

Patch-Based Training of Fully Convolutional Network for Hyperspectral Image Classification With Sparse Point Labels

Xueliang Zhang¹, Member, IEEE, Zixian Zheng, Pengfeng Xiao², Senior Member, IEEE, Zhenshi Li, and Guangjun He³

I. INTRODUCTION

Abstract—Fully convolutional network (FCN), which has excellent capability for capturing spatial context, was introduced to improve the performance of hyperspectral image classification (HSIC). However, training FCN usually requires a huge amount of pixel-level labels, which is difficult to obtain for HSIC in practical applications. How to train an FCN effectively with the supervision of limited sparse point labels has attracted the attention. The patch-free training pattern with sparse point labels was proved to be effective for HSIC task. Then, as a general training mode for remote sensing image semantic segmentation, is patch-based training of FCN also effective for HSIC with sparse point labels? To answer this question, a patch-based training framework with a novel fully convolutional network is proposed for HSIC in this article. First, cropped hyperspectral image (HSI) patches with sparse labels are input for training. Second, considering the limitation of supervision with sparse points for training, a lightweight network on the basis of an encoder-decoder structure with shallow channels is specially designed for HSIC with the aid of residual connections in the encoder and the integration of multiple attention modules to fully exploit the spectral-spatial information of HSI. Third, conditional random field-loss is adopted as a prior complement to the point supervision for further excavation of spatial context information. The performance of the proposed method is quantitatively evaluated on three HSI datasets and achieves state-of-the-art performance in comparison with other representative methods, demonstrating the effectiveness of the patch-based training framework for HSIC.

Index Terms—Convolutional neural network, fully convolutional network, hyperspectral image classification, patch-based training, point label.

COMPOSED of hundreds of contiguous and narrow spectral bands, hyperspectral images (HSIs) are able to capture more abundant spectral signatures about the ground surface than multispectral images [1], [2], [3] and have been widely used for earth observation missions, including land cover mapping [4], precise crop identification [5], [6], mineral mapping [7], and environmental monitoring [8], [9]. Hyperspectral image classification (HSIC), which aims to assign a label to every pixel in an image, is a challenging task due to the limited labeled data and the nonlinear relation, feature redundancy, and the curse of dimensionality in high spectral resolution [1], [2].

For the past few decades, plentiful methods have been proposed for HSIC, which can be roughly categorized into spectral-based classification and spectral-spatial classification [10]. Spectral-based methods, such as support vector machine (SVM) [11], random forest [12], and dynamic or random subspace [13], [14], aim to explore the spectral signature difference in terms of pixels. However, adjacent pixels can provide vital spatial information for improving the performance of HSIC [15]. Hence, an increasing number of spectral-spatial classification methods integrating spectral signatures and spatial contexts emerged, including Gabor filter [16], [17], extended morphological profiles [18], [19], and multiple kernel [20], [21]. Although the abovementioned methods were proven to be effective for the HSIC task, they heavily relied on the handcrafted feature descriptors. Furthermore, the setting of the parameters always depends on expert knowledge [22].

With the rapid development of deep learning, deep learning-based methods have been widely used for HSIC due to their prominent ability to automatically extract discriminative features in a hierarchical learning way [23], [24], [25]. The stacked autoencoder (SAE) was first introduced to HSIC in [26]. Chen et al. [27] proposed a deep belief network (DBN) to learn the restricted Boltzmann machine network layer by layer. However, due to the fully connected layers in both SAE and DBN, these two models have a large number of parameters, which leads to a longer training process. Considering the characteristics of HSIs, recurrent neural network (RNN) models were used for HSIC to better understand the spectral signatures by processing the spectral information as time sequences [28]. Mou et al. [29] first proposed an RNN framework with modified gated recurrent

Manuscript received 25 June 2022; revised 28 August 2022 and 19 September 2022; accepted 1 October 2022. Date of publication 12 October 2022; date of current version 20 October 2022. This work was supported in part by the National Natural Science Foundation of China under Grant 42071297, Grant 41871235, and Grant 41871326, in part by the Fundamental Research Funds for the Central Universities under Grant 020914380095, and in part by the High-level Innovation and Entrepreneurship Talents Introduction Program of Jiangsu Province of China. (Corresponding author: Xueliang Zhang.)

Xueliang Zhang, Zixian Zheng, Pengfeng Xiao, and Zhenshi Li are with the Jiangsu Provincial Key Laboratory of Geographic Information Science and Technology, Key Laboratory for Land Satellite Remote Sensing Applications of Ministry of Natural Resources, School of Geography and Ocean Science, Nanjing University, Nanjing 210023, China (e-mail: zxl@nju.edu.cn; zhengzx95@gmail.com; xiaopf@nju.edu.cn; lzshshi@outlook.com).

Guangjun He is with the State Key Laboratory of Space-Ground Integrated Information Technology, Space Star Technology Co., Ltd., Beijing 100095, China (e-mail: hgjun_2006@163.com).

Digital Object Identifier 10.1109/JSTARS.2022.3212983

units and proved the effectiveness of deep recurrent networks for HSIC. After that, multiple RNN-based networks were proposed for improving the performance of HSIC. In addition to making full use of spectral signatures, some researchers attempted to combine spatial context in RNN-based structures. For example, Zhou et al. [30] adopted a two-branch long short-term memory (LSTM) network integrating spectral and spatial branches, and Sharma et al. [31] designed a patch-based RNN to capture spatial information from adjacent pixels.

Convolutional neural network (CNN), which reduces the numbers of parameters due to the local receptive fields and sharing weights and has hierarchical spectral-spatial feature representation ability [32], is a powerful technique for geographical information extraction [33], [34], and has been widely used for HSIC. HSI patches cropped with the centric pixel and its surrounding adjacent pixels with a fixed size are the input of CNN structures. Since the aim for HSIC is to assign a unique label for every single pixel, the CNN-based networks for HSIC can be summarized into two categories: classification networks and fully convolutional networks (FCNs).

Classification network aims at classifying the centric pixel of the target image patch. The image patches and their centric pixel labels are utilized for training the network. At the inference stage, given an input image patch, the category of its centric pixel can be predicted by the trained classification network. Numerous CNN-based classification networks were developed to mine spectral-spatial information for HSIC. One kind of methods focuses on proposing more effective spectral-spatial modules. For example, three-dimensional convolutional layers were introduced into a CNN network to extract effective features [35], [36]. A CNN structure was proposed to obtain high-level spectral-spatial features [37]. The capsule units were redefined to become spectral-spatial units for accurate classification [38]. The other kind of methods develops the fusion of spectral and spatial features for better performance. Two-branch CNNs were designed for extracting spectral and spatial information separately and an integration module for feature fusion [39], [40]. In [41], a fusion module was proposed for multilevel feature integration.

The size of the HSI patch (patch size) determines the local receptive field of the network, which influences the extraction of spatial contexts. In theory, a larger patch size brings more spatial information about the surrounding spatial dependence. For classification networks, however, since the HSI patch represents the spectral-spatial information for the central pixel, the neighborhood pixels are required to have spatial homogeneity as similar as possible [42]. Unfortunately, HSI patches may inevitably contain pixels from different land-cover categories when they are located near the edges of land-cover regions [43]. As a result, small patches have been used to maintain a high spatial homogeneity in most classification networks to avoid the negative effect from other categories of pixels [44], which will affect the receptive field.

In order to enlarge the receptive field and bring in more spatial information for CNN models, the FCNs were introduced for HSIC in recent years. Compared with classification networks, FCN forms as an end-to-end structure by jointly learning feature

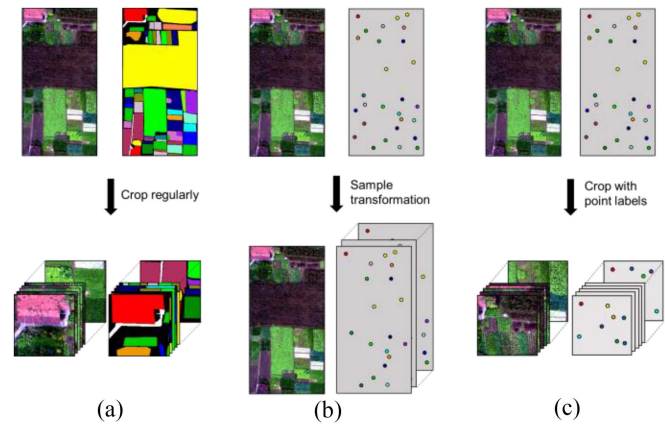


Fig. 1. Different forms of training samples for fully convolutional network. (a) Cropped training patches with pixel-wise labels. (b) Whole image with sparse point labels. (c) Cropped training patches with sparse point labels.

extraction from original input data to final output with great generalization, and is trained with pixelwise labels, as shown in Fig. 1(a). With the benefit of the consecutive down-sampling operations in convolutional forward stage, the extracted features contain rich context and work well for pixel-level classification. Furthermore, the training and inference processes of FCNs are performed whole-image-at-a-time by dense feedforward computation and back propagation [45]. Hence, FCNs have no limitation on the input of images and could greatly improve the efficiency for inference.

However, it requires a huge number of pixel-level labels to train the FCNs, and there are few pixel-level labels available for HSIC in practical applications. How to train an FCN effectively with the supervision of limited sparse point labels has aroused the concern. Since the prior information provided by the sparse point labels is inadequate, the key of training FCN for HSIC is to ensure the convergence of the network with limited constraint.

Several works have successfully trained FCNs with sparse point labels. In [46], an end-to-end fully convolutional and fully supervised network was proposed for HSIC with a fine pixelwise label style. Instead of labeling all pixels for the HSI patch, an HSI patch generation method was used to generate fully pixel-level labels with existing training pixels. However, the HSI patches for training are generated, which cannot reveal the complexity of realistic imagery. The other works adopted patch-free training style of FCNs. As shown in Fig. 1(b), patch-free framework takes the whole hyperspectral image as input and the sparse point training samples as labels. A spectral-spatial FCN was proposed for HSIC with a mask matrix to assist back-propagation in the training stage considering the sparsity of point training samples [47]. A global stochastic stratified sampling strategy was proposed to guarantee the convergence of the patch-free FCN with a spectral attention-based encoder and a lightweight decoder [48]. A spectral-spatial dependent global learning framework was introduced [49], based on global convolutional LSTM and a global joint attention mechanism with a hierarchically balanced sampling strategy and weighted Softmax loss for insufficient and imbalanced HSIC. However, the sparsity of the training samples

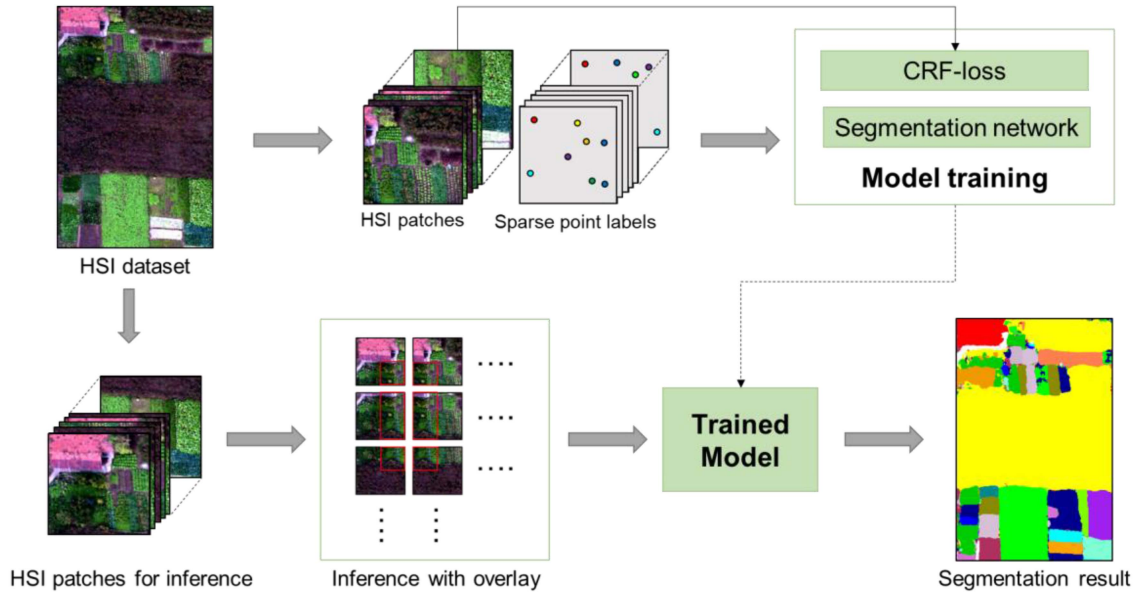


Fig. 2. Flowchart of the patch-based framework for HSIC.

has to be considered when applying patch-free methods for HSIC. Besides, it is hard for patch-free methods to applications of large-scale images due to the limitation of computational memory.

In this article, instead of the patch-free training style, we propose a patch-based training framework in supervision of sparse point labels, aiming at providing a new way of applying FCN to HSIC. As shown in Fig. 1(c), the training framework takes the cropped patch as input. Considering the limitation of supervision with sparse points for training, a novel lightweight fully convolutional network (LFCNet) is designed on the basis of an encoder-decoder structure of shallow channels with the aid of residual connections in the encoder and the integration of multiple attention modules to fully exploit the abundant spatial-spectral information of HSIs, where attention gates (AGs) are applied to make the network pay attention to more important areas in each level of layers, and gated channel transformation is utilized to weight the different significance of abundant channels. Furthermore, considering the limitation of the prior information provided by the sparse point labels, we adopt conditional random field (CRF) loss as a complementary constraint for further excavation of the spatial context information of HSI. The main contributions of this article are as follows.

- 1) Patch-based training of FCN with sparse point labels is proposed and demonstrated to hold the advantages of capturing spatial context and inference efficiency compared with classification network, and be less sensitive to the sparsity of training samples in comparison with patch-free training of FCN, which provides a novel way for HSIC.
- 2) A lightweight and effective fully convolutional network, LFCNet, is designed for HSIC with patch-based training, consisted of AG and gated channel transformation (GCT) modules for better exploitation of spectral-spatial information, as well as the CRF-loss for further excavation of the spatial context information of HSI.

- 3) The proposed method achieves state-of-the-art performance through a large number of experiments on three HSI datasets, and LFCNet is proved to be appropriate and universal for HSIC.

II. METHOD

A patch-based training framework on sparse point labels is proposed for HSIC with the LFCNet, which takes HSI patches as input and is designed on the basis of a shallow-layer encoder-decoder structure with residual connections and multiple attention modules to fully exploit the spectral-spatial information under limited point supervision. Furthermore, CRF-loss is adopted as a complement for capturing spatial information between pixels.

A. Patch-based Training Framework for HSIC

The overall framework of the patch-based training is shown in Fig. 2. In this framework, HSI patches for training and for inference are generated in different ways. According to a certain number of randomly selected sample points for each category, HSI patches for training are generated by taking each labeled pixel as a centric pixel and cropped in a fixed patch size $s \times s$ with its surrounding adjacent pixels, which makes sure that each training patch has at least one-point label. Patches for inference are cropped in order from left to right and top to bottom with an overlay ratio of 50% to make use of information from all its surrounding patches. The fully convolutional network is trained by HSI patches and sparse point labels with CRF-loss. When inferencing, cropped patches with overlay are predicted by the trained network to prevent context limitation and are stitched together by averaging the probability of land cover category for each pixel in overlapped areas.

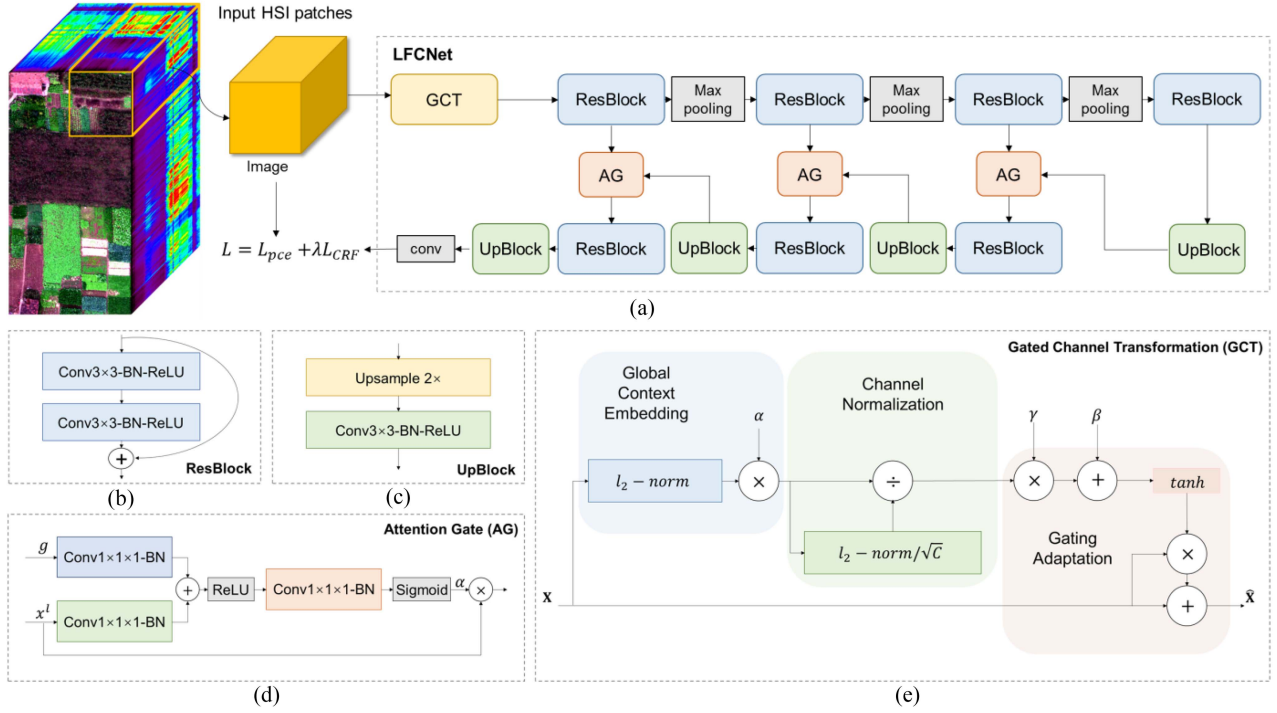


Fig. 3. Structure of the LFCNet. (a) Overall framework of the LFCNet. (b) Structure of the ResBlock. (c) Structure of the UpBlock. (d) Structure of the AG. (e) Structure of the GCT.

B. Lightweight Fully Convolutional Network (LFCNet)

Since the FCN is trained with sparse point labels that provide limited supervision and prior information for training, a lightweight and efficient structure is designed to ensure the convergence and effectiveness of the network. As shown in Fig. 3(a), LFCNet is designed on the basis of a lightweight encoder-decoder structure with shallow channels for each block. In both the encoder and decoder parts, the residual block (ResBlock) is adopted as a basic unit to extract features. Due to the limited number of training samples and the high depth of the network structure, the degradation problem would arise for HSIC [4]. The degradation of accuracy can be alleviated by bringing in shortcut connections in convolutional layers as the successful usage by ResNet [50], [51]. Hence, a shortcut connection is adopted in our ResBlock to facilitate the model training process [36]. As shown in Fig. 3(b), our ResBlock consists of two convolutional layers with a batch normalization (BN) layer and a ReLU activation function after each convolution to speed up the convergence of the network. In the decoder part, the UpBlock plays the role of gradually improving the spatial resolution to restore the detailed information, which is made up of an up-sampling layer and a convolutional layer with BN and ReLU, as illustrated in Fig. 3(c).

The encoder-decoder structure can extract high-level context information but leads to the loss of detailed information, which is of great importance for accurate segmentation [52]. Hence, skip connections between the encoder and decoder are adopted to combine multilevel features from shallow convolutional layers to leverage detailed spatial features. As illustrated in Fig. 3(a), an AG is inserted into the encoder-decoder structure with skip

connections to precisely supplement the spatial details. Furthermore, considering the high spectral resolution and correlation of HSI, GCT is applied as a stem block in the front end of the encoder part to learn the spectral relationships. The detailed structures of AG and GCT are described below.

1) *Attention Gate*: Since the down-sampling operations in the encoder part capture a large receptive field for high-level context information, the loss of detail information comes into being, which does harm to precise segmentation performance. A skip connection between the encoder and decoder in shallow layers can recover detailed spatial information to a certain extent. However, it is vital to use low-level features accurately as a supplement for segmentation. Hence, we adopt an AG in the skip connection to progressively stress effective attention for features of each level [53].

The structure of AG is illustrated in Fig. 3(d). Here, g represents feature maps from the encoder part, x^l represents feature maps from the decoder part, and α represents the attention coefficient. First, g and x^l go through a convolutional layer with BN and add together. Second, the added feature orderly goes through the ReLU activation function, a convolutional layer with BN, and a sigmoid activation function to obtain the attention coefficient, which contains low-level feature responses. The output of the AG is the elementwise multiplication of the input x^l and the attention coefficient. By inserting AG into skip connections, the network tends to pay attention to more important areas in each level of layers.

2) *Gated Channel Transformation*: Due to the high spectral resolution and complexity of HSI, how to fully extract the useful

spectral information to model spectral correlation is crucial for HSIC. We adopt GCT as a stem block to process the original HSI patches. GCT is a lightweight channel normalization layer that employs a normalization method to create competition or cooperation relationships among channels [54]. As illustrated in Fig. 3(e), GCT consists of three parts as follows.

First, a global context embedding module is proposed with the $\mathcal{L}_2$ -norm to aggregate global context information in the channel dimension. The model is expressed as

$$s_c = \alpha_c \|x_c\|_2 = \alpha_c \left\{ \left[\sum_{i=1}^H \sum_{j=1}^W (x_c^{i,j})^2 \right] + \epsilon \right\}^{\frac{1}{2}} \quad (1)$$

where c represents a channel of the input, α is a trainable embedding weight to evaluate the different significance of different channels, and ϵ is a small constant to avoid the problem of derivation at the zero point.

Second, $\mathcal{L}_2$ normalization is adopted to realize channel normalization, which can be expressed as

$$\hat{s}_c = \frac{\sqrt{C} s_c}{\|s\|_2} = \frac{\sqrt{C} s_c}{\left[\left(\sum_{c=1}^C s_c^2 \right) + \epsilon \right]^{\frac{1}{2}}} \quad (2)$$

where $\sqrt{C}$ is used to normalize the scale of $\hat{s}_c$.

Finally, a gating adaptation is proposed to adapt the original feature for competition or cooperation of the channels. The gating function is expressed as

$$\hat{x}_c = x_c [1 + \tanh(\gamma_c \hat{s}_c + \beta_c)] \quad (3)$$

where γ represents the weight, β represents the bias, and $1 + \tanh(\gamma_c \hat{s}_c + \beta_c)$ represents the gating weight for the original feature.

In essence, GCT is a channel attention module to learn channel relationships by weighting the significance of different channels in each location. Hence, we adopt it as a stem block in the front end of the network to explore the spectral correlations of the original HSI patches.

C. CRF-Loss as a Complement Training Constraint

An effective loss function plays a critical role for training FCN to prevent the gradient from disappearing. However, supervision with only the sparse point labels is insufficient for training, which leads to imprecise segmentation [55], [56]. We adopt CRF-loss as an auxiliary loss for further constraint to promote the performance of the network by making full use of spatial context information in HSI [57]. Hence, we train LFCNet with the composite loss function, which is defined as

$$L = L_{pce} + \lambda L_{CRF} \quad (4)$$

where L_{pce} represents partial cross-entropy loss for the sparse labeled pixels, L_{CRF} represents CRF-loss for extra unlabeled pixels, and λ represents the weight of CRF-loss. The detailed principle of CRF-loss is shown as follows.

A quadratic relaxation version of the standard Potts/CRF model adapted for FCN is adopted as the energy formulation

[54]

$$E(i, j) = \sum_{\substack{i, j \in \Phi \\ c \in C}} G(i, j) p_c(i) (1 - p_c(j)) \quad (5)$$

where Φ represents the set of pixels in the HSI patch, C represents the number of categories, p is the Softmax output of the network, and G is a matrix of pairwise discontinuity costs, which is expressed as

$$G(i, j) = W \cdot \exp \left(-\frac{|d_i - d_j|^2}{2\theta_\alpha^2} - \frac{|I_i - I_j|^2}{2\theta_\beta^2} \right) \quad (6)$$

where $G(i, j)$ makes use of the appearance kernel by [58], W represents the normalized weight, which can be found by combining high-dimensional filtering and expectation maximization [56], d is the spatial position of the pixel, I is the spectral value of the pixel, and θ_α and θ_β are hyperparameters that control the scale of the Gaussian kernels.

Hence, the CRF-loss calculation is expressed as

$$L_{CRF} = \sum_{\substack{i, j \in \Phi \\ i \neq j}} E(i, j) \quad (7)$$

III. EXPERIMENTS

A. Dataset Description

The experiments are conducted on three publicly available benchmark HSI datasets. The chosen datasets contain different types of landscape patterns to ensure the richness of the validation area.

1) *Wuhan UAV-Borne Hyperspectral Image (WHU-Hi) Dataset*: The dataset was acquired in farming areas with various crop types in Hubei Province, China, via a Headwall Nano-Hyperspec sensor mounted on a UAV platform [5]. WHU-Hi-HanChuan (WHHC) and WHU-Hi-HongHu (WHHH) were selected from WHU-Hi for the experiment. WHHC was acquired in Hanchuan with an 8-mm focal length imaging sensor equipped on a Leica Aibot X6 UAV V1 platform. The study area contains seven crop species: strawberry, cowpea, soybean, sorghum, water spinach, watermelon, and greens. There are 274 bands from 400–1000 nm, and the size of the imagery is 1217×303 pixels. The spatial resolution of the data is approximately 0.109 m. WHHH was acquired in Honghu City with a 17-mm focal length imaging sensor equipped on a DJI Matrice 600 Pro UAV platform. The study area is typical of the regions affected by fragmentation and is planted with 17 crop types, including cotton, rape, and cabbage. Notably, the region is planted with different cultivars of the same crop type, for example, Chinese cabbage/cabbage and Brassica chinensis/small Brassica chinensis. The size of the imagery is 940×475 pixels, there are 270 bands from 400–1000 nm, and the spatial resolution of the data is approximately 0.043 m.

2) *Pavia Center (PC) Dataset*: PC was acquired by the PO-SIS sensor during a flight campaign over Pavia, northern Italy. The imagery of the PC dataset has 1096×715 pixels with 102

TABLE I
NUMBER OF TRAINING AND TEST SAMPLES FOR THE
WHU-HI-HANCHUAN DATASET

Class	Class name	Training samples	Test samples
1	Strawberry	100	44635
2	Cowpea	100	22653
3	Soybean	100	10187
4	Sorghum	100	5253
5	Water spinach	100	1100
6	Watermelon	100	4433
7	Greens	100	5803
8	Trees	100	17878
9	Grass	100	9369
10	Red roof	100	10416
11	Gray roof	100	16811
12	Plastic	100	3579
13	Bare soil	100	9016
14	Road	100	18460
15	Bright object	100	1036
16	Water	100	75301

bands and contains nine classes, including water, tree, meadows, bricks, bare soil, asphalt, bitumen, tile, and shadows. The spatial resolution of the data is approximately 1.3 m.

B. Experimental Settings

- 1) *Computing Environment*: The proposed method is implemented using the PyTorch platform. The hardware of the equipment is a computer with a GeForce RTX2080Ti GPU, an Intel Core i7-9700KF CPU, and 64 GB memory.
- 2) *Evaluation Metrics*: To evaluate the performance of the proposed method, we chose the accuracy of each class, the overall accuracy (OA), the average accuracy (AA), and the Kappa coefficient (Kappa) as the evaluation metrics.
- 3) *Data Preparation*: We crop the three HSI datasets into patches with a patch size of 200×200 pixels without overlay for training but with 50% overlay in the inference stage to overcome the phenomenon of seam stitching. For both WHHC and WHHH dataset, we randomly select 100 samples of each class for training and the remaining samples for testing. Tables I and II list the number of training and test samples for each category of WHHC and WHHH. For the *PC Dataset*, 100 samples of each class are randomly selected for training and the remaining samples for testing. Table III lists the number of training and test samples for each category of PC.
- 4) *Network Settings*: We adopt the batched stochastic gradient descent optimizer with momentum of 0.9 and weight decay of 0.0005 to train the network. The “poly” learning rate policy is utilized with the initial learning rate setting as 0.001. The network is trained for 50 epochs, and the batch size is set as 3 for all the experiments. The setting of the parameters for the CRF-loss follows [59]. The detailed settings of the network, such as the output channels of each ResBlock, are listed in Table IV.

C. Results

1) *Comparison With Representative Methods*: In order to validate the effectiveness of the proposed method, we

TABLE II
NUMBER OF TRAINING AND TEST SAMPLES FOR THE WHU-HI-HONGHU
DATASET

Class	Class name	Training samples	Test samples
1	Red roof	100	13941
2	Road	100	3412
3	Bare soil	100	21721
4	Cotton	100	163185
5	Cotton firewood	100	6118
6	Rape	100	44457
7	Chinese cabbage	100	24003
8	Pakchoi	100	3954
9	Cabbage	100	10719
10	Tuber mustard	100	12294
11	Brassica parachinensis	100	10915
12	Brassica chinensis	100	8854
13	Small Brassica chinensis	100	22407
14	Lactuca sativa	100	7256
15	Celtuce	100	902
16	Film covered lettuce	100	7162
17	Romaine lettuce	100	2910
18	Carrot	100	3117
19	White radish	100	8612
20	Garlic sprout	100	3386
21	Broad bean	100	1228
22	Tree	100	3940

TABLE III
NUMBER OF TRAINING AND TEST SAMPLES FOR THE PAVIA CENTER DATASET

Class	Class name	Training samples	Test samples
1	Water	100	65871
2	Tree	100	7498
3	Meadows	100	2990
4	Bricks	100	2585
5	Bare soil	100	6484
6	Asphalt	100	9148
7	Bitumen	100	7187
8	Tile	100	42726
9	Shadows	100	2763

TABLE IV
CONFIGURATION DETAILS OF THE LFCNET

LFCNet		
Encoder	ResBlock1	3×3 conv, 64
	ResBlock2	3×3 conv, 96
	ResBlock3	3×3 conv, 128
	ResBlock4	3×3 conv, 192
Decoder	UpBlock4	3×3 conv, 128
	ResBlock3'	3×3 conv, 128
	UpBlock3	3×3 conv, 96
	ResBlock2'	3×3 conv, 96
	UpBlock3	3×3 conv, 64
	ResBlock1'	3×3 conv, 64
	UpBlock1	3×3 conv, 64
	conv	1×1 conv, C

compare the LFCNet results on the three HSI datasets with seven representative methods, including five classification methods, and two FCN-based methods with patch-free training pattern. Specifically, the classification methods include SVM [11] and four state-of-the-art CNN-based classification methods, i.e., DHCNet [60], SSRN [37], CNNCRF [5], and SPRN [61]. The patch-free FCN methods for comparison are FPGA [48] and SSDGL [49]. It is noted that all the experiments share the

TABLE V
ACCURACY COMPARISON BETWEEN LFCNET AND OTHER METHODS ON THE WHHC DATASET

Class	Classification methods					FCN methods		
	SVM	DHCNet	SSRN	CNN CRF	SPRN	FPGA	SSDGL	LFCNet
1	70.38	88.25	83.86	95.01	94.81	98.52	99.08	99.48
2	47.30	86.89	87.18	93.95	94.50	97.74	98.30	98.91
3	75.10	94.75	86.27	99.13	99.60	99.55	98.19	99.43
4	95.39	98.72	89.26	99.28	99.49	99.39	99.28	100.00
5	83.18	99.73	92.73	99.91	100.00	100.00	100.00	100.00
6	50.96	74.26	81.91	83.28	93.71	99.41	95.94	98.74
7	87.83	86.96	84.59	99.05	96.00	98.98	99.98	99.67
8	60.77	65.85	87.50	77.60	90.80	96.46	97.24	98.33
9	65.40	90.70	80.07	89.09	95.91	94.34	95.04	97.68
10	90.82	98.42	97.61	97.49	97.29	98.16	98.12	96.63
11	95.27	94.30	98.23	84.16	97.59	99.23	98.55	99.50
12	62.36	95.53	94.89	98.83	99.75	100.00	100.00	100.00
13	56.57	84.74	87.37	78.59	88.03	95.42	94.12	92.77
14	64.98	89.47	82.92	95.01	94.68	96.34	97.93	98.66
15	78.38	93.05	92.86	92.66	99.90	95.66	98.84	99.71
16	91.35	98.02	98.80	99.94	99.55	98.62	99.18	99.68
OA (%)	75.98	90.68	90.64	93.95	96.38	98.05	98.38	98.94
Kappa	0.7238	0.8914	0.891	0.929	0.9576	0.9772	0.9811	0.9876
AA (%)	73.50	89.98	89.13	92.69	96.35	97.99	98.11	98.7

TABLE VI
ACCURACY COMPARISON BETWEEN LFCNET AND OTHER METHODS ON THE WHHH DATASET

Class	Classification methods					FCN methods		
	SVM	DHCNet	SSRN	CNN CRF	SPRN	FPGA	SSDGL	LFCNet
1	88.23	98.04	97.67	98.32	98.09	98.16	96.54	98.29
2	80.66	95.25	95.75	95.93	98.36	97.71	98.01	98.33
3	74.97	90.24	93.36	95.66	95.81	95.38	97.37	97.94
4	79.38	94.41	93.43	94.76	97.72	99.15	99.27	99.89
5	77.79	95.65	95.73	99.75	98.92	99.75	100.00	99.80
6	78.28	91.51	91.22	92.54	96.03	95.88	95.83	98.36
7	62.50	77.27	85.15	90.46	93.93	95.51	92.54	95.79
8	39.53	74.43	87.25	81.36	95.83	98.66	99.92	99.32
9	90.52	96.75	97.91	95.90	99.07	99.18	98.30	97.92
10	50.94	82.43	88.43	95.31	95.05	96.78	97.03	96.58
11	41.30	75.46	91.06	93.08	95.79	94.65	94.49	95.03
12	57.71	86.55	91.10	84.20	96.11	96.58	96.51	96.39
13	59.41	84.08	84.75	83.80	93.96	93.85	94.33	93.87
14	64.90	92.31	95.69	96.51	97.64	97.15	98.58	98.46
15	87.58	99.45	100.00	100.00	99.78	99.67	100.00	100.00
16	77.93	97.81	96.31	96.34	98.69	99.22	98.77	99.80
17	81.20	97.84	98.66	98.87	99.93	100.00	100.00	100.00
18	72.99	94.74	97.66	98.17	98.04	99.65	100.00	99.13
19	77.86	92.56	96.11	95.65	98.04	98.33	97.39	97.56
20	80.42	98.55	99.17	98.64	99.73	99.91	100.00	99.97
21	70.52	99.76	98.62	100.00	100.00	99.19	99.92	99.76
22	78.40	98.00	99.39	100.00	99.90	100.00	100.00	99.92
OA (%)	74.12	91.20	92.51	93.74	96.94	97.69	97.63	98.45
Kappa	0.687	0.89	0.9066	0.9217	0.9614	0.9708	0.9701	0.9804
AA (%)	71.23	91.50	94.29	94.78	97.56	97.92	97.95	98.28

same training and test samples. The weight parameter λ of the composite loss is set to 0.02. The accuracies of the three HSI datasets from the above methods are given in Tables V–VII. The classification maps of the three HSI datasets are shown in Figs. 4–6.

As illustrated in the tables, the performance of the proposed LFCNet outperforms the other methods on all three datasets. LFCNet surpasses the second-highest SSDGL method by approximately 0.5% in OA, 0.006 in Kappa, and 0.6% in AA for the WHHC dataset and 0.7% in OA, 0.01 in Kappa, and 0.3% in AA for the WHHH dataset. For the PC dataset, LFCNet outperforms other FCN methods in a relatively large margin, surpassing the second highest method FPGA by approximately 0.5% in OA, 0.007 in Kappa, and 0.6% in AA. The accuracy superiority comparing with several classification methods is inapparent, e.g.,

TABLE VII
ACCURACY COMPARISON BETWEEN LFCNET AND OTHER METHODS ON THE PC DATASET

Class	Classification methods				FCN methods		
	SVM	DHCNet	SSRN	SPRN	FPGA	SSDGL	LFCNet
1	88.48	99.99	99.73	100.00	99.58	98.87	100.00
2	98.29	97.85	99.57	98.93	97.38	98.88	98.32
3	44.92	96.54	97.67	98.77	98.80	95.40	99.39
4	50.58	99.03	99.96	100.00	100.00	100.00	100.00
5	69.59	95.61	99.54	96.76	95.88	97.24	100.00
6	72.65	98.19	99.81	99.92	99.92	97.55	99.22
7	56.70	98.41	99.23	99.36	99.46	99.86	99.86
8	53.67	99.96	99.98	100.00	99.58	95.29	99.97
9	88.54	100.00	100.00	99.93	100.00	96.37	99.72
OA (%)	73.93	99.40	99.73	99.74	99.32	97.63	99.83
Kappa	0.6558	0.9915	0.9962	0.9963	0.9904	0.9666	0.9976
AA (%)	69.22	98.40	99.50	99.30	98.96	97.72	99.61

SPRN, because the accuracies of these classification methods on this dataset have already attained a high level (larger than 99.7% in terms of OA). Besides, we speculate another reason is that larger receptive field do not bring in more spatial information due to the large-area blanks in the labels of this dataset.

Through the visualization of the classification results from the above methods, we can go step further to understand and show the advantage of the proposed patch-based FCN. Overall, FCN methods have better and smoother visual performance than classification methods owing to the improvement of the obvious salt and pepper noise. For the WHHH dataset, which is characterized with class imbalance, LFCNet is able to precisely capture the large-area categories and delicately extract small-area categories. For the WHHC dataset characterized by a large number of edges, the results from LFCNet possess clear boundaries. For the PC dataset in which a lot of unlabeled areas exist, LFCNet is able to extract abundant features and obtain accurate results. To sum up, for the three FCN methods, the results of the proposed LFCNet have clearer edges and show a better performance in capturing thin and broken areas.

2) *Sensitivity Analysis of LFCNet*: To demonstrate the effectiveness of each component of the patch-based FCN framework, numbers of experiments are conducted for sensitivity analysis.

In the proposed framework, we designed an efficient and effective network for HSIC with the aid of AG and GCT to explore the abundant spectral signatures and spatial context of HSI. To verify the effectiveness of the modules in LFCNet, we set the network without AG and GCT as the baseline and compare the performance of the networks with AG and with both AG and GCT to the baseline. Moreover, we compare the performance of LFCNet with and without CRF-loss to demonstrate the effectiveness of further constraints. All the abovementioned experiments are conducted on the WHHC dataset. It is noted that the weight of CRF-loss is set as 0.02 here, and the overlay inference is not adopted. As shown in Table VIII, both the application of AG and GCT could improve the performance. When only AG is used, all the evaluation metrics increase approximately 0.3% compared to the baseline. Furthermore, the addition of GCT further boosts the performance, which results in an OA improvement from 98.18%–98.3% and a Kappa improvement from 0.9787–0.9801. Hence, the application of AG and GCT is proven to be effective

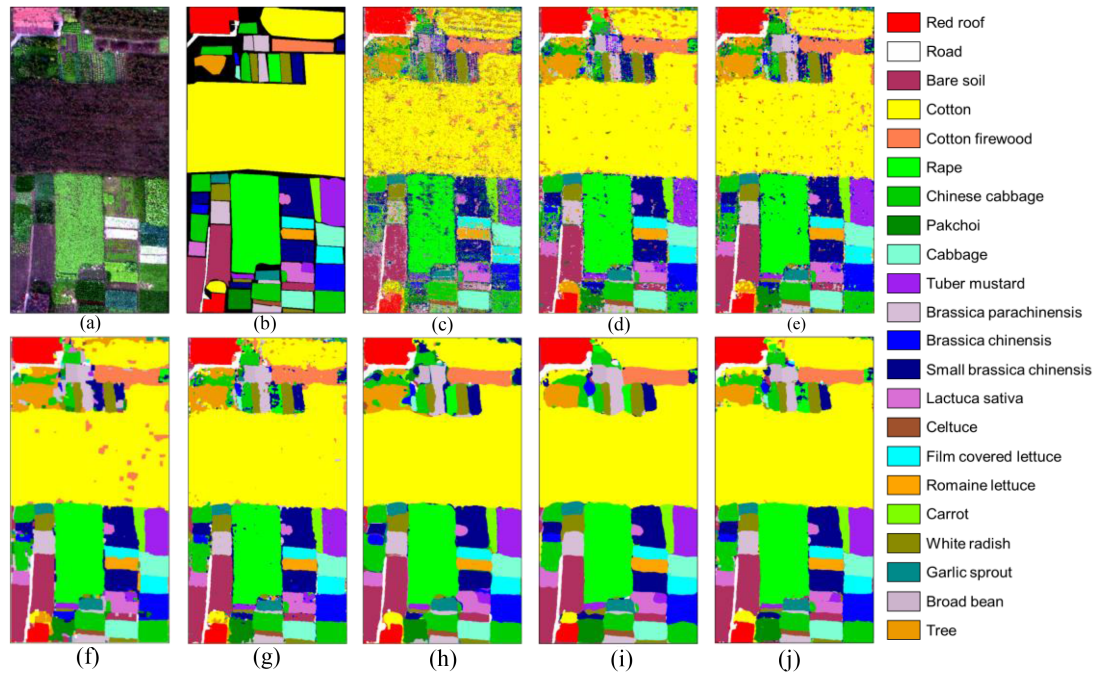


Fig. 4. Visual comparison of classification maps between LFCNet and other methods on the WHHH dataset. (a) Three-band true color composite. (b) Ground truth. (c) SVM. (d) DHCNet. (e) SSRN. (f) CNNCRF. (g) SPRN. (h) FPGa. (i) SSDGL. (j) LFCNet.

TABLE VIII
ACCURACY COMPARISON OF DIFFERENT NETWORKS ON THE WHHC DATASET

Method	AG	GCT	CRF-loss	OA (%)	Kappa	AA (%)
Baseline				97.87	0.9751	97.77
Baseline with AG	✓			98.18	0.9787	98.00
Baseline with AG and GCT	✓	✓		98.30	0.9801	98.13
LFCNet	✓	✓	✓	98.69	0.9847	98.31

for strengthening useful low-level features and modeling the interdependence from abundant spectral signatures from the perspective of accuracy. The use of CRF-loss for training improves the accuracy by approximately 0.4 percentage point in OA and Kappa and 0.2 percentage point in AA, which demonstrates that CRF-loss can supplement spatial information on the basis of sparse point label constraints.

The classification maps on WHHC are presented in Fig. 7 to visually illustrate the effect of each component. Overall, with the addition of different components, the classification maps gradually perform better visually. As shown in the red rectangle in the WHHC, which has a complex distribution of different land cover, the addition of AG brings more accurate capture of fragmented and sparse small regions. In addition, GCT and CRF-loss further boost the performance. Finally, the classification map trained with CRF-loss has a more complete result and accurate shape than other network structures.

Since the classification maps are obtained by patch-based inference, the phenomenon of seam stitching occurs, as shown in the black circle of Fig. 8(a). To eliminate the effect of seam stitching, an overlay strategy for inference is adopted. Here, we set the overlay ratio of the inference patch as 50%. Experiments on all three datasets with and without overlay inference are

TABLE IX
QUANTITATIVE COMPARISON OF THE IMPACT OF INFERENCE OVERLAY STRATEGIES ON THE WHHC, WHHH, AND PC DATASETS

Dataset	Inference overlay	OA (%)	Kappa	AA (%)
WHHC	Wo overlay	98.69	0.9847	98.31
	Wi overlay	98.94	0.9876	98.70
WHHH	Wo overlay	98.26	0.9779	97.99
	Wi overlay	98.45	0.9804	98.28
PC	Wo overlay	99.83	0.9976	99.62
	Wi overlay	99.83	0.9976	99.61

conducted to verify the effectiveness of the inference overlay strategy. As given in Table IX, the usage of the inference overlay brings an improvement of 0.3 and 0.25 percentage point in the Kappa of WHHC and WHHH, respectively. However, it does not improve the accuracies of PC, which can be attributed to the large unlabeled areas of PC so that it does not result in an obvious accuracy change.

The composite loss function influences the optimization of the network. CRF-loss brings in further constraint on training by taking use of spatial connection between pixels, which is proven to be effective from abovementioned experiments on WHHC. As given in Table X, the addition of CRF-loss for

TABLE X
ACCURACY COMPARISON OF THE IMPACT OF THE WEIGHT OF CRF-LOSS ON THE WHHC, WHHH, AND PC DATASETS

Dataset	λ	0	0.01	0.02	0.03	0.04	0.05	0.06
WHHC	OA (%)	98.48	98.68	98.94	98.87	98.78	98.83	98.82
	Kappa	0.9822	0.9845	0.9876	0.9867	0.9857	0.9863	0.9861
	AA (%)	98.35	98.53	98.70	98.57	98.36	98.45	98.46
WHHH	OA (%)	98.31	98.32	98.45	98.25	98.47	98.12	98.45
	Kappa	0.9763	0.9785	0.9804	0.9778	0.9805	0.9762	0.9804
	AA (%)	98.28	98.28	98.28	98.18	98.04	98.03	98.22
PC	OA (%)	99.37	99.85	99.83	99.77	99.80	99.82	99.75
	Kappa	0.9911	0.9979	0.9976	0.9967	0.9971	0.9975	0.9964
	AA (%)	99.16	99.71	99.61	99.33	99.55	99.54	99.44

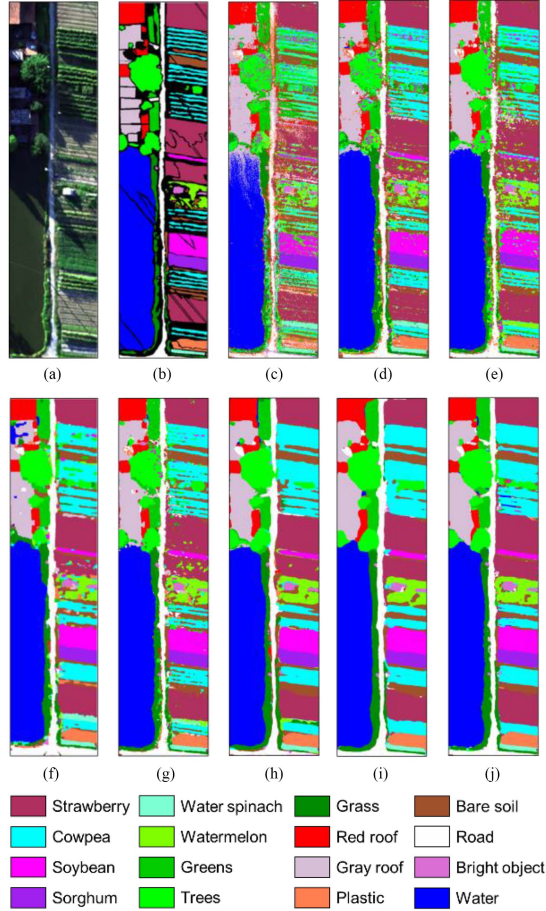


Fig. 5. Visual comparison of classification maps between LFCNet and other methods on the WHHC dataset. (a) Three-band true color composite. (b) Ground truth. (c) SVM. (d) DHCNet. (e) SSRN. (f) CNNCRF. (g) SPRN. (h) FPGA. (i) SSDGL. (j) LFCNet.

training also improves the performance in the WHHH and PC datasets, which illustrates the effectiveness and universality of the composite loss. Specifically, the parameter λ is introduced to control the proportion of cross entropy loss and CRF-loss. Hence, experiments with different weights are conducted to find the most appropriate proportion for the two kinds of loss. Table X shows the performance of three HSI datasets in different weights, where the weight ranges from 0–0.06 with 0.01 intervals. For the WHHC and PC datasets, all the evaluation metrics achieve the highest values when the weight is set as 0.02 and 0.01,

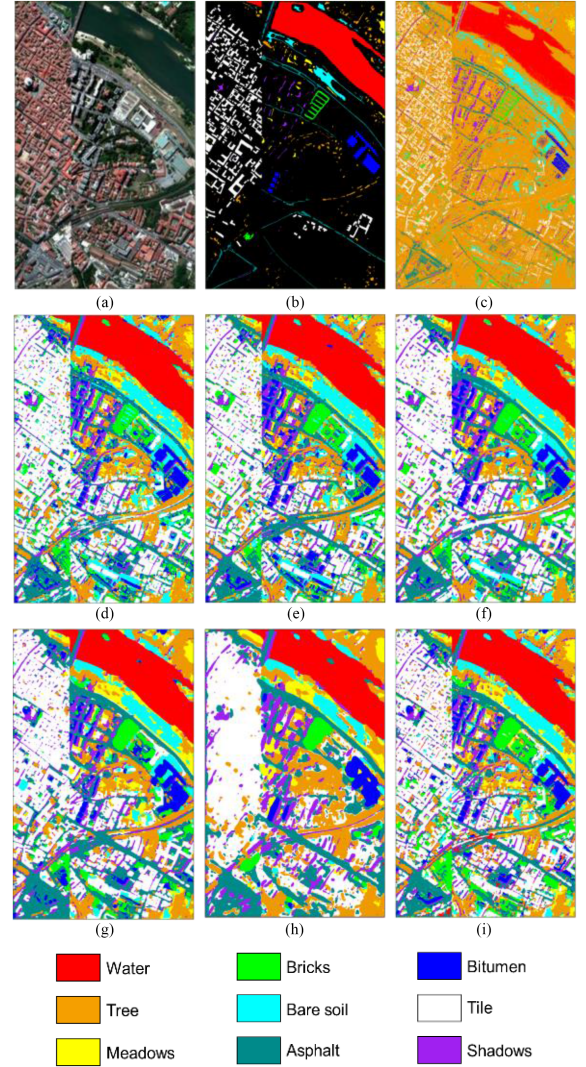


Fig. 6. Visual comparison of classification maps between LFCNet and other methods on the PC dataset. (a) Three-band true color composite. (b) Ground truth. (c) SVM. (d) DHCNet. (e) SSRN. (f) SPRN. (g) FPGA. (h) SSDGL. (i) LFCNet.

respectively. For the WHHH dataset, AA achieves the highest value when the weight is set as 0.02, but OA and Kappa achieve the best values when the weight is set as 0.04. However, the difference between the accuracy when the weight is set as 0.02 and the best accuracy of the WHHH and PC is very small. It can

TABLE XI
ACCURACY COMPARISON OF THE IMPACT OF THE PATCH SIZE FOR TRAINING ON THE WHHC DATASET

Patch size	50×50	100×100	150×150	200×200	250×250	300×300
OA (%)	92.07	97.49	98.23	98.94	98.87	98.56
Kappa	0.9072	0.9705	0.9792	0.9876	0.9876	0.9831
AA (%)	92.10	97.29	98.06	98.70	98.42	98.09

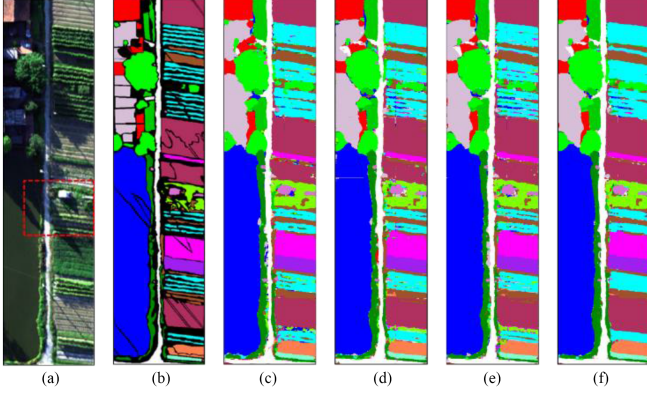


Fig. 7. Visual comparison of classification maps from different network structures on the WHHC dataset. (a) Three-band true color composite. (b) Ground truth. (c) Baseline. (d) Baseline with AG. (e) Baseline with AG and GCT. (f) LFCNet.

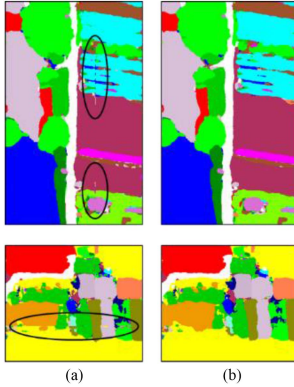


Fig. 8. Visual comparison of partially enlarged classification maps inferred with and without the overlay strategy on the WHHC and WHHH datasets. (a) Inference without overlay. (b) Inference with overlay.

be concluded from the table that a weight of 0.02 performs well on all three datasets. Hence, we uniformly set the weight of λ as 0.02 for all three datasets.

Since we adopt patch-based samples to train the network, experiments with different patch sizes on the WHHC dataset are conducted to study the influence of patch size. The patch size is set from 50×50 to 300×300 in intervals of 50. It should be noted that the patch size for inference is equal to the training patch size. As shown in Table XI, the performance of LFCNet improves as the patch size increases from 50×50 to 200×200 , but the accuracy begins to decline when the patch size surpasses 200. Hence, we set the patch size to 200×200 .

In order to study the impact of the number of training samples on the proposed framework, we conducted extensive experiments on the WHHC dataset. The results from four methods,

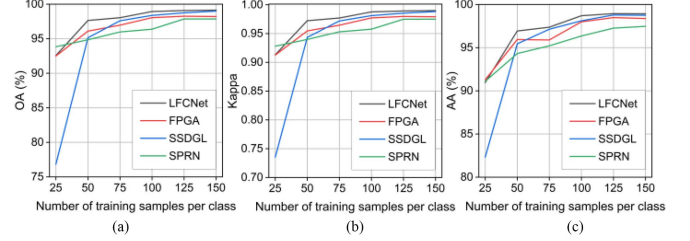


Fig. 9. Sensitivity of the training samples per class for training different methods on the WHHC dataset. (a) Impact on the OA of different number settings. (b) Impact on the Kappa of different number settings. (c) Impact on the AA of different number settings.

including LFCNet, FPGA, SSDGL, and SPRN, are shown in Fig. 9, and the training samples per class are set from 25–150 in an interval of 25. Overall, the performance of four methods improves with the increasing number of training samples per class. It can be concluded from the figure that the increasing range of accuracy decreases with the increase in the number of training samples. After the training samples exceed 100, the increase in evaluation metrics becomes very small. Generally, LFCNet outperforms the other three methods in terms of different levels of training sample number, especially for small numbers, e.g., 50, demonstrating its effectiveness. When the training samples per class are set to 50, which is only approximately 0.6% of the total samples, the proposed LFCNet also achieves a high accuracy at 97.64% of OA, 0.9723 in Kappa, and 96.93% in AA. Hence, it suggests that the proposed framework is also robust when only using limited training samples.

D. Discussion

1) *Classification Network or Fully Convolutional Network?*
Through the comparison in Section III-C, the FCN methods achieve higher accuracies than the classification methods. To further understand the difference between the FCN and classification network, the classification method with the best performance on all three datasets, SPRN, is used for comparison with the proposed LFCNet. The partially enlarged classification maps of WHHH and WHHC are shown in Fig. 10 for intuitive visual comparison. In addition to the higher accuracy, the results of LFCNet are smoother and have more precise boundaries compared to SPRN. As shown in Fig. 10, unlike SPRN, which has considerable salt-and-pepper noise within a region, LFCNet maintains unity within a region. This can be attributed to capturing more spatial context by FCN since the input patch sizes of FCN are much larger than those of the classification network.

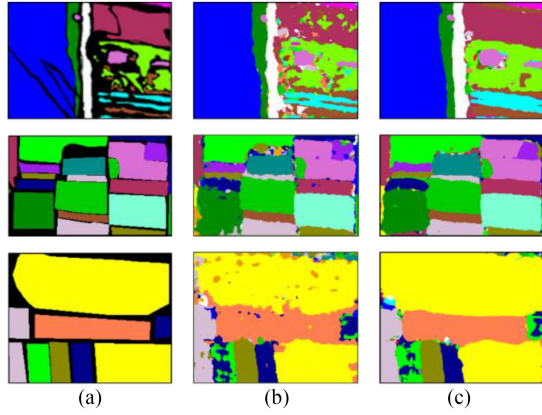


Fig. 10. Visual comparison of partially enlarged classification maps on the WHHC and WHHH datasets. (a) Ground truth. (b) SPRN. (c) LFCNet.

TABLE XII

COMPARISON OF INFERENCE SPEED FOR DIFFERENT METHODS ON THE WHHC, WHHH, AND PC DATASETS

Inference speed (s)	SPRN	LFCNet
WHHC	478.96	13.62
WHHH	543.78	12.63
PC	390.13	19.87

Furthermore, the comparison of inference speed between SPRN and LFCNet is conducted to show the efficiency difference. In theory, the classification networks need to be inferred pixel by pixel, but FCNs predict the result of one patch at a time. Table XII lists the comparison results of inference speed on the WHHC, WHHH, and PC datasets. The inference speed of LFCNet is approximately 40 times faster than that of SPRN on three datasets, which proves the superiority of FCN in time efficiency.

2) *Patch-Free or Patch-Based Network?*: As presented in Section III-C, the proposed patch-based FCN surpasses the patch-free FCNs in accuracy on all three datasets. Essentially, the patch-free pattern changes the point label for every training iteration but holds the spatial context of images, while the patch-based method is just the opposite. For a fair comparison, we conduct experiments on different networks with different training patterns to show the effectiveness of the proposed patch-based method on the WHHC, WHHH, and PC datasets. Here, the fully convolutional network proposed in FPGA, named FreeNet [48], is trained in patch-free and patch-based training framework, respectively, for comparison with the proposed patch-based training framework with LFCNet. As shown in Table XIII, on all three datasets, training FreeNet with a patch-based pattern has better performance than training FreeNet with a patch-free pattern but is inferior to the patch-based training for LFCNet, which further demonstrates the effectiveness of the patch-based training framework on different FCNs. Furthermore, many strategies and skills from remote sensing image semantic segmentation could be used in the future since the universality of the patch-based pattern.

3) *Applicability*: The effectiveness of the proposed patch-based training framework has been proved qualitatively and

TABLE XIII
ACCURACY COMPARISON OF THE IMPACT OF THE TRAINING PATTERN WITH DIFFERENT TRAINING NETWORK ON THE WHHC DATASET

	Training pattern	Network	OA (%)	Kappa	AA (%)
WHHC	Patch-free	FreeNet	98.05	0.9972	97.99
	Patch-based	FreeNet	98.87	0.9867	98.39
	Patch-based	LFCNet	98.94	0.9876	98.70
WHHH	Patch-free	FreeNet	97.69	0.9708	97.92
	Patch-based	FreeNet	97.92	0.9737	97.93
	Patch-based	LFCNet	98.45	0.9804	98.28
PC	Patch-free	FreeNet	99.32	0.9904	98.96
	Patch-based	FreeNet	99.35	0.9907	99.15
	Patch-based	LFCNet	99.83	0.9976	99.61

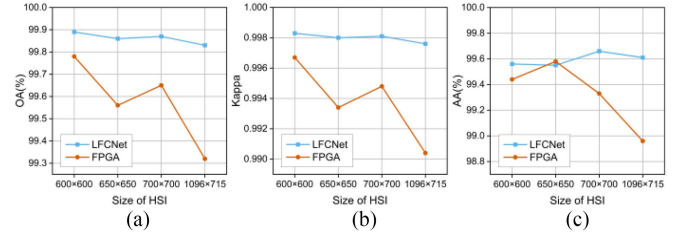


Fig. 11. Accuracy comparison of LFCNet and FPGA on different image sizes of cropped PC datasets. (a) OA. (b) Kappa. (c) AA.

quantitatively in the study. To reveal the differences in applicability between patch-based and patch-free FCN methods, experiments on PCs with different crop sizes are conducted. Taking the upper left corner of the PC dataset as the starting point, PCs are cut with patch sizes of 600, 650, and 700 for comparison with the original PC. It is noted that the numbers of training samples per class are all set to 100, and the accuracies are calculated based on the corresponding image size. As shown in Fig. 11, with the change of image size, OA of the proposed method is relatively stable, but the performance of FPGA fluctuates greatly. The accuracy gap between the two methods shows an upward trend with the increasing image size. According to the analysis of the sample distribution, as the size of HSI increases, the sparsity of the samples relative to the whole image increases under an unchanged sample number, which may lead to the degradation of the patch-free method for HSI with large size. By contrary, the patch-based method is not sensitive to the sparsity of the samples given an unchanged sample number. Hence, when applying patch-free method for HSIC, we would need to consider the sparsity of the training samples, which relates to both the number of training samples and the image size. When applying patch-based method for HSIC, we would not consider the sparsity of the training samples but need to consider whether the number of training samples is enough.

IV. CONCLUSION

In this article, a novel patch-based training framework of FCN on sparse point labels with a lightweight and efficient network is proposed for HSIC. As a general training mode for remote sensing image semantic segmentation, the training framework is improved in network structure and training constraint in order to achieve good classification result under the limited supervision of sparse point. A lightweight network which guarantees the

convergence of training is proposed with the aid of AGs and gated channel transformation on the basis of shallow-channel encoder–decoder with residual connections to excavate spectral-spatial information of HSI. CRF-loss is adopted as a complementary constraint for further excavation of spatial context to make up for the insufficiency of prior information from sparse point labels in training stage. Extensive experiments conducted on three HSI datasets achieve state-of-the-art performance compared with other types of methods, which demonstrates the effectiveness of the patch-based training framework of FCN with sparse point labels. In addition to the classification accuracy, the advantages of the proposed method were also demonstrated in terms of the efficiency and the stability to deal with images of various sizes.

In the future, we would like to explore more universal network structures for HSIC in patch-based training patterns with limited point labels. Furthermore, we hope that the proposed method can be used in the practical application of hyperspectral classification and even extended to multispectral remote sensing images.

ACKNOWLEDGMENT

The authors would like to thank Prof. Y. Zhong et al. with Wuhan University for providing the WHU-Hi datasets and Prof. P. Gamba et al. with Pavia University for providing the Pavia Center dataset. The authors also appreciate the anonymous reviewers for their constructive comments.

REFERENCES

- [1] S. Li, W. Song, L. Fang, Y. Chen, P. Ghamisi, and J. A. Benediktsson, "Deep learning for hyperspectral image classification: An overview," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 9, pp. 6690–6709, Sep. 2019, doi: [10.1109/TGRS.2019.2907932](#).
- [2] B. Rasti et al., "Feature extraction for hyperspectral imagery: The evolution from shallow to deep: Overview and toolbox," *IEEE Geosci. Remote Sens. Mag.*, vol. 8, no. 4, pp. 60–88, Dec. 2020, doi: [10.1109/MGRS.2020.2979764](#).
- [3] 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, doi: [10.1109/MGRS.2018.2854840](#).
- [4] 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, doi: [10.1109/TGRS.2016.2584107](#).
- [5] 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. no. 112012, doi: [10.1016/j.rse.2020.112012](#).
- [6] L. Liang et al., "Estimation of crop LAI using hyperspectral vegetation indices and a hybrid inversion method," *Remote Sens. Environ.*, vol. 165, pp. 123–134, 2015, doi: [10.1016/j.rse.2015.04.032](#).
- [7] N. Yokoya, J. Chan, and K. Segl, "Potential of resolution-enhanced hyperspectral data for mineral mapping using simulated EnMAP and Sentinel-2 images," *Remote Sens.*, vol. 8, no. 3, Feb. 2016, Art. no. 172, doi: [10.3390/rs8030172](#).
- [8] X. Yang and Y. Yu, "Estimating soil salinity under various moisture conditions: An experimental study," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 5, pp. 2525–2533, May 2017, doi: [10.1109/TGRS.2016.2646420](#).
- [9] B. Koch, "Status and future of laser scanning, synthetic aperture radar and hyperspectral remote sensing data for forest biomass assessment," *ISPRS J. Photogramm. Remote Sens.*, vol. 65, no. 6, pp. 581–590, Nov. 2010, doi: [10.1016/j.isprsjprs.2010.09.001](#).
- [10] N. Audebert, B. L. Saux, and S. Lefevre, "Deep learning for classification of hyperspectral data: A comparative review," *IEEE Geosci. Remote Sens. Mag.*, vol. 7, no. 2, pp. 159–173, Jun. 2019, doi: [10.1109/MGRS.2019.2912563](#).
- [11] F. Melgani and L. Bruzzone, "Classification of hyperspectral remote sensing images with support vector machines," *IEEE Trans. Geosci. Remote Sens.*, vol. 42, no. 8, pp. 1778–1790, Aug. 2004, doi: [10.1109/TGRS.2004.831865](#).
- [12] P. O. Gislason, J. A. Benediktsson, and J. R. Sveinsson, "Random forests for land cover classification," *Pattern Recognit. Lett.*, vol. 27, no. 4, pp. 294–300, Mar. 2006, doi: [10.1016/j.patrec.2005.08.011](#).
- [13] B. Du and L. Zhang, "Random-selection-based anomaly detector for hyperspectral imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 49, no. 5, pp. 1578–1589, May 2011, doi: [10.1109/TGRS.2010.2081677](#).
- [14] B. Du and L. Zhang, "Target detection based on a dynamic subspace," *Pattern Recognit.*, vol. 47, no. 1, pp. 344–358, Jan. 2014, doi: [10.1016/j.patcog.2013.07.005](#).
- [15] 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, doi: [10.1109/TGRS.2017.2765364](#).
- [16] W. Li and Q. Du, "Gabor-filtering-based nearest regularized subspace for hyperspectral image classification," *IEEE J. Sel. Topics Appl. Earth Obs. Remote Sens.*, vol. 7, no. 4, pp. 1012–1022, Apr. 2014, doi: [10.1109/JSTARS.2013.2295313](#).
- [17] K.-K. Huang, C.-X. Ren, H. Liu, Z.-R. Lai, Y.-F. Yu, and D.-Q. Dai, "Hyperspectral image classification via discriminant gabor ensemble filter," *IEEE Trans. Cybern.*, vol. 52, no. 8, pp. 8352–8365, Aug. 2022, doi: [10.1109/TCYB.2021.3051141](#).
- [18] J. A. Benediktsson, J. A. Palmason, and J. R. Sveinsson, "Classification of hyperspectral data from urban areas based on extended morphological profiles," *IEEE Trans. Geosci. Remote Sens.*, vol. 43, no. 3, pp. 480–491, Mar. 2005, doi: [10.1109/TGRS.2004.842478](#).
- [19] J. Li, P. R. Marpu, A. Plaza, J. M. Bioucas-Dias, and J. A. Benediktsson, "Generalized composite kernel framework for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 51, no. 9, pp. 4816–4829, Sep. 2013, doi: [10.1109/TGRS.2012.2230268](#).
- [20] 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, doi: [10.1109/LGRS.2005.857031](#).
- [21] M. Fauvel, J. Chanussot, and J. A. Benediktsson, "A spatial-spectral kernel-based approach for the classification of remote-sensing images," *Pattern Recognit.*, vol. 45, no. 1, pp. 381–392, Jan. 2012, doi: [10.1016/j.patcog.2011.03.035](#).
- [22] X. He, Y. Chen, and P. Ghamisi, "Heterogeneous transfer learning for hyperspectral image classification based on convolutional neural network," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 5, pp. 3246–3263, May 2020, doi: [10.1109/TGRS.2019.2951445](#).
- [23] G. E. Hinton, S. Osindero, and Y.-W. Teh, "A fast learning algorithm for deep belief nets," *Neural Computation*, vol. 18, no. 7, pp. 1527–1554, Jul. 2006, doi: [10.1162/neco.2006.18.7.1527](#).
- [24] X. X. Zhu et al., "Deep learning in remote sensing: A comprehensive review and list of resources," *IEEE Geosci. Remote Sens. Mag.*, vol. 5, no. 4, pp. 8–36, Dec. 2017, doi: [10.1109/MGRS.2017.2762307](#).
- [25] L. Ma, Y. Liu, X. Zhang, Y. Ye, G. Yin, and B. A. Johnson, "Deep learning in remote sensing applications: A meta-analysis and review," *ISPRS J. Photogramm. Remote Sens.*, vol. 152, pp. 166–177, Jun. 2019, doi: [10.1016/j.isprsjprs.2019.04.015](#).
- [26] Y. Chen, Z. Lin, X. Zhao, G. Wang, and Y. Gu, "Deep learning-based classification of hyperspectral data," *IEEE J. Sel. Topics Appl. Earth Obs. Remote Sens.*, vol. 7, no. 6, pp. 2094–2107, Jun. 2014, doi: [10.1109/JSTARS.2014.2329330](#).
- [27] Y. Chen, X. Zhao, and X. Jia, "Spectral-spatial classification of hyperspectral data based on deep belief network," *IEEE J. Sel. Topics Appl. Earth Obs. Remote Sens.*, vol. 8, no. 6, pp. 2381–2392, Jun. 2015, doi: [10.1109/JSTARS.2015.2388577](#).
- [28] M. E. Paoletti, J. M. Haut, J. Plaza, and A. Plaza, "Scalable recurrent neural network for hyperspectral image classification," *J. Supercomputing*, vol. 76, no. 11, pp. 8866–8882, Nov. 2020, doi: [10.1007/s11227-020-03187-0](#).
- [29] 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, doi: [10.1109/TGRS.2016.2636241](#).

- [30] F. Zhou, R. Hang, Q. Liu, and X. Yuan, "Hyperspectral image classification using spectral-spatial LSTMs," *Neurocomputing*, vol. 328, pp. 39–47, Feb. 2019, doi: [10.1016/j.neucom.2018.02.105](#).
- [31] A. Sharma, X. Liu, and X. Yang, "Land cover classification from multi-temporal, multi-spectral remotely sensed imagery using patch-based recurrent neural networks," *Neural Netw.*, vol. 105, pp. 346–355, Sep. 2018, doi: [10.1016/j.neunet.2018.05.019](#).
- [32] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems*, vol. 25. Cambridge, MA, USA: MIT Press, 2012.
- [33] Q. Zhu et al., "Knowledge-guided land pattern depiction for urban land use mapping: A case study of Chinese cities," *Remote Sens. Environ.*, vol. 272, Apr. 2022, Art. no. 112916, doi: [10.1016/j.rse.2022.112916](#).
- [34] Q. Zhu et al., "Land-Use/Land-Cover change detection based on a Siamese global learning framework for high spatial resolution remote sensing imagery," *ISPRS J. Photogrammetry Remote Sens.*, vol. 184, pp. 63–78, Feb. 2022, doi: [10.1016/j.isprsjprs.2021.12.005](#).
- [35] Y. Li, H. Zhang, and Q. Shen, "Spectral-spatial classification of hyperspectral imagery with 3d convolutional neural network," *Remote Sens.*, vol. 9, no. 1, Jan. 2017, Art. no. 67, doi: [10.3390/rs9010067](#).
- [36] 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, doi: [10.1109/TGRS.2017.2755542](#).
- [37] 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, Jun. 2015, doi: [10.1080/2150704X.2015.1047045](#).
- [38] M. E. Paoletti et al., "Capsule networks for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 4, pp. 2145–2160, Apr. 2019, doi: [10.1109/TGRS.2018.2871782](#).
- [39] J. Yang, Y.-Q. Zhao, and J. C.-W. Chan, "Learning and transferring deep joint spectral-spatial features for hyperspectral classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 8, pp. 4729–4742, Aug. 2017, doi: [10.1109/TGRS.2017.2698503](#).
- [40] X. Xu, W. Li, Q. Ran, Q. Du, L. Gao, and B. Zhang, "Multisource remote sensing data classification based on convolutional neural network," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 2, pp. 937–949, Feb. 2018, doi: [10.1109/TGRS.2017.2756851](#).
- [41] 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, doi: [10.1109/TGRS.2018.2794326](#).
- [42] 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, doi: [10.1109/TGRS.2020.2994057](#).
- [43] H. Sun, X. Zheng, X. Lu, and S. Wu, "Spectral-spatial attention network for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 5, pp. 3232–3245, May 2020, doi: [10.1109/TGRS.2019.2951160](#).
- [44] 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 Obs. Remote Sens.*, vol. 11, no. 11, pp. 4141–4155, Nov. 2018, doi: [10.1109/JSTARS.2018.2844873](#).
- [45] E. Shelhamer, J. Long, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 3431–3440.
- [46] H. Sun, X. Zheng, and X. Lu, "A supervised segmentation network for hyperspectral image classification," *IEEE Trans. Image Process.*, vol. 30, pp. 2810–2825, Feb. 2021, doi: [10.1109/TIP.2021.3055613](#).
- [47] Y. Xu, B. Du, and L. Zhang, "Beyond the patchwise classification: Spectral-spatial fully convolutional networks for hyperspectral image classification," *IEEE Trans. Big Data*, vol. 6, no. 3, pp. 492–506, Sep. 2020, doi: [10.1109/TBDATA.2019.2923243](#).
- [48] Z. Zheng, Y. Zhong, A. Ma, and L. Zhang, "FPGA: Fast patch-free global learning framework for fully end-to-end hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 8, pp. 5612–5626, Aug. 2020, doi: [10.1109/TGRS.2020.2967821](#).
- [49] Q. Zhu et al., "A spectral-spatial-dependent global learning framework for insufficient and imbalanced hyperspectral image classification," *IEEE Trans. Cybern.*, May 2021, doi: [10.1109/TCYB.2021.3070577](#).
- [50] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 770–778.
- [51] K. He, X. Zhang, S. Ren, and J. Sun, "Identity mappings in deep residual networks," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 630–645.
- [52] Z. Zheng, X. Zhang, P. Xiao, and Z. Li, "Integrating gate and attention modules for high-resolution image semantic segmentation," *IEEE J. Sel. Topics Appl. Earth Obs. Remote Sens.*, vol. 14, pp. 4530–4546, Apr. 2021, doi: [10.1109/JSTARS.2021.3071353](#).
- [53] Z. Yang, L. Zhu, Y. Wu, and Y. Yang, "Gated channel transformation for visual recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2020, pp. 11794–11803.
- [54] O. Oktay et al., "Attention U-Net: Learning where to look for the pancreas," May 2018, *arXiv:1804.03999*.
- [55] M. Tang, F. Perazzi, A. Djelouah, I. B. Ayed, C. Schroers, and Y. Boykov, "On regularized losses for weakly-supervised CNN segmentation," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 507–522.
- [56] A. Obukhov, S. Georgoulis, D. Dai, and L. Van Gool, "Gated CRF loss for weakly supervised semantic image segmentation," Jun. 2019, *arXiv:906.04651*.
- [57] Z. Li, X. Zhang, P. Xiao, and Z. Zheng, "On the effectiveness of weakly supervised semantic segmentation for building extraction from high-resolution remote sensing imagery," *IEEE J. Sel. Topics Appl. Earth Obs. Remote Sens.*, vol. 14, pp. 3266–3281, Mar. 2021, doi: [10.1109/JSTARS.2021.3063788](#).
- [58] P. Krähenbühl and V. Koltun, "Efficient inference in fully connected crfs with gaussian edge potentials," in *Proc. Adv. Neural Inf. Process. Syst.*, 2011, pp. 109–117.
- [59] B. Zhang, J. Xiao, Y. Wei, M. Sun, and K. Huang, "Reliability does matter: An end-to-end weakly supervised semantic segmentation approach," in *Proc. AAAI Conf. Artif. Intell.*, Feb. 2020, pp. 12765–12772.
- [60] J. Zhu, L. Fang, and P. Ghamisi, "Deformable convolutional neural networks for hyperspectral image classification," *IEEE Geosci. Remote Sens. Lett.*, vol. 15, no. 8, pp. 1254–1258, Aug. 2018, doi: [10.1109/LGRS.2018.2830403](#).
- [61] X. Zhang, S. Shang, X. Tang, J. Feng, and L. Jiao, "Spectral partitioning residual network with spatial attention mechanism for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, Jan. 2022, Art. no. 5507714, doi: [10.1109/TGRS.2021.3074196](#).



Xueliang Zhang (Member, IEEE) received the B.S. degree in geographical information system and the Ph.D. degree in remote sensing of resources and environment from Nanjing University, Nanjing, China, in 2010 and 2015, respectively.

From 2014 to 2015, he was a visited student with the Information Institute, University of Missouri, Columbia, USA. From 2016 to 2018, he was an Associate Researcher with the Department of Geographic Information Science, Nanjing University. He is currently an Associate Professor with the Department of Geographic Information Science, Nanjing University. His research interests include high-resolution remote sensing image analysis, semantic segmentation, and deep learning for remote sensing.



Zixian Zheng received the B.S. degree in geographic information science from Sun Yat-sen University, Guangzhou, China, in 2019, and the M.S. degree in cartography and geographic information system from Nanjing University, Nanjing, China, in 2022.

Her research interests include semantic segmentation and deep learning for remote sensing.



Pengfeng Xiao (Senior Member, IEEE) received the B.M. degree in land resource management from Hunan Normal University, Changsha, China, in 2002, and the Ph.D. degree in cartography and geographical information system from Nanjing University, Nanjing, China, in 2007.

From 2007 to 2009, he was a Lecturer with the School of Geography and Ocean Science, Nanjing University, where he was an Associate Professor from 2010 to 2018. Since 2019, he has been a Professor with Nanjing University. He was a Visiting Scholar with the Department of Geography, University of Giessen, Giessen, Hesse, Germany, from 2011 to 2012, and the Department of Environmental Science, Policy, and Management, University of California at Berkeley, Berkeley, CA, USA, from 2014 to 2015. He has authored 4 books and more than 160 articles. His current research interests include high-resolution remote sensing image analysis, remote sensing of snow cover, and land use and land cover change.



Guangjun He was born in Hubei, China, in 1987. He received the Ph.D. degree in remote sensing of resources and environment from Nanjing University, Nanjing, China, in 2015.

From 2015 to 2017, he was an Assistant Researcher Fellow with the State Key Laboratory of Space-Ground Integrated Information Technology, CAST, Beijing, China, where he has been an Associate Researcher Fellow, since 2018. He has authored more than 40 articles. His research interests include remote sensing image processing, multisensor information fusion, and machine learning.



Zhenshi Li received the B.S. degree in geographic information science from Hohai University, Nanjing, China, in 2019, and the M.S degree in cartography and geographic information system in 2022 from Nanjing University, Nanjing, China, where he is currently working toward the Ph.D. degree in cartography and geographic information system.

His research interests include semantic segmentation and weakly supervised deep learning for remote sensing.