

FPGA: Fast Patch-Free Global Learning Framework for Fully End-to-End Hyperspectral Image Classification

Zhuo Zheng, *Student Member, IEEE*, Yanfei Zhong, *Senior Member, IEEE*, Ailong Ma,
and Liangpei Zhang, *Fellow, IEEE*,

Abstract—Deep learning techniques have provided significant improvements in hyperspectral image (HSI) classification. The current deep learning based HSI classifiers follow a patch-based learning framework by dividing the image into overlapping patches. As such, these methods are local learning methods, which have a high computational cost. In this paper, a fast patch-free global learning (FPGA) framework is proposed for HSI classification. The proposed framework consists of three main parts: 1) a designed sampling strategy; 2) an encoder-decoder based fully convolutional network (FCN); and 3) lateral connections between the encoder and decoder. In FPGA, an encoder-decoder based FCN is utilized to consider the global spatial information by processing the whole image, which results in fast inference. However, it is difficult to directly utilize the encoder-decoder based FCN for HSI classification as it always fails to converge due to the insufficiently diverse gradients caused by the limited training samples. To solve the divergence problem and maintain the FCN’s abilities of fast inference and global spatial information mining, a global stochastic stratified (GS^2) sampling strategy is first proposed by transforming all the training samples into a stochastic sequence of stratified samples. This strategy can obtain diverse gradients to guarantee the convergence of the FCN in the FPGA framework. For a better design of FCN architecture, FreeNet, which is a fully end-to-end network for HSI classification, is proposed to maximize the exploitation of the global spatial information and boost the performance via a spectral attention based encoder and a lightweight decoder. A lateral connection module is also designed to connect the encoder and decoder, fusing the spatial details in the encoder and the semantic features in the decoder. The experimental results obtained using three public benchmark datasets suggest that the FPGA framework is superior to the patch-based framework in both speed and accuracy for HSI classification. Code has been made available at: <https://github.com/Z-Zheng/FreeNet>.

Index Terms—Patch-free global learning, fully convolutional network, feature fusion, hyperspectral image classification

I. INTRODUCTION

This work was supported by National Key Research and Development Program of China under Grant No. 2017YFB0504202, National Natural Science Foundation of China under Grant Nos. 41771385; and the National Natural Science Foundation of China under Grant NO. 41801267, in part by the China Postdoctoral Science Foundation under Grant 2017M622522. (*Corresponding authors:* Yanfei Zhong, Ailong Ma)

The authors are with the State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing, Wuhan University, Wuhan 430079, China, with the Hubei Provincial Engineering Research Center of Natural Resources Remote Sensing Monitoring, Wuhan University, Wuhan 430079, China. (e-mail: zhengzhuo@whu.edu.cn; zhongyanfei@whu.edu.cn; maailong007@whu.edu.cn; zlp62@whu.edu.cn)

HYPERSPECTRAL imaging, as a particularly important technique, is able to obtain abundant spectral information about the ground surface [1–3]. As a result, it is widely applied in the fields of geology, agriculture, forestry, and environmental monitoring [4–6]. Hyperspectral image classification with the goal to assign a unique semantic label to each pixel in a hyperspectral image (HSI) [7], is a fundamental but challenging part of hyperspectral remote sensing (HRS).

For HSI classification, the spectral feature-based methods, such as support vector machine (SVM) [8], random forest (RF) [9], rotation forest (RoF) [10], canonical correlation forest (CCF) [11, 12] and multinomial logistic regression (MLR) [13], are the traditional classifiers for HSIs. To further improve the accuracy of HSI classification, spatial information has been integrated into the existing pipelines [14–16]. Thereby, spectral-spatial feature based methods, such as the gray level cooccurrence matrix [17], wavelet transform [18], the Gabor filter [19], etc., have been proposed to improve the discrimination of the features. Extended morphological profiles (EMPs) [20, 21] has also been proposed to leverage the spatial context with multiple morphological operations for HSI classification. However, these spectral-spatial features are handcrafted, and they are strongly reliant on the prior information and empirical hyperparameters [22, 23].

To automatically obtain more general spectral-spatial features, deep learning technology [24], as a data-driven automatic feature learning framework, has now been introduced into HSI classification. Among the deep learning based methods, convolutional neural networks (CNNs), as hierarchical spectral-spatial feature representation learning frameworks, have been widely used in HSI classification [25–29], significantly boosting the accuracy when compared with the traditional methods. More importantly, CNN-based methods can act as an end-to-end training feature extractor and classifier for global optimization to obtain a better accuracy. These CNN-based methods follow a patch-based local learning framework [22, 23, 25, 29–32], where patch generation is first performed to obtain a dense set of patches with a fixed size $S \times S$ and then patchwise classification is applied to each patch.

However, these methods usually have a high computational complexity under the patch-based local learning framework. This is because these methods first generate overlapping image patches and then assign semantic labels obtained by the CNN to the corresponding central pixels to obtain complete classification map. However this results in redundant

computation since the image patches generated by adjacent pixels overlap with each other. This seriously constrains the speed of the methods under the patch-based local learning framework. Meanwhile, the limited patch size constrains the spatial context, making it difficult for the CNN to model long-range dependency.

In this work, a fast patch-free global learning (FPGA) framework is proposed for HSI classification. The FPGA framework includes a sampling strategy, an encoder-decoder based FCN, and lateral connections between the encoder and decoder. To share the computation in the spatial dimension and leverage the global spatial information, an encoder-decoder based fully convolutional network (FCN) is introduced to end-to-end HSI classification. However, training an FCN for HSI classification is difficult since it always fails to converge. The main reason for this is the insufficiently diverse gradients caused by the limited training samples during the backward computation. To guarantee the convergence of the training of the FCN, a global stochastic stratified (GS^2) sampling strategy is proposed to obtain diverse gradients during back-propagation. Furthermore, FreeNet, which is a novel network architecture for HSI classification, is proposed to maximize the exploitation of the global spatial information and further boost the performance. FreeNet consists of a spectral attention based encoder and a lightweight decoder. In addition, a lateral connection is applied to fuse the spatial details in the encoder and the semantic features in the decoder, for better exploitation of the encoder-decoder structural characteristics, which can help to recover more clear edges of objects in the classification map.

The main contributions of our study are summarized as follows:

- 1) A fast patch-free global learning (FPGA) framework is proposed for HSI classification. The FPGA framework includes a sampling strategy, an encoder-decoder based FCN, and lateral connections between the encoder and decoder, which can achieve faster patch-free inference and learn from the global spatial information, for a better accuracy.
- 2) To guarantee the convergence of the training of the FCN, the GS^2 sampling strategy is designed to assist with the training of the FCN. GS^2 strategy transforms all the training samples into a stochastic sequence of stratified samples, to obtain diverse gradients during back-propagation, for more effective parameter updating.
- 3) To further boost the performance, a novel network architecture, FreeNet, is proposed for HSI classification through exploiting the global spatial information. FreeNet consists of a spectral attention based encoder and a lightweight decoder. Spectral attention involves modeling the interdependencies of the feature maps, using the global spatial context to guide the importance of the feature maps. This ensures sufficient exploitation of the redundant spectral information and the global spatial information. A lightweight decoder is responsible for the progressive recovery of the classification map, with less burden on optimization.
- 4) To make full use of the encoder-decoder structural

characteristics, a lateral connection between the encoder and decoder is designed for the fusion of the spatial details in the encoder with the semantic information in the decoder. This refines the semantic features with the spatial detail features to obtain a clearer classification map.

The rest of this paper is organized as follows. Section II briefly introduces the handcrafted feature based and CNN based HSI classifiers via the patch-based local learning framework. Section III then describes the details of the proposed unified patch-free learning framework. Section IV describes the comparative results obtained on three HSI classification benchmark datasets, and further analyzes the proposed modules and the introduced hyperparameters. Finally, Section VI concludes this paper.

II. RELATED WORKS

A. Deep Learning Based HSI Classifiers Under a Patch-Based Local Learning Framework

The dominant methods in modern HSI classification are based on deep networks and follow a patch-based local learning framework. The deep networks include stacked autoencoders (SAEs), deep belief networks (DBNs), CNNs, recurrent neural networks (RNNs) and generative adversarial networks (GANs), which have all been explored in HSI classification [22, 23, 26, 27, 31–35]. Among these methods, CNN-based classifiers, which are regarded as the natural spectral-spatial classification methods, have obvious advantages in accuracy. To conveniently extract features and learn a classifier using CNNs, HSI patches are first generated from the original image by a window with a fixed size $S \times S$ (e.g. 7×7 or 28×28). HSI classification always involves modeling a patch classification task [1, 3], which involves learning a mapping $f : R^{S \times S} \rightarrow R$, as shown in Fig. 1.

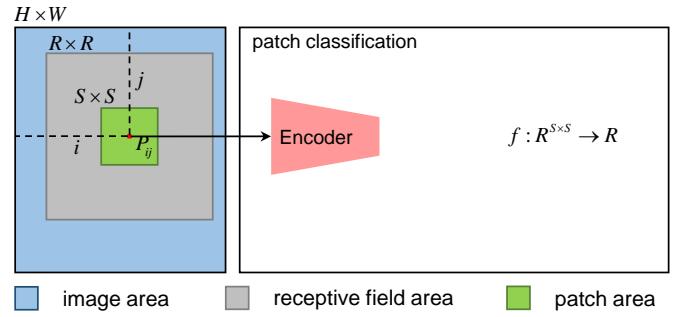


Fig. 1. The patch-based local learning framework for HSI classification. For simplification, the mapping considers only the spatial dimension.

Under the patch-based local learning framework, the main difference between the CNN-based classifiers is in the different deep CNN designs. A simple deep CNN was employed for HSI classification in [26]. To utilize the spatial context information, a contextual CNN was proposed in [36], further improving the classification accuracy. A CNN with pixel-pair features (CNN-PPF) was proposed in [37] to enhance the original CNN by using deep pixel-pair features. A spectral-spatial feature based classification framework was proposed in [34],

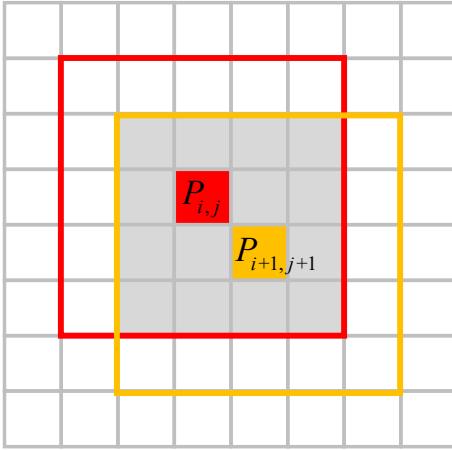


Fig. 2. The overlap of 5×5 patches under the patch-based local learning framework. The gray cell presents the overlap area. $P_{i,j}$ and $P_{i+1,j+1}$ are the central pixels of the two patches with red and yellow borders, respectively.

combining a balanced local discriminant embedding algorithm used as the spectral feature extractor with the CNN used as a spectral-spatial feature extractor for HSI classification. To obtain more discriminative features, the Siamese convolutional neural network (S-CNN) was proposed in [38] to learn low intraclass and high interclass features via a two-branch network, supervised by a margin ranking loss. The deformable HSI classification networks (DHCNet) method [30] uses an adaptive spatial context modeling method to capture the complex spatial context in the HSIs and boost the performance. However, the overfitting issue gradually emerges as the model complexity increases. To alleviate this issue, Gabor-CNN [39] combines Gabor filters with convolutional filters to reduce the feature extraction burden of the CNN. The deep feature fusion network (DFFN) was proposed in [40] as a multi-layer feature fusion method that adopts residual learning to mitigate the overfitting brought by the introduction of more convolutional layers, significantly improving the classification accuracy for HSIs. Although these patch-based methods have achieved remarkable HSI classification accuracies, obtaining a fast inference speed remains a challenge, which limits the further application of HSI classifiers. The main reason for this is the redundant computation in the overlapping areas between patches, as shown in Fig. 2. The pixels in the gray area will take part in multiple computations since these pixels are the neighbors of multiple central pixels.

Although the DMS³FE-classifier [41] utilizes a pretrained FCN [42] to extract features, its own FCN is not trained. A convolution-deconvolution (conv-deconv) network with an optimized extreme learning machine (ELM) method was proposed in [43] for HSI classification with an FCN, but the method is not an end-to-end classifier.

Neither of these methods are end-to-end trainable FCNs since they only utilize the FCN to extract features, and apply separate classifiers to label the pixels, which means that the whole pipeline cannot be globally optimized and the global spatial information cannot be exploited sufficiently. To overcome the aforementioned issues, we propose the FPGA

framework for HSI classification.

III. FPGA: FAST PATCH-FREE GLOBAL LEARNING FRAMEWORK FOR HSI CLASSIFICATION

We investigated the main speed bottleneck of the patch-based methods and concluded that the redundant computation on the highly overlapping areas between patches is the central cause. To address this issue, we propose a fast patch-free global learning (FPGA) framework and a variant of an FCN (FreeNet) as a fundamental classification model in the FPGA framework through sharing computation in the spatial dimension.

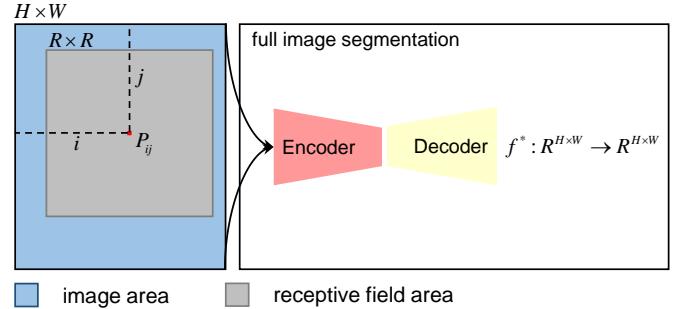


Fig. 3. The patch-free global learning framework for HSI classification. For simplification, the mapping only considers the spatial dimension.

The FPGA framework aims to learn a mapping $f^*: R^{H \times W} \rightarrow R^{H \times W}$ for full image segmentation, as shown in Fig. 3. In the FPGA framework, there are three core components: 1) the GS² sampler; 2) the encoder-decoder based FCN; and 3) the lateral connections. The GS² sampler ensures the convergence of the end-to-end trained FCN based model, and the encoder-decoder based FCN is responsible for one-shot forward computation by sharing the computation in the spatial dimension. The lateral connections are designed to effectively fuse the spatial detail features in the encoder and the semantic features in the decoder. The model architecture follows the classical encoder-decoder framework [42, 44–46] with semantic-spatial fusion (SSF). The encoder is used to transform spatially finer features to semantically stronger features by progressively learning higher-dimensional feature embedding. The decoder is used to recover the spatial information of the semantic features with the high-dimension feature embedding learned by the encoder for the full classification map. The lateral connection based SSF forwards spatially finer features from the encoder to the decoder, which is beneficial for the recovery of the spatial details of the semantic feature maps.

A. Patch-Free Global Learning

The core idea of patch-free global learning is to replace the explicit patching with the implicit receptive field of the model, to avoid redundant computation on the overlapping areas and obtain a wider latent spatial context.

Given an HSI $X \in R^{C \times H \times W}$, the predicted probability cube $\hat{Y}_i \in R^{\#class \times H \times W}$ is formulated as:

$$\hat{Y}_i = f^*(X) \quad (1)$$

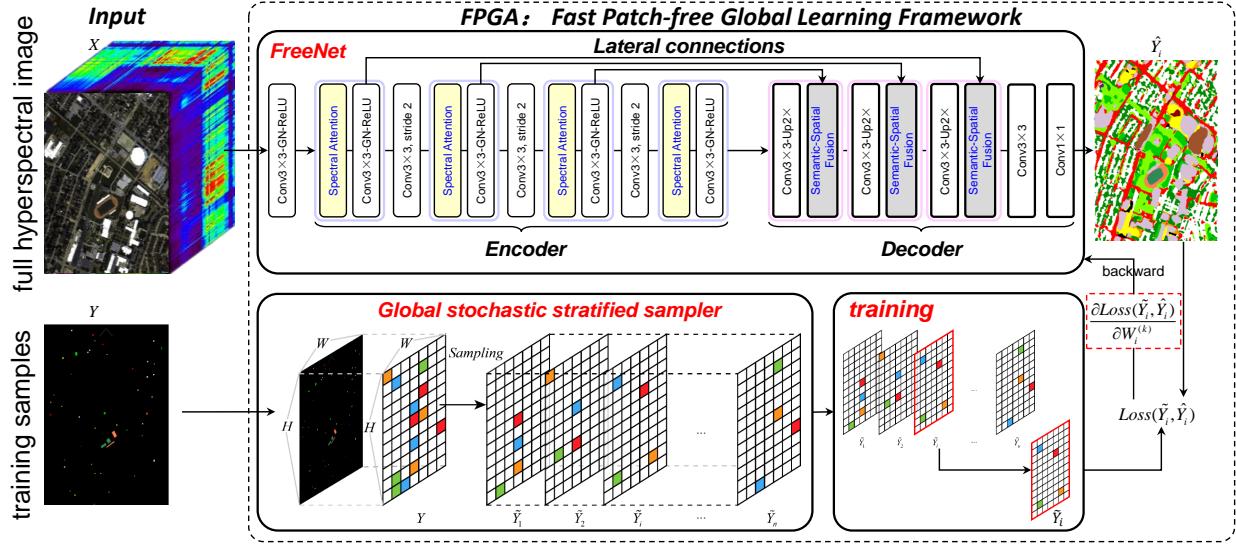


Fig. 4. The overview of the fast patch-free global learning framework. The FPGA framework includes three core components: the GS² sampler, the encoder-decoder based FCN and lateral connections.

where the mapping $f^* : R^{C \times H \times W} \rightarrow R^{\#class \times H \times W}$ is modeled as the classifier without patching, and C is the number of bands of X .

Stochastic gradient descent (SGD) [47] is used to minimize the classification loss l (e.g., cross-entropy loss) over the sampled positions \mathcal{R} . The sampling algorithm is described in Section III-B.

For the i -th iteration, the k -th weight of the model can be updated as follows:

$$W_{i+1}^{(k)} = W_i^{(k)} - \eta \frac{1}{n} \sum_{p \in \mathcal{R}_i} \frac{\partial l(\tilde{Y}_i(p), \hat{Y}_i(p))}{\partial W_i^{(k)}} \quad (2)$$

where p is the 2-D spatial position in \mathcal{R}_i , $n = |\mathcal{R}_i|$, η is the learning rate and \tilde{Y}_i is the sampled ground truth map.

The main difference with patch-based local learning is that all the pixels take part in the forward computation during the training, but only the sampled position can obtain supervised signals for every iteration. In this way, the model inference is consistent during the training and testing, which are both one-shot forward computations. The one-shot forward computation significantly boosts the speed of the model inference for HSI classification. Meanwhile, the patch-free global learning allows the model to leverage the spatial context as much as possible. It thus provides more potential to boost the accuracy of the model by following the patch-free global learning framework.

B. Global Stochastic Stratified Sampling Strategy

The global stochastic stratified (GS²) sampling strategy is proposed to ensure the convergence of the end-to-end trained FCN based model. The GS² sampling strategy is formally described in Algorithm 1. The key idea of the GS² sampler is to transform all the training samples into a stochastic sequence of stratified samples. In this way, the GS² sampler ensures the class-balanced distribution of the training samples and simulates the behavior of the mini-batch sampler to obtain

Algorithm 1: Global Stochastic Stratified Sampling

```

Input:  $G = \{g_i\}_{i=1}^M$ : a set of labels for training
       $N$ : the number of classes
       $\alpha$ : mini-batch per class
Output:  $T$ : a list of sets of stratified labels
 $R \leftarrow []$  // an empty list
for  $k = 0$  to  $N$  do
   $I_k \leftarrow \{j | g_j = k, g_j \in G\}$ 
   $I_k \leftarrow \text{shuffle}(I_k)$ 
   $R[k] \leftarrow []$ 
  while  $|I_k| > \alpha$  do
    | fetch  $\alpha$  samples from  $I_k$ ,  $t \leftarrow I_k.pop(\alpha)$ 
    |  $R[k].push(t)$ 
  end
   $R[k].push(I_k)$ 
end
 $T \leftarrow []$ 
 $c \leftarrow 0$ 
while  $\text{any}(R[i] > 0, i = 1, 2, \dots, N - 1)$  do
   $T[c] \leftarrow []$ 
  for  $k = 0$  to  $N$  do
    | if  $|R[k]| > 0$  then
      | | fetch 1 element from  $R[k]$ ,
      | |  $t_k = R[k].pop(1)$ 
      | |  $T[c] \leftarrow T[c] \cap t_k$ 
    | end
  end
   $T.push(T[c])$ 
   $c \leftarrow c + 1$ 
end

```

stable yet diverse gradients. During the training, this sequence is randomly shuffled to ensure stochasticity of the gradients, to prevent overfitting.

Firstly, in more detail, we split all the training samples to

obtain a list R , where the index of R is the class label, and the element is the training samples of each class. During the splitting, the order of the training samples for each class needs to be shuffled to keep the stochasticity of the combination. Stratification is then performed on the training samples of each class to obtain T , which is a list of the sets of stratified labels.

In this sampling strategy, the hyperparameter α (the mini-batch per class) is introduced, which is of great significance for the training stability. The smaller the value of α , the greater the number of gradient orientations that can be obtained when optimizing the network, which makes it possible to train an FCN using limited training samples for HSI classification. A more detailed analysis of parameter α is provided in Section V-A.

C. FreeNet in FPGA

FreeNet is a simple, unified network made up of an *encoder* network and a *decoder* network. The encoder is responsible for computing the hierarchical convolutional feature maps over an entire input HSI. The decoder recovers the spatial dimension of the coarsest convolutional feature map progressively via lateral connection based SSF, outputting a classification probability map of the same spatial size as the input image. To improve the FreeNet compactness, we introduce a compression factor (β) to control the number of feature maps in the whole network, achieving a trade-off between speed and accuracy. FreeNet is a lightweight FCN designed for faster and more accurate HSI classification, as shown in Fig. 5. Each component of FreeNet is described in the following.

1) **Encoder Network Architecture:** The encoder network follows a modular design, made up of a stem block and four hybrid blocks, all of which contain the basic module. The basic module of the encoder network is a 3×3 convolutional layer followed by group normalization [48] and rectified linear unit (ReLU) activation. Under the FPGA framework, the batch size is always equal to 1, and the iterative inputs are the same image, because of the entire HSI being used as input. In this case, the error of the batch normalization (BN) increases rapidly due to the inaccurate batch statistics estimation. Therefore, we adopt group normalization (GN) as an alternative to BN, which is independent of batch size and can obtain a comparable performance to BN.

Due to the different numbers of bands for HSIs, we first introduce a stem block to transform the variable channels of the input to a fixed 64 channels. The stem block is simply implemented by a basic module. The four hybrid blocks share the same network topology, unless otherwise specified. The hybrid block is made up of a spectral attention module, as described in Section III-C2, a basic module and an optional downsampling module. For the downsampling module, we use a 3×3 convolutional layer with a stride of 2 followed by ReLU activation to replace the commonly used 3×3 maxpool with a stride of 2 to align the projected spatial location with its receptive field center for more robust HSI classification. Block#1 ~ #3 are the hybrid blocks with downsampling modules and block#4 is the block without a downsampling module.

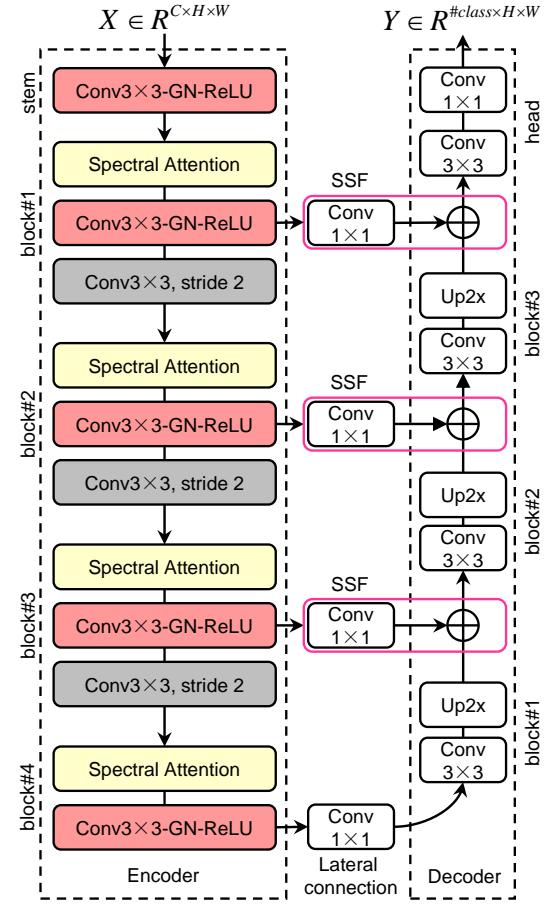


Fig. 5. The FreeNet network architecture designed for HSI classification.

2) **Spectral Attention:** Spectral attention models the interdependencies of the feature maps with the global spatial context. The encode function is simply implemented by 3×3 convolution. This module reweights the feature maps via global context guiding and highlights the more important feature maps to improve the accuracy of the model.

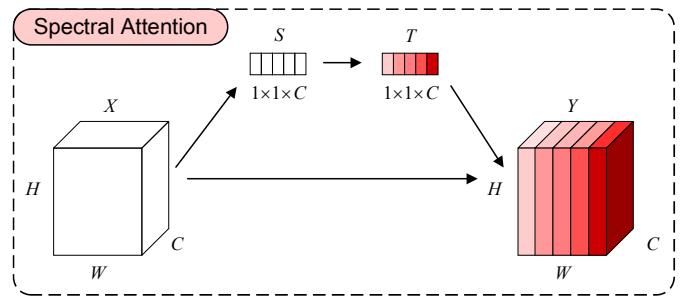


Fig. 6. The network architecture of the spectral attention module.

The spectral attention is an in-place module where there is no dimension change between the input and output, which is a similar implementation to the SE-Block [49]. Given the input tensor $X \in \mathbb{R}^{C \times H \times W}$, we first compute the global context

embedding vector $S \in \mathbb{R}^{C \times 1 \times 1}$, where

$$S(k, :, :) = \frac{1}{H \cdot W} \sum_{i=1}^H \sum_{j=1}^W X(k, i, j) \quad (3)$$

where k is the index of the channel dimension. In order to model the interdependencies between feature maps, we use two thin fully connected layers (nonlinear transformation), followed by a sigmoid gating function to compute the channel scaling coefficient vector $T \in \mathbb{R}^{C \times 1 \times 1}$, where

$$T = \text{sigmoid}(W_2 \delta(W_1 S)) \quad (4)$$

where δ denotes the ReLU function, $W_1 \in \mathbb{R}^{\frac{C}{r} \times C}$, and $W_2 \in \mathbb{R}^{C \times \frac{C}{r}}$. The reduction ratio r is introduced to balance the capacity of the model and the computational cost. We find $r = 16$ works well in practice, so this setting was used in most of the experiments. The output $Y \in \mathbb{R}^{C \times H \times W}$ of the spectral attention module is simply computed by:

$$Y(k, i, j) = T(k, :, :) \cdot X(k, i, j) \quad (5)$$

3) Decoder Network Architecture: The decoder network also follows a modular design, for simplicity, which is made up of a refinement module for progressive spatial feature refinement and a head subnetwork for pixel classification. The refinement module contains multiple refinement stages, which are simply implemented by stacking the upsampling modules and inserting SSF after each upsampling module. The progressive refinement first upsamples the input feature maps with stronger semantic information and then aggregates the feature map with finer spatial information from the encoder to recover the spatial details of the input. The upsampling module is a 3×3 convolutional layer followed by nearest neighbor upsampling with a factor of 2. The SSF receives two features from the blocks in the encoder and decoder, respectively, and aggregates the features into a new enhanced feature that is forwarded to the next upsampling module. The head subnetwork is used to perform pixel classification with the feature from the top layer of the decoder, and is made up of a 3×3 convolutional layer followed by a 1×1 convolutional layer with N filters. N is the number of categories.

4) Lateral Connection Based SSF: Lateral connection based SSF leverages the spatial detail features of the shallow convolutional layers to enhance the deep semantic features of the convolutional layers and boost the performance. The lateral connection is implemented by a 1×1 convolutional layer, which passes more precise locations of features from the encoder to the decoder. The aggregation function is pointwise addition. Lateral connection based SSF can be formulated as follows:

$$q_{i+1} = q_i + \text{conv}(p_{4-i}), i = 1, 2, 3 \quad (6)$$

where q_i is the feature map of refinement stage# i in the decoder, and p_{4-i} is the feature map of block# $4 - i$ in the encoder. q_{i+1} is the output of SSF, which is forwarded to the next block in the decoder. The design of the lateral connection based SSF follows that of residual learning [50]. The first item q_i is a baseline item obtained by nearest neighbor interpolation and $\text{conv}(p_{4-i})$ is the residual item that needs to be learned.

In this case, the gradients are lossless to flow into the shallow layers, making the optimization easier.

D. Fully End-to-End HSI Classification Using FreeNet in FPGA

After convergence of the trained FreeNet, HSI classification can be implemented by one-shot forward computation using FreeNet in FPGA. Compared with patchwise classification for HSIs, FreeNet can perform faster patch-free inference over the whole HSI through sharing the computation in the spatial dimension. Due to the introduction of three $2 \times$ upsampling blocks in FreeNet, the input size in the spatial dimension of the HSI should be multiples of $2^3 = 8$. To ensure that the raw HSI is unchanged, a “padding-crop” trick is used for the inference, which pads the original input with zeros into a new size with multiples of 8 before the inference. After the inference, the final classification map is obtained by cropping the output using a box with the original input size.

IV. EXPERIMENTAL RESULTS AND ANALYSIS

Extensive experiments were conducted to validate the effectiveness of the proposed method on three benchmark datasets: the ROSIS-03 Pavia University dataset, the Salinas dataset and the Compact Airborne Spectrographic Imager (CASI) University of Houston dataset. Four patch-based HSI classifiers were compared with the proposed FreeNet and its variants. The patch-based classifiers used in the comparison were classical SVM [8] and four state-of-the-art deep learning based methods (S-CNN [38], Gabor-CNN [39], DFFN [40], 3D-GAN [51]), following [3]. All the experiments were performed with an NVIDIA Tesla P100 GPU accelerator (with 16GB GPU memory).

A. Experimental Settings

1) Network Architecture: The hyperparameter settings of the standard FreeNet, namely FreeNet ($\beta = 1.0$), such as the output channels of the layers and the reduction ratio r in the spectral attention module, are listed in Table. I. For a fair comparison, this architecture setting was used for all three benchmark datasets, and there was no specific tuning for the dataset.

2) Optimization: For all the experiments, the patch-free methods were trained for 1k iterations using SGD with a “poly” learning rate policy, where the initial learning rate was set to 0.0001 and multiplied by $(1 - \frac{\text{iter}}{\text{max_iter}})^{\text{power}}$ with $\text{power} = 0.9$. The momentum was set to 0.9 and the weight decay was set to 0.0001. We did not use any data augmentation strategy. Unless otherwise specified, α was set to 20 for the GS² sampler.

3) Metrics: To evaluate the performance of the proposed methods, four common metrics are adopted, which are the accuracy of each class, the overall accuracy (OA), the average accuracy (AA), and the Kappa coefficient (Kappa).

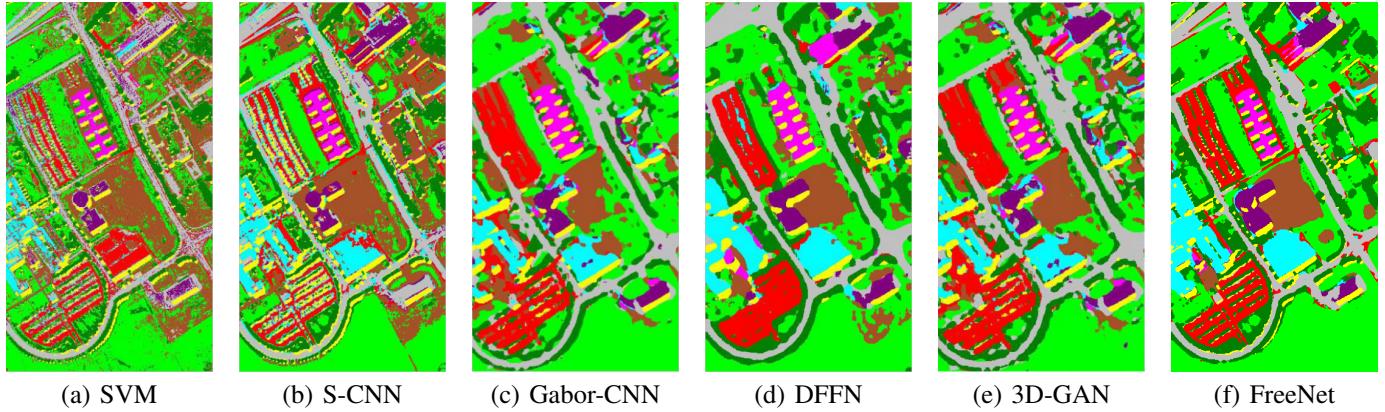


Fig. 7. Visualization of the classification maps for the ROSIS-03 Pavia University dataset. (a) SVM. (b) S-CNN. (c) Gabor-CNN. (d) DFFN. (e) 3D-GAN. (f) FreeNet.

TABLE I
THE CONFIGURATION DETAILS OF THE STANDARD FREE NET (FREE NET WITH $\beta = 1.0$).

FreeNet ($\beta = 1.0$)		
Encoder	stem	3×3 conv, 64
	block#1	spectral attention, $r = 16$ 3×3 conv, 64 3×3 conv, 128, stride 2
	block#2	spectral attention, $r = 16$ 3×3 conv, 128 3×3 conv, 192, stride 2
	block#3	spectral attention, $r = 16$ 3×3 conv, 192 3×3 conv, 256, stride 2
	block#4	spectral attention, $r = 16$ 3×3 conv, 256
Lateral	lateral 4-1	1×1 conv, 128
	lateral 3-1	1×1 conv, 128
	lateral 2-2	1×1 conv, 128
	lateral 1-3	1×1 conv, 128
Decoder	block#1	3×3 conv, 128 upsample, 2
	block#2	3×3 conv, 128 upsample, 2
	block#3	3×3 conv, 128 upsample, 2
	head	3×3 conv, 128 1×1 conv, N

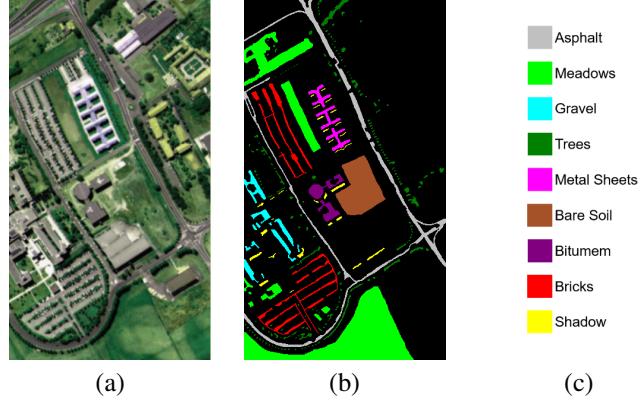


Fig. 8. The ROSIS-03 Pavia University dataset. (a) Three-band false color composite. (b) Ground-truth map (c) Legend

TABLE II
THE NUMBER OF TRAINING SAMPLES AND TEST SAMPLES FOR THE
ROSIS-03 PAVIA UNIVERSITY DATASET

Class	Class name	#Training	#Test	#Total
C1	Asphalt	200	6431	6631
C2	Meadows	200	18449	18649
C3	Gravel	200	1899	2099
C4	Trees	200	2864	3064
C5	Metal Sheets	200	1145	1345
C6	Bare Soil	200	4829	5029
C7	Bitumem	200	1130	1330
C8	Bricks	200	3482	3682
C9	Shadow	200	747	947
Total	-	1800	40976	42776

B. Experiment 1: ROSIS-03 Pavia University Dataset

The HSI of this dataset contains 610×340 pixels and 103 spectral bands with a spatial resolution of 1.3 m per pixel. This dataset contains nine urban land-cover types. Fig. 8 shows the false-color composite of the image and the corresponding ground truth.

Table. II lists the number of training and test samples for

each class. The training samples were randomly chosen from the ground truth with a fixed random seed and the remaining samples were used to evaluate the accuracy.

For the HSI classification task, the visual performance is of importance for the classifier. Fig. 7 shows the classification maps of the compared methods. As can be seen in Fig. 7 (b)-(f), the CNN based classifiers have a better visual performance

TABLE III

THE CLASSIFICATION RESULTS OF SVM [8], S-CNN[38], GABOR-CNN [39], DFFN [40], 3D-GAN [51] AND FREEFNET ON THE ROSIS-03 PAVIA UNIVERSITY DATASET.

Class	Patch-based					Patch-free	
	SVM	S-CNN	Gabor-CNN	DFFN	3D-GAN	FreeNet	
Accuracy(%)	C1	85.49	95.47	99.53	99.53	99.18	99.58
	C2	92.12	98.71	98.21	97.71	98.86	99.88
	C3	85.77	97.32	89.74	99.89	94.94	99.95
	C4	96.41	97.72	93.02	97.88	90.15	99.27
	C5	98.60	100	99.42	99.48	99.49	100
	C6	92.52	97.67	98.77	99.69	98.56	100
	C7	93.79	98.36	98.82	100	92.74	100
	C8	86.56	95.56	94.12	98.59	97.18	99.83
	C9	97.97	100	97.91	99.61	98.51	100
OA(%)		90.78	97.93	97.33	98.57	97.81	99.81
AA(%)		92.14	97.88	96.62	99.16	96.65	99.83
Kappa		0.8813	0.9743	0.9662	0.9808	0.9697	0.9974

than the classical SVM classifier due to the strongly discriminative deep features. It is clear that the patch-free method is superior to the patch-based methods in spatial detail when comparing Fig. 7 (f) and Fig. 7 (b)-(e). This can be attributed to the introduction of more global spatial context. Among the different methods, the edges of the classification map of FreeNet are smoother than those of the other methods. This indicates that FreeNet can capture finer spatial detail by the lateral connection based SSF and the better design of decoder structure.

Table. III lists the results of the state-of-the-art patch-based methods and the proposed patch-free methods. When constraining the number of training samples (200 samples for each class), DFFN achieves the best accuracy (OA of 98.57%, AA of 99.16%, and Kappa of 0.9808) among the patch-based methods. However, the patch-free method obtains more accurate results than DFFN. FreeNet obtains a higher OA of 99.81%, exceeding DFFN by $\sim 1\%$. Meanwhile, we can observe that the accuracies for each class with the patch-free method are higher than those of the patch-based methods. In fact, the accuracy of the patch-free method has reached saturation. This suggests that the patch-free global learning framework is superior to the patch-based local learning framework with this benchmark dataset.

C. Experiment 2: Salinas Dataset

To further evaluate the effectiveness of the FPGA framework, we also performed experiments on the Salinas dataset. The HSI of Salinas dataset has 512×217 pixels and 204 spectral bands with a spatial resolution of 3.7 m. The ground-truth map covers 16 classes of interest. Fig. 9 shows the three-band false-color composite image and the corresponding ground-truth map. The numbers of training and test samples are listed in Table. IV. The selection of samples was random selection with the same random seed as used in Experiment 1.

Fig. 10 shows the visual performance of the different methods. We can observe that the classification maps in Fig. 10(d)-

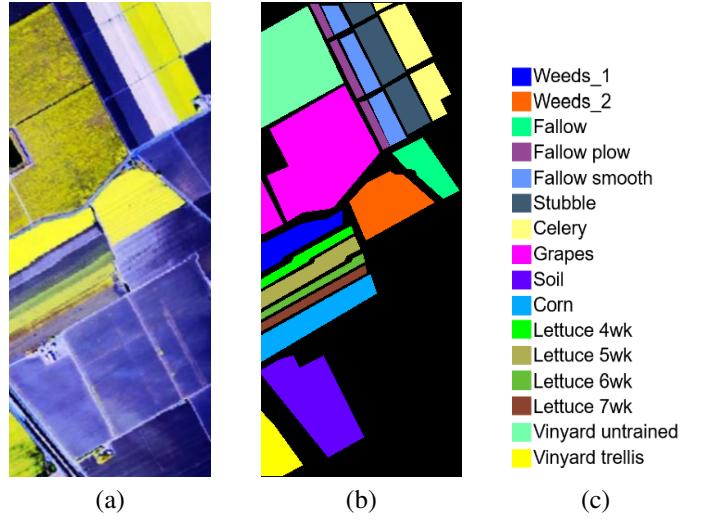


Fig. 9. The Salinas dataset. (a) Three-band false color composite. (b) Ground truth map (c) Legend

TABLE IV
THE NUMBER OF TRAINING SAMPLES AND TEST SAMPLES FOR THE SALINAS DATASET

Class	Class name	#Training	#Test	#Total
C1	Brocoli_green_weeds_1	200	1809	2009
C2	Brocoli_green_weeds_2	200	3526	3726
C3	Fallow	200	1776	1976
C4	Fallow_rough_plow	200	1194	1394
C5	Fallow_smooth	200	2478	2678
C6	Stubble	200	3759	3959
C7	Celery	200	3379	3579
C8	Grapes_untrained	200	11071	11271
C9	Soil_vinyard_develop	200	6003	6203
C10	Corn_senesced_green_weeds	200	3078	3278
C11	Lettuce_romaine_4wk	200	868	1068
C12	Lettuce_romaine_5wk	200	1727	1927
C13	Lettuce_romaine_6wk	200	716	916
C14	Lettuce_romaine_7wk	200	870	1070
C15	Vinyard_untrained	200	7068	7268
C16	Vinyard_vertical_trellis	200	1607	1807
Total	-	3200	50929	54129

(f) (DFFN, 3D-GAN and FreeNet) are better than those in Fig. 10 (a)-(c). Although DFFN and FreeNet obtain a similar accuracy, as shown in Table. V, the result of FreeNet contains less noise in the classification map, achieving a better visual performance. This suggests that the global spatial information is important for HSI classification. The results of the CNN-based methods show the over-smoothing problem, whereas, surprisingly, the SVM method can obtain sharper edges. We speculate that this is because these methods model the spatial context information. This causes the label of a pixel to be not only dependent on the center pixel, but also the neighboring pixels. Thus, the pixels near the edges of objects usually have different labels but a highly similar spatial context, which

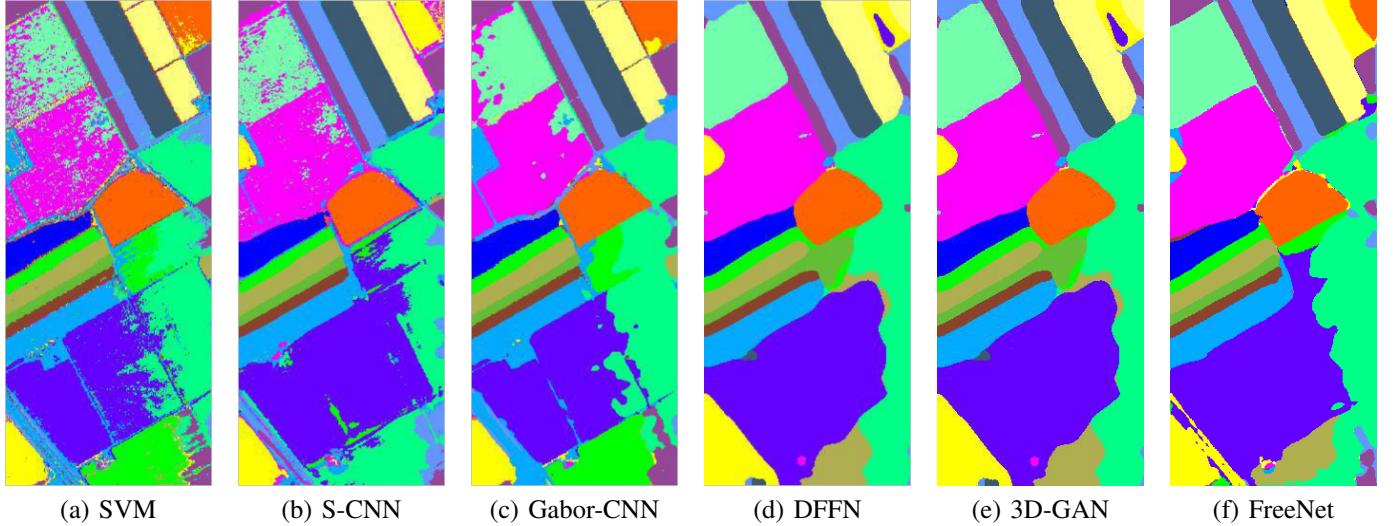


Fig. 10. Visualization of the classification maps for the Salinas dataset. (a) SVM. (b) S-CNN. (c) Gabor-CNN. (d) DFFN. (e) 3D-GAN. (f) FreeNet.

results in the over-smoothing problem.

TABLE V
THE CLASSIFICATION RESULTS OF SVM [8], S-CNN[38], GABOR-CNN [39], DFFN [40], 3D-GAN [51] AND FREEFNET ON THE SALINAS DATASET

Class	Patch-based					Patch-free
	SVM	S-CNN	Gabor-CNN	DFFN	3D-GAN	
C1	99.61	100	100	99.99	75.35	100
C2	99.69	97.75	88.06	99.94	98.49	100
C3	99.56	98.88	99.25	100	100	100
C4	99.41	100	100	100	99.76	99.92
C5	98.72	99.93	98.88	99.11	100	99.11
C6	99.77	89.48	100	99.95	99.97	100
C7	99.52	99.30	98.94	99.43	99.45	100
C8	76.74	98.69	99.48	99.56	98.30	99.94
C9	99.37	99.34	97.27	100	99.95	100
C10	95.13	100	99.21	99.79	99.79	99.77
C11	99.32	99.93	100	99.48	100	100
C12	99.70	99.89	100	99.84	100	100
C13	99.15	88.62	100	99.96	100	100
C14	98.38	88.72	99.79	99.95	100	100
C15	75.56	90.62	94.31	99.45	99.52	99.99
C16	99.23	99.96	93.38	99.96	90.25	99.88
OA(%)	90.92	97.62	97.63	99.71	98.22	99.92
AA(%)	96.18	96.94	98.04	99.78	97.75	99.91
Kappa	0.8985	0.9510	0.9734	0.9967	0.9793	0.9991

Table. V lists the accuracies of the patch-based methods and the patch-free method, where it can be seen that FreeNet performs much better than most of the patch-based methods in terms of OA, AA, and Kappa. Compared to the state-of-the-art patch-based method of DFFN, FreeNet obtains a slightly higher OA of 99.92%, exceeding DFFN by 0.2%. The proposed FreeNet achieves the highest accuracy among all the methods, slightly surpassing DFFN in all of the metrics. FreeNet benefits from the proposed semantic-spatial feature fusion module (lateral based SSF), which is based on residual learning. Furthermore, FreeNet fuses the features from the

shallow layer in the encoder and the deep layer in the decoder to achieve the fusion of semantic information and spatial details. It is interesting that the lateral based SSF in FreeNet and the multi-layer fusion in DFFN both belong to cross-layer feature fusion. This indicates that cross-layer feature fusion is an important component of HSI classification.

D. Experiment 3: CASI University of Houston Dataset

The proposed patch-free method achieved saturation on the ROSIS-03 Pavia University dataset and the Salinas dataset. Although the proposed FreeNet surpasses the state-of-the-art methods on these two simple benchmark datasets, the performance difference is fairly small due to the saturation of the accuracy. Therefore, we chose the CASI University of Houston dataset for Experiment 3, which is a relatively difficult benchmark dataset, to clearly present the performance difference.

The CASI University of Houston dataset was published as part of the 2013 IEEE Geoscience and Remote Sensing Society (GRSS) data fusion contest. The HSI has 349×1905 pixels with 144 spectral bands and a spatial resolution of 2.5 m, the wavelength of which ranges from 0.38 to 1.05 μm .

Fig. 11 (a) shows the false-color composite of the hyperspectral image. The dataset provides a ground-truth map of 15 classes, as shown in Fig. 11 (b) and (c). The corresponding legend is shown in Fig. 11 (d). Table. VI lists the number of training and test samples per class. Differing from the first two datasets, the CASI University of Houston dataset provides officially predefined training and test samples. Thus, the results obtained with this benchmark dataset can be considered as more reliable and stable.

Fig. 12 shows the classification map of the compared methods for the visual performance estimation. It can be clearly observed that the maps in Fig. 12 (c)-(f) are clearer and contain less noise than those in Fig. 12 (a)-(b). For the Road, Highway, and Railway classes, these three classes in the classification map of FreeNet show better connectivity. Meanwhile, the accuracy for these three classes is higher than

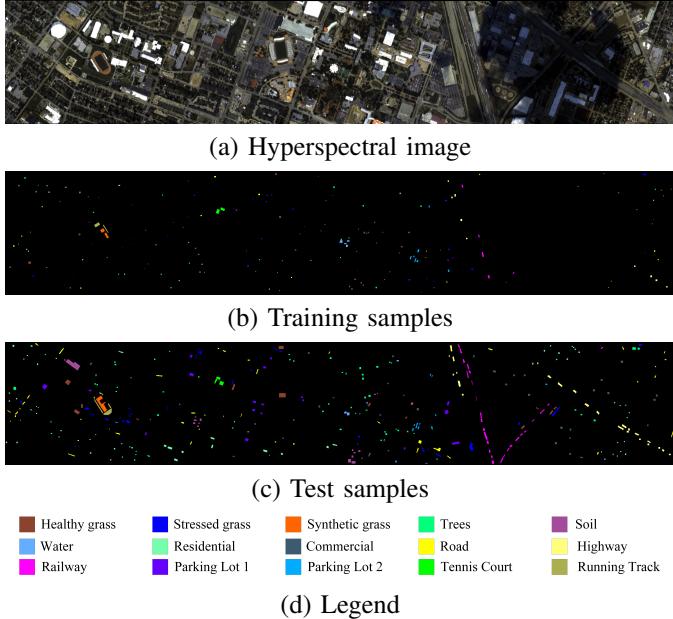


Fig. 11. The CASI University of Houston data set. (a) Color composite representation of the hyperspectral data using bands of 70, 50, and 20, as red, green, and blue, respectively; (b) Training samples; (c) Test samples; (d) Legend

TABLE VI
THE NUMBERS OF TRAINING SAMPLES AND TEST SAMPLES FOR THE
CASI UNIVERSITY OF HOUSTON DATASET

Class	Class name	#Training	#Test	#Total
C1	Grass–Healthy	198	1053	1251
C2	Grass–Stressed	190	1064	1254
C3	Grass–Synthetic	192	505	697
C4	Tree	188	1056	1244
C5	Soil	186	1056	1242
C6	Water	182	143	325
C7	Residential	196	1072	1268
C8	Commercial	191	1053	1244
C9	Road	193	1059	1252
C10	Highway	191	1036	1227
C11	Railway	181	1054	1235
C12	Parking Lot 1	192	1041	1234
C13	Parking Lot 2	184	285	469
C14	Tennis Court	181	247	428
C15	Running Track	187	473	660
Total	-	2832	12179	15011

for the other methods as shown in Table. VII C9-C11. We speculate that the increased spatial context of these classes enhances the discriminative ability for each class.

Table. VII lists the classification results obtained on this dataset. FreeNet achieves a state-of-the-art result and achieves a $\sim 2\%$ improvement over the best result (DFFN) of the patch-based methods. This suggests that the patch-free global learning framework performs even better than the patch-based local learning framework. FreeNet benefits from the increased

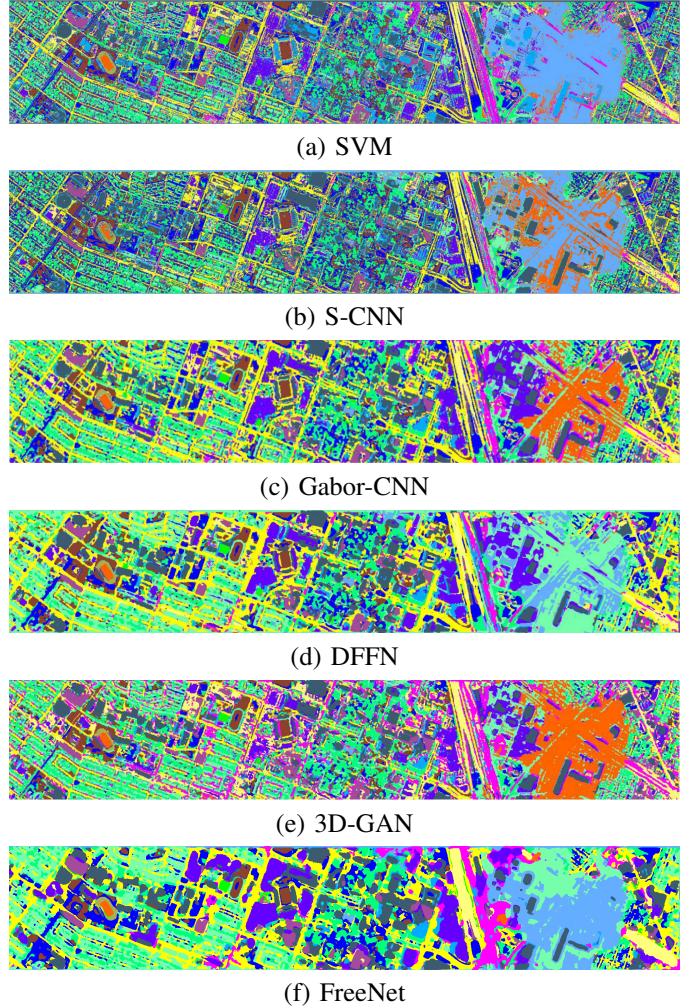


Fig. 12. Visualization of the classification maps for the CASI University of Houston dataset. (a) SVM. (b) S-CNN. (c) Gabor-CNN. (d) DFFN. (e) 3D-GAN. (f) FreeNet.

global spatial information, including the global spatial context and spatial details. By replacing the patch with a larger receptive field, the model under the patch-free global learning framework is able to obtain more global spatial context, which boosts the classification performance. FreeNet makes full use of the global spatial context via the spectral attention module and the lateral connection based SSF. The spectral attention module utilizes the global spatial context embedding vector to re-weight the feature maps for modeling the interdependencies between feature maps, which aims to estimate the importance of the different feature maps by the global spatial context. This benefits the classification of HSIs with redundant spectral information. Furthermore, the lateral connection based SSF leverages the finer spatial detail features to refine semantically stronger but spatially coarser deep features, which are aggregated to enhanced features with finer spatial details and stronger semantic information.

The accuracy obtained on this dataset is clearly lower than the accuracies obtained on the other two datasets using the same method. To explore the reason for this, we drew the confusion matrix for FreeNet on this dataset for a quantitative

TABLE VII
THE CLASSIFICATION RESULTS OF SVM [8], S-CNN[38], GABOR-CNN [39], DFFN [40], 3D-GAN [51] AND FREEFNET ON THE CASI UNIVERSITY OF HOUSTON DATASET.

Class	Patch-based					Patch-free
	SVM	S-CNN	Gabor-CNN	DFFN	3D-GAN	FreeNet
Accuracy(%)	C1	82.05	83.00	82.30	77.41	81.58
	C2	80.55	83.27	84.24	81.39	79.74
	C3	100	98.66	95.61	94.59	97.42
	C4	92.52	93.50	92.39	87.82	93.36
	C5	98.11	96.91	99.80	96.05	99.71
	C6	95.10	94.12	96.47	96.15	95.08
	C7	75.00	79.95	84.58	80.25	89.90
	C8	40.17	68.19	73.09	77.78	70.52
	C9	74.88	76.39	78.14	84.66	54.89
	C10	51.64	48.10	58.01	64.63	49.89
	C11	78.37	74.64	73.35	88.61	77.36
	C12	68.40	85.49	86.74	98.57	60.46
	C13	69.47	88.58	91.16	83.09	81.71
	C14	100	99.19	100	99.72	95.14
	C15	98.10	96.40	77.73	81.90	65.35
OA(%)	76.88	82.34	84.32	84.56	78.16	86.61
AA(%)	80.29	84.43	84.17	86.18	79.98	88.44
Kappa	0.7513	0.8052	0.8114	0.8328	0.7616	0.8555

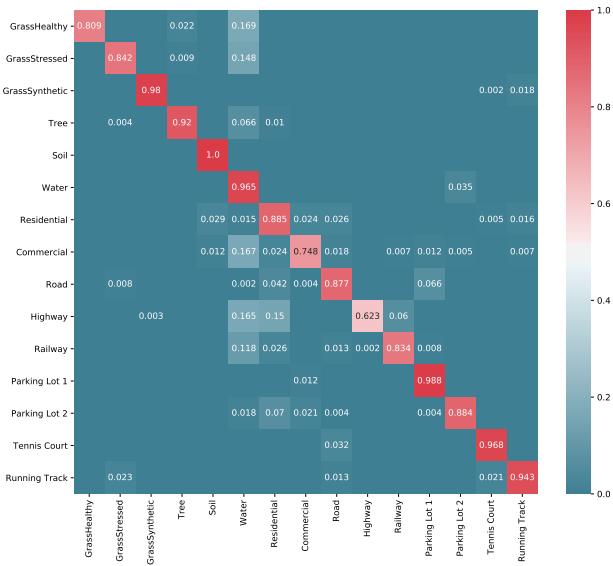


Fig. 13. Confusion matrix for FreeNet on the CASI University of Houston dataset.

analysis, as shown in Fig. 13. We can see that the other categories tend to be wrongly classified as Water. For example, a large amount of GrassHealthy, Commercial, and Railway pixels are wrongly classified as Water. This is because these three categories are distributed in the shadow area of the HSI, as shown in Fig. 11 (a). The shadow significantly influences the spectral information, which misleads the classification. Qualitatively, from the visualization result (Fig. 12 (e)), the shadow area is occupied by water. This indicates that the spectral information is sensitive to the observation conditions, which have a great impact on the qualitative and quantitative

performance of HSI classification.

V. FPGA SENSITIVITY ANALYSIS

To clearly understand the effectiveness of each component in the proposed FreeNet under the patch-free global learning framework, we conducted extensive module analysis experiments. All the component analysis experiments were performed on the CASI University of Houston dataset as this dataset has official training and test samples. We chose the CASI University of Houston dataset for more stable, reliable and reproducible analysis results. The baseline methods shown in Table.VIII (a) are encoder-decoder FCNs with $\beta = 0.75$ and 1.0, respectively, which were trained by directly using all the training samples, without any sampling strategy.

A. The GS² Sampling Strategy

Table. VIII (b) presents the results of the baseline methods with the GS² sampling strategy. The results indicate that the GS² sampling strategy improves the OA for both $\beta = 0.75$ (from 15.23% to 65.12%) and $\beta = 1.0$ (from 12.49% to 65.16%). A model with an OA of ~10% on this dataset means that this model fails to converge. The GS² sampling strategy effectively addresses this issue for the baseline encoder-decoder FCN. The GS² sampling strategy splits the original training samples into many mini-batch training samples with the same spatial size to obtain diverse gradients and partially supervised signals. These diverse gradients and partially supervised signals make it easier to skip local minimum points. This suggests that it is very important for the end-to-end trainable FCNs used in HSI classification to obtain more diverse gradients.

The hyperparameter α is introduced in the GS² sampling strategy to control the number of samples in the mini-batch per class. The classification results of FreeNet ($\beta = 1.0$) with different value of α from 10 to 200 with an interval of 10 are shown in Fig. 14. We can observe that the proposed FreeNet shows a stable performance when α is set to a value within 30% of the number of total training samples. This indicates the robustness of the GS² sampling strategy with respect to α when the stochasticity is sufficient. When α is set to a value larger than 55% of the number of total training samples, the classification accuracy of FreeNet drops rapidly, to even lower than the SVM baseline. The reason for this is that α indirectly controls the stochasticity of the sampling. A smaller α means greater stochasticity of the sampling. Meanwhile, the stochasticity of the sampling directly influences the diversity of the gradients. Therefore, FreeNet with a smaller α always obtains a higher accuracy, while a larger α brings a worse performance.

B. Lateral Connection Based SSF

Lateral connection based SSF is described in Section III-C4. Table. VIII (c) presents the effectiveness of the lateral connection based SSF. When applying FreeNet without the lateral connection (LC) and spectral attention (SA) modules (Table. VIII (b)), the classification performance shows a significant reduction. The addition of the lateral connection and

TABLE VIII

HSI CLASSIFICATION RESULTS EVALUATED ON THE CASI UNIVERSITY OF HOUSTON DATASET. STARTING FROM OUR ENCODER-DECODER BASELINE, THE GS² SAMPLING STRATEGY, LATERAL CONNECTION AND SPECTRAL ATTENTION ARE GRADUALLY ADDED IN FREE NET FOR THE MODULE ANALYSIS. SA DENOTES SPECTRAL ATTENTION AND LC DENOTES LATERAL CONNECTION BASED SSF.

Compression factor	Method	GS ² sampling strategy	Lateral connection	Spectral attention	OA	AA	Kappa
$\beta = 0.75$	(a) Baseline	-	-	-	15.23	19.54	0.0972
	(b) FreeNet w/o LC and SA	✓	-	-	65.12	67.45	0.6225
	(c) FreeNet w/o SA	✓	✓	-	84.23	85.61	0.8293
	(d) FreeNet	✓	✓	✓	85.49	86.64	0.8423
$\beta = 1.0$	(a) Baseline	-	-	-	12.49	13.89	0.0657
	(b) FreeNet w/o LC and SA	✓	-	-	65.16	65.2	0.6212
	(c) FreeNet w/o SA	✓	✓	-	84.91	86.25	0.8363
	(d) FreeNet	✓	✓	✓	86.61	88.44	0.8555

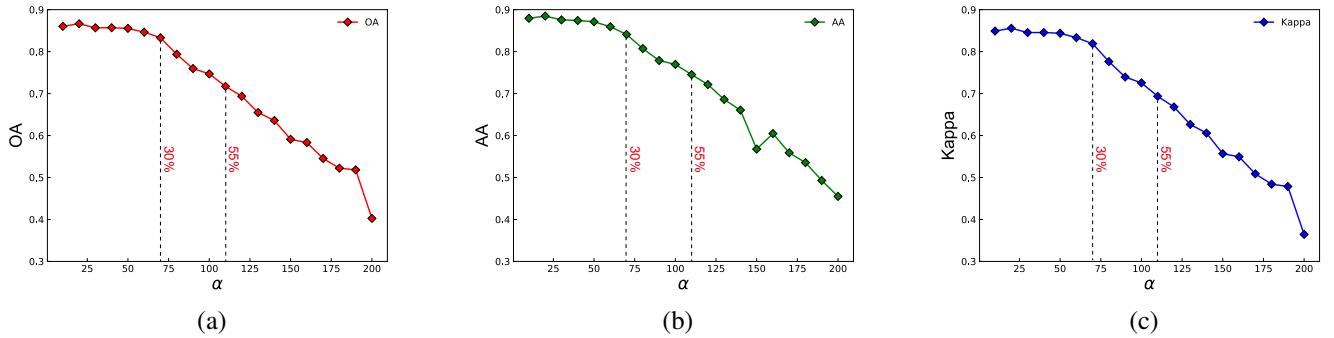


Fig. 14. Sensitivity of the mini-batch size per class (α) in the GS² sampling strategy on the CASI University of Houston dataset. (a) The impact on the OA of different α settings. (b) the impact on the AA of different α settings. (c) the impact on the Kappa of different α settings.

spectral attention modules to FreeNet ($\beta = 0.75$) results in an OA improvement from 65.12% to 84.23%, and The addition of the lateral connection and spectral attention modules to FreeNet ($\beta = 1.0$) results in an OA improvement from 65.16% to 84.91%, achieving a similar performance to the state-of-the-art patch-based HSI classifiers. With the help of the lateral connection based SSF, the spatial detail features of the shallow convolutional layers can be passed on to the decoder to progressively refine the spatial detail of the deep semantic features, thus obtaining spatially finer and semantically stronger fused features. Meanwhile, the aggregation function of pointwise addition can alleviate the gradient vanishing problem, making the optimization easier. This suggests that lateral connection based SSF is important for HSI classification when using an encoder-decoder architecture under the patch-free global learning framework.

C. Spectral Attention

Table. VIII (d) presents the effectiveness of the spectral attention module. Based on FreeNet without the spectral attention module (Table. VIII (c)), the spectral attention module brings an additional improvement to FreeNet ($\beta = 0.75$) (84.23% to 85.49) and FreeNet ($\beta = 1.0$) (84.91% to 86.61%). The spectral attention module models the interdependencies of the feature maps in the encoder of FreeNet via the global spatial context, boosting the classification performance. This indicates that it is valuable to further exploit the raw redundant

spectral features guided by spatial context for HSI classification.

D. Model Complexity and Inference Speed Analysis

For a fair comparison of the model complexity and inference speed between the patch-based and patch-free methods, we directly used the encoder of FreeNet followed by a 1×1 convolutional layer as the classifier, to represent the patch-based method. Three variants ($\beta = 0.5, 0.75$ and 1.0) of FreeNet and the corresponding encoders were used as a comparison. During the inference, the batch size of the patch-based methods was set to 1024 for parallel computation. The image used for the benchmarking was from the CASI University of Houston dataset, which was converted to a 3-D float32 tensor of shape [349, 1905, 144].

We adopted the number of parameters (# params) to measure the model complexity and the giga floating-point operations per second (GFLOPs) to measure the theoretical computational overhead. Meanwhile, the GPU time cost was used to measure the actual efficiency of the models.

Table. IX lists the measured results. The results suggest that the patch-free methods are much faster than the patch-based methods, in both theory and practice, by avoiding redundant computation on the overlapping area. Note that the practical speedup ratio (~ 560) is always larger than the theoretical speedup ratio (~ 480) because we ignore the influence of the parallel computation. Although the patch-based methods

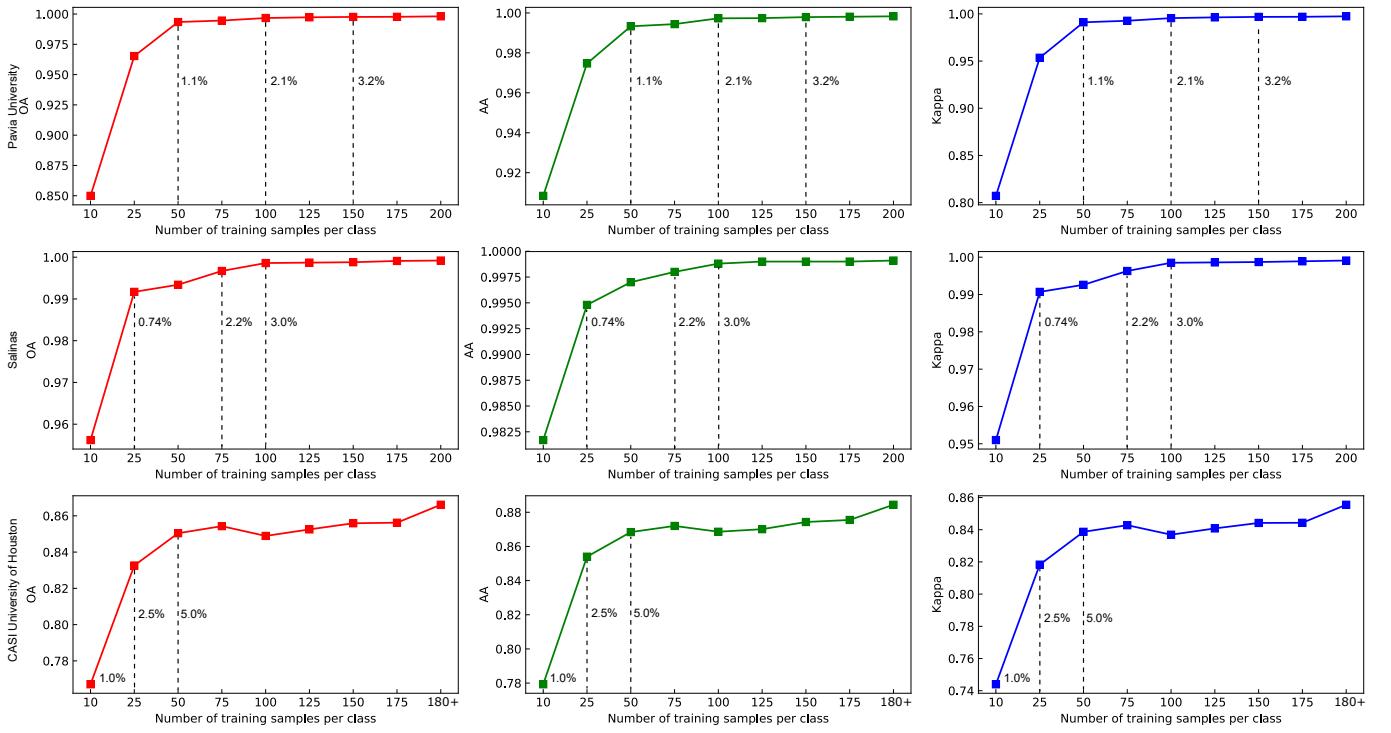


Fig. 15. Performance versus the number of training samples per class. These three rows represent the performance on Pavia University, Salinas and CASI University of Houston dataset, respectively. The percentage near the dashed line is the percentage of training samples.

TABLE IX

THE MODEL COMPLEXITY AND INFERENCE SPEED OF PATCH-BASED AND PATCH-FREE METHODS. ENCODER (β) IS SIMPLY IMPLEMENTED BY ENCODER IN CORRESPONDING FREE NET WITH SAME β AND A 1×1 CONVOLUTIONAL LAYER.

Methods	Model complexity		Inference speed	
	#Params (M)	GFLOPs	GPU (s)	
<i>Patch-based</i>				
Encoder ($\beta = 0.5$)	0.554	53644.8	55.221	
Encoder ($\beta = 0.75$)	1.191	107289.6	69.319	
Encoder ($\beta = 1.0$)	2.070	167640.0	84.106	
<i>Patch-free</i>				
FreeNet ($\beta = 0.5$)	0.724	112.37	0.094	
FreeNet ($\beta = 0.75$)	1.575	220.82	0.122	
FreeNet ($\beta = 1.0$)	2.749	364.11	0.146	

have fewer parameters than the patch-free methods, the actual computation of the patch-based methods is slower than that of the patch-free methods. Therefore, for real-time applications, such as HSI classification on unmanned aerial vehicle (UAV) or satellite imagery, the patch-free method is a better proposal than the patch-based method.

We also list the detailed running time costs of FreeNet for each dataset in Table X. It can be seen that the inference time cost is much smaller than the training time cost for each HSI. This suggests that FreeNet is suitable for the application scenarios which allow offline learning and online inference, such as training FreeNet on ground station data and performing

TABLE X
DETAILED RUNNING TIME COSTS OF FREE NET

Dataset	Training time (s)	Test time (s)
ROSIS-03 Pavia University	202	0.039
Salinas	158	0.030
CASI University of Houston	523	0.146

inference on UAV or satellite data.

E. The Impact of the Number of Training Samples

To study the impact of the number of training samples on FreeNet, we conducted extensive experiments on the Pavia University, Salinas, and CASI University of Houston benchmark datasets. The results are plotted in Fig. 15. Overall, reducing the training samples results in a drop in the performance of FreeNet. For the Pavia University dataset and Salinas dataset, the decreases in OA, AA, and Kappa are relatively small when using 1% ~ 3% of the labeled samples. It suggests that FreeNet is also robust, only using limited training sample number. In order to further explore the effectiveness of FreeNet, we also trained FreeNet using only 10 samples per class. Compared with the Salinas dataset, the decrease in performance on the Pavia dataset is much more significant, at 14.82% of OA, 8.99% of AA, and 0.19 of Kappa. Meanwhile the decrease in accuracy for the Salinas dataset is 4.3% of OA, 1.74% of AA, and 0.048 of Kappa. This indicates that classification on the Salinas dataset with FreeNet is easier than for the Pavia University dataset. We speculate

that FreeNet benefits from more spectral information via the spectral attention module. This is because the spatial resolution of these two datasets is high, so that the contribution of the spatial information to the performance has achieved saturation. However the Salinas dataset has more bands, which may be the core reason for the easier classification. For the CASI University of Houston dataset, the performance is steady until the training samples are reduced to 5% of the labeled samples. This dataset has fewer labeled samples than the other two datasets. Thus, the entries with a high percentage still have minimal training samples, which causes the performance to drop rapidly.

VI. CONCLUSION

In this paper, we have proposed a fast patch-free global learning (FPGA) framework, pushing HSI classification further on both speed and accuracy. In the FPGA framework, the GS² sampling strategy is proposed to ensure encoder-decoder based FCN training convergence by transforming the entire training samples into a stochastic sequence of class-stratified samples, to obtain stable and diverse gradients. After ensuring the convergence of the training, FreeNet is proposed, which is a simple and unified encoder-decoder based FCN. FreeNet directly inputs the entire HSI without requiring any dimension reduction and outputs the classification map. Therefore, FreeNet is a fully end-to-end trainable HSI classifier that does not require dimension reduction or any post-processing technology. FreeNet avoids the redundant computation on the overlapping areas between patches, which significantly boosts its inference speed. To maximize the exploitation of the global spatial context and details, a spectral attention module and lateral connection based SSF are proposed. The spectral attention module models the interdependencies of the feature maps guided by the global spatial context to effectively boost the performance of FreeNet. The lateral connection based SSF progressively refines the semantic features with the global spatial detail of the features from the shallow layers. Meanwhile, lateral connection based SSF follows the residual learning approach to fuse the features by pointwise addition, which can alleviate the gradient vanishing problem, thereby, significantly improving the performance of FreeNet.

In the future, we will further explore memory-efficient HSI classifiers, which will be important for the real-time classification of HSIs from satellite and airborne platforms. We hope that the proposed method will serve as a strong baseline and aid future research in HSI classification.

ACKNOWLEDGEMENTS

The authors would like to thank the Editor, Associate Editor, and anonymous reviewers for their helpful comments and suggestions that improved this article.

REFERENCES

- [1] P. Ghamisi, E. Maggiori, S. Li, R. Souza, Y. Tarabla, G. Moser, A. De Giorgi, L. Fang, Y. Chen, M. Chi *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 Geoscience and Remote Sensing Magazine*, vol. 6, no. 3, pp. 10–43, 2018.
- [2] Y. Zhong, X. Wang, Y. Xu, S. Wang, T. Jia, X. Hu, J. Zhao, L. Wei, and L. Zhang, “Mini-uav-borne hyperspectral remote sensing: From observation and processing to applications,” *IEEE Geoscience and Remote Sensing Magazine*, vol. 6, no. 4, pp. 46–62, 2018.
- [3] S. Li, W. Song, L. Fang, Y. Chen, P. Ghamisi, and J. A. Benediktsson, “Deep learning for hyperspectral image classification: An overview,” *IEEE Transactions on Geoscience and Remote Sensing*, 2019, doi: 10.1109/TGRS.2019.2907932.
- [4] M. Govender, K. Chetty, and H. Bulcock, “A review of hyperspectral remote sensing and its application in vegetation and water resource studies,” *Water Sa*, vol. 33, no. 2, 2007.
- [5] E. Adam, O. Mutanga, and D. Rugege, “Multispectral and hyperspectral remote sensing for identification and mapping of wetland vegetation: a review,” *Wetlands Ecology and Management*, vol. 18, no. 3, pp. 281–296, 2010.
- [6] B. Koch, “Status and future of laser scanning, synthetic aperture radar and hyperspectral remote sensing data for forest biomass assessment,” *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 65, no. 6, pp. 581–590, 2010.
- [7] G. Camps-Valls, D. Tuia, L. Bruzzone, and J. A. Benediktsson, “Advances in hyperspectral image classification: Earth monitoring with statistical learning methods,” *IEEE Signal Processing Magazine*, vol. 31, no. 1, pp. 45–54, 2014.
- [8] F. Melgani and L. Bruzzone, “Classification of hyperspectral remote sensing images with support vector machines,” *IEEE Transactions on geoscience and remote sensing*, vol. 42, no. 8, pp. 1778–1790, 2004.
- [9] P. O. Gislason, J. A. Benediktsson, and J. R. Sveinsson, “Random forests for land cover classification,” *Pattern Recognition Letters*, vol. 27, no. 4, pp. 294–300, 2006.
- [10] J. Xia, P. Du, X. He, and J. Chanussot, “Hyperspectral remote sensing image classification based on rotation forest,” *IEEE Geoscience and Remote Sensing Letters*, vol. 11, no. 1, pp. 239–243, 2013.
- [11] T. Rainforth and F. Wood, “Canonical correlation forests,” *arXiv preprint arXiv:1507.05444*, 2015.
- [12] J. Xia, N. Yokoya, and A. Iwasaki, “Hyperspectral image classification with canonical correlation forests,” *IEEE Transactions on Geoscience and Remote Sensing*, vol. 55, no. 1, pp. 421–431, 2016.
- [13] B. Krishnapuram, L. Carin, M. A. Figueiredo, and A. J. Hartemink, “Sparse multinomial logistic regression: Fast algorithms and generalization bounds,” *IEEE transactions on pattern analysis and machine intelligence*, vol. 27, no. 6, pp. 957–968, 2005.
- [14] M. Fauvel, J. Chanussot, J. A. Benediktsson, and J. R. Sveinsson, “Spectral and spatial classification of hyperspectral data using svms and morphological profiles,” in *2007 IEEE International Geoscience and Remote*

- Sensing Symposium.* IEEE, 2007, pp. 4834–4837.
- [15] Y. Tarabalka, J. A. Benediktsson, and J. Chanussot, “Spectral–spatial classification of hyperspectral imagery based on partitional clustering techniques,” *IEEE Transactions on Geoscience and Remote Sensing*, vol. 47, no. 8, pp. 2973–2987, 2009.
 - [16] J. Li, J. M. Bioucas-Dias, and A. Plaza, “Spectral–spatial classification of hyperspectral data using loopy belief propagation and active learning,” *IEEE Transactions on Geoscience and remote sensing*, vol. 51, no. 2, pp. 844–856, 2012.
 - [17] M. Pesaresi, A. Gerhardinger, and F. Kayitakire, “A robust built-up area presence index by anisotropic rotation-invariant textural measure,” *IEEE Journal of selected topics in applied earth observations and remote sensing*, vol. 1, no. 3, pp. 180–192, 2008.
 - [18] C. Zhu and X. Yang, “Study of remote sensing image texture analysis and classification using wavelet,” *International Journal of Remote Sensing*, vol. 19, no. 16, pp. 3197–3203, 1998.
 - [19] W. Li and Q. Du, “Gabor-filtering-based nearest regularized subspace for hyperspectral image classification,” *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 7, no. 4, pp. 1012–1022, 2014.
 - [20] J. A. Benediktsson, J. A. Palmason, and J. R. Sveinsson, “Classification of hyperspectral data from urban areas based on extended morphological profiles,” *IEEE Transactions on Geoscience and Remote Sensing*, vol. 43, no. 3, pp. 480–491, 2005.
 - [21] J. Li, P. R. Marpu, A. Plaza, J. M. Bioucas-Dias, and J. A. Benediktsson, “Generalized composite kernel framework for hyperspectral image classification,” *IEEE transactions on geoscience and remote sensing*, vol. 51, no. 9, pp. 4816–4829, 2013.
 - [22] Y. Chen, Z. Lin, X. Zhao, G. Wang, and Y. Gu, “Deep learning-based classification of hyperspectral data,” *IEEE Journal of Selected topics in applied earth observations and remote sensing*, vol. 7, no. 6, pp. 2094–2107, 2014.
 - [23] Y. Xu, L. Zhang, B. Du, and F. Zhang, “Spectral–spatial unified networks for hyperspectral image classification,” *IEEE Transactions on Geoscience and Remote Sensing*, no. 99, pp. 1–17, 2018.
 - [24] Y. LeCun, Y. Bengio, and G. Hinton, “Deep learning,” *nature*, vol. 521, no. 7553, p. 436, 2015.
 - [25] J. Yue, W. Zhao, S. Mao, and H. Liu, “Spectral–spatial classification of hyperspectral images using deep convolutional neural networks,” *Remote Sensing Letters*, vol. 6, no. 6, pp. 468–477, 2015.
 - [26] W. Hu, Y. Huang, L. Wei, F. Zhang, and H. Li, “Deep convolutional neural networks for hyperspectral image classification,” *Journal of Sensors*, vol. 2015, 2015.
 - [27] 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 Transactions on Geoscience and Remote Sensing*, vol. 54, no. 10, pp. 6232–6251, 2016.
 - [28] S. Yu, S. Jia, and C. Xu, “Convolutional neural networks for hyperspectral image classification,” *Neurocomputing*, vol. 219, pp. 88–98, 2017.
 - [29] M. Paoletti, J. Haut, J. Plaza, and A. Plaza, “A new deep convolutional neural network for fast hyperspectral image classification,” *ISPRS journal of photogrammetry and remote sensing*, vol. 145, pp. 120–147, 2018.
 - [30] J. Zhu, L. Fang, and P. Ghamisi, “Deformable convolutional neural networks for hyperspectral image classification,” *IEEE Geoscience and Remote Sensing Letters*, vol. 15, no. 8, pp. 1254–1258, 2018.
 - [31] Z. Gong, P. Zhong, Y. Yu, W. Hu, and S. Li, “A cnn with multiscale convolution and diversified metric for hyperspectral image classification,” *IEEE Transactions on Geoscience and Remote Sensing*, vol. 57, no. 6, pp. 3599–3618, 2019, doi: 10.1109/TGRS.2018.2886022.
 - [32] R. Hang, Q. Liu, D. Hong, and P. Ghamisi, “Cascaded recurrent neural networks for hyperspectral image classification,” *IEEE Transactions on Geoscience and Remote Sensing*, 2019, doi: 10.1109/TGRS.2019.2899129.
 - [33] Y. Chen, X. Zhao, and X. Jia, “Spectral–spatial classification of hyperspectral data based on deep belief network,” *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 8, no. 6, pp. 2381–2392, 2015.
 - [34] W. Zhao and S. Du, “Spectral–spatial feature extraction for hyperspectral image classification: A dimension reduction and deep learning approach,” *IEEE Transactions on Geoscience and Remote Sensing*, vol. 54, no. 8, pp. 4544–4554, 2016.
 - [35] P. Zhou, J. Han, G. Cheng, and B. Zhang, “Learning compact and discriminative stacked autoencoder for hyperspectral image classification,” *IEEE Transactions on Geoscience and Remote Sensing*, 2019, doi: 10.1109/TGRS.2019.2893180.
 - [36] H. Lee and H. Kwon, “Contextual deep cnn based hyperspectral classification,” in *2016 IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*. IEEE, 2016, pp. 3322–3325.
 - [37] W. Li, G. Wu, F. Zhang, and Q. Du, “Hyperspectral image classification using deep pixel-pair features,” *IEEE Transactions on Geoscience and Remote Sensing*, vol. 55, no. 2, pp. 844–853, 2016.
 - [38] B. Liu, X. Yu, P. Zhang, A. Yu, Q. Fu, and X. Wei, “Supervised deep feature extraction for hyperspectral image classification,” *IEEE Transactions on Geoscience and Remote Sensing*, vol. 56, no. 4, pp. 1909–1921, 2017.
 - [39] Y. Chen, L. Zhu, P. Ghamisi, X. Jia, G. Li, and L. Tang, “Hyperspectral images classification with gabor filtering and convolutional neural network,” *IEEE Geoscience and Remote Sensing Letters*, vol. 14, no. 12, pp. 2355–2359, 2017.
 - [40] W. Song, S. Li, L. Fang, and T. Lu, “Hyperspectral image classification with deep feature fusion network,” *IEEE Transactions on Geoscience and Remote Sensing*, vol. 56, no. 6, pp. 3173–3184, 2018.
 - [41] L. Jiao, M. Liang, H. Chen, S. Yang, H. Liu, and X. Cao, “Deep fully convolutional network-based spatial distribution prediction for hyperspectral image classification,”

- IEEE Transactions on Geoscience and Remote Sensing*, vol. 55, no. 10, pp. 5585–5599, 2017.
- [42] J. Long, E. Shelhamer, and T. Darrell, “Fully convolutional networks for semantic segmentation,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 3431–3440.
 - [43] J. Li, X. Zhao, Y. Li, Q. Du, B. Xi, and J. Hu, “Classification of hyperspectral imagery using a new fully convolutional neural network,” *IEEE Geoscience and Remote Sensing Letters*, vol. 15, no. 2, pp. 292–296, 2018.
 - [44] O. Ronneberger, P. Fischer, and T. Brox, “U-net: Convolutional networks for biomedical image segmentation,” in *International Conference on Medical image computing and computer-assisted intervention*. Springer, 2015, pp. 234–241.
 - [45] V. Badrinarayanan, A. Kendall, and R. Cipolla, “Segnet: A deep convolutional encoder-decoder architecture for image segmentation,” *IEEE transactions on pattern analysis and machine intelligence*, vol. 39, no. 12, pp. 2481–2495, 2017.
 - [46] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, “Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs,” *IEEE transactions on pattern analysis and machine intelligence*, vol. 40, no. 4, pp. 834–848, 2017.
 - [47] H. Robbins and S. Monro, “A stochastic approximation method,” *The annals of mathematical statistics*, pp. 400–407, 1951.
 - [48] Y. Wu and K. He, “Group normalization,” in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 3–19.
 - [49] J. Hu, L. Shen, and G. Sun, “Squeeze-and-excitation networks,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 7132–7141.
 - [50] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.
 - [51] L. Zhu, Y. Chen, P. Ghamisi, and J. A. Benediktsson, “Generative adversarial networks for hyperspectral image classification,” *IEEE Transactions on Geoscience and Remote Sensing*, vol. 56, no. 9, pp. 5046–5063, 2018.