Spatial–Spectral ConvNeXt for Hyperspectral Image Classification

Yimin Zhu, Kexin Yuan, Wenlong Zhong, and Linlin Xu, *Member, IEEE*

***Abstract*—Hyperspectral image (HSI) classification is a difficult task due to the heterogeneous spatial–spectral information, high-dimensionality, and noise effect in the HSI. Lately, an enhanced convolutional approach, i.e., ConvNeXt, has demonstrated a stronger feature representation capability than the popular vision transformer approaches. This article presents a spatial–spectral ConvNeXt approach, called SS-ConvNeXt, for hyperspectral classification. To better learn the spatial and spectral information in the HSI, the Spatial-ConvNeXt block, Spectral-ConvNeXt block, and spectral projection module are, respectively, designed. The depthwise and pointwise convolutions are adopted to reduce the model size and prevent vanishing gradient. The proposed model is evaluated against 14 other state-of-the-art methods on four different HSI datasets. Moreover, extensive ablation studies are conducted to investigate the roles of building blocks in the proposed model. The results demonstrate that the proposed method not only can achieve a high classification accuracy but also can better preserve class boundaries and reduce within-class noise.**

***Index Terms*—ConvNeXt, convolutional neural networks, deep learning, hyperspectral image classification (HSIC), spatial–spectral ConvNeXt (SS-ConvNeXt).**

I. Introduction

HYPERSPECTRAL image classification (HSIC) aims to estimate the semantic class labels for each pixel on a hyperspectral image (HSI) [1]. It is one of the most important HSI processing tasks, and has been widely used to support various applications, e.g., land cover and crop mapping [2], [3], urban monitoring [4], minerals mapping [5], etc. Despite its importance, the HSIC is a challenging task due to the heterogeneous spectral–spatial information, high dimensionality, and noise effect in an HSI, which make it very difficult to extract discriminative features from the HSI [6].

Traditional feature extractors have been used to support the HSIC [7], [8], [9], e.g., principal component analysis (PCA) [10], local binary pattern (LBP) [11], morphological profile (MP) [12], and extended multiattribute MP (EAMP) [13]. However, these feature extractors are mostly knowledge driven, and as such, cannot effectively adapt to the data characteristics for the enhanced HSIC. On the other hand, data-driven feature learning approaches, represented by deep convolutional neural networks (CNNs), have led to an improved spatial–spectral feature extraction capability that greatly boost the HSIC performance; e.g., see [14], [15], [16], [17], [18], [19], [20], [21], and [22]. Recently, the transformer [23] model, originally designed for natural language processing (NLP), has proven stronger feature learning capability than CNNs by using the attention mechanism and has been successfully used for the HSIC [24], [25], [26], [27], [28]. Nevertheless, transformer models have a quadratic complexity with respect to the input size, leading to a high computational cost and the risk of overfitting given limited training samples. To overcome these limitations, more recent transformer approaches, e.g., Swin transformer [29], tend to reuse key CNN features, such as local windows and weight sharing mechanisms. Transformer-based architectures become increasingly like CNNs [30]. Therefore, for enhancing the HSIC, it is essential to explore enhanced CNN approaches that avoids transformer's limitations.

Recently, to compare with transformer models, i.e.,vision transformer (ViT) [31] and Swin transformer [29], the ConvNeXt method [32] is proposed to improve the traditional CNN approaches. The ConvNeXt introduces the Swin transformer design concepts to modernize standard residual neural networks (ResNet), leading to a better performance than transformer-based architectures on ImageNet classification [33], object detection, and semantic segmentation tasks on COCO [34]. ConvNeXt's success is owing to rethink and redesign the key CNN components. From the macrodesign perspective, the ConvNeXt has four main characteristics. First, the ConvNeXt adopts a four stages architecture, and changes the stage compute ratio into (3, 3, 9, 3), which represents the number of blocks in each stage and likely to be the optimal distribution of computation. Second, the ConvNeXt uses a 4×4 nonoverlapping convolution to aggressively downsample the input images at the network's beginning. Third, following the strategy proposed in the ResNeXt, the ConvNeXt uses depthwise convolution and 1×1 convolution to separate the mixed spatial and channel information. Fourth, motivated by the transformer block, the ConvNeXt also uses inverted bottlenecks and revisits the use of large-sized convolutions. At the microscale, fewer activation functions and normalization layers are adopted in the ConvNeXt, which is the same with transformer models. Moreover, the ConvNeXt

Manuscript received 10 April 2023; revised 28 May 2023; accepted 1 June 2023. Date of publication 5 June 2023; date of current version 26 June 2023. *(Corresponding author: Linlin Xu.)*

Yimin Zhu, Kexin Yuan, and Wenlong Zhong are with the Department of Land Science and Technology, China University of Geosciences, Beijing 100083, China (e-mail: 2012210018@email.cugb.edu.cn; 2112210056@email.cugb.edu.cn; zwlong@cugb.edu.cn).

Linlin Xu is with the Department of Systems Design Engineering, University of Waterloo, Waterloo, ON N2L 3G1, Canada, and also with the Department of Land Science and Technology, China University of Geosciences, Beijing 100083, China (e-mail: linlinxu618@gmail.com).

The codes of this work will be available at https://github.com/ZHUYMGeo/SS-ConvNeXt for the sake of reproducibility.

Digital Object Identifier 10.1109/JSTARS.2023.3282975

performs better by replacing ReLU with GELU and substituting batch normalization (BN) with layer normalization (LN). The aforementioned macro and microdesigns make the ConvNeXt a cutting-edge model. Given the advantages of the ConvNeXt, the critical research question is how to adapt ConvNeXt under the consideration of effectively extracting spatial and spectral feature separately, one of the prime challenges in HSIC. This article, therefore, presents a new spatial–spectral ConvNeXt approach, called SS-ConvNeXt, for hyperspectral classification, with the following characteristics.

1) To better learn the discriminative spatial and spectral information in the HSI, we design a new Spatial-ConvNeXt (Spa-cv) block and a Spectral-ConvNeXt (Spe-cv) block. The Spa-cv and Spa-cv blocks are used to implement a four-stage architecture, with the number of blocks being (3, 3, 9, and 3), respectively. The Spa-cv block is used to implement the first two stages, and Spe-cv is used to implement the last two stages.
2) We use both depthwise and pointwise convolutions to reduce the model size and prevent vanishing gradient. To decouple spatial and spectral information learning, instead of using depthwise and pointwise convolutions together in all blocks, we use depthwise convolution in Spa-cv for spatial information learning and use pointwise convolution in Spe-cv for spectral information learning.
3) To better learn the rich spectral information in the HSI, instead of performing downsampling in the spatial domain using 4×4 and 2×2 convolution layers, as conducted in the ConvNeXt, we perform projection in the spectral domain using the pointwise convolution layer to enhance discriminative features in stages regularly, which we call the spectral projection module (SPM).

The aforementioned characteristics of the proposed model enable efficient discriminative spatial–spectral feature learning, leading to an enhanced HSIC approach that can better address the key HSI challenges. We qualitatively and quantitatively evaluate the classification performance of the proposed methods on four HSI datasets. The results demonstrate that the proposed model not only can achieve high classification accuracy but also can better preserve class boundaries and reduce within-class noise. The proposed model approach shows significant improvement over the original ConvNeXt (i.e., ConvNeXt-T [32]) approach and various state-of-the-art (SOTA) CNNs-based and transformer-based backbone networks.

The remainder of this article is organized as follows. Section II introduces the proposed model. Section III presents the relevant experimental results and highlights the comparison of our results with other results. Section IV draws the discussion. Finally, Section V concludes this article.

II. Proposed Framework

A. Problem Formulation

The HSI is denoted by X, where the ith pixel x_i is extracted as a 3-D cube of size $W \times W \times P$, with W being the patch size and P being the number of bands in the HSI. The class labels of x_i is denoted by y_i, which takes discrete values, i.e., $y_i \in \{1, 2, \ldots, C\}$, where C is the total number of classes. The task of the HSIC aims to estimate the labels of all pixels, i.e., $Y = \{y_i | i \in T\}$, where T is a total number of pixels. Deep-learning-based approaches solve this task by mapping x_i to y_i using a neural network model $y_i = g(x_i, \Theta)$, and estimating the model parameters Θ using training samples. Once $g(x_i, \Theta)$ is established, it can be used to predict all pixels in X and generate classification maps.

B. Overall Architecture

Fig. 1 shows the overall architecture of the proposed SS-ConvNeXt model. As we can seen in the top row of Fig. 1, the proposed model consists of four stages, where the first two stages are implemented by the Spa-cv block and the last two stages implemented by Spe-cv block. The Spa-cv block uses depthwise convolutions for spatial information learning, whereas the Spe-cv block uses pointwise convolutions for spectral information learning. Different stages are connected by SPM via a pointwise convolution layer. The adaptive average pooling (AAP) layer and a fully connected layer are used to generate the class label. Mathematically, the proposed model can be formulated as

$$\begin{aligned} g(x_i, \Theta) = \text{FC}(\text{AAP}(\text{GELN}(\text{Spe-cv}(\text{PC}_{1\times1}\text{LN} \\ \times (\text{Spe-cv}(\text{GELN}(\text{PC}_{1\times1} \\ \times (\text{Spa-cv}(\text{PC}_{1\times1}\text{LN}(\text{Spa-cv}(\text{GELN} \\ \times (\text{PC}_{1\times1}(x_i)))))))))))))) \end{aligned} \tag{1}$$

where $\text{PC}_{1\times1}$ is the pointwise convolution layer, i.e., a convolution layer with kernel size being 1. GELN represents the coactivation function of GELU and LN layer.

C. Spa-cv Module

As indicated in Fig. 1, we design a new Spa-cv module to implement the first two stages in the proposed model, where Spa-cv consists, sequentially, of a depthwise convolution layer, a layerwise convolution (LN) layer, an expansion linear layer, a GELU activation layer, another linear layer, and a dropout and scaling layer. The residual learning approach is also adopted by using a skip connection operation. The use of depthwise convolution in the Spa-cv module encourages Spa-cv to focus on learning the spatial information in the HSI. Moreover, with less parameters, depthwise convolution also reduces the size of the proposed model.

The Spa-cv module in (1) can be expressed as

$$\begin{aligned} \text{Spa-cv}(\text{input}) = \text{input} + \text{LayerScaleDrop} \\ \times (\text{FC}(\text{GELU}(\text{FC}(\text{LN} \\ \times (\text{DConv}_{3\times3}(\text{input})))))) \end{aligned} \tag{2}$$

where $\text{DConv}_{3\times3}$ is depthwise convolution with a total of 64 convolution filters of size $3 \times 3 \times 1$ in the first Spa-cv module and 128 same-sized convolution filters in the second Spa-cv module.

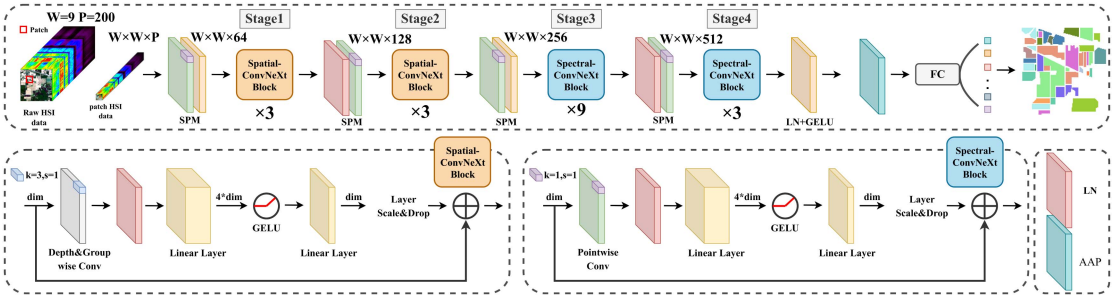

Fig. 1. Architecture overview of the proposed SS-ConvNeXt approach. The top row indicates a four-stage architecture, with the first two stages implemented by the Spa-cv block and the last two stages implemented by the Spe-cv block. Depthwise convolutions are used in Spa-cv for spatial information learning, and pointwise convolutions used in Spe-cv for spectral information learning. Different stages are connected by the SPM via a pointwise convolution layer. W is the spatial size of patch. P is the number of band. LN denotes layerwise normalization. AAP denotes adaptive average pooling layer.

D. Spe-cv Module

As indicated in Fig. 1, we design a new Spe-cv module to implement the last two stages in the proposed model, where Spe-cv consists, sequentially, of a pointwise convolution layer, a layerwise convolution (LN) layer, an expansion linear layer, a GELU activation layer, another linear layer, and a dropout and scaling layer. Similar to Spa-cv, the residual learning approach is adopted by using a skip connection operation. The use of pointwise convolution in Spe-cv encourages the Spe-cv module to focus on learning the spectral information in the HSI in an efficient manner.

The Spe-cv module in (1) can be expressed as

$$\begin{aligned}\text{Spe-cv}(\text{input}) = \text{input} + \text{LayerScaleDrop} \\ \times (\text{FC}(\text{GELU}(\text{FCLN}(\\ \times (\text{PConv}_{1\times1}(\text{input})))))) \end{aligned} \tag{3}$$

where $\text{PConv}_{1\times1}$ is pointwise convolution with a total of 256 1 $\times$ 1 convolution filters in the first Spe-cv module and 512 filters in the second Spe-cv module, respectively.

E. Spectral Projection Module (SPM)

As indicated in Fig. 1, we design SPM to connect different stages using pointwise convolution layer. By applying the SPM to patch-wise samples, more discriminative features in spectral domain can be established in stages regularly, instead of performing spatial downsampling as in the ConvNeXt model. In detail, before the first and third stages, we insert a pointwise convolutional layer, an LN layer, and a GELU layer. Before the second and the last stage, we add LN layer and pointwise convolution layer.

III. Experiment Results and Analysis

A. Data Description

To evaluate the performance of the proposed method, four classical HSI datasets are adopted, i.e., Indian Pines (IN),[1] Pavia University (PU), WHU-Hi-HongHu (WHHH), and WHU-Hi-HanChuan (WHHC).[2]

1) IN Data: IN data were collected in 1992 by the Airborne Visible/Infrared Imaging Spectrometer sensor over Northwestern Indiana, USA. The HSI consists of 145 $\times$ 145 pixels at a ground sampling distance (GSD) of 20 m and 220 spectral bands covering the wavelength range of 400–2500 nm with a 10-m spectral resolution. In the experiment, 24 water-absorption bands and noise bands were removed, and 200 bands were selected. There are 16 mainly investigated categories in this studied scene.

2) PU Data: PU data were acquired by the Reflective Optics System Imaging Spectrometer (ROSIS) sensor over PU and its surroundings, Pavia, Italy. This dataset has 103 spectral bands ranging from 430 to 860 nm. Its spatial resolution (SR) is 1.3 m, and its image size is 610 $\times$ 340. Nine land-cover categories are covered.

3) WHHC Data: WHHC data were captured by Headwall Nano-Hyperspec imaging sensor equipped on a Leica Aibot X6 UAV V1 platform on June 17, 2016, in Hanchuan, Hubei province, China. It contains 1217 $\times$ 303 pixels, with an SR of 0.109 m, and 274 bands from 400 to 1000 nm. There are 16 classes in this studied scene.

4) WHHH Data: WHHH data were acquired on November 20, 2017, by the Headwall Nano-Hyperspec imaging sensor equipped on a DJI Matrice 600 Pro UAV platform over the area of Honghu City, Hubei province, China, with an SR of 0.043 m, and image size of 940 $\times$ 475, and 270 bands in the range of from 400 to 1000 nm. Twenty-two land-cover categories are covered.

B. Experimental Setting

1) Evaluation Metrics: To quantitatively evaluate the proposed method and other compared methods, we choose three commonly used metrics, i.e., *overall classification accuracy* (OA), *average classification accuracy* (AA), *category accuracy* (CA), and *Kappa coefficient* (k).

[1][Online]. Available: https://www.ehu.eus/ccwintco/index.php/Hyperspectral_Remote_Sensing_Scenes

[2][Online]. Available: http://rsidea.whu.edu.cn/resource_WHUHi_sharing.htm

2) Implementation and Training Details: Our proposed SS-ConvNeXt model is implemented on the PyTorch 1.10.2 platform using a workstation with Intel(R) Xeon(R) CPU E5-2640 v4, 256-GB RAM, and an NVIDIA GeForce RTX 2080 Ti 11-GB GPU. Training, validation, and test samples are extracted as 3-D cubes.

To train the proposed model, the CrossEntropy loss function is adopted and the gradient descent approach, i.e., the Adam optimizer [35], is used to estimate the unknown parameters in the proposed model. The ExponentialLR scheduler is adopted, and the initial learning rate is set to be 0.0001 and decayed by multiplying a factor of 0.9 after each one-tenth of the total 400 epochs. We set batch size of 16, 32, 64, and 64 and patch size of 9, 9, 13, and 13 for IN, PU,WHHC, and WHHH, respectively, to allow a better computational efficiency.

C. Compared Methods

1) Methods Implemented: A total of eight other state-of-the-art deep learning models are selected and implemented for comparison, i.e., SVM, 1D-CNN [14], 2D-CNN [36], 3D-CNN [37], SSRN [38], HybridSN [17], A^{2}S^{2}K-ResNet [39], SSFTT [40], SSRes [41], and SSTN [42]. For fair comparison, we use the same model settings that were described in corresponding articles.

2) Methods With Published Results: In addition, published results of a total of six advanced methods are used to further verify the effectiveness of the proposed method, including few-shot learning-based approaches (i.e., S-DMM [43], UM^{2}L [44], and DCFSL [45]), CRF-based approaches (i.e., CNNCRF [46]), and transformer-based model (i.e., SpectralFormer [26] and SST-FA [28]).

3) Methods for Ablation Studies: Moreover, to investigate the performance gain of the proposed SS-ConvNeXt, five variants of SS-ConvNeXt (i.e., SS-ConvNeXt(E), SS-ConvNeXt(D), SS-ConvNeXt(F), Spa-ConvNeXt, and Spe-ConvNeXt) as well as the ConvNeXt-T [32] model are also compared in the ablation studies. Fig. 2 shows the architecture design of variants of the SS-ConvNeXt. In Fig. 2, the SS-ConvNeXt(E) is the same with the SS-ConvNeXt, except that it exchanges the location of Spa-cv and Spe-cv modules in the SS-ConvNeXt. The difference between the SS-ConvNeXt(D) and SS-ConvNeXt is that the SS-ConvNeXt(D) replaces the SPM in the SS-ConvNeXt with the 2×2 spatial downsampling layer before stages 2 and 4. The SS-ConvNeXt(F) fuses the Spa-cv and Spe-cv modules in a branch manner, with three feature fusion methods, i.e., point-wise multiplication, point-wise addition, and concatenation. The Spa-ConvNeXt and Spe-ConvNeXt only use the spatial and spectral modules, respectively.

D. Numerical Evaluation

We conduct experiments on these four datasets to investigate the classification accuracy performance of the SS-ConvNeXt and other compared algorithms under a different number of labeled samples; four errorbar plots are drawn based on the OA. The proportion of training samples, fixed training samples, and training samples for per class for the IN dataset is in the set {1%, 3%, 5%}, {100, 150, 200, 250, 300}, and {5, 10, 15}, respectively. For PU dataset, the proportion of training samples, fixed training samples, and training samples for per class is in the set {1%, 3%, 5%}, {100, 150, 200, 250, 300}, and {5, 10, 25}, respectively. The proportion of training samples, fixed training samples, and training samples for per class for the WHHH dataset is in the set {0.1%, 0.3%, 0.5%}, {200, 300, 500}, and {10, 25, 50}, respectively, so is the WHHC dataset. Each experiment was repeated ten times and the results were averaged.

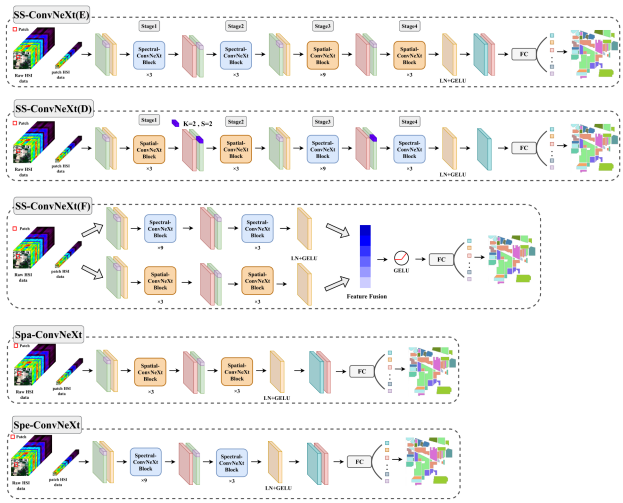

Fig. 2. Variants of the SS-ConvNeXt used for ablation analysis. Zoom in for best.

The results are shown in Fig. 3(a)–(d). In general, the classification accuracy of each algorithm increases as the number of training samples increases. Moreover, deep learning models exhibit better stability in their classification results, as reflected by their lower standard deviation compared to the traditional classification methods (i.e., SVM and 1D-CNN); as anticipated, the results unequivocally demonstrate that the proposed SS-ConvNeXt surpasses other methods with superior OA values on all four datasets. Since the labeling process of HSI data samples is time consuming, the classification performance in the case of small samples can better test the quality of the algorithm. For example, in the PU dataset, under 1% training proportion, our SS-ConvNeXt's OA can reach 98.50%; and in the WHHH dataset, under 0.1% training ratio, our SS-ConvNeXt's OA can reach 92.8%, which is much higher than other algorithms.

Tables I–IV also shows the numerical results of four datasets.

1) For IN dataset, in Table I, the proposed SS-ConvNeXt model achieves an OA of 94.34% with only 200 labeled training samples, which is 4% higher than the second best method, i.e., SSTN. In Fig. 3(a), the bar of the proposed method is much higher than the other methods, regardless of the number of training samples. With the increase of training samples, the OAs obtained by the proposed method increase very fast, but its standard deviations decreases.
2) For PU dataset, in Table II, the SS-ConvNeXt achieves an OA of 96.83% with only 200 labeled training samples,

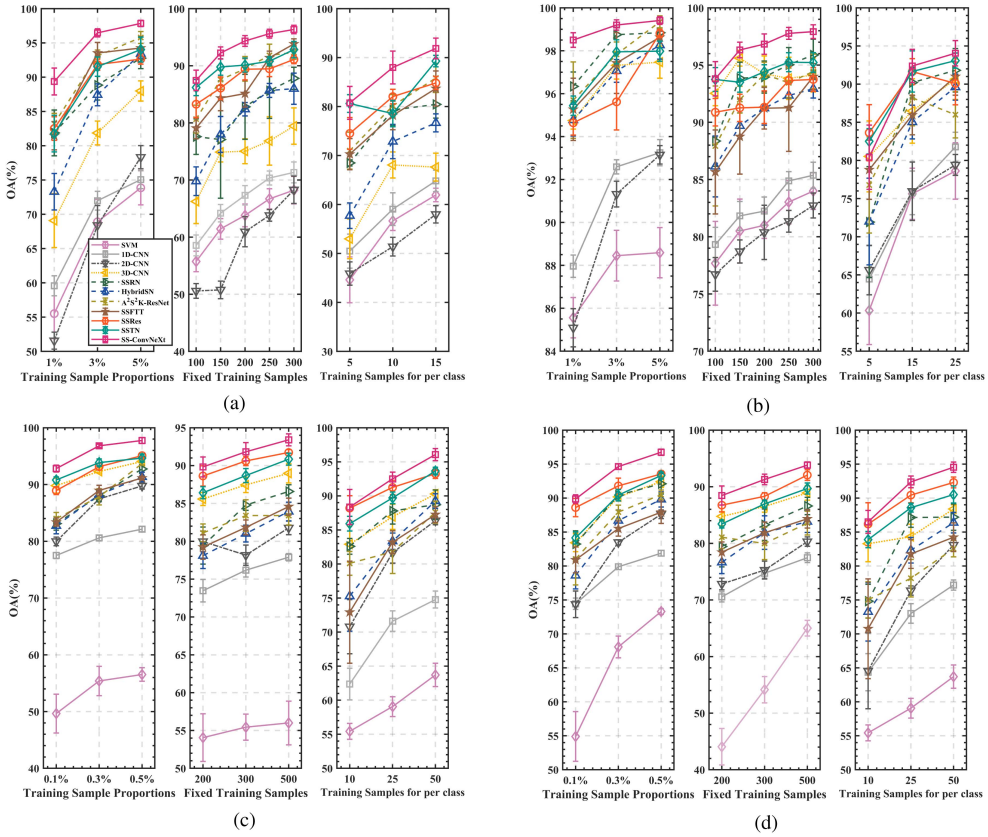

Fig. 3. Errorbar plots of OA (averaged by ten runs) achieved by other networks and the proposed model, SS-ConvNeXt, on four datasets with different training samples. Globally, as is shown in the picture, the proposed model, SS-ConvNeXt, achieves the best classification accuracy compared with other representative networks. (a) IN dataset. (b) PU dataset. (c) WHHH dataset. (d) WHHC dataset.

TABLE I

CLASSIFICATION ACCURACIES *AVERAGED BY TEN EXPERIMENTS* (VALUES ± STANDARD DEVIATION) ON THE IN DATASET USING *200 TRAINING SAMPLES* (OPTIMAL RESULTS (OA,AA,CA,k) ARE COLORED SHADOW AND OPTIMAL STANDARD DEVIATION IN BOLD)

Class no.	Color	Method									
		SVM	1D-CNN	2D-CNN	3D-CNN	SSRN	HybridSN	A²S²K-ResNet	SSFTT	SSRes	SSTN	***SS-ConvNeXt***
1		57.63±19.04	68.91±11.09	57.83±9.39	81.82±13.65	91.58±7.21	86.32±16.49	96.84±**3.68**	85.87±14.43	91.74±8.80	89.95±6.87	93.48±5.23
2		57.46±9.20	57.25±6.43	49.94±4.77	70.64±8.66	76.19±16.59	80.55±6.74	87.81±6.92	78.27±8.58	86.77±5.08	90.92±5.99	93.91±**1.89**
3		45.64±6.43	46.01±7.72	41.71±8.57	56.39±12.29	75.40±17.76	73.26±7.84	85.67±8.99	83.39±14.25	88.70±5.06	73.16±7.44	95.35±**4.89**
4		25.79±14.51	25.95±11.52	27.97±**5.81**	62.33±16.50	52.24±21.03	53.71±13.08	76.43±14.67	64.14±13.93	90.00±9.38	78.76±7.10	84.18±8.94
5		68.54±7.88	59.69±13.48	36.31±9.28	63.06±11.04	82.01±15.47	84.96±**4.55**	87.35±9.64	82.90±6.63	72.84±8.75	87.65±5.00	86.09±7.47
6		89.44±6.09	87.78±5.14	67.34±6.53	61.75±10.40	97.02±4.78	97.53±1.49	97.78±2.78	94.81±3.48	94.32±2.32	98.00±**0.70**	98.16±0.96
7		85.45±14.81	88.574.46	68.57±12.76	91.32±14.30	94.29±9.20	99.05±2.01	100.00±**0.00**	96.79±5.94	97.50±3.78	96.36±10.00	98.93±3.39
8		88.81±5.15	88.16±6.72	89.44±3.17	76.81±10.59	96.48±2.65	94.92±4.78	97.83±3.54	96.09±2.25	98.24±2.70	95.27±4.85	99.67±**0.67**
9		59.29±10.13	77.00±9.54	89.00±8.00	82.47±18.17	89.23±19.59	90.00±13.59	98.46±4.87	92.00±11.35	98.50±2.42	91.43±7.38	99.00±**2.11**
10		58.26±8.40	61.60±7.31	56.63±7.13	74.67±8.43	69.76±9.75	72.12±5.74	82.44±6.09	82.03±8.64	83.79±6.00	88.41±**3.91**	88.80±3.94
11		63.94±3.39	75.94±3.28	70.50±5.96	81.10±5.16	88.15±3.95	84.01±5.34	89.66±3.86	88.58±5.66	92.09±3.06	93.82±4.35	95.72±**2.62**
12		36.86±**4.72**	30.96±10.67	33.78±7.44	64.28±6.87	61.73±20.27	63.77±13.46	81.96±7.93	73.95±9.75	76.24±11.74	85.43±8.28	92.14±7.15
13		88.84±5.84	86.63±6.39	64.00±12.10	57.48±12.73	95.17±5.01	97.72±6.24	95.72±4.99	94.05±8.60	95.02±6.24	91.06±3.43	99.02±**0.56**
14		86.05±5.65	91.00±3.50	88.33±1.85	87.65±5.51	97.06±2.76	95.17±3.52	96.99±3.08	91.84±6.07	97.11±**1.14**	97.84±1.42	98.70±1.59
15		27.26±**3.88**	41.32±4.73	41.42±6.20	59.89±10.22	73.33±20.11	70.29±14.33	94.88±8.54	76.06±12.79	86.24±13.18	73.25±14.88	91.27±7.10
16		84.20±4.86	87.85±4.46	78.60±13.60	71.89±16.41	92.25±8.49	88.75±20.03	96.00±6.48	89.46±11.26	97.42±4.22	94.02±7.12	96.34±**4.16**
OA(%)		63.81±1.99	67.32±1.70	60.95±2.64	72.67±3.73	82.81±5.69	82.39±1.17	89.46±2.16	85.13±2.36	89.39±1.86	90.12±1.23	94.34±**0.94**
AA(%)		63.97±2.36	67.16±2.01	60.09±2.64	71.47±4.09	83.24±6.93	83.26±1.47	90.99±1.64	85.56±2.95	90.41±1.83	89.08±2.16	94.31±**1.31**
k*100		59.49±2.24	62.41±1.98	55.28±3.01	69.39±4.16	80.66±6.72	80.36±1.41	88.27±2.34	83.06±2.72	87.91±2.11	88.90±1.37	93.55±**1.07**

TABLE II

CLASSIFICATION ACCURACIES *AVERAGED BY TEN EXPERIMENTS* (VALUES ± STANDARD DEVIATION) ON THE PU DATASET USING *200 TRAINING SAMPLES* (OPTIMAL RESULTS (OA,AA,CA,k) ARE COLORED SHADOW AND OPTIMAL STANDARD DEVIATION IN BOLD)

Class no.	Color	Method										
		SVM	1D-CNN	2D-CNN	3D-CNN	SSRN	HybridSN	A²S²K-ResNet	SSFTT	SSRes	SSTN	***SS-ConvNeXt***
1		82.16±4.97	83.30±3.26	87.30±**1.76**	94.72±4.53	94.16±4.85	94.05±4.19	96.28±2.67	74.24±12.48	91.25±3.73	93.09±1.79	96.92±2.06
2		84.44±4.11	94.35±2.39	90.26±1.71	98.72±4.53	98.61±0.89	98.37±1.35	98.78±**0.66**	98.84±1.24	93.19±1.84	93.62±1.37	98.93±1.25
3		56.02±10.74	45.32±8.93	47.33±7.81	72.14±13.52	69.03±13.74	76.06±8.08	74.04±8.58	57.98±14.60	66.07±12.47	86.33±5.65	91.01±**2.86**
4		83.45±6.07	77.44±5.42	91.45±6.79	89.60±3.64	91.16±2.66	96.60±5.15	93.51±3.19	86.07±4.93	94.92±**1.22**	92.52±1.51	93.22±4.01
5		97.39±2.76	98.48±6.53	95.42±4.88	99.36±0.48	99.50±0.54	99.98±**0.05**	99.20±2.08	99.69±0.65	99.51±0.44	99.69±0.59	99.95±0.06
6		79.15±6.89	48.63±7.10	53.96±8.60	96.39±3.40	91.60±4.89	75.59±6.83	86.39±6.40	92.33±4.37	91.72±4.45	99.66±**0.57**	94.47±3.41
7		58.39±18.25	70.20±9.18	58.30±11.83	779.82±25.75	83.07±10.99	94.31±3.24	87.89±11.12	94.43±6.65	86.10±7.05	98.34±**1.28**	95.05±4.57
8		73.61±5.00	83.95±5.10	70.90±10.89	89.25±9.66	90.42±6.22	80.07±6.90	89.49±6.87	92.18±4.71	90.29±5.12	95.41±**1.78**	95.49±1.88
9		99.78±**0.18**	99.30±1.29	88.98±5.46	95.12±1.74	98.88±1.21	89.13±9.29	98.00±1.85	82.75±7.61	96.86±3.18	98.29±0.98	95.44±2.26
OA(%)		81.00±1.15	82.24±1.23	80.38±2.37	94.23±1.57	93.95±1.40	91.23±1.48	94.20±1.05	90.30±1.91	91.32±1.43	94.41±**0.62**	96.83±0.89
AA(%)		79.37±2.07	77.88±1.53	75.99±2.53	90.57±3.06	90.71±2.79	88.24±2.06	91.50±1.87	86.50±1.91	89.99±1.64	95.22±0.69	95.61±**0.48**
k*100		75.21±1.36	75.94±1.67	74.72±2.22	92.63±2.09	91.99±1.86	88.30±1.97	92.30±1.40	87.16±2.51	88.62±1.86	92.71±**0.78**	95.80±1.18

TABLE III
CLASSIFICATION ACCURACIES *AVERAGED BY TEN EXPERIMENTS* (VALUES ± STANDARD DEVIATION) ON THE WHHH DATASET USING *0.5% TRAINING SAMPLES* (OPTIMAL RESULTS (OA,AA,CA, AND k) ARE COLORED SHADOW AND OPTIMAL STANDARD DEVIATION IN BOLD)

Class no.	Color	Method										
		SVM	1D-CNN	2D-CNN	3D-CNN	SSRN	HybridSN	A^2S^2K-ResNet	SSFTT	SSRes	SSTN	***SS-ConvNeXt***
1		77.08±2.66	90.63±2.04	96.77±**0.56**	94.74±1.98	95.62±1.83	95.67±2.06	97.42±1.17	95.73±0.86	96.44±0.73	96.02±2.48	97.82±1.06
2		76.40±4.38	67.11±8.58	84.01±4.36	78.06±**1.72**	81.43±5.64	78.05±5.84	78.44±6.84	73.60±8.40	83.97±7.75	90.60±3.46	90.36±4.88
3		74.02±6.66	88.00±3.20	91.6±1.09	91.99±3.25	93.25±3.15	92.65±1.44	94.6±**0.79**	92.50±2.38	94.02±2.66	93.05±2.65	96.18±1.61
4		70.25±4.56	98.25±0.53	99.15±0.24	99.05±0.60	99.43±**0.21**	99.16±0.30	99.3±0.22	99.35±0.44	99.26±0.54	99.10±0.68	99.70±0.43
5		70.25±5.77	39.86±8.06	79.04±5.19	89.30±4.37	85.57±6.11	81.29±5.52	89.58±4.16	69.83±6.59	89.30±4.08	89.38±4.54	95.30±**1.37**
6		62.28±3.44	89.68±1.96	94.73±1.16	97.88±**0.64**	97.27±1.04	96.81±1.03	96.99±0.78	96.06±0.94	98.06±0.75	98.13±1.19	99.22±0.69
7		38.28±3.84	72.70±3.64	79.23±2.92	90.52±3.43	88.68±4.02	86.69±2.65	90.63±1.72	88.58±2.83	92.24±2.24	87.89±5.07	97.36±**1.39**
8		15.69±8.06	6.24±2.88	19.27±**2.72**	64.13±8.50	42.82±17.20	50.64±7.80	50.20±6.91	47.34±5.26	54.45±5.95	62.63±14.67	81.23±8.20
9		65.41±3.14	92.89±1.13	96.46±1.17	97.50±**0.66**	97.11±1.68	96.54±1.56	96.62±1.30	96.81±1.00	96.77±1.84	96.82±2.57	98.59±1.98
10		9.87±5.77	50.70±6.03	74.16±3.32	88.81±5.33	82.26±7.71	80.68±4.73	84.66±2.89	79.62±3.35	92.48±3.05	86.94±6.40	95.95±**2.41**
11		7.20±3.21	40.55±6.54	64.36±2.89	81.24±8.80	79.69±9.65	77.46±5.15	82.49±2.97	72.49±4.54	88.20±3.76	85.22±6.35	96.13±**1.47**
12		48.22±8.64	47.71±6.80	69.34±4.31	77.83±6.71	70.75±9.56	60.87±4.68	75.57±**3.36**	65.98±5.24	82.91±3.79	79.86±5.80	90.98±3.43
13		43.06±7.38	64.28±4.76	79.51±3.52	87.83±2.39	83.09±3.32	79.14±4.18	85.85±2.36	77.48±3.05	89.81±**1.39**	91.91±2.03	95.29±1.63
14		46.21±3.52	59.82±4.90	78.67±3.74	88.35±3.10	88.15±3.07	81.45±5.73	87.95±4.91	85.11±3.75	93.70±1.89	93.20±3.79	96.36±**1.19**
15		38.63±10.11	34.39±14.21	23.05±10.57	77.28±13.58	68.18±27.50	78.84±11.24	78.96±9.19	73.25±8.29	62.87±15.22	97.13±**3.08**	87.37±5.23
16		56.88±12.80	78.33±6.97	92.73±2,88	98.24±0.97	94.58±5.28	90.67±3.05	94.40±1.56	90.32±2.56	98.31±1.14	94.44±3.56	98.78±**0.97**
17		42.47±18.24	53.39±12.91	76.85±6.61	89.93±8.90	86.32±6.92	79.84±7.08	85.96±4.93	83.19±7.40	90.10±**1.14**	96.58±1.88	97.13±2.20
18		38.59±4.72	34.83±11.49	58.80±6.28	84.26±8.10	76.9±20.82	79.34±8.76	81.58±8.97	81.72±6.51	85.79±6.10	89.71±3.68	94.18±**2.57**
19		11.35±4.27	74.04±3.93	85.62±2.77	86.10±4.81	86.95±5.45	84.15±3.73	88.14±2.85	86.99±**2.20**	92.96±2.29	90.27±3.81	95.88±2.70
20		48.67±11.57	59.79±6.96	86.03±4.44	91.78±**1.20**	78.53±15.03	64.21±9.59	83.48±6.38	60.11±5.67	89.60±4.18	90.71±3.33	94.50±2.89
21		44.36±14.52	7.12±**4.72**	29.44±8.04	62.99±23.32	45.78±23.14	48.33±13.98	64.75±9.11	32.82±8.45	44.17±17.11	62.90±13.52	79.39±10.32
22		58.03±8.04	34.50±8.32	74.28±4.90	91.06±**3.12**	82.39±8.38	72.71±4.88	85.30±6.03	76.99±4.08	91.85±4.33	85.48±6.47	95.60±4.59
OA(%)		58.28±1.76	82.14±0.47	89.75±0.55	94.06±0.26	92.74±2.12	91.33±0.84	93.68±**0.26**	91.14±0.33	95.09±0.30	94.66±0.72	97.75±0.53
AA(%)		47.30±1.05	58.40±1.67	74.23±1.66	86.77±5.94	82.03±6.16	79.78±1.82	85.13±1.17	78.44±**1.02**	86.69±1.39	89.00±1.44	94.24±1.31
k*100		50.50±1.65	77.21±0.61	87.02±0.70	92.52±0.43	90.86±2.69	89.09±1.05	92.04±**0.34**	88.76±0.41	93.82±0.38	93.28±0.90	97.17±0.66

TABLE IV
CLASSIFICATION ACCURACIES *AVERAGED BY TEN EXPERIMENTS* (VALUES ± STANDARD DEVIATION) ON THE WHHC DATASET USING *0.5% TRAINING SAMPLES* (OPTIMAL RESULTS (OA,AA,CA, AND k) ARE COLORED SHADOW AND OPTIMAL STANDARD DEVIATION IN BOLD)

Class no.	Color	Method										
		SVM	1D-CNN	2D-CNN	3D-CNN	SSRN	HybridSN	A^2S^2K-ResNet	SSFTT	SSRes	SSTN	***SS-ConvNeXt***
1		82.01±2.22	92.30±0.97	95.75±0.98	95.41±0.77	94.60±4.30	93.39±1.96	94.77±1.11	89.78±7.23	97.55±0.81	95.99±1.81	98.62±**0.56**
2		58.63±1.73	71.80±2.71	87.67±**1.14**	92.27±1.77	93.09±1.45	90.11±2.42	87.97±2.03	88.68±1.70	91.79±3.56	93.01±2.26	97.29±1.19
3		59.47±3.43	72.31±3.46	70.94±5.13	92.32±3.09	87.91±5.29	80.85±4.34	83.95±3.78	76.78±4.02	90.72±4.05	90.48±3.16	97.20±**2.17**
4		75.30±4.70	90.68±2.43	72.43±7.98	96.03±**0.85**	95.11±1.97	94.66±2.22	94.95±2.02	95.17±1.24	96.25±1.60	95.17±2.59	96.69±1.82
5		10.73±**5.56**	15.01±6.91	50.77±9.22	86.75±7.27	51.14±17.86	36.15±13.01	39.97±12.21	38.10±12.80	83.32±11.86	81.67±12.45	88.15±8.86
6		20.73±4.06	11.45±4.25	24.47±4.02	54.53±11.63	59.71±10.28	53.68±6.37	60.19±**3.42**	47.10±11.08	63.33±7.25	62.94±4.61	79.82±10.30
7		56.09±4.69	69.55±4.24	84.08±4.49	86.86±2.37	84.79±6.45	82.23±4.14	83.83±6.14	80.33±7.51	85.75±4.45	90.32±2.67	94.24±**2.44**
8		49.79±**1.88**	68.15±2.41	74.41±2.32	88.98±2.51	88.18±2.89	85.68±2.31	85.19±2.01	81.87±3.69	90.41±3.18	82.98±6.39	94.30±3.06
9		38.99±4.22	56.97±5.66	66.95±3.56	85.38±5.46	86.95±5.41	74.99±6.29	77.81±3.37	76.67±5.42	87.72±6.01	90.40±**2.38**	94.05±3.50
10		81.00±2.70	90.32±1.42	88.71±2.73	96.29±1.77	96.03±2.30	93.79±4.37	93.43±2.56	93.10±4.36	94.65±2.68	98.16±**0.83**	97.82±2.52
11		82.22±2.80	86.47±2.48	94.25±1.39	95.40±1.22	96.27±1.50	94.54±2.36	94.87±2.17	91.91±3.35	97.64±2.03	96.53±1.29	98.17±**1.21**
12		21.63±3.09	22.65±4.96	65.70±8.89	77.34±5.33	55.90±12.83	49.76±6.85	55.79±12.73	44.20±9.15	78.32±9.39	60.90±9.29	93.93±**2.78**
13		35.86±**2.24**	48.81±4.50	66.86±2.62	75.78±8.10	69.67±4.19	67.19±3.10	69.42±4.49	62.52±8.99	77.48±4.15	82.11±4.58	85.33±3.72
14		62.12±4.43	79.48±3.29	87.77±2.26	85.26±6.32	91.12±3.28	86.60±2.14	88.60±2.45	85.78±3.63	90.02±2.02	90.71±2.60	94.83±**1.48**
15		40.38±10.22	47.80±13.49	47.24±9.61	70.97±8.52	60.27±13.18	55.17±10.32	56.27±12.39	52.48±8.30	75.06±8.15	79.97±8.00	72.40±**4.68**
16		96.91±0.44	97.89±0.40	99.01±0.26	98.54±1.32	99.14±0.53	99.28±0.36	99.16±0.50	98.92±0.51	98.72±0.98	99.54±**0.20**	99.56±0.31
OA(%)		73.29±**0.42**	81.88±**0.42**	87.61±0.64	92.41±0.85	92.10±1.50	89.80±1.12	90.50±0.68	87.92±1.65	93.55±0.46	93.32±0.52	96.74±0.46
AA(%)		54.49±**0.69**	63.85±1.06	73.56±1.43	91.15±0.98	81.87±3.37	77.39±1.88	79.14±2.09	75.21±2.25	87.42±1.31	86.93±1.44	92.65±1.17
k*100		68.94±**0.48**	78.74±0.49	85.48±0.75	86.13±0.41	90.81±1.74	88.13±1.29	88.95±0.78	85.87±1.91	92.47±0.54	92.21±0.61	96.20±0.54

which is about 2% higher than the second best method, i.e., SSTN. Moreover, in Fig. 3(b), the bar of the SS-ConvNeXt is higher than the other methods in all cases, with the only exception when there is five labeled samples per class. In Fig. 3(b), the standard deviation of the SS-ConvNeXt is lower than the other methods.

3) For WHHH dataset, in Table III, the SS-ConvNeXt achieves an OA of 97.75% with only 0.5% training samples, which is about 2.6% higher than the second best method, i.e., SSRes. Fig. 3(c) indicates that the proposed methods can outperform all methods in all cases.
4) For WHHC dataset, in Table IV, the SS-ConvNeXt achieves an OA of 96.74% with only 0.5% training samples, which is 3% higher than the transformer-based model SSTN. In Fig. 3(d), with the increase in training numbers, the SS-ConvNeXt significantly performs better than other networks.

Table V compares the proposed method with the published results of another advanced methods, which indicates that the proposed approach performs the best in most cases.

E. Visual Evaluation

Figs. 4–7 show classification maps of different methods on four datasets. Region of interests are used to highlight differences. Overall, on all datasets, the proposed SS-ConvNeXt offers better classification maps that are closest to the ground-truth map. Moreover, referring to RGB composite image, the SS-ConvNeXt shows less inner class misclassification, more accurate class boundaries and edges, and finer details with a less oversmoothing phenomenon.

The conventional approach, as exemplified by the SVM and 1D-CNN models, yields classification maps that are noisy and exhibit discontinuous land cover blocks, resulting in rough classification outcomes. Classic backbone networks, as exemplified by 2D-CNN and 3D-CNN models, and HybridSN, show better classification maps with less noise. The method based on the residual network, as exemplified by SSRN, A^2S^2K-ResNet, and SSRes, has strong feature extraction ability, which improves the classification accuracy to a certain extent. A transformer-based network, represented by SSFTT and SSTN, performs better because of the attention mechanism.

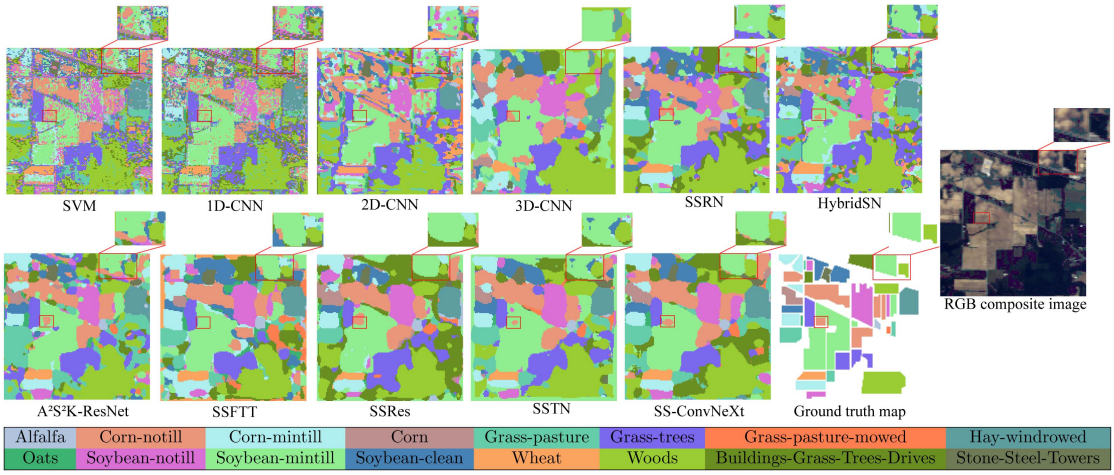

Fig. 4. Classification maps obtained by other models and SS-ConvNeXt on the IN dataset with *200 training samples*. A local area (red square) is demarcated and zoomed for easy observation.

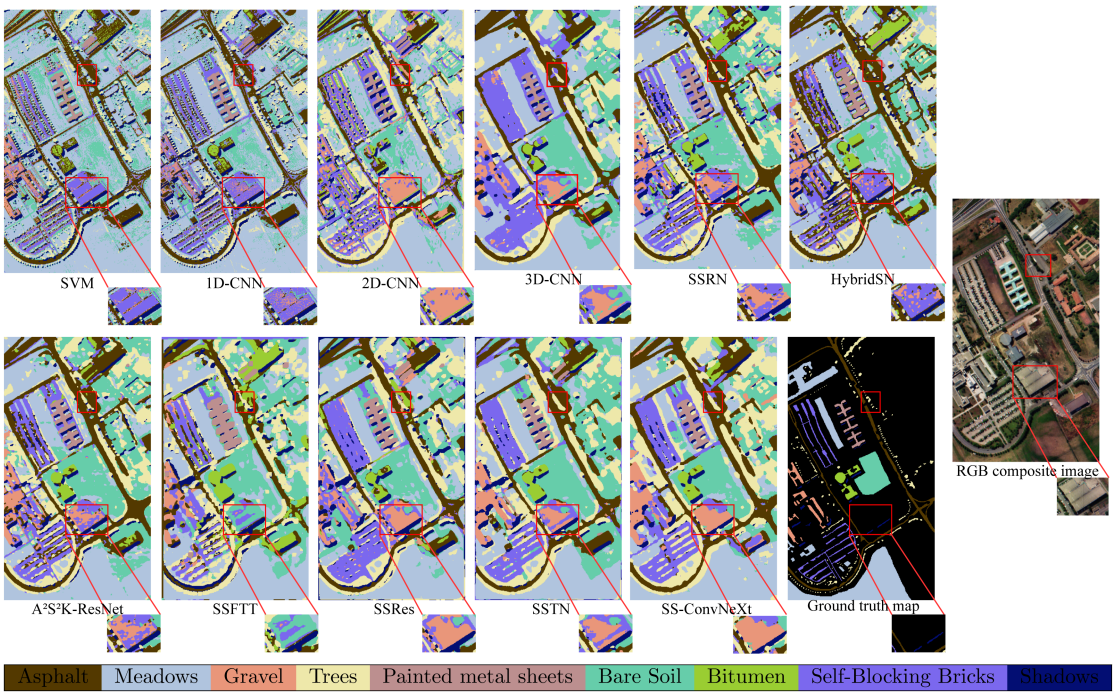

Fig. 5. Classification maps obtained by other models and SS-ConvNeXt on the PU dataset with *200 training samples*. A local area (red square) is demarcated and zoomed for easy observation.

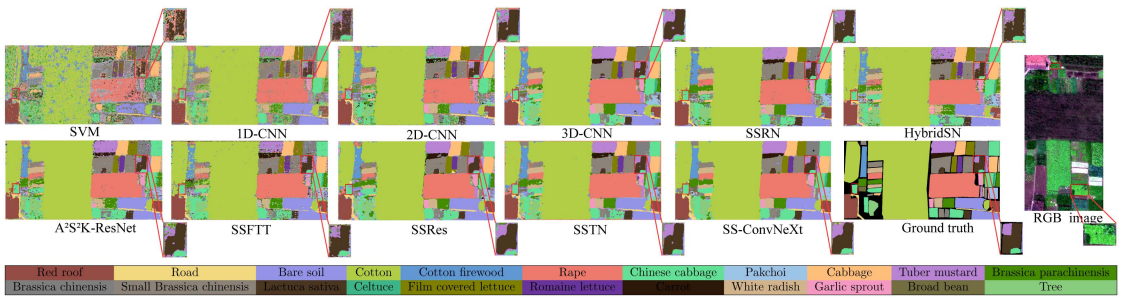

Fig. 6. Classification maps obtained by other models and SS-ConvNeXt on the WHHH dataset with *0.5% training samples*. A local area (red square) is demarcated and zoomed for easy observation (best view in zoom in).

TABLE V
COMPARISON BETWEEN SS-CONVNEXT AND SIX OTHER ADVANCED METHODS ON DIFFERENT DATASETS (OPTIMAL RESULTS (OA,AA,k) ARE COLORED SHADOW AND OPTIMAL STANDARD DEVIATION IN BOLD)

HSI Dataset	Indian Pines (IN)				Pavia University (PU)				
Compare with few-shot learning methods	***SS-ConvNeXt***	S-DMM [43]	UM²-L [44]	DCFSL [45]	***SS-ConvNeXt***		S-DMM	UM²-L	DCFSL
Sample size per category	5	-	5	5	5	10	10	5	5
OA(%)	80.69±3.40	-	72.09±**2.20**	66.81±2.37	81.81±4.36	88.01±2.78	87.59±**2.43**	80.92±3.55	83.65±**1.77**
AA(%)	88.50±1.91	-	64.84±3.17	77.89±**0.86**	86.45±1.89	91.80±1.28	92.89±1.30	77.77±3.19	83.77±**1.74**
Kappa	0.78±0.04	-	0.69±0.02	0.63±**0.01**	0.76±0.05	0.84±0.03	0.84±0.03	0.76±0.04	0.79±**0.02**
Training time(s)	506	-	-	-	300	567	-	-	-
Prediction time(s)	14	-	-	-	138	130	-	-	-

HSI dataset	Indian Pines (IN)				Pavia University (PU)	
Compare with transformer-based methods	***SS-ConvNeXt***		SpectralFormer [26]	SST-FA [28]	***SS-ConvNeXt***	SST-FA
Training samples number	15 per class	total 200	50 per class	total 200	total 200	total 200
OA(%)	91.87±**2.11**	94.34±**0.94**	81.76	88.98±1.96	96.83±**0.89**	93.37±1.96
AA(%)	95.40±**1.11**	94.31±1.31	87.81	68.15±**1.06**	**95.61±0.48**	85.01±3.78
Kappa	0.91±**0.02**	0.94±**0.01**	0.79	0.87±0.01	0.96±**0.01**	0.92±0.02
Training time(s)	1203	1399	-	-	1222	-
Prediction time(s)	13	13	-	-	120	-

HSI dataset	WHU-Hi-HongHu (WHHH)		WHU-Hi-HanChuan (WHHC)	
Compare with CRF-based methods	***SS-ConvNeXt***	CNNCRF [46]	***SS-ConvNeXt***	CNNCRF
Sample size per category	100	100	100	100
OA(%)	97.53±**0.41**	93.74	96.44±**0.51**	93.95
AA(%)	97.75±**0.52**	94.78	96.74±**0.38**	94.78
Kappa	0.97±**0.01**	0.92	0.96±**0.01**	0.93
Training time(s)	2784	502	2166	480
Prediction time(s)	346	811	370	412

* The CAs of SpectralFormer, SST-FA, S-DMM, UM²L, DCFSL, SST-FA and CNNCRF are directly quoted from references [43], [44], [45], [26], [28] and [46] respectively.

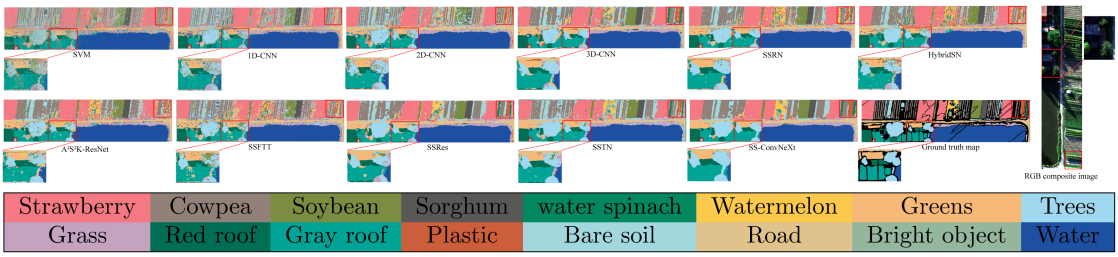

Fig. 7. Classification maps obtained by other models and SS-ConvNeXt on the WHHC dataset with *0.5% training samples*. A local area (red square) is demarcated and zoomed for easy observation (best view in zoom in).

TABLE VI
ABLATION ANALYSIS OF THE PROPOSED SS-CONVNEXT WITH A COMBINATION OF DIFFERENT MODULES AND ORIGINAL CONVNEXT-T ON THE IN DATASET WITH 200 TRAINING SAMPLES (OPTIMAL RESULTS (OA,AA,CA, AND k) ARE COLORED SHADOW AND OPTIMAL STANDARD DEVIATION IN BOLD)

Case	Component			Indices			blocks	dims[1]	#param	Training Time	Prediction Time
	Spa-cv	Spe-cv	Fusion Module	OA	AA	k*100					
ConvNeXt-T²[32]	✗	✗	✗	59.67±1.24	54.70±1.38	55.41±2.09	[3,3,9,3]	[96,192,384,768]	28.14M	1058s	10s
SS-ConvNeXt(D)[3]	✓	✓	✗	86.98±**0.67**	83.59±1.29	85.58±**0.74**	[3,3,9,3]	[64,128,256,512]	14.64M	2662s	10s
SS-ConvNeXt(E)	✓	✓	✗	92.92±0.98	93.18±1.17	92.07±1.05	[3,3,9,3]	[64,128,256,512]	12.94M	1571s	8s
Spa-ConvNeXt	✓	✗	✗	89.85±1.22	89.26±2.20	88.66±1.34	[3,3]	[64,128]	1.41M	543s	2s
Spe-ConvNeXt	✗	✓	✗	91.71±2.24	88.96±4.56	90.68±2.57	[9,3]	[64,128]	1.41M	1074s	3s
SS-ConvNeXt(F)	✓	✓	mul	89.40±0.89	87.02±3.14	88.18±0.98	Spa:[3,3],Spe:[9,3]	[64,128]	1.39M	1619s	4s
			add	90.15±2.41	88.21±4.13	88.89±2.77				1632s	4s
			cat	91.52±1.68	90.63±1.85	90.50±1.85				1626s	4s
SS-ConvNeXt	✓	✓	✗	93.56±1.43	93.73±1.68	92.82±1.61	[1,1,3,1]	[64,128,256,512]	4.60M	835s	6s
				93.37±1.22	93.27±1.98	92.64±1.38	[2,2,6,2]		8.91M	1253s	9s
				94.03±0.83	94.31±**1.17**	93.34±0.94	[3,3,9,3]		13.12M	1241s	12s

[1] Here, "dim" means the output dimension of Spa-cv and Spe-cv modules respectively. For example, [64,128,256,512] means that the output dimensions of Spa-cv module are 64 and 128, the output dimensions of Spe-cv module are 256 and 512.

The figure indicates that the SS-ConvNeXt outperforms other methods in identifying most areas for the Corn-notill class (the red box on the left in Fig. 4) in the IN dataset, while also maintaining more precise class boundaries and edges as shown in the area that is zoomed in. The SS-ConvNeXt has also accurately classified the building boundary on the PU dataset. The WHHH and WHHC datasets demonstrate that the SS-ConvNeXt has better performance in terms of clearer delineation, although the distribution structure of the ground cover of these two agricultural scenes is very large and complex.

F. Ablation Analysis

Table VI shows the results achieved by variants of the proposed SS-ConvNeXt model, whose architecture is illustrated in Fig. 2. The ConvNeXt-T is also included for comparison. As we can see in Table VI, the SS-ConvNeXt outperforms all its variants and the original ConvNeXt-T method. In Table VI, the SS-ConvNeXt implementations with different number of blocks (i.e., [1,1,3,1], [2,2,6,2,], and [3,3,9,3]) achieve the similar classification performance. In this article, we use the number of block in [3,3,9,3] as shown in Fig. 1. The poor performance

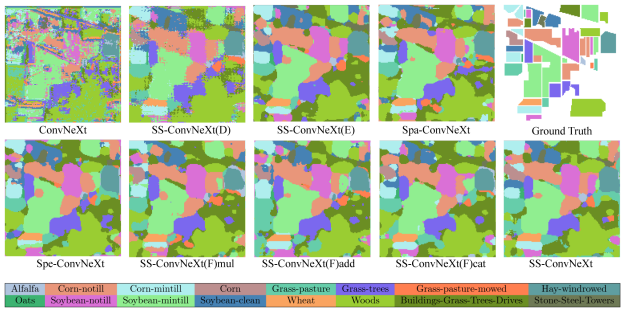

Fig. 8. Classification maps obtained by different variants of the proposed SS-ConvNeXt model, as well as the original ConvNeXt-T approach on the IN dataset with 200 training samples.

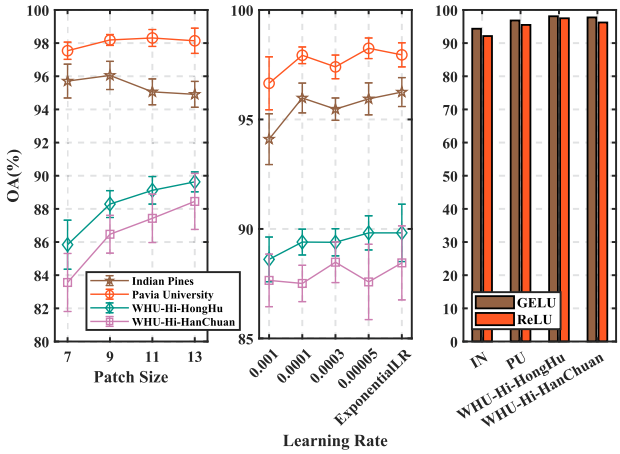

Fig. 9. Errorbar plots of the OA (averaged by ten runs) achieved by different hyperparameter settings (i.e., patch size and learning rate) and different activation functions.

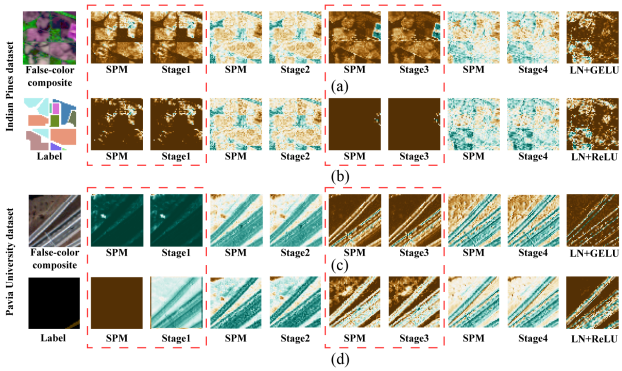

Fig. 10. Feature maps achieved by different activation functions (i.e., GELU and ReLU).

of the ConvNeXt is probably due to the fact that it is not designed for the HSIC, and thereby, cannot efficiently capture the discriminative spectral and spatial information. Accuracies of the SS-ConvNeXt is more than 1% higher than that of the SS-ConvNeXt(E).

The difference between the SS-ConvNeXt(D) and SS-ConvNeXt is that the SS-ConvNeXt(D) replaces the SPM in the SS-ConvNeXt with a 2×2 spatial downsampling layer before stages 2 and 4. The better performance of the SS-ConvNeXt and SS-ConvNeXt(D) demonstrates the superiority of the proposed SPM over spatial downsampling. The SS-ConvNeXt(F) with a concatenation operation shows better results. We observe that the parallel SS-ConvNeXt(F) performs worse than the serial SS-ConvNeXt, which might be because different brunches of the parallel SS-ConvNeXt(F) are concentrated in a manner that is insufficient for interactions between the spatial and spectral branches, whereas in serial SS-ConvNeXt, spatial and spectral information is extracted by stages, allowing more efficient extraction of both low- and high-level features in a hierarchical manner.

Fig. 8 shows classification maps of different variants of the SS-ConvNeXt. Overall, the SS-ConvNeXt provides better preserved class boundaries with less within-class artifacts and noise. Direct application of the original ConvNeXt-T model to HSIC gives the worst results.

G. Hyperparameter Sensitivity Analysis and Feature Map Visualization

Fig. 9 shows performance variation of the SS-ConvNeXt with changes of patch size, learning rate, and different activation functions (i.e., GELU and ReLU) on four datasets. Except for IN dataset, the accuracy indicator increases with patch size on the remaining three datasets. Additionally, the SR of these four datasets is quite different. IN has an SR of only 20 m, nevertheless, PU, WHHH, and WHI-Hi-HanChuan have an SR of 1.3, 0.043, 0.109 m, respectively. This influence of the window size can be interpreted as the smaller patch size containing insufficient spatial information on the high SR HSI dataset, and the larger patch size is not conducive to extracting key information on the low SR HSI dataset. Based on this observation, we set the patch size of the IN and PU datasets to be 9, and that of the WHHH and WHHC datasets to be 13.

We also conduct ablation study on the learning rate strategy. As we can see in Fig. 9(middle), the ExponentialLR strategy enables a higher performance than all other fixed learning-rate-based approaches. We, therefore, adopt ExponentialLR to train our model.

As shown in the red box of Fig. 10, feature maps achieved by the GELU function can better perceive detailed information than the ReLU function. SS-ConvNeXt with ReLU fails to perceive the boundary between different object types. The OA is slightly improved with the use of GELU as shown in Fig. 9(right). So, we use the GELU function in the SS-ConvNeXt.

IV. Discussion

A. "Serial structure (Spatial–Spectral)" or "Parallel scheme (Spatial and Spectral)"

We propose a spatial–spectral ConvNeXt approach for HSIC. The architecture of SSRN, SSTN, and SSRes are serial structure,

which means extracting spatial information first, and then, spectral information. The ablation analysis shows that serial structure SS-ConvNeXt performs worth than the parallel scheme SS-ConvNeXt(F), which combines the spatial and spectral branches separately. At the same time, it is proved that only extracting spatial or spectral information cannot solve the HSIC problem well.

B. How the Window Size Affects the Accuracy?

The spatial size of input data is one of the main factors that influence the HSIC performance. Based on the observation in ablation analysis, smaller spatialized input contains insufficient spatial information on high SR HSI dataset (e.g., WHHH dataset), and the larger spatialized input is not conducive to extracting key information on low SR HSI dataset (e.g., IN dataset). Consequently, to make a fair comparison, ensuring the consistency of the window size of the same dataset is a fair guarantee.

V. Conclusion

This article has presented a new spatial–spectral convolution neural network model, called SS-ConvNeXt, for the HSIC. This new model was inspired by the recent ConvNeXt model, which has demonstrated stronger feature representation capability than the popular ViT approaches. The proposed SS-ConvNeXt was tailor designed to the characteristics of HSIs, and thereby, can efficiently learn discriminative spatial–spectral information for the enhanced HSIC. To better learn the spatial and spectral information in the HSI, the Spa-cv and Spe-cv blocks were, respectively, designed. The depthwise and pointwise convolutions were adopted to reduce the model size and prevent vanishing gradient. The proposed model was evaluated against 14 other state-of-the-art methods on four different datasets. Moreover, extensive ablation studies were conducted to investigate the roles of building blocks in the proposed model. The results demonstrated that the proposed SS-ConvNeXt not only can achieve a high classification accuracy but also can better preserve class boundaries and reduce within-class noise.

References

[1] P. Ghamisi et al., "New frontiers in spectral-spatial hyperspectral image classification: The latest advances based on mathematical morphology, Markov random fields, segmentation, sparse representation, and deep learning," *IEEE Geosci. Remote Sens. Mag.*, vol. 6, no. 3, pp. 10–43, Sep. 2018.

[2] X. Li and G. Shao, "Object-based urban vegetation mapping with high-resolution aerial photography as a single data source," *Int. J. Remote Sens.*, vol. 34, no. 3, pp. 771–789, 2013, doi: 10.1080/01431161.2012.714508.

[3] B. B. Damodaran and R. R. Nidamanuri, "Dynamic linear classifier system for hyperspectral image classification for land cover mapping," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 7, no. 6, pp. 2080–2093, Jun. 2014.

[4] X. Tong, H. Xie, and Q. Weng, "Urban land cover classification with airborne hyperspectral data: What features to use?," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 7, no. 10, pp. 3998–4009, Oct. 2014.

[5] R. J. Murphy, S. Schneider, and S. T. Monteiro, "Consistency of measurements of wavelength position from hyperspectral imagery: Use of the ferric iron crystal field absorption at 900 nm as an indicator of mineralogy," *IEEE Trans. Geosci. Remote Sens.*, vol. 52, no. 5, pp. 2843–2857, May 2014.

[6] L. He, J. Li, C. Liu, and S. Li, "Recent advances on spectral–spatial hyperspectral image classification: An overview and new guidelines," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 3, pp. 1579–1597, Mar. 2018.

[7] 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. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0924271617303416

[8] P. Ghamisi, J. Plaza, Y. Chen, J. Li, and A. J. Plaza, "Advanced spectral classifiers for hyperspectral images: A review," *IEEE Geosci. Remote Sens. Mag.*, vol. 5, no. 1, pp. 8–32, Mar. 2017.

[9] J. Gualtieri and S. Chettri, "Support vector machines for classification of hyperspectral data," in *Proc. IEEE Int. Geosci. Remote Sens. Symp. Taking Pulse Planet, Role Remote Sens. Manag. Environ.*, 2000, pp. 813–815.

[10] M. Ye, C. Ji, H. Chen, L. Lei, H. Lu, and Y. Qian, "Residual deep PCA-based feature extraction for hyperspectral image classification," *Neural Comput. Appl.*, vol. 32, no. 18, pp. 14287–14300, Sep. 2020, doi: 10.1007/s00521-019-04503-3.

[11] P. Sidike, C. Chen, V. Asari, Y. Xu, and W. Li, "Classification of hyperspectral image using multiscale spatial texture features," in *Proc. 8th Workshop Hyperspectral Image Signal Process., Evol. Remote Sens.*, 2016, pp. 1–4.

[12] G. Camps-Valls, L. Gomez-Chova, J. Munoz-Mari, J. Vila-Frances, and J. Calpe-Maravilla, "Composite kernels for hyperspectral image classification," *IEEE Geosci. Remote Sens. Lett.*, vol. 3, no. 1, pp. 93–97, Jan. 2006.

[13] M. D. Mura, J. A. Benediktsson, B. Waske, and L. Bruzzone, "Extended profiles with morphological attribute filters for the analysis of hyperspectral data," *Int. J. Remote Sens.*, vol. 31, no. 22, pp. 5975–5991, 2010, doi: 10.1080/01431161.2010.512425.

[14] W. Hu, Y. Huang, L. Wei, F. Zhang, and H. Li, "Deep convolutional neural networks for hyperspectral image classification," *J. Sensors*, vol. 2015, Jul. 2015, Art. no. 258619, doi: 10.1155/2015/258619.

[15] L. Zhu, Y. Chen, P. Ghamisi, and J. A. Benediktsson, "Generative adversarial networks for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 9, pp. 5046–5063, Sep. 2018.

[16] X. Yang, Y. Ye, X. Li, R. Y. K. Lau, X. Zhang, and X. Huang, "Hyperspectral image classification with deep learning models," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 9, pp. 5408–5423, Sep. 2018.

[17] 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," *IEEE Geosci. Remote Sens. Lett.*, vol. 17, no. 2, pp. 277–281, Feb. 2020.

[18] 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.

[19] J. Yue, W. Zhao, S. Mao, and H. Liu, "Spectral–spatial classification of hyperspectral images using deep convolutional neural networks," *Remote Sens. Lett.*, vol. 6, no. 6, pp. 468–477, 2015, doi: 10.1080/2150704X.2015.1047045.

[20] C. Yu, R. Han, M. Song, C. Liu, and C.-I. Chang, "A simplified 2D-3D CNN architecture for hyperspectral image classification based on spatial–spectral fusion," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 13, pp. 2485–2501, Apr. 27, 2020.

[21] A. Qin, Z. Shang, J. Tian, Y. Wang, T. Zhang, and Y. Y. Tang, "Spectral–spatial graph convolutional networks for semisupervised hyperspectral image classification," *IEEE Geosci. Remote Sens. Lett.*, vol. 16, no. 2, pp. 241–245, Feb. 2019.

[22] Y. Wang, K. Li, L. Xu, Q. Wei, F. Wang, and Y. Chen, "A depthwise separable fully convolutional resnet with convcrf for semisupervised hyperspectral image classification," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 4621–4632, Apr. 15, 2021.

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

[24] L. Mou and X. X. Zhu, "Learning to pay attention on spectral domain: A spectral attention module-based convolutional network for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 1, pp. 110–122, Jan. 2020.

[25] 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.

[26] D. Hong et al., "SpectralFormer: Rethinking hyperspectral image classification with transformers," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5518615.

[27] B. Liu, A. Yu, K. Gao, X. Tan, Y. Sun, and X. Yu, "DSS-TRM: Deep spatial–spectral transformer for hyperspectral image classification," *Eur. J. Remote Sens.*, vol. 55, no. 1, pp. 103–114, 2022, doi: 10.1080/22797254.2021.2023910.
[28] X. He, Y. Chen, and Z. Lin, "Spatial-spectral transformer for hyperspectral image classification," *Remote Sens.*, vol. 13, no. 3, 2021, Art. no. 498. [Online]. Available: https://www.mdpi.com/2072-4292/13/3/498
[29] Z. Liu et al., "Swin transformer: Hierarchical vision transformer using shifted windows," *IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 9992–10 002.
[30] 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.
[31] A. Dosovitskiy et al., "An image is worth 16x16 words: Transformers for image recognition at scale," in *Proc. Int. Conf. Learn. Representations*, 2021.
[32] Z. Liu, H. Mao, C.-Y. Wu, C. Feichtenhofer, T. Darrell, and S. Xie, "A convNet for the 2020s," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 11966–11976.
[33] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2009, pp. 248–255.
[34] T. Lin et al., "Microsoft COCO: Common objects in context," in *Proc. Eur. Conf. Comput. Vis.*, Sep. 2014.
[35] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. 3rd Int. Conf. Learn. Representations*, 2015.
[36] 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.
[37] A. Ben Hamida, A. Benoit, P. Lambert, and C. Ben Amar, "3-D deep learning approach for remote sensing image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 8, pp. 4420–4434, Aug. 2018.
[38] 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.
[39] S. K. Roy, S. Manna, T. Song, and L. Bruzzone, "Attention-based adaptive spectral-spatial kernel ResNet for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 9, pp. 7831–7843, Sep. 2021.
[40] L. Sun, G. Zhao, Y. Zheng, and Z. Wu, "Spectral–spatial feature tokenization transformer for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, Jan. 18, 2022, Art. no. 5522214.
[41] K. Li et al., "Depthwise separable ResNet in the MAP framework for hyperspectral image classification," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, Nov. 4, 2020, Art. no. 5500305.
[42] Z. Zhong, Y. Li, L. Ma, J. Li, and W.-S. Zheng, "Spectral–spatial transformer network for hyperspectral image classification: A factorized architecture search framework," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, Oct. 8, 2021, Art. no. 5514715.
[43] B. Deng, S. Jia, and D. Shi, "Deep metric learning-based feature embedding for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 2, pp. 1422–1435, Feb. 2020.
[44] K. Gao, B. Liu, X. Yu, and A. Yu, "Unsupervised meta learning with multiview constraints for hyperspectral image small sample set classification," *IEEE Trans. Image Process.*, vol. 31, pp. 3449–3462, May 5, 2022.
[45] Z. Li, M. Liu, Y. Chen, Y. Xu, W. Li, and Q. Du, "Deep cross-domain few-shot learning for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, Feb. 17, 2021, Art. no. 5501618.
[46] Y. Zhong, X. Hu, C. Luo, X. Wang, J. Zhao, and L. Zhang, "WHU-Hi: UAV-borne hyperspectral with high spatial resolution (H2) benchmark datasets and classifier for precise crop identification based on deep convolutional neural network with CRF," *Remote Sens. Environ.*, vol. 250, 2020, Art. 112012. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0034425720303825

Yimin Zhu received the B.Eng. degree in geomatic engineering from Suqian University, Suqian, China, in 2021. He is currently working toward the M.Sc. degree in survey and mapping with the School of Land Science and Technology, China University of Geoscience, Beijing, China.

His research interests include hyperspectral image processing, artificial intelligence algorithm, and its application in remote sensing image.

Kexin Yuan received the B.Eng. degree in geographic information science from the Heilongjiang Institute of Technology, Harbin, China, in 2020. She is currently working toward the M.Sc. degree in survey and mapping with the School of Land Science and Technology, China University of Geosciences, Beijing, China.

Her research interests include remote sensing image processing and remote sensing image application.

Wenlong Zhong received the B.Eng. degree in surveying and mapping engineering from the East China University of Technology, Nanchang, China, in 2020. He is currently working toward the M.Sc. degree in surveying and mapping engineering with the China University of Geosciences, Beijing, China.

His research interests include hyperpectral image processing, land cover mapping, artificial intelligence algorithm, and its application in remote sensing image.

Linlin Xu (Member, IEEE) received the B.Eng. and M.Sc. degrees in geomatics engineering from the China University of Geosciences, Beijing, China, in 2007 and 2010, respectively, and the Ph.D. degree in remote sensing from the Department of Geography and Environmental Management, University of Waterloo, Waterloo, ON, Canada, in 2014.

He is currently a Research Assistant Professor with the Department of Systems Design Engineering, University of Waterloo. He has authored and co-authored various papers on high-impact remote sensing journals and conferences. His research interests include hyperspectral and synthetic aperture radar data processing and their applications in various environmental applications.