order to evaluate the performance of our proposed method MSDBFA, we select eight classification methods to compare with our model. The eight methods are support vector machine (SVM) with radial basis function kernel, multinomial logistic regression (MLR), random forest (RF), spectral CNN (CNN-1D), spatial CNN with 2-D kernels, Hybrid-SN [55], SSRN [40], A2S2KResnet [56]. Among them, SVM, RF and MLR are classical machine learning classification methods. They complete the classification using the spectral dimension of HSIs, which has been widely used in previous classification research but has low accuracies. In order to highlight the progressiveness of our method, we also employ a variety of deep learning-based methods. 1-D CNN is an early neural network model using spectral dimension to classify hyperspectral images; 2-D CNN classifies based on spatial features; SSRN is a classical spatial-spectral classification model that incorporates residual connections to mitigate gradient disappearance and shorten training time. Hybrid-SN creatively utilizes the 3-D and 2-D convolution to explore the shallow and deep features respectively in HSIs. A2S2kResnet introduces an adaptive spectral-spatial kernel improved residual network with spectral attention for the purpose of capturing discriminative spectral-spatial features in HSIs.

TABLE IX
TRAINABLE PARAMETERS, FLOPS AND TRAINING TIMES OF DIFFERENT MODELS FOR IP DATA SET

Methods	SSRN	Hybird-SN	A2S2KResnet	MSDBFA
TTPs(M)	36.4	512.2	37.1	46.5
FLOPS(G)	7.022	7.971	5.454	4.642
Training Time(s)	702.5	164.6	1149.6	366.2

In order to ensure the fairness of experiments, the spatial size and the number of principal components are set to 15×15 and 30 for all DL methods respectively. Due to the fact that SSRN does not follow the PCA as described in the original paper, so we do not apply PCA operation to the SSRN model. Other parameters of the network are configured according to their papers. Our proposed method outperforms the other methods by approximately 0.51%-26.05% in terms of OA, 0.5%-40.78% in terms of AA, and 0.58% -30.27% in terms of Kappa in IP dataset. The sample distribution is extremely unbalanced across the IP dataset's various classes. The class of Alfalfa, Grass-p-m, and Oats, for example, have only 46, 28, and 20 samples per class, respectively. That is a great challenge for the HSIs classification resulting in problems with unbalanced sample training. Notably, our proposed method achieves 100% overall accuracies on the grass-pasture, grass-t, grass-p-m, hay-w, and oats classes. By comparison, SVM, RF, and MLR have relatively poor classification performance. To be precise, the SVM classifier has the highest overall accuracy among the three machine learning methods. The MLR classifier's values fall precipitously when dealing with small sample sizes. While RF classifier performs the worst (OA=76.09%). Comparatively, some DL classification methods have superior performance; for example, the

A2S2Resnet method achieves the best classification results with 98.63 % value of OA among all the comparative methods.

For the SA and Botswana datasets, our proposed model MSDBFA also achieves the highest value of OA. Among all the classification method, RF model performs the worst again, indicating that the random forest algorithm cannot deal well with the complex spatial-spectral features in HSIs. At the same time, some simple deep learning methods such as 1-D CNN and 2-D CNN, the classification accuracy has been significantly improved compared with the conventional machine learning methods. However, they still have their own limitations. For example, 1-D CNN utilizes the redundant spectral information to complete the classification, which is bounded to be affected by the Hughes phenomenon. 2-D CNN relies on the spatial distribution and characteristics of ground objects to classify, whereas ignoring the characteristics of rich spectral of HSIs. For many recent deep learning methods that employ spatial-spectral feature fusion strategies, including SSRN, Hybrid-SN, A2S2KResnet, they generally outperform the former methods (1D-CNN,2D-CNN), especially when the number of training samples is relatively small. It's demonstrated that in the case of a limited number of training samples, the hierarchical fusion mechanism can combine the complementary and relevant information from the output of distinct convolutional layers, making the extracted features more effective for classification. Additionally, to evaluate the computational cost and complexity of the proposed model. Table IX summarizes the total trainable parameters (TTP), floating-point operations (FLOPs), and training times for various models with the IP data set. As can be seen, Hybrid-SN has the largest parameters and highest FLOPs compared with other methods, owning to its large kernel filters and batch size. A2S2KResNet has approximately the same number of parameters as SSRN, but SSRN takes higher FLOPs due to the fact that it does not use PCA to reduce the spectral dimension of the HSIs as conventional methods do. Instead, it utilizes a 3-D kernel filter to squeeze the dimension hierarchy. Our method has the lowest FLOPs due to the lightweight multiscale extraction module and effective shuffle attention block. In terms of training time, our method is only slightly longer than the Hybrid-SN, but achieves significantly better classification performance in all three datasets. On the whole, the MSDBFA model outperforms other methods in terms of both classification performance and computational cost.

IV. Conclusion

In this article, a novel multiscale dual-branch feature fusion and attention network has been proposed. Specifically, we propose a multiscale feature extraction module (MSFE) by constructing multiple residual-like connections, thus the structure of the module can obtain multiscale features at a granular level. Moreover, we design the dual-branch feature fusion interactive module (DBFM) to complete the deep fusion of spatial-spectral features via concatenative and additive links, which can not only enhance the feature reuse at shallow level but also explore new discriminative information from the fused spatial-spectral features. In addition, we introduce a novel shuffle attention block to improve performance over the network by creatively altering the conventional weight distribution method in channel and spatial dimensions, thereby enhancing the representation ability of the feature map. The obtained results on three HSIs datasets reveal that our proposed MSFDBA model provides competitive results compared to the other state-of-the-art approaches for classification performance.

References

[1] X. Zhang, Y. Sun, K. Shang, L. Zhang, and S. Wang, "Crop classification based on feature band set construction and object-oriented approach using hyperspectral images," IEEE J. Sel. Topics Appl. Earth Observat., Remote Sens., vol. 9, no. 9, pp. 4117–4128, Sep. 2016.

[2] X. Yang and Y. Yu, "Estimating soil salinity under various moisture conditions: An experimental study," IEEE Trans. Geosci. Remote Sens., vol. 55, no. 5, pp. 2525–2533, May 2017.

[3] Q. Li, and Z. Liu, "Tongue color analysis and discrimination based on hyperspectral images," Computerized Medical Imaging and Graphics, vol. 33, no. 3, pp. 217-221, 2009.

[4] Z. Liu, J. Yan, D. Zhang, and Q. Li, "Automated tongue segmentation in hyperspectral images for medicine," Applied Optics, vol. 46, no. 34, pp. 8328-8334, 2007.

[5] A. Brook, E. Ben-Dor and R. Richter, "Fusion of hyperspectral images and LiDAR data for civil engineering structure monitoring", 2nd Workshop Hyperspectral Image and Signal Processing: Evolution in Remote Sensing (WHISPERS), pp. 1-5, 2010-Jun.

[6] N. Acito and M. Diani, "Unsupervised Atmospheric Compensation of airborne hyperspectral images in the VNIR spectral range", IEEE Trans. Geosci. Remote Sens., vol. 56, no. 7, pp. 3924-3940, 2018.

[7] J. Behmann, J. Steinrücken and L. Plümer, "Detection of early plant stress responses in hyperspectral images", ISPRS J. Int. Soc. Photo. Remote Sens., vol. 93, pp. 98-111, July 2014.

[8] Y. Tarabalka, M. Fauvel, J. Chanussot and J. A. Benediktsson, "SVM and MRF-based method for accurate classification of hyperspectral images", IEEE Geosci. Remote Sens. Lett., vol. 7, no. 4, pp. 736-740, Oct. 2010.

[9] J. Xia, P. Ghamisi, N. Yokoya and A. Iwasaki, "Random forest ensembles and extended multiextinction profiles for hyperspectral image classification", IEEE Trans. Geosci. Remote Sens., vol. 56, pp. 1-216, 2018.

[10] J. Li, J. M. Bioucas-Dias and A. Plaza, "Semisupervised hyperspectral image classification using soft sparse multinomial logistic regression", IEEE Geosci. Remote Sens. Lett., vol. 10, no. 2, pp. 318-322, Mar. 2013.

[11] M. C. Alonso, J. A. Malpica and A. Martínez de Agirre, "Consequences of the Hughes phenomenon on some classification Techniques", Annu. Conf. (ASPRS), May 2011.

[12] J. Ham, D. Lee, S. Mika and B. Schölkopf, "A Kernel View of the Dimensionality Reduction of Manifolds", Proc. Int'l Conf. Machine Learning, pp. 47-54, 2004.

[13] F. Palsson, J. R. Sveinsson, M. O. Ulfarsson and J. A. Benediktsson, "Model-based fusion of multi- and hyperspectral images using PCA and wavelets", IEEE Trans. Geosci. Remote Sens., vol. 53, no. 5, pp. 2652-2663, May 2015.

[14] T. V. Bandos, L. Bruzzone, and G. Camps-Valls, "Classification of hyperspectral images with regularized linear discriminant analysis," IEEE Trans. Geosci. Remote Sens., vol. 47, no. 3, pp. 862–873, Mar. 2009.

[15] D. Hong, N. Yokoya, J. Chanussot and X. X. Zhu, "An augmented linear mixing model to address spectral variability for hyperspectral unmixing", IEEE Trans. Image Process., vol. 28, no. 4, pp. 1923-1938, 2019.

[16] Y. Li, H. Zhang, X. Xue, Y. Jiang and Q. Shen, "Deep learning for remote sensing image classification: A survey", WIREs Data Mining Knowl. Discovery, vol. 8, no. 6, pp. e1264-e1280, May 2018.

[17] D. Zhang, J. Han, G. Cheng, Z. Liu, S. Bu and L. Guo, "Weakly supervised learning for target detection in remote sensing images", IEEE Geosci. Remote Sens. Lett., vol. 12, no. 4, pp. 701-705, Apr. 2015.

[18] R. Pierdicca, M. Paolanti, F. Matrone, M. Martini, C. Morbidoni, E. S. Malinverni, et al., "Point cloud semantic segmentation using a deep learning framework for cultural heritage", Remote Sens., vol. 12, no. 6, pp. 1005, Mar. 2020.

[19] Y. Chen, Z. Lin, Z. Xing, W. Gang, and Y. Gu, “Deep learning-based classification of hyperspectral data,” IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens., vol. 7, no. 6, pp. 2094–2107, Jun. 2014.
[20] D. Hong, W. He, N. Yokoya, J. Yao, L. Gao, L. Zhang, and X. X. Zhu, “Interpretable Hyperspectral AI: When Non-Convex Modeling meets Hyperspectral Remote Sensing”. IEEE Geoscience and Remote Sensing Magazine, 2021, DOI: 10.1109/MGRS.2021.3064051
[21] D. Hong et al., "More diverse means better: Multimodal deep learning meets remote-sensing imagery classification", IEEE Trans. Geosci. Remote Sens., Aug. 2020.
[22] D. Hong, L. Gao, J. Yao, B. Zhang, A. Plaza and J. Chanussot, "Graph convolutional networks for hyperspectral image classification", IEEE Trans. Geosci. Remote Sens., Aug. 2020.
[23] L. Gao, D. Hong, J. Yao, B. Zhang, P. Gamba and J. Chanussot, "Spectral Superresolution of multispectral imagery with joint sparse and low-rank learning", IEEE Trans. Geosci. Remote Sens., Jun. 2020.
[24] Z. Lei, Y. Zeng, P. Liu and X. Su, "Active Deep Learning for Hyperspect ral Image Classification With Uncertainty Learning," in IEEE Geoscienc e and Remote Sensing Letters, doi: 10.1109/LGRS.2020.3045437.
[25] Sellami A, Farah I R. Spectra-spatial graph-based deep restricted boltzmann networks for hyperspectral image classification[C]//2019 PhotonIcs & Electromagnetics Research Symposium-Spring (PIERS-Spring). IEEE, 2019: 1055-1062.
[26] Y. Chen, X. Zhao, and X. Jia, “Spectral–spatial classification of hyperspectral data based on deep belief network,” IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens., vol. 8, no. 6, pp. 2381–2392, Jun. 2015.
[27] X. Yuan, B. Huang, Y. Wang, C. Yang, and W. Gui, “Deep learning-based feature representation and its application for soft sensor modeling with variable-wise weighted SAE,” IEEE Trans. Ind. Informat., vol. 14, no. 7, pp. 3235–3243, Jul. 2018.
[28] K. Makantasis, K. Karantzalos, A. Doulamis and N. Doulamis, "Deep supervised learning for hyperspectral data classification through convolutional neural networks", pp. 4959-4962, 2015.
[29] J. Li, P. R. Marpu, A. Plaza, J. M. Bioucas-Dias, and J. A. Benediktsson, “Generalized composite kernel framework for hyperspectral image classification,” IEEE Trans. Geosci. Remote Sens., vol. 51, no. 9, pp. 4816–4829, Sep. 2013.
[30] L. Fang, S. Li, W. Duan, J. Ren, and J. A. Benediktsson, “Classification of hyperspectral images by exploiting spectral–spatial information of superpixel via multiple kernels,” IEEE Trans. Geosci. Remote Sens., vol. 53, no. 12, pp. 6663–6674, Dec. 2015.
[31] F. Luo, B. Du, L. Zhang, L. Zhang, and D. Tao, “Feature learning using spatial–spectral hypergraph discriminant analysis for hyperspectral image,” IEEE Trans. Cybern., vol. 49, no. 7, pp. 2406–2419, Jul. 2019.
[32] W. Li, G. Wu, F. Zhang, and Q. Du, “Hyperspectral image classification using deep pixel-pair features,” IEEE Trans. Geosci. Remote Sens., vol. 55, no. 2, pp. 844–853, Feb. 2017.
[33] B. Pan, Z. Shi and X. Xu, "MugNet: Deep learning for hyperspectral image classification using limited samples", ISPRS J. Photogramm. Remote Sens., vol. 145, pp. 108-119, Nov. 2018.
[34] X. Cao, J. Yao, Z. Xu and D. Meng, "Hyperspectral image classification with convolutional neural network and active learning", IEEE Trans. Geosci. Remote Sens., vol. 58, no. 7, pp. 4604-4616, Jul. 2020.
[35] A. J. X. Guo and F. Zhu, “A CNN-based spatial feature fusion algorithm for hyperspectral imagery classification,” IEEE Trans. Geosci. Remote Sens., vol. 57, no. 9, pp. 7170–7181, Sep. 2019.
[36] R. Li, S. Zheng, C. Duan, Y. Yang and X. Wang, "Classification of hyperspectral image based on double-branch dual-attention mechanism network", Remote Sens., vol. 12, no. 3, pp. 582, Feb. 2020.
[37] Q. Liu, F. Zhou, R. Hang and X. Yuan, "Bidirectional-convolutional LSTM based spectral-spatial feature learning for hyperspectral image classification", Remote Sens., vol. 9, no. 12, pp. 1330, 2017.
[38] J. Nickolls and W. J. Dally, "The GPU computing era", Micro IEEE, vol. 30, no. 2, pp. 56-69, Mar.-Apr. 2010.
[39] K. He, X. Zhang, S. Ren and J. Sun, "Deep residual learning for image recognition", arXiv:1512.03385, 2015.
[40] Z. Zhong, J. Li, Z. Luo and M. Chapman, "Spectral–spatial residual network for hyperspectral image classification: A 3-D deep learning framework", IEEE Trans. Geosci. Remote Sens., vol. 56, no. 2, pp. 847-856, Feb. 2018.
[41] H.Lee and H.Kwon,“Going deeper with contextual CNN for hyperspectral image classification,” IEEE Trans. Image Process., vol. 26, no. 10, pp. 4843–4855, Oct. 2017.
[42] W. Song, S. Li, L. Fang, and T. Lu, “Hyperspectral image classification with deep feature fusion network,” IEEE Trans. Geosci. Remote Sens., vol. 56, no. 6, pp. 3173–3184, Jun. 2018.
[43] M. E. Paoletti, J. M. Haut, J. Plaza and A. Plaza, "Deep&dense convolutional neural network for hyperspectral image classification", Remote Sens., vol. 10, no. 9, pp. 1454, 2018, [online] Available: http://www.mdpi.com/2072-4292/10/9/1454.
[44] R. Xu, Y. Tao, Z. Lu and Y. Zhong, "Attention-mechanism-containing neural networks for high-resolution remote sensing image classification", Remote Sens., vol. 10, no. 10, pp. p. 1602, 2018.
[45] R. A. Jarvis, "A perspective on range finding techniques for computer vision", IEEE Trans. Pattern Anal. Mach. Intell., vol. PAMI-5, no. 2, pp. 122-139, Mar. 1983.
[46] N. Komodakis, N. Paragios and G. Tziritas, "MRF energy minimization and beyond via dual decomposition", IEEE Trans. Pattern Anal. Mach. Intell., vol. 33, no. 3, pp. 531-552, Jan. 2011.
[47] Z. Deng, H. Sun, S. Zhou, J. Zhao, L. Lei and H. Zou, "Multi-scale object detection in remote sensing imagery with convolutional neural networks", ISPRS J. Photogramm. Remote Sens., vol. 145, pp. 3-22, Nov. 2018.
[48] K. He, X. Zhang, S. Ren and J. Sun, "Spatial pyramid pooling in deep convolutional networks for visual recognition", Proc. 13th Eur. Conf. Comput. Vis., pp. 346-361, 2014.
[49] T.-Y. Lin, P. Dollar, R. Girshick, K. He, B. Hariharan and S. Belongie, "Feature pyramid networks for object detection", Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), pp. 2117-2125, Jul. 2017.
[50] G. Huang, Z. Liu, K. Q. Weinberger and L. van der Maaten, "Densely connected convolutional networks", Proc. IEEE Conf. Comput. Vis. Pattern Recognit., pp. 4700-4708, 2017.
[51] J. Hu, L. Shen and G. Sun, "Squeeze-and-excitation networks", Proc. Conf. Comput. Vis. Pattern Recognit. (CVPR), pp. 7132-7141, 2018.
[52] S. Woo, J. Park, J.-Y. Lee and I. S. Kweon, "CBAM: Convolutional block attention module", Proc. Eur. Conf. Comput. Vis. (ECCV), pp. 3-19, Sep. 2018.
[53] Y. Cao, J. Xu, S. Lin, F. Wei and H. Hu, "GCNet: Non-local networks meet squeeze-excitation networks and beyond", arXiv:1904.11492, Apr. 2019, [online] Available: https://arxiv.org/abs/1904.11492.
[54] Q. Zhang and Y. Yang. Sa-net: "Shuffle attention for deep convolutional neural networks. In International Conference on Acoustics", Speech and Signal Processing, 2021.
[55] S. K. Roy, G. Krishna, S. R. Dubey and B. B. Chaudhuri, "HybridSN: Exploring 3-D–2-D CNN Feature Hierarchy for Hyperspectral Image Classification," in IEEE Geoscience and Remote Sensing Letters, vol. 17, no. 2, pp. 277-281, Feb. 2020, doi: 10.1109/LGRS.2019.2918719.
[56] S. K. Roy, S. Manna, T. Song and L. Bruzzone, "Attention-Based Adaptive Spectral-Spatial Kernel ResNet for Hyperspectral Image Classification," in IEEE Transactions on Geoscience and Remote Sensing, doi: 10.1109/TGRS.2020.3043267.

Hongmin Gao (M’21) received the Ph.D. degree in computer application technology from Hohai University, Nanjing, China, in 2014.

He is currently a Professor with the College of Computer and Information, Hohai University. His research interests include deep learning, information fusion, and image processing in remote sensing.

Yiyan Zhang (S’21) received the B.S. degree in Internet of Things engineering from Jiangsu University of Technology, Changzhou, China, in 2020.

He is a Graduate Student with the College of Computer and Information, Hohai University. His research interests include deep learning and image processing.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JSTARS.2021.3103176, IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing

Zhonghao Chen (S'21) received the B.S. degree in electronics and information engineering from West Anhui University, Luan, China, in 2019.

He is a Graduate Student with the College of Computer and Information, Hohai University. His research interests include deep learning and image processing.

Chenming Li received the B.S., M.S., and Ph.D. degrees in computer application technology from Hohai University, Nanjing, China, in 1993, 2003, and 2010, respectively.

He is a Professor and the Deputy Dean of the College of Computer and Information, Hohai University. His research interests include information processing systems and applications, system modeling and simulation, multisensor systems, and information processing.

Dr. Li is a Senior Member of the China Computer Federation and the Chinese Institute of Electronics.