Spatial–Spectral Split Attention Residual Network for Hyperspectral Image Classification

Zhenqiu Shu, Zigao Liu, Jun Zhou, Songze Tang, Zhengtao Yu, and Xiao-Jun Wu

***Abstract*—In the past few years, many convolutional neural networks (CNNs) have been applied to hyperspectral image (HSI) classification. However, many of them have the following drawbacks: they do not fully consider the abundant band spectral information and insufficiently extract the spatial information of HSI; all bands and neighboring pixels are treated equally, so CNNs may learn features from redundant or useless bands/pixels; and a significant amount of hidden semantic information is lost when a single-scale convolution kernel is used in CNNs. To alleviate these problems, we propose a spatial–spectral split attention residual networks (S^3ARN) for HSI classification. In S^3ARN, a split attention strategy is used to fuse the features extracted from multireceptive fields, in which both spectral and spatial split attention modules are composed of bottleneck residual blocks. Thanks to the bottleneck structure, the proposed method can effectively prevent overfitting, speeds up the model training, and reduces the network parameters. Moreover, the spectral and spatial attention residual branches aim to generate the attention masks, which can simultaneously emphasize useful bands and neighbor pixels and suppress useless ones. Experimental results on three benchmark datasets demonstrate the effectiveness of the proposed model for HSI classification.**

***Index Terms*—Attention masks, bottleneck residual, channel attention, hyperspectral image classification (HSIC), spatial attention, split attention.**

I. Introduction

THE development of sensor technology has enabled the collection of abundant spectral information in images. The resulting hyperspectral image (HSI) can be represented as a 3-D tensor with high number of spectral bands, which can be used to predict the class of ground objects. HSI classification (HSIC) aims at predicting the class label for each pixel in the image. It has been widely used in remote sensing applications, such as agricultural monitoring, meteorological analysis, and data mining [1], [2], [3], [4], [5].

In the past few decades, many spectral-based classification approaches have been proposed for HSI. In traditional HSIC approaches, such as support vector machines (SVMs) [6], random forest (RF) [7], and *K*-nearest neighbors (KNN) [8], the spectral data are directly used as the input to the classifiers. More recent, deep neural networks have been developed for HSIC due to their effectiveness in image classification. They can automatically extract the high-level semantic features of each pixel from the spectral information. Some classic spectral-based deep neural network models, such as recurrent neural networks (RNN) [9], stacked autoencoders (SAE) [10], 1-D convolutional neural networks (CNN) [11], and deep belief network (DBN) [12] have been applied for HSIC. However, all of these methods only utilize the spectral information of HSI and missed the spatial information.

To address this problem, spectral–spatial CNNs have been developed for HSIC, which can effectively extract both spectral and spatial features from HSI. However, given the central pixels, some pixels in the neighborhood may be useless or even reduce the classification performance in some cases [13]. Therefore, the neighboring pixels in the patch should be assigned different weights to improve the classification performance. Although CNNs can automatically learn the contribution of neighboring pixels with the convolution kernels [14], the convolution kernel may not sufficiently learn the spatial information due to limited receptive fields. A consensus is to use multilayer convolution to learn the spatial features. Therefore, several 2-D and 3-D CNNs have been proposed [15], [16], [17], [18], [19], [20], [21], which process 3-D image patches directly and extract the spatial context information more effectively. Feng et al. [22] introduced self-supervised learning into generative adversarial network to exploit the rich information in unlabeled samples. Besides, multibranch generators and discriminators alleviate the pattern collapse problem and improve its classification ability. Recently, the spatial–spectral CNNs show a powerful ability in HSI feature extraction and have been widely applied to HSIC. Some studies [23], [24] show that the classification performance of the neural network can be improved by adopting the metric learning and expanding its depth. In general, the classification results become better with an increase level of semantic

Manuscript received 10 July 2022; revised 4 September 2022 and 31 October 2022; accepted 28 November 2022. Date of publication 1 December 2022; date of current version 15 December 2022. This work was supported in part by the National Natural Science Foundation of China under Grant 61603159, Grant 62162033, and Grant U21B2027; in part by Yun nan Provincial Major Science and Technology Special Plan Projects under Grant 202002AD080001 and Grant 202103AA080015; in part by Yun nan Foundation Research Projects under Grant 202101AT070438 and Grant 202101BE070001-056; and in part by the Qing Lan Project of Higher Education of Jiangsu Province. *(Corresponding author: Zhenqiu Shu.)*

Zhenqiu Shu, Zigao Liu, and Zhengtao Yu are with the Faculty of Information Engineering and Automation, Kunming University of Science and Technology, Kunming 650032, China (e-mail: shuzhenqiu@163.com; 1322502591@qq.com; ztyu@hotmail.com).

Jun Zhou is with the School of Information and Communication Technology, Griffith University, Nathan, QLD 4111, Australia (e-mail: jun.zhou@griffith.edu.au).

Songze Tang is with the Department of Criminal Science and Technology, Nanjing Forest Police College, Nanjing 210042, China (e-mail: tangsz@nfpc.edu.cn).

Xiao-Jun Wu is with the Jiangsu Provincial Engineering Laboratory of Pattern Recognition and Computational Intelligence, Jiangnan University, Wuxi 214126, China (e-mail: wu_xiaojun@jiangnan.edu.cn).

Digital Object Identifier 10.1109/JSTARS.2022.3225928

information in CNNs [25]. However, the semantic information hidden in HSI may still be distorted with the increase of stacking layers, resulting in the loss of detailed information on target objects. Moreover, the training of the deep network is very computationally intensive and leads to network degradation [26], [27]. To address these issues, He et al. [28] introduced the residual network (ResNet) framework, which brings the network to an unprecedented depth while avoiding network degradation. In the past few years, Some variants of the ResNet [29], [30], [31] have been proposed to deal with the classification of HSI.

The aforementioned methods believe that features from different spatial coordinates and different bands make the same contribution to the classification task in HSIC. Motivated by the human visual theory, the attention mechanism was proposed to selectively learn significant targets in image sample [32], [33]. In HSIC tasks, it selects the spectral and spatial information that is useful for classification and discards the unnecessary parts. Recently, Xiang et al. [34] designed a coordinate attention mechanism, in which the generated mask is encoded separately into a pair of horizontal-wise and vertical-wise attention vectors that represent the weights of the corresponding positions. Yu et al. [35] produced the attention map in a feedback manner. The squeeze-and-excitation (SE) module [36], [37] introduced the global pooling to derive the channel descriptors of a feature map, and then, added the nonlinear information between those features, which better accommodated the complex correlation between channels. In the spectral grouping and the integration module, the spectral attention mask aimed to integrate the group features [38]. Unlike attention mechanisms, Zhang et al. [39] proposed the domain feature enhancement network (DFENet), which builds the dependencies between channels and spaces to calibrate global and local features using a dynamic self-gating mechanism. Feng et al. [40] proposed an end-to-end CNN framework based on bandwise-independent convolution and hard thresholding for band selection of HSIs.

Previous studies show that the features extracted by convolution operations with different receptive fields are usually complementary to each other [41]. Inception combines the results of several convolutional operations to achieve excellent accuracy in image classification [42], [43]. To extract the spatial information from the multireceptive fields, Li et al. [44] proposed a hybrid dilated convolution stacked with various dilation rates. Gao et al. [31] introduced a multiscale residual network (MSRN) method to extract multiscale spatial features. Tan et al. [45] combined multisize kernels into a single convolutional operation using mixed depth-wise convolution operations. However, these methods add the features of multiple receptive fields directly, rather than integrate them selectively. Split-attention network (ResNeSt) [46] was developed to adaptively select different network branch features using channel-wise attention. It explored the cross-channel feature correlations and learned diverse representations, and demonstrated potential for neural architecture search and classification.

In this article, we propose a novel spatial–spectral split attention residual network (S^3ARN) for HSIC. By integrating the split attention mechanism (SAM), our S^3ARN model can adaptively select and fuse the features extracted by the kernels with different receptive fields. The main contributions of this article can be summarized as follows.

1) We propose a two-branch spectral–spatial attention network for HSIC. Compared with the existing CNNs, our S^3ARN model adopts three attention mechanisms, such as split attention, spectral attention, and spatial attention, to effectively capture the spatial–spectral feature information of HSI.
2) The proposed S^3ARN model integrates the spatial and spectral split attention residual blocks (spatial block and spectral block) to comprehensively learn the feature information by incorporating SAM into each bottleneck residual unit. It can extract the high-level semantic features accurately by combining different receptive field features, and can effectively improve the classification performance. Compared to the existing selective kernel attention modules, our spatial and spectral split attention residual block can effectively capture the difference between feature channels, and thus, models multiscale spatial–spectral features at a finer level.
3) Two attention residual branches integrate the spatial coordinate attention mechanism and the spectral channels attention mechanism, respectively, which makes it concentrate on the spectral and spatial information that are beneficial to the classification task and suppress other unnecessary parts, simultaneously. An adaptive weight method is used to merge the results from two branches during the fusion phase. Unlike the existing spatial–spectral attention structures, the proposed attention residual branches recalibrate the data and can link back to the trunk in a conservative residual approach, which is beneficial for preserving intrinsic properties of trunk information.
4) We conducted experiments on three benchmark datasets to compare the proposed S^3ARN method with several state-of-the-art methods. The results show the advantages of our method over the alternatives.

The remainder of this article is organized as follows. In Section II, we overview the relative works. Section III illustrates the details of the proposed method. Experiment results are shown in Section IV. Finally, Section V concludes this article.

II. Related Works

A. Residual Network (ResNet)

The ResNet [28] was proposed to address the network degradation issue with the increase of network depth [47]. The ResNet introduces an identity shortcut connection in a deep neural network and allows the deep network to learn improved feature representations. The idea of the ResNet makes the neural network deeper and more efficient. Fig. 1 shows the residual unit and the bottleneck residual unit, which are the two most representative components in residual networks.

The output feature map of the residual unit can be obtained by the following formula:

$$y = f(x) + x \tag{1}$$

where x and y represent the input and the output of the residual unit, respectively. The mapping pattern of the shortcut

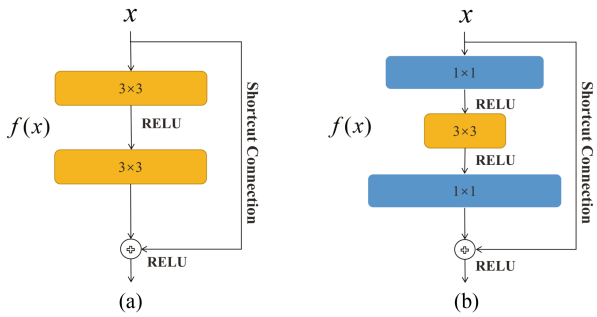

Fig. 1. Structures of residual unit and bottleneck residual unit. (a) Residual Unit. (b) Bottleneck Residual Unit.

connection is identity mapping, which directly transports the features learned in the shallower network. $f(x)$ denotes the residual mapping learning operation, which reflects the residuals information between the truth and x. Using residual learning, we may learn features that have not been extracted by the shallow network. Obviously, we can obtain the feature map y by adding the elements of both the residual features and the shortcut connection. Therefore, we can construct a deeper network than traditional deep networks by explicitly constructing the identity connection.

In the ResNet, the main parts are composed of residual convolution units, pooling layers, and fully connected layers. A series of experimental results have shown that the addition of residual connections can accelerate the convergence of the model and delay the network degradation. With the depth growth of the ResNet, its classification accuracy can be also improved in most cases. However, the ResNet needs to become larger in terms of the number of feature channels to learn more data patterns. Therefore, the computational complexity of the ResNet also increases exponentially with the increase of feature channels. Fig. 1(b) shows the bottleneck residual unit, which significantly reduces the number of parameters without the performance loss. This setting of the bottleneck residual unit includes the following three steps:

1) downscale the features of samples;
2) perform feature extraction;
3) project the features back into the original dimension.

As a result, the number of large convolution kernels can be reduced significantly, resulting in a lightweight unit. Therefore, the ResNet can construct a network structure with more than 1000 layers based on the bottleneck residual units [28].

B. HSIC Based on the ResNet

In the past few years, the ResNet has been widely applied to various HSI tasks since its residual structure can effectively reduce overfitting and speed up network training [48]. To reduce overfitting in HSIC, Cao et al. [49] introduced a multiresidual network embedding within the 3-D–2-D framework that successfully reduces overfitting and achieves superior classification results. Another advantage of the ResNet is that the residual connections can enhance the network information flow and prevent network degradation. Therefore, Zhong et al. [50] proposed to add the residual connection into each 3-D feature module, which makes the gradient backpropagation process smoother and alleviate accuracy degradation, simultaneously. Similarly, Huang et al. [51] designed an adaptive residual convolutional neural network (ARCNN) for HSIC. It successfully mitigates the network degradation by connecting each convolutional layer with a residual connection. To deal with the "small-sample problem," Feng et al. [52] proposed a residual HybridSN (R-HybridSN) method for HSIC. In R-HybridSN, the deep hierarchical spatial–spectral features can be effectively captured by using the limited training samples. Zhang et al. [53] established a deep feature residual network (DFRN) model. It aggregates the features by the summation operation and achieves satisfactory performances when the training samples are limited. As you know, the lower level features have more detailed information and the higher level features are quite abstract. For aggregating more semantic information, Song et al. [54] proposed a deep feature fusion network (DFFN) model to greatly expand the depth of the network and reflect the correlation of the features between different residual layers. Li et al. [55] further used the fused features for HSIC by adaptively learning the channel weights of the features from different residual layers. To construct more economical residual structures, Paoletti et al. [56] adopted the bottleneck residual unit to balance the workload between different units while maintaining the time complexity of each layer.

III. Proposed Method

In this section, we introduce our proposed S^{3}ARN model in detail. Specifically, we first present the overall framework of our proposed model, and then, describe the spectral block and the spatial block, respectively. Finally, the spectral attention residual branch and the spatial attention residual branch are introduced, respectively.

A. Overall Framework of Our S^{3}ARN Model

Fig. 2 shows the framework structure of the proposed S^{3}ARN model. It mainly consists of a trunk branch and two attention mask branch. The former consists of three 3-D spectral split attention blocks and three 2-D spatial split convolution attention blocks for semantic features extraction. In addition, to introduce the attention mechanism into features extracted at the spatial and spectral levels, the S^{3}ARN contains a spectral attention residual branch and a spatial attention residual branch. Moreover, we design an end-to-end learning framework to effectively extract the feature of HSI, in which the spatial and spectral blocks are applied directly to the original HSI cube.

Instead of the more aggressive direct element-wise multiplication of the attention masks and the features, we adopt a conservative residual learning approach, so the mask branches ensure that the performance degradation of the proposed model can be avoided. Let $X \in \mathbb{R}^{S\times S\times B}$ be the original patches of HSI, where $S \times S$ is the spatial dimensionality of the patches and B is the number of bands in HSI. The spectral attention residual branch connects the original patches and the first spectral block to generate the spectral attention vector mask. Assume that H^k_{spc}

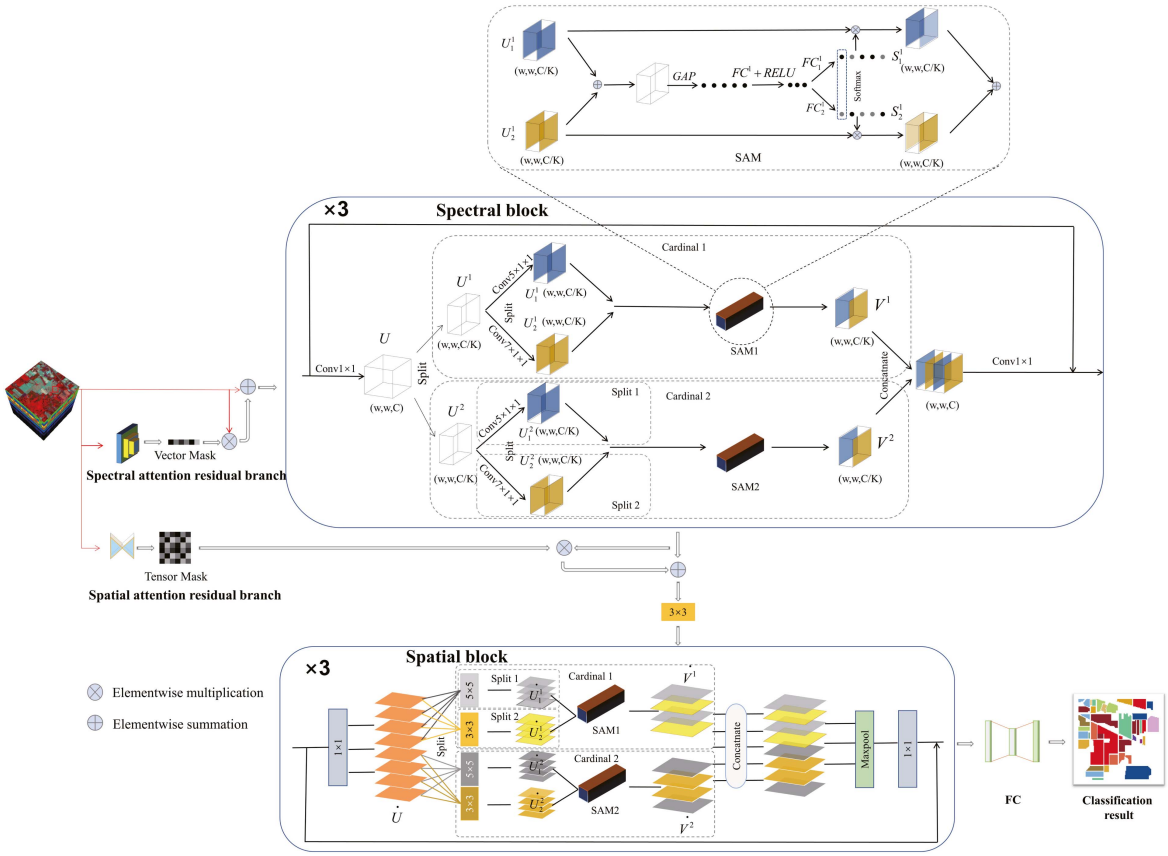

Fig. 2. Framework structure of the proposed S^{3}ARN method.

and H^k_{spa} represent the output of the kth spectral block and spatial block, respectively. The input of the first spectral block $\hat{H}^1_{\text{spc}}$ is given as follows:

$$\hat{H}^1_{\text{spc}} = (1 + M) \otimes X \tag{2}$$

where $\otimes$ denotes the elementwise multiplication and $M \in \mathbb{R}^{1\times1\times B}$ stands for the output vector mask of the spectral bands attention branch. Similarly, the spatial attention residual branch connects the original patches and the first spatial blocks to generate the spatial attention tensor mask. Therefore, the input of the first spatial block $\hat{H}^1_{\text{spa}}$ is represented as follows:

$$\hat{H}^1_{\text{spa}} = (1 + T) \otimes H^3_{\text{spc}} \tag{3}$$

where $T \in \mathbb{R}^{S\times S\times1}$ denotes the spatial attention tensor mask. The output of the last spatial module is fed into a classifier composed of two fully connected layers to obtain the classification result.

B. Spectral Split Attention Residual Block (Spectral Block)

Previous studies demonstrated that bottleneck residual units need far fewer parameters than the normal residual units [28]. In addition, convolutional kernel attention has been successfully explored in selective kernel networks [57]. Based on the successful implementation of the aforementioned structures, we integrate bottleneck residual unit, SAM, and 3-DCNN to construct our spectral block. The spectral block is intended to extract and integrate the spectral features from kernels with the different receptive fields, which consists of convolutional branches and split attention structure. Fig. 2 plots an overview of a spectral block. Unlike the bottleneck residual unit, the feature map $U \in \mathbb{R}^{w\times w\times C}$ obtained from the first 1×1 convolution layer is split into several cardinal groups $U^k \in \mathbb{R}^{w\times w\times C/K}$, where $w \times w$ and C are the spatial dimensionality and the channel number of U, respectively. Hyperparameter K represents the number of the cardinal groups. In general, each radix within a cardinal group has a unique receptive field size, and thus, can only extract one kind of receptive field features. To integrate different receptive field features into a cardinal group, each cardinal group is split into R radix groups, in which each radix group contains a convolutional branch. Therefore, R kinds of receptive field features extracted from R radix groups can be integrated within a cardinal group. Thus, it can be see that the total number of the radix groups is $G = KR$. By grouping the feature mappings, we can obtain more discriminative fusion features. Additionally, we append the BN function [58] to each convolution kernel to improve the classification performance and try to regularize the training process.

For each radix group, there is a separate 3-D convolutional operation F^k_r, where $r \in \{1, \ldots, R\}$ and $k \in \{1, \ldots, K\}$ are the index numbers of the radix groups and cardinal groups, respectively. To unify the dimensionality of the feature within the spectral block, we sum the output of each convolution kernel in each F^k_r using elementwise summation. Here, the output of

the kth radix group within the rth cardinal group is calculated as follows:

$$U_r^k = F_r^k\left(U^k\right) \tag{4}$$

where $U_r^k \in \mathbb{R}^{w\times w\times C/K}$. Afterwards, we fuse the feature maps within cardinal group using elementwise summation. Then, a global average-pooling layer (GAP) and two 1-D fully connection layers (FC) are employed to generate the channel-wise weight vector mask $S \in \mathbb{R}^{C/K}$, which is used to adaptively select the features from different radix groups. The aforementioned operations produce a mask S for each radix group. Therefore, the optimization process is expressed as follows:

$$S_r^k = FC_r^k\left(\delta\left(FC^k\left(\text{GAP}\left(\sum_{r=1}^{R} U_r^k\right)\right)\right)\right) \tag{5}$$

where S_r^k and FC_r^k denote the vector mask and FC operation corresponding to the rth radix group within kth cardinal group, respectively. δ denote the ReLU function. Within a cardinal group, the cross-vector softmax is used to balance the value of the corresponding position of the mask. The cth component of S_r^k can be updated by the following formula:

$$S_{r_c}^k = \frac{\exp\left(S_{r_c}^k\right)}{\sum_{r=1}^{R}\exp\left(S_{r_c}^k\right)}. \tag{6}$$

The kth cardinal reorganized feature map V^k is obtained by the weighted sum of each split feature. Then, the cth channel of V^k is calculated as follows:

$$V_c^k = \sum_{r=1}^{R} S_{r_c}^k \cdot U_{r_c}^k. \tag{7}$$

The calculation process of (5)–(7) is referred to SAM. By adopting the SAM, the features of different scales are reweighted and reorganized. For convenience, the operations of SAM are described as follows:

$$V^k = F_{SAM}\left(U_1^k, U_2^k, \ldots, U_R^k\right). \tag{8}$$

Finally, the features of the cardinal groups are concatenated along the channel dimension. Therefore, the output of the spectral blocks is given as follows:

$$H_{\text{spc}} = \text{conv}^3\left(\text{concat}\left(V^1, V^2, \ldots, V^K\right)\right) + I_{\text{spc}} \tag{9}$$

where $\text{conv}^3(.)$ represents the 1×1 convolution of the third layer within the spectral block, and I_{spc} is the identity mapping of spectral block. It is noted that the kernel number of each layer in the spectral block needs to meet the setting of the bottleneck structure. In this article, we set the values of both radix and cardinal to 2. Convolutional receptive fields in the two branches are set to $1\times 1\times 5$ and $1\times 1\times 7$, respectively.

C. Spatial Split Attention Residual Block (Spatial Block)

Similar to the spectral block, the spatial block can filter the spatial features from different receptive fields, because it adopts the same structure as the spectral block. Fig. 2 gives an overview of a spatial block. Unlike the spectral block, the spatial block employs pure 2-D convolutional operation $\dot{F}_r^k$ in each radix group to extract the spatial features. The split result of the first 1×1 convolutional layer is the input $\dot{U}^k$ of each cardinal group. Thus, the output of the rth radix group within kth cardinal is given as follows:

$$\dot{U}_r^k = \dot{F}_r^k\left(\dot{U}^k\right). \tag{10}$$

In the spatial block, each cardinal grouping employs the separate SAM to explore the correlations between the features with different receptive fields in different channels. Therefore, the output of the kth cardinal group is expressed as follows:

$$\dot{V}^k = F_{SAM}\left(\dot{U}_1^k, \dot{U}_2^k, \ldots, \dot{U}_R^k\right). \tag{11}$$

Following the concatenation of the output in each cardinal group, the result is then passed through the residual structure of the spatial block. Thus, the output of the spatial block is given as follows:

$$H_{\text{spa}} = \dot{\text{conv}}^3\left(\phi\left(\text{concat}\left(\dot{V}^1, \dot{V}^2, \ldots, \dot{V}^K\right)\right)\right) + I_{\text{spa}} \tag{12}$$

where $\dot{\text{conv}}^3(.)$ and I_{spa} represents 1×1 convolution within the third layer and the identity mapping of the spatial block, respectively, and ϕ denotes the maximum pooling operation. The radix and cardinal of the spatial block are set as the same as the spectral block in our experiments. The convolutional receptive fields of the radix are set to 5×5 and 3×3, respectively.

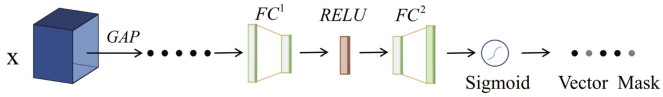

Fig. 3. Spectral attention residual branch.

D. Spectral Attention Residual Branch

Recently, some advanced network models have incorporated the soft-attention-masked branch [59]. The spectral attention residual branch aims to generate a weight vector, which reflects the discriminative ability of different spectral bands. In addition, it improves the robustness of the model by suppressing useless features and enhancing discriminative ones, which can prevent the classification features from being contaminated by mixed pixels and noise. Fig. 3 provides the structure of the spectral attention residual branch.

To generate spectral-wise summary statistics, the global average pooling is a cross-spatial operation of each spectral feature. Then, two fully connected layers FC^2 and FC^1 are used to learn the correlations between different spectral bands. The channels attention mask M can be formulated as follows:

$$M = \sigma\left(\text{FC}^2\left(\delta\left(\text{FC}^1\left(\text{GAP}(X)\right)\right)\right)\right) \tag{13}$$

where σ denote the sigmoid function.

E. Spatial Attention Residual Branch

The spatial attention branch aims to generate a spatial attention tensor mask that represents the weights of the features with different coordinates. The spatial features recalibrated by this

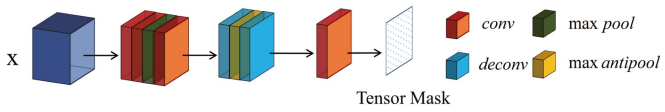

Fig. 4. Spatial attention residual branch.

spatial coordinate attention mechanism can also effectively improve the classification results. Fig. 4 shows the spatial attention residual branch. We can see that three convolutions operations and two deconvolution operations are applied to encode the correlation between pixels at different coordinates. Moreover, the interspersed maximum pooling and maximum antipooling operations enhance the robustness of the model. Then a single kernel convolution operation summarizes the spatial relationships in the patch data.

Therefore, the spatial attention mask tensor T can be expressed as follows:

$$T = \mathrm{conv}_4(\mathrm{deconv}_2(\varphi(\mathrm{deconv}_1(\mathrm{conv}_3(\phi (\mathrm{conv}_2(\mathrm{conv}_1(X))))))) \quad (14)$$

where $\mathrm{conv_i}(.)$ and $\mathrm{deconv_j}(.)$ represents the convolution and deconvolution operation, for $i \in \{1, \ldots, 4\}$ and $j \in \{1, \ldots, 2\}$. φ denote the maximum anti-pooling operation. In particular, except for $\mathrm{conv}_1(.)$ and $\mathrm{conv}_4(.)$, the other convolution and deconvolution operations in the branch adopt the RELU and BN functions.

IV. Experiments and Results

In this section, we report some classification experiments on different HSI datasets, and then, provide a detailed analysis of the experimental results.

A. Experimental Datasets

In this experiment, three benchmark HSI datasets were used to verify the performance of our proposed S^{3}ARN model.

The Indian Pines dataset was gathered by the AVIRIS sensor over the Indian Pines test site in North-Western Indiana and consists of 145 $\times$ 145 pixels. There are 16 classes of ground truth available and they are not mutually exclusive. By removing bands covering the water absorption region, the number of bands is reduced to 200.

The Loukia dataset is a subdataset of the HyRANK hyperspectral dataset, which has been developed in the framework of the ISPRS Scientific Initiatives [60]. The HyRANK dataset contains satellite hyperspectral data from the Hyperion sensor (EO-1 and USGS). The spatial size of the Loukia dataset is 249 $\times$ 945 pixels, which contains 176 spectral channels. There are 16 different land cover classes in the HyRANK benchmark datasets, and the selected Loukia dataset covers 14 classes.

The Houston 2013 dataset was collected by the ITRES CASI-1500 sensor in June 2012, which was provided by IEEE GRSS Data Fusion Competition 2013. [61]. Houston 2013 has 144 spectral bands with a spatial resolution of 349 $\times$ 1905 pixels. There are 15 land-cover classes in the Houston 2013 dataset.

TABLE I
Sample Size of the Indian Pines Dataset

No.	Class	Trainning	Validation	Testing
1	Alfalfa	5	5	36
2	Corn-notill	25	25	1378
3	Corn-mintill	25	25	780
4	Corn	25	25	187
5	Grass-pasture	25	25	433
6	Grass-trees	25	25	680
7	Grass-pasture-mowed	5	5	18
8	Hay-windrowed	25	25	428
9	Oats	5	5	10
10	Soybean-notill	25	25	922
11	Soybean-mintill	25	25	2405
12	Soybean-clean	25	25	543
13	Wheat	25	25	155
14	Woods	25	25	1215
15	Building-Grass-Trees-Drives	25	25	336
16	Stone-Steel-Towers	10	10	73
Total		325	325	9599

TABLE II
Sample Size of the Loukia Dataset

No.	Class	Training	Validation	Testing
1	Dense Urban Fabric	20	20	248
2	Mineral Extraction Sites	10	10	47
3	Non-Irrigated Arable Land	20	20	502
4	Fruit Trees	10	10	59
5	Olive Groves	50	50	1301
6	Broad-leaved Forest	20	20	183
7	Coniferous Forest	20	20	460
8	Mixed Forest	50	50	972
9	Dense Sclerophyllous Vegetation	50	50	3693
10	Sparce Sclerophyllous Vegetation	50	50	2703
11	Sparcely Vegetated Areas	20	20	364
12	Rocks and Sand	20	20	447
13	Water	50	50	1293
14	Coastal Water	50	50	351
Total		440	440	12623

B. Experimental Settings and Evaluation Metrics

We first eliminated unlabelled pixels from the processed dataset. Then, all available labeled pixels were randomly divided into training, validation, and testing sets. To avoid the overlap of test samples and training samples, fewer samples were utilized as the training and validation sets, and more samples were divided into the testing set. Tables I–III report the details of three HSI datasets.

The hardware configuration for the experiment was as follows: Nvidia GeForce RTX 3090 GPU, Intel i7-12700kf CPU, and 64-GB DRAM. In addition, the software configuration included Windows 11, Python 3.7, Pytorch 1.11.0, Scikit-learn 1.0.2, and Cuda 11.3.

The size of the input patches was empirically set to 15 $\times$ 15, and the batch size of all datasets was set to 64. In addition, the maximal training epoch was set to 100. We utilized the momentum optimizer and learning rate decay strategy to accelerate the convergence. Specifically, the initial learning rate was set to 0.01 and the decay rate was set to 0.2. The values of the momentum and the weight decay in S^{3}ARN were empirically set to 0.9 and 0.0001, respectively.

TABLE III
SAMPLE SIZE OF THE HOUSTON 2013 DATASET

No.	Class	Training	Validation	Testing
1	Healthy grass	40	40	1171
2	Stressed grass	40	40	1174
3	Synthetic grass	10	10	677
4	Tree	40	40	1164
5	Soil	40	40	1162
6	Water	10	10	305
7	Residential	40	40	1188
8	Commercial	40	40	1164
9	Road	40	40	1172
10	Highway	40	40	1147
11	Railway	40	40	1155
12	Parking lot1	40	40	1153
13	Parking lot2	10	10	449
14	Tennis court	10	10	408
15	Running track	10	10	640
Total		450	450	14129

Some well-known evaluation metrics, such as overall accuracy (OA), average accuracy (AA), and Kappa coefficient (Kappa) were used to evaluate the performances of different methods. In general, the higher the value of these three indicators, the better the classification performance of the model. Each model was run 100 epochs, and the one with the largest OA value on the validation set was chosen as the best-trained model. Then, we tested the generalization ability of the best-trained model on the testing set. Finally, we predicted the labels of all pixels using our best-trained model to create a feature map.

C. Classification Performance

For a fair comparison, some well-known neural network models (e.g., 2D-CNN [15], 3D-CNN [20], DFFN [54], SSRN [50], RSSAN [48], SSAN [62], MLNet-A [63], and MLNet-B [63]) were used as the comparison methods. In addition, all models adopted the same training dataset and test dataset. Tables IV–VI recorded the classification results of both our S^{3}ARN model and the comparison models on the Indian Pines, Loukia, and Houston 2013 datasets.

From the classification results, we can obtain the following observations.

1) It can be found that the proposed S^{3}ARN model achieves the highest classification accuracy among all models on all datasets. As shown in Table IV, our proposed S^{3}ARN achieves the highest OA (89.51%) and AA (88.98%) and kappa (87.95%) among all experimental models. Specifically, the OA values of the proposed S^{3}ARN model improves by 37.31% (2DCNN), 20.33% (3DCNN), 18.23% (DFFN), 19.39% (RSSAN), 5.07% (SSRN), 13.69% (SSAN), 3.72% (MLNet-A), and 5.99% (MLNet-B) on the Indian Pines datasets. Table V displays the performances of different models on Loukia dataset, it is easy to see that our S^{3}ARN model achieves the highest OA (78.47%), AA (77.87%) and Kappa (74.62%), and the improvement of OA values obtained by the proposed S^{3}ARN modal are 20.31% (2DCNN), 16.13% (3DCNN), 7.55% (DFFN), 5.90% (RSSAN), 6.19% (SSRN), 16.64% (SSAN), 8.45% (MLNet-A), and 9.61% (MLNet-B). Table VI shows the classification results of different methods on the Houston 2013 dataset. Our S^{3}ARN model achieves the highest OA (95.16%), AA (95.78%), and Kappa (94.76%) on Houston 2013 dataset. In terms of OA metric, our proposed S^{3}ARN model has increased by 34.84% (2DCNN), 28.73% (3DCNN), 12.01% (DFFN), 10.27% (RSSAN), 8.12% (SSRN),20.78% (SSAN), 2.72% (MLNet-A), and 1.77% (MLNet-B).
2) Notably, proposed S^{3}ARN approach achieves superior performance in classifying small sample datasets. The main reason is that our proposed S^{3}ARN model has powerful multireceptive field feature extraction capability.
3) Figs. 5–7 show the classification maps obtained by different models on three HSI datasets. The classification maps is consistent with the results reported in Tables IV–VI. Specifically, the classification maps of the 2DCNN and 3DCNN models mix with a lot of noise and a large proportion of edge pixel misclassification. The other comparison models produce smoother edges, but there are still some noises in some categories. The classification map of our S^{3}ARN model not only contains less noise than other methods but also detects more accurate edges. Therefore, these experimental results further demonstrate the superiority of our proposed S^{3}ARN model.

D. Ablation Study

To analyze the influence of each component of S^{3}ARN, we constructed some variants of S^{3}ARN with different settings. The first variant is called S^{3}ARN-WSAM, which does not activate split attention mechanism. The remaining two variants are called S^{3}ARN-WSPC and S^{3}ARN-WSPA, which remove the spectral attention branch and the spatial attention branch, respectively. Table VII illustrates the classification performances of these three variants on the three HSI datasets. It is clear to see that S^{3}ARN can achieve the more performances than its variants, which demonstrates the effectiveness of each component of our proposed method.

E. Effectiveness Analysis of the Bottleneck Stucture

To verify the effectiveness of bottleneck structure in the proposed model, a variant of S^{3}ARN, named S^{3}ARN-WBS, is constructed by removing the bottleneck structure of S^{3}ARN. The results are shown in Table VIII. It can be seen that the OA values, the number of parameters and the training time are greatly improved by embedding the bottleneck structure. It also demonstrates the effectiveness of our proposed method.

F. Analysis of Cardinal and Radix

The radix indicates the number of receptive fields in the spectral and spatial blocks within S^{3}ARN, and the cardinal represents the degree to which the feature extraction operation is split along the channel dimension. Here, we investigated the performances of S^{3}ARN varied with different radix and cardinal to determine the optimal settings for various datasets. Figs. 8 and 9 shows the performances of S^{3}ARN with different values of the cardinal and radix.

TABLE IV
CLASSIFICATION RESULTS OF DIFFERENT MODELS ON THE INDIAN PINES DATASET

No.	Class	S^3ARN	2D-CNN	3D-CNN	DFFN	RSSAN	SSRN	SSAN	MLNet-A	MLNet-B
1	Alfalfa	85.29	0.00	90.48	93.33	48.65	0.00	100.00	62.79	96.88
2	Corn-notill	87.84	41.13	51.94	57.39	59.75	75.76	63.99	85.64	79.26
3	Corn-mintill	88.58	35.76	69.17	73.71	67.77	88.66	59.58	84.61	80.77
4	Corn	84.16	40.96	49.03	96.81	68.64	73.23	67.42	76.23	67.39
5	Grass-pasture	99.02	50.22	85.87	83.57	61.50	92.26	81.88	94.08	96.62
6	Grass-trees	96.24	66.52	97.69	79.16	82.62	94.35	78.58	97.26	96.22
7	Grass-pasture-mowed	100.00	0.00	40.91	85.71	13.16	100.00	100.00	90.00	78.26
8	Hay-windrowed	96.80	37.81	97.46	99.76	99.29	96.15	88.18	100.00	100.00
9	Oats	83.33	0.00	30.77	20.00	34.78	0.00	0.00	33.33	32.26
10	Soybean-notill	78.78	49.70	57.61	44.22	53.23	65.92	77.21	77.41	73.75
11	Soybean-mintill	88.01	71.43	68.55	70.16	76.70	88.02	82.87	87.18	85.89
12	Soybean-clean	89.02	23.97	41.74	96.88	52.00	75.83	72.79	65.75	61.06
13	Wheat	82.45	63.79	71.76	83.15	81.48	77.89	53.82	75.61	75.61
14	Woods	97.08	87.17	97.36	94.68	90.76	97.16	91.28	96.94	97.21
15	Building-Grass-Trees-Drives	89.49	55.58	54.17	68.13	67.11	74.14	58.55	76.44	73.19
16	Stone-Steel-Towers	77.53	4.76	64.00	90.79	88.61	85.53	58.62	75.00	76.19
	OA(%)	**89.51**	52.20	69.18	71.28	70.12	84.44	75.82	85.79	83.52
	AA(%)	**88.98**	39.30	66.78	77.34	65.38	74.05	70.92	79.89	79.41
	KAPPA(%)	**87.95**	46.89	64.99	67.37	66.14	82.18	72.38	83.76	81.17

TABLE V
CLASSIFICATION RESULTS OF DIFFERENT MODELS ON THE LOUKIA DATASET

No.	Class	S^3ARN	2D-CNN	3D-CNN	DFFN	RSSAN	SSRN	SSAN	MLNet-A	MLNet-B
1	Dense urban fabric	75.59	45.83	46.84	38.26	58.63	46.61	33.46	42.07	53.11
2	Mineral extraction sites	100.00	89.58	97.87	100.00	100.00	100.00	100.00	100.00	100.00
3	Non-irrigated arable land	80.74	69.16	79.45	65.71	82.21	86.11	50.88	72.07	72.34
4	Fruit trees	54.79	0.00	16.67	20.65	48.84	0.00	20.59	28.89	28.12
5	Olive groves	82.92	55.46	50.64	62.38	65.80	69.59	46.54	76.55	76.42
6	Broad-leaved forest	43.17	0.00	13.25	36.25	48.13	63.16	21.43	15.40	12.97
7	Coniferous forest	75.32	50.00	51.94	68.97	74.01	66.16	76.40	74.92	64.80
8	Mixed forest	58.74	28.93	45.06	61.86	53.34	57.26	39.20	53.39	49.45
9	Dense Sclerophyllous vegetation	79.58	63.71	66.05	74.15	75.61	73.31	68.56	70.54	69.34
10	Sparse Sclerophyllous vegetation	75.57	59.02	61.80	70.99	70.66	69.43	65.67	69.69	65.97
11	Sparsely vegetated areas	68.39	21.30	36.86	52.46	50.82	53.89	39.49	56.59	45.51
12	Rocks and sand	95.72	63.42	75.15	81.96	91.91	97.57	73.78	86.82	89.16
13	Water	100.00	99.85	100.00	100.00	100.00	100.00	100.00	100.00	100.00
14	Coastal water	99.71	75.27	95.36	100.00	99.71	85.33	100.00	99.97	98.86
	OA(%)	**78.47**	58.16	62.34	70.92	72.57	72.28	61.83	70.02	68.86
	AA(%)	**77.87**	51.54	59.78	66.69	72.83	69.17	59.71	67.62	66.15
	KAPPA(%)	**74.62**	50.88	55.95	65.67	67.59	67.00	55.41	64.92	63.53

TABLE VI
CLASSIFICATION RESULTS OF DIFFERENT MODELS ON THE HOUSTON 2013 DATASET

No.	Class	S^3ARN	2D-CNN	3D-CNN	DFFN	RSSAN	SSRN	SSAN	MLNet-A	MLNet-B
1	Healthy grass	96.08	82.67	86.77	97.84	84.34	91.73	87.00	86.94	90.16
2	Stressed grass	97.49	80.59	91.20	87.43	93.20	96.31	83.97	90.86	93.94
3	Synthetic grass	98.24	0.00	93.51	96.23	98.01	99.40	87.57	98.83	98.97
4	Tree	96.79	77.99	87.83	95.54	96.13	82.84	85.32	96.97	95.96
5	Soil	92.57	64.78	76.46	84.02	86.43	99.32	82.11	96.02	96.26
6	Water	99.19	96.27	71.49	88.61	91.28	100.00	93.75	99.57	100.00
7	Residential	96.07	79.38	63.05	81.48	89.63	83.89	68.16	89.82	90.54
8	Commercial	99.79	68.03	63.74	73.61	79.14	99.76	86.88	99.13	99.58
9	Road	91.98	70.22	68.11	82.89	76.26	81.98	55.72	88.36	87.45
10	Highway	90.08	39.08	37.78	87.05	67.59	72.24	55.08	80.08	83.38
11	Railway	89.67	25.65	36.71	81.66	83.63	95.89	69.97	94.39	94.84
12	Parking lot1	97.08	35.16	36.59	60.76	79.58	66.07	69.73	95.27	95.52
13	Parking lot2	95.32	0.00	0.00	60.62	100.00	0.00	84.04	98.34	97.67
14	Tennis court	97.83	65.00	100.00	93.13	93.32	100.00	95.69	100.00	100.00
15	Running track	98.76	91.26	95.65	89.79	94.87	97.39	99.64	95.74	95.75
	OA(%)	**95.16**	60.32	66.43	83.15	84.89	87.04	74.38	92.44	93.39
	AA(%)	**95.78**	58.41	67.26	84.04	87.56	84.46	80.31	94.02	94.67
	KAPPA(%)	**94.76**	56.81	63.58	81.76	83.64	85.96	72.22	91.83	92.86

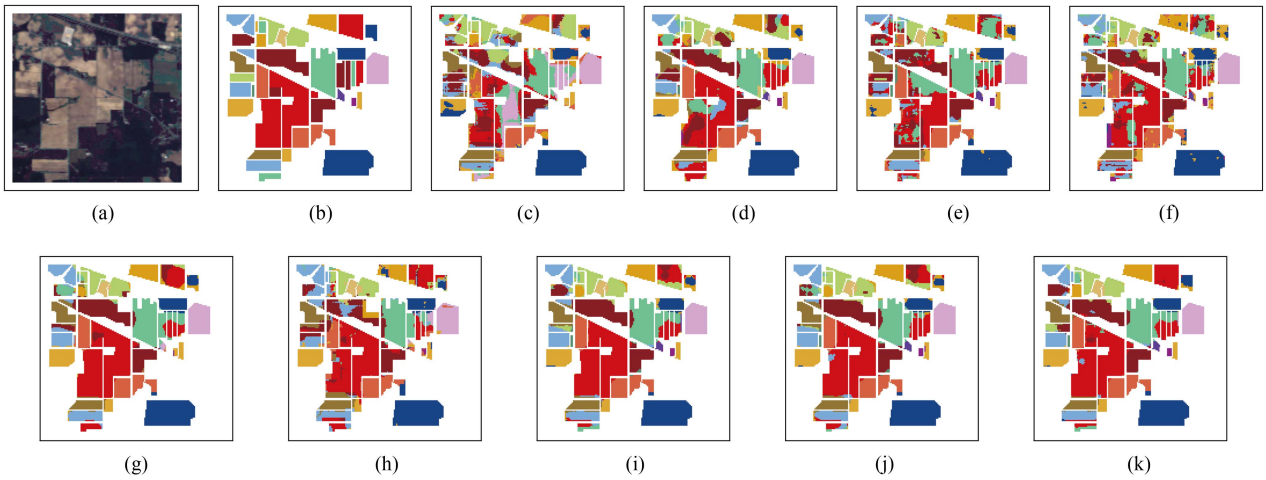

Fig. 5. Classification maps of different models on the Indian Pines dataset. (a) RGB. (b) GT. (c) 2DCNN. (d) 3DCNN. (e) DFFN. (f) RSSAN. (g) SSRN. (h) SSAN. (i) MLNet-A. (j) MLNet-B. (k) S^3ARN.

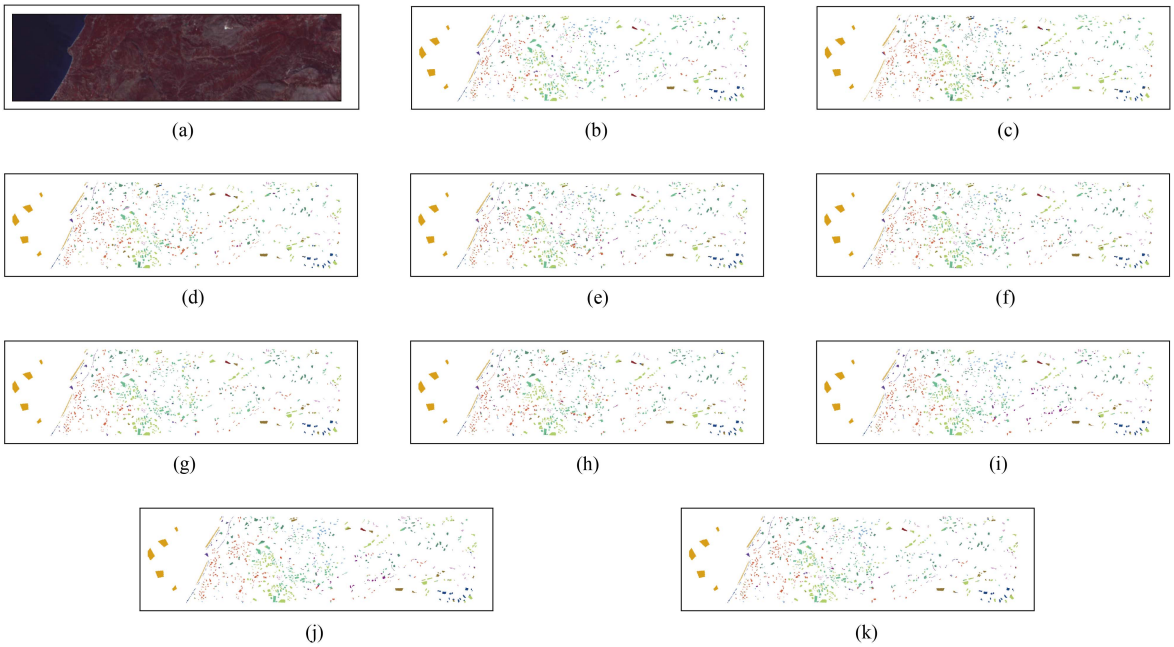

Fig. 6. Classification maps of different models on the Loukia dataset. (a) RGB. (b) GT. (c) 2DCNN. (d) 3DCNN. (e) DFFN. (f) RSSAN. (g) SSRN. (h) SSAN. (i) MLNet-A. (j) MLNet-B. (k) S^3ARN.

TABLE VII
THE ABLATION STUDY OF EACH ATTENTION MECHANISM

	Indian Pines	Loukia	Houston 2013
S^3ARN	89.51	78.47	95.16
S^3ARN-WSPC	88.74	76.60	94.40
S^3ARN-WSPA	89.44	76.75	94.82
S^3ARN-WSAM	87.76	76.67	94.68

On the Indian Pines dataset, our proposed model achieves the highest performance in the OA metric when the cardinal and the radix were set to 2 and 4, respectively. On the Loukia dataset, the proposed model can obtain the best OA value by setting the cardinal and the radix to 2 and 2, respectively. On the Houston 2013 dataset, the highest OA of our proposed method occurs when the cardinal and the radix are equal to 2 and 4, respectively.

TABLE VIII
THE RESULTS OF S^3ARN AND S^3ARN-WBS

		S^3ARN	S^3ARN-WBS
Indian Pines	OA(%)	89.51	82.37
	Paramters(M)	1.16	8.34
	Training Time(s)	26.98	57.34
Loukia	OA(%)	78.47	65.66
	Paramters(M)	1.12	8.33
	Training Time(s)	31.35	65.40
Houston 2013	OA(%)	95.16	86.31
	Paramters(M)	1.00	7.99
	Training Time(s)	34.05	63.38

It can be noticed that the optimal values of radix and cardinal are different on different datasets due to the differences in data distribution and sample size.

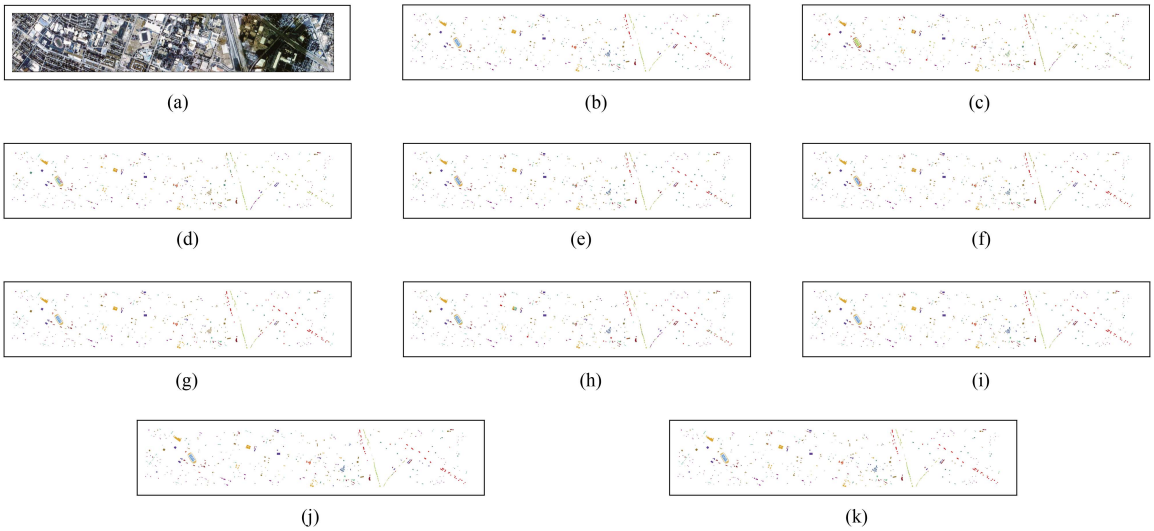

Fig. 7. Classification maps of different models on the Houston 2013 dataset. (a) RGB. (b) GT. (c) 2DCNN. (d) 3DCNN. (e) DFFN. (f) RSSAN. (g) SSRN. (h) SSAN. (i) MLNet-A. (j) MLNet-B. (k) S^{3}ARN.

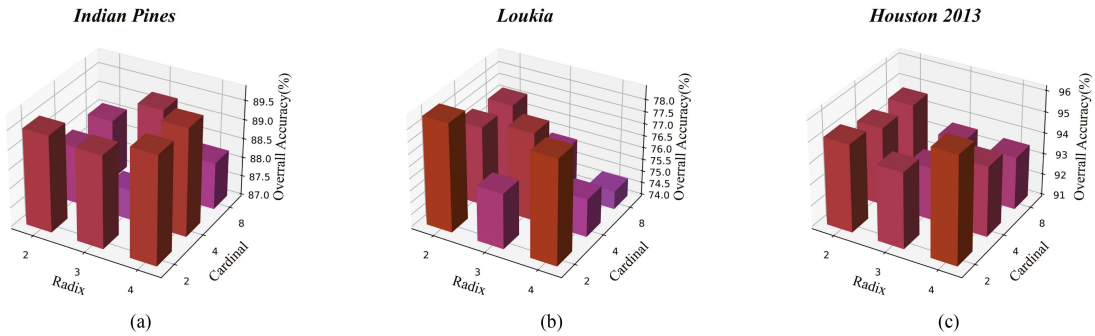

Fig. 8. Classification performances of S^{3}ARN with different settings on three HSI datasets.

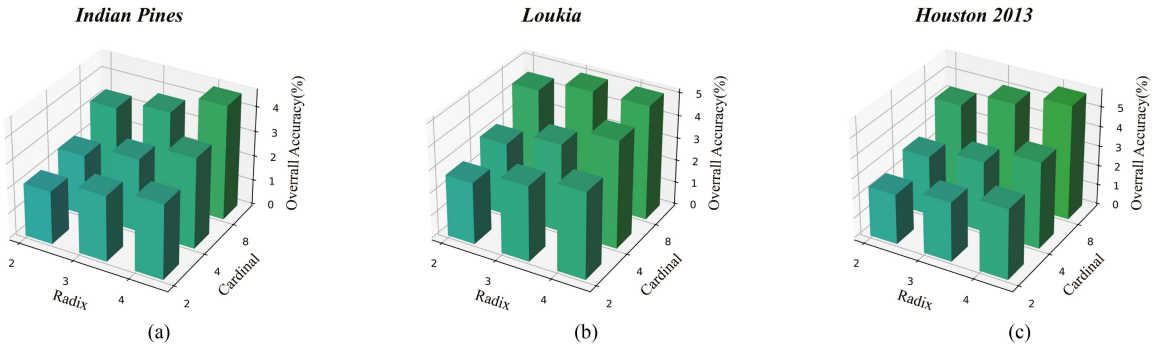

Fig. 9. Test time of S^{3}ARN with different settings on thee HSI datasets.

Noteworthy, as the radix increases, the test time of our proposed S^{3}ARN model rises dramatically. On the one hand, the increase of radix leads to the increase in convolutional branches. On the other hand, the increase of radix also leads to the increase in the number of fully connected layers. Additionally, as the cardinal increases, the number of convolution kernels of each radix group will decrease, but the number of radix groups will increase significantly. Therefore, the test time increases with the increase of the cardinal.

V. Conclusion

In this article, we have introduced a novel S^{3}ARN model for HSIC. Our proposed S^{3}ARN model integrates three attention mechanisms, including split attention, spatial attention, and spectral attention. Using the split attention, the proposed S^{3}ARN model seeks to extract multireceptive fields cross features information from HSIs. In addition, our proposed model is constructed based on the framework of the ResNet, in which the

two attention branches focus on extracting more discriminative spectral–spatial features based on the attention masks, making it more suitable for HSIC. Experimental results on three datasets have demonstrated the superiority of our proposed S^3ARN approach.

REFERENCES

[1] P. Ghamisi et al., "Advances in hyperspectral image and signal processing: A comprehensive overview of the state of the art," *IEEE Geosci. Remote Sens. Mag.*, vol. 5, no. 4, pp. 37–78, Dec. 2017.

[2] Z. Yan et al., "HD-CNN: Hierarchical deep convolutional neural networks for large scale visual recognition," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2015, pp. 2740–2748.

[3] X. Ma, A. Fu, J. Wang, H. Wang, and B. Yin, "Hyperspectral image classification based on deep deconvolution network with skip architecture," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 8, pp. 4781–4791, Aug. 2018.

[4] S. Jia, X. Zhang, L. Deng, and Z. Shu, "An l 1/2 regularized low-rank representation for hyperspectral imagery classification," in *Proc. IEEE Int. Conf. Image Process.*, 2015, pp. 1777–1780.

[5] Z. Shu, J. Zhou, L. Tong, X. Bai, and C. Zhao, "Multilayer manifold and sparsity constrainted nonnegative matrix factorization for hyperspectral unmixing," in *Proc. IEEE Int. Conf. Image Process.*, 2015, pp. 2174–2178.

[6] J. A. Gualtieri and R. F. Cromp, "Support vector machines for hyperspectral remote sensing classification," in *Proc. 27th AIPR Workshop Adv. Comput.-Assist. Recognit.*, 1999, pp. 221–232.

[7] V. Jain and A. Phophalia, "Exponential weighted random forest for hyperspectral image classification," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, 2019, pp. 3297–3300.

[8] Y. Chen, Z. Lin, and X. Zhao, "Riemannian manifold learning based k-nearest-neighbor for hyperspectral image classification," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, 2013, pp. 1975–1978.

[9] L. Mou, P. Ghamisi, and X. X. Zhu, "Deep recurrent neural networks for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 7, pp. 3639–3655, Jul. 2017.

[10] Y. Chen, Z. Lin, X. Zhao, G. Wang, 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.

[11] W. Hu, Y. Huang, L. Wei, F. Zhang, and H. Li, "Deep convolutional neural networks for hyperspectral image classification," *J. Sensors*, vol. 2015, 2015, Art. no. 258619.

[12] T. Guofeng, L. Yong, C. Lihao, and J. Chen, "A DBN for hyperspectral remote sensing image classification," in *Proc. 12th IEEE Conf. Ind. Electron. Appl.*, 2017, pp. 1757–1762.

[13] X. Zhang, Y. Sun, K. Jiang, C. Li, L. Jiao, and H. Zhou, "Spatial sequential recurrent neural network for hyperspectral image classification," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 11, no. 11, pp. 4141–4155, Nov. 2018.

[14] K. Makantasis, K. Karantzalos, A. Doulamis, and N. Doulamis, "Deep supervised learning for hyperspectral data classification through convolutional neural networks," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, 2015, pp. 4959–4962.

[15] Y. Chen, H. Jiang, C. Li, X. Jia, and P. Ghamisi, "Deep feature extraction and classification of hyperspectral images based on convolutional neural networks," *IEEE Trans. Geosci. Remote Sens.*, vol. 54, no. 10, pp. 6232–6251, Oct. 2016.

[16] W. Zhao and S. Du, "Spectral–Spatial feature extraction for hyperspectral image classification: A dimension reduction and deep learning approach," *IEEE Trans. Geosci. Remote Sens.*, vol. 54, no. 8, pp. 4544–4554, Aug. 2016.

[17] 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.

[18] B. Pan, Z. Shi, and X. Xu, "MugNet: Deep learning for hyperspectral image classification using limited samples," *ISPRS J. Photogrammetry Remote Sens.*, vol. 145, pp. 108–119, 2018.

[19] X. Cao, F. Zhou, L. Xu, D. Meng, Z. Xu, and J. Paisley, "Hyperspectral image classification with Markov random fields and a convolutional neural network," *IEEE Trans. Image Process.*, vol. 27, no. 5, pp. 2354–2367, May 2018.

[20] Y. Li, H. Zhang, and Q. Shen, "Spectral–Spatial classification of hyperspectral imagery with 3D convolutional neural network," *Remote Sens.*, vol. 9, no. 1, 2017, Art. no. 67.

[21] S. Mei, J. Ji, Y. Geng, Z. Zhang, X. Li, and Q. Du, "Unsupervised spatial–Spectral feature learning by 3D convolutional autoencoder for hyperspectral classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 9, pp. 6808–6820, Sep. 2019.

[22] J. Feng, N. Zhao, R. Shang, X. Zhang, and L. Jiao, "Self-supervised divide-and-conquer generative adversarial network for classification of hyperspectral images," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, Aug. 2022, Art. no. 5536517.

[23] L. Li, X. Yao, G. Cheng, and J. Han, "AIFS-DATASET for few-shot aerial image scene classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, Feb. 2022, Art. no. 5618211.

[24] 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.

[25] M. D. Zeiler and R. Fergus, "Visualizing and understanding convolutional networks," in *Proc. Eur. Conf. Comput. Vis.*, 2014, pp. 818–833.

[26] K. He and J. Sun, "Convolutional neural networks at constrained time cost," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 5353–5360.

[27] R. K. Srivastava, K. Greff, and J. Schmidhuber, "Training very deep networks," *Adv. Neural Inf. Process. Syst.*, vol. 28, 2015.

[28] 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.

[29] L. Wang, J. Peng, and W. Sun, "Spatial–Spectral squeeze-and-excitation residual network for hyperspectral image classification," *Remote Sens.*, vol. 11, no. 7, 2019, Art. no. 884.

[30] L. Dang, P. Pang, and J. Lee, "Depth-wise separable convolution neural network with residual connection for hyperspectral image classification," *Remote Sens.*, vol. 12, no. 20, 2020, Art. no. 3408.

[31] H. Gao, Y. Yang, C. Li, L. Gao, and B. Zhang, "Multiscale residual network with mixed depthwise convolution for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 4, pp. 3396–3408, Apr. 2021.

[32] L. Itti, C. Koch, and E. Niebur, "A model of saliency-based visual attention for rapid scene analysis," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 20, no. 11, pp. 1254–1259, Nov. 1998.

[33] M.-H. Guo et al., "Attention mechanisms in computer vision: A survey," *Comput. Vis. Media*, vol. 8, 2022, pp. 331–368,.

[34] J. Xiang, C. Wei, M. Wang, and L. Teng, "End-to-end multilevel hybrid attention framework for hyperspectral image classification," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, Nov. 2021, Art. no. 5511305.

[35] C. Yu, R. Han, M. Song, C. Liu, and C.-I. Chang, "Feedback attention-based dense CNN for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, Feb. 2021, Art. no. 5501916.

[36] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 7132–7141.

[37] J. Wang, J. Zhou, and W. Huang, "Attend in bands: Hyperspectral band weighting and selection for image classification," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 12, no. 12, pp. 4712–4727, Dec. 2019.

[38] D. Liu, J. Li, and Q. Yuan, "A spectral grouping and attention-driven residual dense network for hyperspectral image super-resolution," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 9, pp. 7711–7725, Sep. 2021.

[39] X. Zhang, X. Yao, X. Feng, G. Cheng, and J. Han, "DFENet for domain adaptation-based remote sensing scene classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, Oct. 2021, Art. no. 5611611.

[40] J. Feng et al., "Convolutional neural network based on bandwise-independent convolution and hard thresholding for hyperspectral band selection," *IEEE Trans. Cybern.*, vol. 51, no. 9, pp. 4414–4428, Sep. 2021.

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

[42] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 2818–2826.

[43] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. A. Alemi, "Inception-v4, inception-ResNet and the impact of residual connections on learning," in *Proc. 31st AAAI Conf. Artif. Intell.*, 2017, pp. 4278–4284.

[44] C. Li, Z. Qiu, X. Cao, Z. Chen, H. Gao, and Z. Hua, "Hybrid dilated convolution with multi-scale residual fusion network for hyperspectral image classification," *Micromachines*, vol. 12, no. 5, 2021, Art. no. 545.

[45] M. Tan and Q. V. Le, "MixConv: Mixed depthwise convolutional kernels," in *Proc. Brit. Mach. Vis. Conf.*, 2019.

[46] H. Zhang et al., "ResNeSt: Split-attention networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit*, 2022, pp. 2736–2746.

[47] X. Glorot and Y. Bengio, "Understanding the difficulty of training deep feedforward neural networks," in *Proc. 13th Int. Conf. Artif. Intell. Statist.*, 2010, pp. 249–256.
[48] 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.
[49] F. Cao and W. Guo, "Deep hybrid dilated residual networks for hyperspectral image classification," *Neurocomputing*, vol. 384, pp. 170–181, 2020.
[50] 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. 847–858, Feb. 2018.
[51] H. Huang, C. Pu, Y. Li, and Y. Duan, "Adaptive residual convolutional neural network for hyperspectral image classification," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 13, pp. 2520–2531, May 2020.
[52] F. Feng, S. Wang, C. Wang, and J. Zhang, "Learning deep hierarchical spatial–spectral features for hyperspectral image classification based on residual 3D-2D CNN," *Sensors*, vol. 19, no. 23, 2019, Art. no. 5276.
[53] C. Zhang et al., "Deep feature aggregation network for hyperspectral remote sensing image classification," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 13, pp. 5314–5325, Sep. 2020.
[54] 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.
[55] G. Li, L. Li, H. Zhu, X. Liu, and L. Jiao, "Adaptive multiscale deep fusion residual network for remote sensing image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 11, pp. 8506–8521, Nov. 2019.
[56] M. E. Paoletti, J. M. Haut, R. Fernandez-Beltran, J. Plaza, A. J. Plaza, and F. Pla, "Deep pyramidal residual networks for spectral–spatial hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 2, pp. 740–754, Feb. 2019.
[57] X. Li, W. Wang, X. Hu, and J. Yang, "Selective kernel networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 510–519.
[58] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *Proc. Int. Conf. Mach. Learn.*, 2015, pp. 448–456.
[59] F. Wang et al., "Residual attention network for image classification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 3156–3164.
[60] K. Karantzalos, C. Karakizi, Z. Kandylakis, and G. Antoniou, *Hyrank Hyperspectral Satellite Dataset I*, Version v001, 2018. [Online]. Available: https://doi.org/10.5281/zenodo
[61] N. Acito, S. Matteoli, A. Rossi, M. Diani, and G. Corsini, "Hyperspectral airborne "Viareggio 2013 trial" data collection for detection algorithm assessment," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 9, no. 6, pp. 2365–2376, Jun. 2016.
[62] X. Mei et al., "Spectral-spatial attention networks for hyperspectral image classification," *Remote Sens.*, vol. 11, no. 8, 2019, Art. no. 963.
[63] Z. Meng, L. Jiao, M. Liang, and F. Zhao, "Hyperspectral image classification with mixed link networks," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 2494–2507, Jan. 2021.

Zhenqiu Shu received the Ph.D. degree in computer applications from the Nanjing University of Science and Technology, Nanjing, China, in 2015.

He was a Visiting Scholar with the School of Information and Communication Technology, Griffith University, Southport, QLD, Australia, from November 2014 to May 2015. In February 2021, he joined the Faculty of Information Engineering and Automation, Kunming University of Science and Technology, Kunming, China, where he is currently an Associate Professor. Before joining the Kunming University of Science and Technology, he worked as a Postdoctoral with the Jiangnan University for four years. He has authored and co-authored more than 60 research papers in refereed international journals and conferences. His research interests include pattern recognition, computer vision, and machine learning.

Zigao Liu is currently working toward the Master degree with the Faculty of Information Engineering and Automation, Kunming University of Science and Technology, Kunming, China.

His current research interests include hyperspectral classification and pattern recognition.

Jun Zhou received the B.S. degree in computer science and the B.E. degree in international business from the Nanjing University of Science and Technology, Nanjing, China, in 1996 and 1998, respectively, the M.S. degree in computer science from Concordia University, Montreal, QC, Canada, in 2002, and the Ph.D. degree in computing science from the University of Alberta, Edmonton, AB, Canada, in 2006.

He was a Research Fellow with the Research School of Computer Science, Australian National University, Canberra, ACT, Australia, and a Researcher with the Canberra Research Laboratory, NICTA, Canberra. In June 2012, he joined the School of Information and Communication Technology, Griffith University, Nathan, QLD, Australia, where he is an Associate Professor. His research interests include pattern recognition, computer vision, and spectral imaging with their applications in remote sensing and environmental informatics.

Songze Tang received the B.S. degree in information and computation science from Anhui Agriculture University, Hefei, China, in 2009, and the Ph.D. degree in control science and engineering from the School of Computer Science and Technology, Nanjing University of Science and Technology, Nanjing, China, in 2015.

In 2016, he joined the Department of Criminal Science and Technology, Nanjing Forest Police College, Nanjing, as a Lecturer. His research interests include image/video processing, computer vision, and artificial intelligence.

Zhengtao Yu received the Ph.D. degree in computer application technology from the Beijing Institute of Technology, Beijing, China, in 2005.

He is currently a Professor with the School of Information Engineering and Automation, Kunming University of Science and Technology, Kunming, China. His current research interests include natural language processing, information retrieval, and machine learning.

Xiao-Jun Wu received the B.Sc. degree in mathematics from Nanjing Normal University, Nanjing, China, in 1991, and the M.S. and Ph.D. degrees in pattern recognition and intelligent systems from the Nanjing University of Science and Technology, Nanjing, in 1996 and 2002, respectively.

From 1996 to 2006, he taught at the School of Electronics and Information, Jiangsu University of Science and Technology, where he was promoted to a Professor. He has been with the School of AI and CS, Jiangnan University, since 2006, where he is a Professor of computer science and technology. He was a Visiting Researcher with the Centre for Vision, Speech and Signal Processing, University of Surrey, Guildford, U.K., from 2003 to 2004. He has authored and co-authored more than 300 research papers in refereed international journals and conferences. His current research interests include pattern recognition and computational intelligence.

Dr. Wu was a Fellow of the International Institute for Software Technology, United Nations University, from 1999 to 2000. He was the recipient of the Most Outstanding Postgraduate Award from the Nanjing University of Science and Technology.