Spectral Feature Fusion Networks with Dual Attention for Hyperspectral Image Classification

Xian Li, *Student Member, IEEE*, Mingli Ding, and Aleksandra Pižurica, *Senior Member, IEEE*

***Abstract*—Recent progress in spectral classification is largely attributed to the use of convolutional neural networks (CNN). While a variety of successful architectures have been proposed, they all extract spectral features from various portions of adjacent spectral bands. In this paper, we take a different approach and develop a deep spectral feature fusion method, which extracts both local and interlocal spectral features, capturing thus also the correlations among non-adjacent bands. To our knowledge, this is the first reported deep spectral feature fusion method. Our model is a two-stream architecture, where an intergroup and a groupwise spectral classifiers operate in parallel. The interlocal spectral correlation feature extraction is achieved elegantly, by reshaping the input spectral vectors to form the so-called non-adjacent spectral matrices. We introduce the concept of groupwise band convolution to enable efficient extraction of discriminative local features with multiple kernels adopting to the local spectral content. Another important contribution of this work is a novel dual-channel attention mechanism to identify the most informative spectral features. The model is trained in an end-to-end fashion with a joint loss. Experimental results on real data sets demonstrate excellent performance compared to the current state-of-the-art.**

***Index Terms*—Spectral feature fusion, attention mechanism, deep learning, hyperspectral image classification.**

I. Introduction

HYPERSPECTRAL image encompasses both rich spectral and spatial information, which offers great potentials for land cover identification [1]. Hence, a range of methods for hyperspectral data processing have been reported recently [2, 3]. Scene classification remains to be one of the most demanding tasks since it is a fundamental processing step in various fields [4–6].

Spectral-spatial classification typically outperforms spectral classification alone [7–10] due to the use of spatial context. However, one has to deal with how to select the input window size for different images [11], since they often present diverse and complex spectral and spatial features due to different spatial resolutions together with various land cover types [12]. Moreover, in hyperspectral image processing, pixel-sharing among the training and testing local regions (windows) is frequently encountered since the training and testing sets are taken from the same image. Unless the centers of the training and testing samples are displaced by at least the window size in either horizontal or vertical direction, the two regions will overlap as shown in Fig. 1. This pixel-sharing increases the classification accuracy in the tests and can thus hinder fair assessment of the methods, as indicated in [13]. Our motivation in this paper is to avoid the pixel-sharing completely while yielding state-of-the-art accuracy.

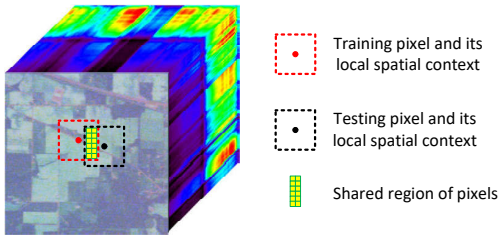

Fig. 1. An illustration of pixel-sharing between the training and testing data. The shared pixels aggravate with a larger window scale or more training data.

With spectral classification alone there is no pixel-sharing and no window scale selection of input data, as the input is simply a vector of pixel values at a given spatial location across all the bands. The main challenge of spectral classification is how to extract the discriminative spectral features from a high-dimensional spectral information under limited training data [14]. The related phenomena are referred to as the curse of dimensionality [15]. A range of spectral classification methods have been proposed [16], building on support vector machines [17], random forests [18], multinomial logistic regression [19] and neural networks [20]. Recent studies demonstrate the success of deep leaning in spectral feature extraction, using e.g., stacked auto-encoder [21], deep belief network [22], CNN [23–25], and recurrent neural network [26, 27]. The first two are fully connected networks that require much more learning parameters, tending to be overfitting due to lack of sufficient training data to fit them. CNN models can reduce hugely the number of learning parameters with the local-connection and shared-weight architecture. Recurrent neural network models handle hyperspectral pixels sequentially because of their powerful learning capability from sequential data.

Although the above described methods made great progress in spectral classification, their classification maps are still noisy due to various degradations including image noise, spectral variability and mixed pixels. A common approach is to employ spatial information in a post-processing stage. Training is done based on spectral information alone and the

This work was partially supported by the China Scholarship Council, and received funding from the AI program of the Flemish Government.

X. Li is with the School of Instrumentation Science and Engineering, Harbin Institute of Technology, 150001 Harbin, China, and also with the Department of Telecommunications and Information Processing, UGent-GAIM, Ghent University, 9000 Ghent, Belgium (e-mail: xianli0511@gmail.com).

M. Ding is with the School of Instrumentation Science and Engineering, Harbin Institute of Technology, 150001 Harbin, China (e-mail: dingml@hit.edu.cn).

A. Pižurica is with the Department of Telecommunications and Information Processing, UGent-GAIM, Ghent University, 9000 Ghent, Belgium (e-mail: Aleksandra.Pizurica@UGent.be)

spatial context is then incorporated posteriorly at the testing stage as refinement [28], typically via some voting strategies [25, 29]. However, the pixel-sharing problem still remains in these post-processing methods. A recent work [30] proposed an adaptive spectral-spatial voting strategy to refine the final label while excluding the training samples at the testing stage.

In this paper, we aim to improve the spectral feature learning capability for hyperspectral image classification. To achieve this, we propose a two-stream spectral feature fusion method based on 1D-CNN, which extracts simultaneously interlocal and discriminative local spectral features in parallel. The first spectral classifier aims to extract interlocal spectral correlation features from a non-adjacent spectral matrix that is reshaped from the input pixel vector. The second spectral classifier comprises an original groupwise band convolution to extract more discriminative local spectral features from groups of sub-bands. Moreover, we develop a novel dual-channel attention mechanism to further boost the spectral feature learning capability of the two classifiers. Then, the two parallel classifiers are integrated adaptively via a decision fusion. At the testing stage, we introduce a novel voting method, which avoids the pixel-sharing effectively and exploits both local and global spatial information posteriorly to make the final prediction.

The main contributions of this paper are as follows:

1) We propose a unified two-stream CNN-based spectral feature fusion method for hyperspectral image classification. We are not aware of any other reported works on spectral feature fusion of hyperspectral data based on deep learning. The main advantage of the proposed method is its powerful spectral feature learning capability, exhibited as a significantly improved performance compared to the state-of-the-art spectral classifiers.
2) We propose a novel intergroup spectral classifier and an original groupwise spectral classifier in a unified model, which extracts interlocal and discriminative local spectral correlation features simultaneously.
3) We develop a novel dual-channel attention method for improving the spectral feature learning capability of the two parallel classifiers based on non-local and global inter-channel correlations. This attention method can be also applied to other feature learning networks.
4) We introduce a decision fusion scheme and a joint loss to train the unified model in an end-to-end fashion. Moreover, we introduce a local and global majority voting method at the testing stage, to make use of spatial information posteriorly while avoiding the pixel-sharing.

The rest of this paper is organized as follows. Section II reviews the spectral classification and attention mechanism. Section III introduces the proposed method. Section IV evaluates the effectiveness of the proposed method on real hyperspectral data sets and Section V draws the conclusion.

II. Related Work

A. Spectral Feature Extraction and Classification

Recent comprehensive reviews on spectral classification include [16, 31]. For instance, support vector machines have been widely adopted since they were suited to separate the high-dimensional hyperspectral data with limited training samples [17]. Lately, extreme learning machines were introduced to increase their nonlinear representation power [32]. Some other advanced spectral classifiers, including random forests [18], neural networks [20] and logistic regression [19], have been proposed to solve various classification problems.

Recent spectral classifiers are often based on deep learning, using e.g., stacked auto-encoders (SAEs) [21, 33], deep belief networks (DBNs) [22], CNNs [23–25] and recurrent neural networks (RNNs) [26, 27]. The CNN models reduce the number of parameters compared to SAE and DBN due to local connections and shared weights. While these CNN models with the local connection mechanism extract local spectral features, they ignore non-adjacent spectral features [27]. From sequential perspective, RNNs can learn non-adjacent spectral features with a band grouping strategy [26] and with a two-stage RNN [27]. Representatives of these CNNs [23–25] and RNNs [26, 27] are chosen for comparison in Section IV-B.

Although the above described spectral classifiers demonstrated huge success, they typically employed a single stream network [23–27], and when employing 1D-CNN (e.g., in [23–25]) those extracted local spectral features only, and none within the attention mechanism. We instead introduce a two-stream spectral feature fusion method based on 1D-CNN, which extracts both interlocal and discriminative local spectral correlation features simultaneously. Moreover, we incorporate a novel attention mechanism into the two parallel streams to further boost their spectral feature leaning capability.

B. Attention Mechanism

Inspired by the human visual system to understand an image by concentrating on informative features, attention mechanism has been incorporated into deep learning to improve the feature learning efficiency [34]. Hu *et al.* [35] proposed a channel attention to learn the global inter-channel correlations from the global averaging spatial information and has drawn much attention. Subsequently, Fu *et al.* [36] introduced a dual-attention method, which consists of a spatial position attention and a channel attention, to learn the global spatial contextual and inter-channel correlation separately. Very recently, Wang *et al.* [37] proposed an efficient channel attention for deep CNNs to capture local inter-channel correlations.

In hyperspectral image processing, the attention mechanism was often utilized to boost the spectral-spatial feature extraction capability. For instance, Mou *et al.* [38] designed a spectral attention method to learn important spectral bands from global spatial information in spectral-spatial classification. Sun *et al.* [39] proposed a spectral-spatial attention network, which concentrated on learning features from homogeneous areas. Very recently, Zhu *et al.* [40] introduced a two-stage attention for a residual CNN model. In the first stage, spectral and spatial attentions in series were designed to emphasize informative spectral and spatial features respectively. The second stage was embedded a spectral-spatial attention into a residual convolutional block to facilitate the training process.

The aforementioned attention approaches were typically designed for 2D or 3D CNN models to learn the global inter-channel correlations [35, 36, 38–40] or local inter-channel

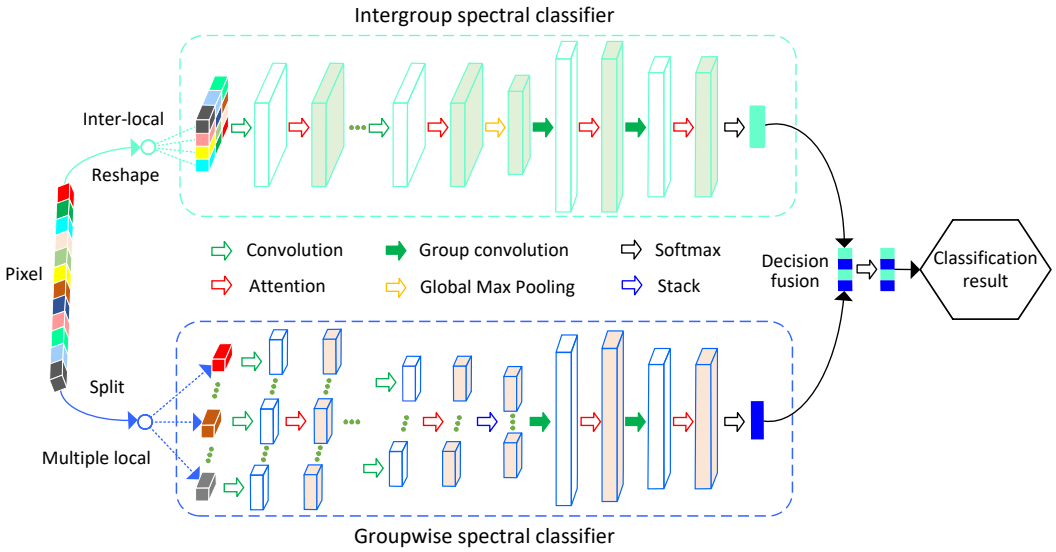

Fig. 2. The overall architecture of the proposed method. The intergroup spectral classifier extracts interlocal spectral correlation features from a matrix that is reshaped from the input pixel vector. The groupwise spectral classifier extracts discriminative local spectral features from groups of sub-bands.

correlations [37], and none within the 1D-CNN nor non-local inter-channel correlations yet. In contrast, we here develop a dual-channel attention for 1D-CNN model that learns simultaneously non-local and global inter-channel correlations via 1D convolution, embedding in the whole network.

III. Methodology

Let $\mathcal{X} \in \mathbb{R}^{H\times W\times B}$ denote a 3D hyperspectral cube with the spatial size of $H \times W$ and with B spectral bands. Among the total number of HW pixels, T pixels are labelled and denoted as training set $\mathcal{T} = \{\mathbf{x}_i, r_i\}_{i=1}^{T}$, where $\mathbf{x}_i \in \mathbb{R}^{B\times 1}$ denotes a spectral vector of one pixel, and r_i is its label from the set $\mathcal{C} = \{1, \cdots, C\}$, where C is the number of classes. For the given $\mathcal{X}$, with the training set $\mathcal{T}$, our goal is to predict the labels $\mathbf{r} = \{r_i\}_{i=T+1}^{HW}$ of the unlabelled pixels $\{\mathbf{x}_i\}_{i=T+1}^{HW} \notin \mathcal{T}$.

A. Overall Architecture

We propose a unified two-stream spectral feature fusion architecture based on 1D-CNN for hyperspectral image classification. The two streams (classifiers) that operate in parallel as shown in Fig. 2, learn simultaneously interlocal and discriminative local spectral correlation features. While the intergroup spectral classifier captures interlocal spectral correlation features, the groupwise spectral classifier learns the discriminative local spectral features. At its core lies a novel concept that we refer to as groupwise band convolution and that we elaborate later on. The proposed attention method is incorporated into the two parallel classifiers to improve the spectral feature learning capability. The outputs of the two classifiers are integrated with a decision fusion method to predict the outputs of labels.

B. Intergroup Spectral Classifier

Recent classification methods based on 1D-CNN [23–25] typically capture local spectral features from adjacent spectral bands. The intergroup spectral classifier in our architecture (see the top of Fig. 2) develops a wider 1D-CNN to capture interlocal spectral correlation features. The main idea is to reshape the input pixel vector into a non-adjacent spectral matrix and then to employ cascaded 1D convolutions for the adjacent and non-adjacent band correlation representations. Because this classifier builds up the correlations among groups of sub-bands, we name it intergroup spectral classifier. This classifier offers two advantages: it captures interlocal spectral correlation features and at the same time avoids inputting hundreds of spectral bands directly, mitigating this way also the associated adverse effects [15]. Fig. 3 illustrates this architecture and we give a formal description next.

The input is a matrix $\mathbf{X} \in \mathbb{R}^{L\times S}$ that we call non-adjacent spectral matrix, obtained by reshaping a spectral vector of one pixel $\mathbf{x} \in \mathbb{R}^{B\times 1}$. Here L is the number of spectral bands in each of the S non-adjacent channels. For a desired L, we set $S = \lceil B/L \rceil$, where $\lceil\cdot\rceil$ denotes the ceil function. The insufficient number of bands (i.e., $LS-B$) is padded with zero value. In this way, the actual input spectral dimensionality to our classifier L is reduced by a factor $S \geq 2$ via a simple reshaping operation. In the special case when $S = 1$, this reduces to feeding all the spectral bands B as e.g. in [23–27].

Local spectral feature extraction: Current methods typically use a regular 1D convolution directly to extract local spectral features [23–25]. Given the vector of pixel values in K adjacent spectral bands $\mathbf{x}_p \in \mathbb{R}^{K\times 1} = [x_p, x_{p+1}, \cdots, x_{p+K-1}]^T$ at position p, the local feature value $f_{p,j}$ in the j-th feature map of the first layer is computed by

$$f_{p,j} = \delta(\mathbf{w}_j \cdot \mathbf{x}_p) \tag{1}$$

where $\cdot$ denotes the dot product ($\mathbf{x}\cdot\mathbf{y} = \sum_i x_i y_i$). $\mathbf{w}_j \in \mathbb{R}^{1\times K} = [w_{j,0}, w_{j,1}, \cdots, w_{j,K-1}]$ is the kernel vector connected to the j-th feature map, and K is the kernel size. δ is the activation function. The bias terms are omitted in this paper. Equation (1) shows that the local spectral feature $f_{p,j}$

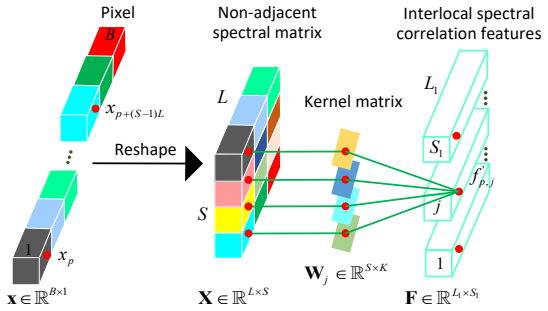

Fig. 3. A detail architecture of the proposed intergroup spectral extractor. The red solid-circles show the process of interlocal spectral correlation extraction.

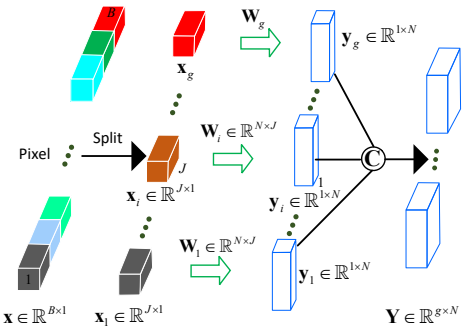

Fig. 4. An illustration of the proposed groupwise band convolution. Ⓒ denotes the concatenation operation.

is extracted from the K adjacent spectral bands with $\mathbf{w}_j$.

Interlocal spectral correlation feature extraction: To be able to capture interlocal spectral features, we transform the input $\mathbf{x} \in \mathbb{R}^{B\times 1}$ into a non-adjacent spectral matrix $\mathbf{X} \in \mathbb{R}^{L\times S}$ and then feed it into a regular 1D convolution. Given the spectral matrix in S non-adjacent spectral vectors $\mathbf{X}_p \in \mathbb{R}^{K\times S} = [\mathbf{x}_p^1, \mathbf{x}_p^2, \cdots, \mathbf{x}_p^S]$ at position p, where $\mathbf{x}_p^s \in \mathbb{R}^{K\times 1} = [x_{p+(s-1)L}, x_{p+(s-1)L+1}, \cdots, x_{p+(s-1)L+K-1}]^T$ is the s-th non-adjacent channel. We now define the interlocal feature value $f'_{p,j}$ at the corresponding position as follows:

$$f'_{p,j} = \delta(\sum_s \mathbf{w}_j^s \cdot \mathbf{x}_p^s) \tag{2}$$

where $\mathbf{w}_j^s \in \mathbb{R}^{1\times K}$ is the s-th row of the kernel matrix $\mathbf{W}_j \in \mathbb{R}^{S\times K} = [\mathbf{w}_j^1, \mathbf{w}_j^2, \cdots, \mathbf{w}_j^S]^T$ connected to the j-th feature map. The interlocal feature $f'_{p,j}$ is extracted from S non-adjacent spectral vectors in each of the K adjacent bands with $\mathbf{W}_j$. We employ multiple kernels to extract different interlocal spectral features. The output of the first layer for the interlocal classifier is $\mathbf{F} \in \mathbb{R}^{L_1\times S_1} = [\mathbf{f}_1, \mathbf{f}_2, \cdots, \mathbf{f}_{S_1}]$, where $\mathbf{f}_j \in \mathbb{R}^{L_1\times 1} = [f_{1,j}, f_{2,j}, \cdots, f_{L_1,j}]^T$, $1 \le j \le S_1$, is the j-th feature map with L_1 spectral bands, and S_1 is the number of kernels.

Similarly, we cascade three convolutional layers to extract deep interlocal spectral features. We then use a global max pooling to reduce the spectral size (i.e., let L reduce to 1). We employ a 1D interleaved group convolution (IGC) [41] to fuse the extracted features, which requires less fusion parameters compared to traditional methods with the fully connected layers [25, 26]. Given the extracted feature vector $\mathbf{q} \in \mathbb{R}^{1\times S_1} = [\mathbf{q}_1, \mathbf{q}_2, \cdots, \mathbf{q}_g]$, where $\mathbf{q}_j \in \mathbb{R}^{1\times S_1/g}$ is the feature sub-vector of the j-th group, the output of the IGC is $\hat{\mathbf{q}} = \mathbf{P}\mathbf{W}^{(2)}\mathbf{P}^T\mathbf{W}^{(1)}\mathbf{q}$. Here $\mathbf{P}$ is the permutation matrix. $\mathbf{W}^{(i)} = \text{diag}(\mathbf{W}_1^{(i)}, \mathbf{W}_2^{(i)}, \cdots, \mathbf{W}_g^{(i)})$, $i = \{1, 2\}$ is a block-diagonal matrix representing the weights of the i group convolution with kernel size of 1. In our experiments, we empirically set $S_1 = 256$ and $g = 8$. The output sizes of the two group convolutions in the IGC are set to 512 and 320, respectively. Finally, we use the softmax layer to predict the probability of each class.

C. Groupwise Spectral Classifier

We devise a novel groupwise spectral classifier (see the bottom of Fig. 2) based on 1D-CNN to extract discriminative local spectral features. Current classifiers based on 1D-CNN [23–25] typically exploit a set of 1D convolutional kernels to extract local spectral features from adjacent spectral bands, and apply them over all spectral bands with the same kernels (shared weight mechanism). This approach ignores however the differences among different spectral bands.

We propose a groupwise spectral classification method to capture discriminative local spectral features. The main idea is to exploit multiple sets of 1D convolutional kernels to convolve in parallel with groups of input sub-bands. This way we aim to extract more discriminative local spectral features adopted to the different portions of the spectral responses, and we concatenate them subsequently. This idea naturally raises the question: how to choose the number of groups in each layer appropriately? Would it be reasonable to reduce this hyper-parameter and to eliminate the redundancy simultaneously?

Groupwise band convolution: We thus introduce an extreme version where the size of convolution kernel is equal to the number of sub-bands in each group. Consequently, the number of groups should be equal to the integer division between the number of input spectral bands and the kernel size. Considering this convolution operates independently over each group of sub-bands (without shared weight mechanism), we name it groupwise band convolution. Fig. 4 illustrates this architecture. For simplicity, let the kernel size be the same for all the layers, and denote it by J. We define the number of groups in the first layer as

$$g = \lfloor B/J \rfloor \tag{3}$$

where $\lfloor \cdot \rfloor$ denotes the floor function. Given the input $\mathbf{x} \in \mathbb{R}^{B\times 1}$, the corresponding g groups are $[\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_g]^T$, where $\mathbf{x}_i \in \mathbb{R}^{J\times 1}$, $1 \le i \le g-1$, is the i-th group of sub-bands, and $\mathbf{x}_g \in \mathbb{R}^{(B-J(g-1))\times 1}$. We define the groupwise band convolution as follows:

$$\begin{bmatrix} \mathbf{y}_1^T \\ \vdots \\ \mathbf{y}_i^T \\ \vdots \\ \mathbf{y}_g^T \end{bmatrix} = \begin{bmatrix} \mathbf{W}_1 & \cdots & \mathbf{0} & \cdots & \mathbf{0} \\ \vdots & \ddots & \vdots & & \vdots \\ \mathbf{0} & \cdots & \mathbf{W}_i & \cdots & \mathbf{0} \\ \vdots & & \vdots & \ddots & \vdots \\ \mathbf{0} & \cdots & \mathbf{0} & \cdots & \mathbf{W}_g \end{bmatrix} \begin{bmatrix} \mathbf{x}_1 \\ \vdots \\ \mathbf{x}_i \\ \vdots \\ \mathbf{x}_g \end{bmatrix} \tag{4}$$

where $\mathbf{W}_i \in \mathbb{R}^{N\times J}$ denotes the kernel matrix of the i-th group, and N is the number of kernels. Observe that

each group performs a regular convolution to extract local spectral features: $\mathbf{y}_i^T \in \mathbb{R}^{N\times 1} = \mathbf{W}_i\mathbf{x}_i$. The groupwise band convolution amounts to extracting g local spectral features: $\mathbf{Y} \in \mathbb{R}^{g\times N} = [\mathbf{y}_1, \cdots, \mathbf{y}_g]^T$. The regular convolution is a special case of the groupwise band convolution for $\mathbf{W}_1 = \cdots = \mathbf{W}_i = \cdots = \mathbf{W}_g$.

To adapt to input different number of spectral bands, the groupwise band convolution can be designed in two versions: with no padding (for reducing the number of bands) as shown in Fig. 4, and with padding (for maintaining the number of bands). Given B input spectral bands, for a groupwise band convolution with no padding, the number of output bands is reduced to $g = \lfloor B/J \rfloor$. For a regular convolution, this number is $(B - J)$. The reduction ratio thus is

$$r = \frac{B-J}{\lfloor B/J \rfloor} \geq \frac{B-J}{B/J} = J\frac{B-J}{B} \approx J \tag{5}$$

where the approximate equality is based on the fact that $B \gg J$. This is why current deep spectral classifiers [23–25] yield a slow band reduction, while our method reduces the number of spectral bands by a factor J exponentially. This is an important asset of the proposed groupwise classifier. We repeat the groupwise band convolution several times to extract deep local spectral features. Similar to the intergroup spectral classifier, we employ the same feature fusion method and softmax layer to predict the probability of each class.

Comparison to depthwise convolution: Depthwise convolution [42] is a spatial convolution, which is different from groupwise band convolution in both the motivation and design. Depthwise convolution aims to extract spatial features with less parameters, while groupwise band convolution extracts discriminative spectral features with more parameters. In terms of design, depthwise convolution is typically used in 2D or 3D convolution [43] and operates on the channel dimension. Differently, groupwise band convolution stems from the 1D convolution and operates over the spectral dimension.

D. Dual-Channel Attention

We shall further boost the spectral feature learning capability of the two aforementioned classifiers by embedding an attention mechanism into both of them. We develop a novel dual-channel attention method based on 1D convolution. It consists of two attention modules, which learn non-local and global inter-channel correlations in parallel. Fig. 5 shows this architecture. Next, we elaborate its details.

Non-local channel attention: The non-local channel attention (see the top of Fig. 5) aims to build the non-local inter-channel correlations. To achieve this, we adopt the reshaping operation to construct non-local channel statistics. Let $\mathbf{F} \in \mathbb{R}^{L_1\times S_1}$ be the input, we first use a global average pooling (GAP) to generate channel-wise statistics $\mathbf{z} \in \mathbb{R}^{1\times S_1}$, as done in [35, 37]. We then reshape $\mathbf{z}$ into the non-local statistics $\mathbf{Z}^{nl} \in \mathbb{R}^{\frac{S_1}{M}\times M}$. Note that the number of channels S_1 is usually set to power of 2, we thus set M to power of 2 for convenience. To limit model complexity, we introduce a 1D convolution with M kernels to self-learn the non-local inter-channel correlation matrix $\mathbf{E}^{nl} \in \mathbb{R}^{\frac{S_1}{M}\times M} = [\mathbf{e}_1^{nl}, \mathbf{e}_2^{nl}, \cdots, \mathbf{e}_M^{nl}]$ from $\mathbf{Z}^{nl}$. We define the j-th correlation vector $\mathbf{e}_j^{nl} \in \mathbb{R}^{\frac{S_1}{M}\times 1}$ as follows:

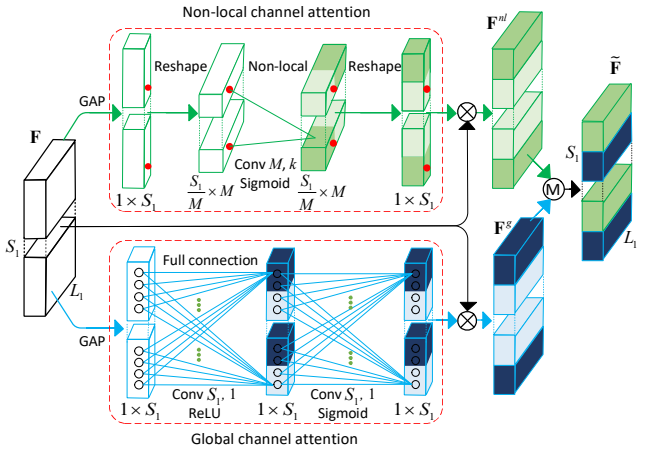

Fig. 5. Diagram of the proposed dual-channel attention. GAP is the global average pooling. $\otimes$ denotes the element-wise product. Ⓜ illustrates the element-wise maximum.

$$\mathbf{e}_j^{nl} = \sigma(\mathbf{W}_j^{nl} * \mathbf{Z}^{nl}) \tag{6}$$

where σ is the sigmoid function. $\mathbf{W}_j^{nl} \in \mathbb{R}^{M\times k}$ denotes the weight matrix and k is the kernel size. To avoid manual tuning of k, we introduce a method to select it adaptively: $k = |\frac{S_1}{M}|_{odd}$, where $|t|_{odd}$ denotes the nearest odd number of t. Finally, we reshape $\mathbf{E}^{nl}$ into a vector as the non-local correlations: $\mathbf{e}^{nl} \in \mathbb{R}^{1\times S_1}$. We define the output of the non-local channel attention $\mathbf{F}^{nl} \in \mathbb{R}^{L_1\times S_1}$ as follows:

$$\mathbf{F}^{nl} = \mathbf{e}^{nl} \odot \mathbf{F} \tag{7}$$

where $\odot$ denotes channel-wise product.

Global channel attention: The global channel attention (see the bottom of Fig. 5) is to learn the global inter-channel correlations. We achieve this by using the 1D convolution in line with the non-local channel attention. Given the input $\mathbf{F} \in \mathbb{R}^{L_1\times S_1}$, GAP is also used to produce the statistics $\mathbf{z}^g \in \mathbb{R}^{1\times S_1}$. Considering the statistics are all in channel dimension, the 1D convolution learns thereby global (full connection) inter-channel correlation $\mathbf{e}^g \in \mathbb{R}^{1\times S_1}$:

$$\mathbf{e}^g = \sigma(\mathbf{W}_2^g * \delta(\mathbf{W}_1^g * \mathbf{z}^g)) \tag{8}$$

where $\mathbf{W}_1^g \in \mathbb{R}^{S_1\times S_1}$ and $\mathbf{W}_2^g \in \mathbb{R}^{S_1\times S_1}$ are the weights of the two convolutional layers, respectively. The output of the global channel attention $\mathbf{F}^g \in \mathbb{R}^{L_1\times S_1}$ is defined as

$$\mathbf{F}^g = \mathbf{e}^g \odot \mathbf{F} \tag{9}$$

Dual-attention aggregation and embedding: To make use of the inter-channel correlations, we aggregate the features from the two attentions. To reduce the memory requirements, we choose the element-wise maximum to identify informative features of the two attentions automatically. Note that our dual-attention is incorporated into all the feature learning layers of the network except for the softmax layers (see the red arrows of Fig. 2), and it can be applied to other networks too.

E. Classification and Post-Processing

Decision fusion classification: Having constructed the two classifiers, we introduce a decision fusion method to integrate them efficiently. Specifically, we employ an element-wise maximum to select informative probabilities from the two classifiers, and then utilize a softmax layer as the decision fusion classifier to make the final prediction. In addition, we use two auxiliary loss functions on the two classifiers to facilitate the training. We define the joint loss as follows:

$$\begin{aligned}\mathcal{L} &= \mathcal{L}_f + \mathcal{L}_{ig} + \mathcal{L}_{gw} \\ &= -\frac{1}{T}\sum_{i=1}^{T}\log p_f^{r_i} - \frac{1}{T}\sum_{i=1}^{T}\log p_{ig}^{r_i} - \frac{1}{T}\sum_{i=1}^{T}\log p_{gw}^{r_i}\end{aligned} \tag{10}$$

where $\mathcal{L}_f$, $\mathcal{L}_{ig}$ and $\mathcal{L}_{gw}$ are the losses for the decision fusion, intergroup and groupwise classifiers, respectively. $p_f^{r_i}$, $p_{ig}^{r_i}$ and $p_{gw}^{r_i}$ are the r_i-th output probabilities of the corresponding classifiers for the i-th training sample $\mathbf{x}_i$. r_i is the label of $\mathbf{x}_i$. Having trained the proposed network with (10) using the mini-batch Adadelta [44], the spectral classification results of the unlabelled pixels are predicted by the trained network.

Local and Global Majority Voting: To further improve the predicted spectral classification results, the local spatial information is often used at the testing stage posteriorly [25, 30]. Due to complex and various spatial context in hyperspectral data, how to choose a robust window size to different data is a major challenge. We design a local and global majority voting method, which integrates local and global spatial information from a small and a large window sizes to adapt to different data. To avoid the pixel-sharing problem between training and testing sets, we exclude all the training pixels and their labels before voting. We then replace each excluded training pixel with the average of its four adjacent unlabelled pixels. This guarantees the integrity of hyperspectral image and makes use of the local spatial information. Because the large window contains more pixels than the small one, we make them have the same importance by weighting. Considering that the predicted probability of each pixel represents its reliability, we thus use it as the weight to make a reliable voting.

Let $\mathbf{x}_j^{L(G)}$ denotes the j-th neighbour of the testing pixel $\mathbf{x}$ in the local (L) or global (G) window, and $\hat{r}_j^{L(G)}$ the predicted label of $\mathbf{x}_j^{L(G)}$. We define the weight for candidate label u as

$$w_u = \lambda_L \sum_{j=1}^{s^2} p_f(\hat{r}_j^L = u) + \lambda_G \sum_{j=1}^{l^2} p_f(\hat{r}_j^G = u) \tag{11}$$

where $p_f(\hat{r}_j^{L(G)} = u)$ is the predicted probability of the decision fusion classifier for $\hat{r}_j^{L(G)} = u$. l^2 and s^2 are the number of pixels in the global and local windows, respectively. λ_L and λ_G are the weighting coefficients. To assign the same importance to the local and global windows, we set $\lambda_L = l^2$ and $\lambda_G = s^2$. With the majority voting mechanism, the testing pixel $\mathbf{x}$ is finally predicted by

$$r_f = \arg\max_u \{w_u\} \tag{12}$$

where $u \in \{1, \cdots, C\}$, and C is the number of classes.

TABLE I
THE NETWORK ARCHITECTURE OF THE PROPOSED METHOD FOR ALL THE DATA SETS. B IS THE NUMBER OF SPECTRAL BANDS.

Classifier	Input shape	Layer	Kernel size	Channel
Intergroup	$B/4 \times 4$	3	9	256
Groupwise	$B \times 1$	3	3	128

IV. Experimental Results and Analysis

We perform experiments on three well-known hyperspectral data sets[1]: Indian Pines, the University of Pavia (denoted as PaviaU) and Salinas. Three objective metrics, overall accuracy (OA), average accuracy (AA), and Kappa coefficient (κ) are used for evaluation. For each experiment, we report the mean and standard deviation of the classification results over ten runs with randomly selected training samples.

A. Data Description and Hyperparameter Setting

The Indian Pines image, captured by the AVIRIS sensor over the agricultural Indian Pines site in northwestern Indiana in 1992, contains 145×145 pixels with a spatial resolution 20 m. After removing the water absorption bands, $B = 220$ out of 224 bands are retained for analysis, with the spectral range from 0.4 to 2.5 μm. It contains 16 classes, out of which we select $C = 8$ large classes as did in [25, 45]. The PaviaU image, acquired by the ROSIS-03 sensor over an urban area surrounding the University of Pavia, Pavia, Italy, consists of 610×340 pixels with $C = 9$ classes and $B = 103$ spectral bands covering the spectral range from 0.43 to 0.86 μm with a spatial resolution of 1.3 m. The Salinas image, collected by the AVIRIS sensor over the area of Salinas Valley, CA, USA, has 512×217 pixels with $C = 16$ classes and with $B = 224$ spectral bands covering the spectral range from 0.4 to 2.5 μm with spatial resolution of 3.7 m.

We randomly select 50 labelled pixels per class for training. The remaining labelled pixels are used as the test set to evaluate the classification performance. Note that the training set is excluded with the proposed voting method at the testing stage, avoiding the pixel-sharing problem. We randomly select 10% of the training set as the validation set to determine the hyperparameters. The network architecture of the proposed method is the same for all the test images and is shown in Table I. The number of training epochs and batch size are empirically set to 100 and 64, respectively. The initial learning rate is empirically set to 10 for the Indian Pines image and 1 for the other two images, and it reduces with a clever strategy[2]. Due to limited training data, we duplicate them three times. The duplicated data is then shuffled before each epoch by setting shuffle=True for the fit method[3]. This strategy makes the learning rate reduce slowly during the training process, which facilitates the training of the proposed network. An ablation study regarding the key hyperparameters is given in Section IV-C.

[1] http://www.ehu.eus/ccwintco/index.php?%20title=Hyperspectral_Remote_Sensing_Scenes

[2] https://keras.io/api/callbacks/reduce_lr_on_plateau/

[3] https://keras.io/api/models/model_training_apis/#fit-method

TABLE II
COMPARISON OF THE CLASSIFICATION ACCURACIES AMONG THE PROPOSED METHOD AND THE BASELINES USING THE INDIAN PINES IMAGE.

Classes	Train/Test	CNNL	DCNN	RNN	CRNN	**TCNN**	MRFCNN	PPFCNN	ANNC	**TCNNS**
Corn-notill	50/1378	64.52±11.91	48.90±5.90	60.28±3.46	61.09±3.93	78.21±2.34	69.75±5.61	80.70±3.40	74.42±3.89	83.66±4.80
Corn-mintill	50/780	74.28±5.34	45.59±3.85	48.20±5.41	61.10±4.13	84.44±1.94	79.22±4.72	79.54±6.60	89.30±5.42	95.14±1.34
Grass-pasture	50/433	93.09±1.98	86.30±3.61	74.92±6.44	85.59±3.68	95.36±0.92	95.20±1.55	92.44±3.02	94.93±0.67	95.50±0.90
Hay-windrowed	50/428	99.79±0.30	97.94±1.18	93.84±2.72	98.83±0.78	99.77±0.33	99.84±0.11	99.98±0.07	100±0	99.95±0.14
Soybean-notill	50/922	69.11±9.30	58.26±4.03	58.20±3.60	74.21±4.13	84.93±2.90	80.38±3.06	75.77±5.17	85.68±6.01	92.73±2.62
Soybean-mintill	50/2405	54.73±10.27	43.50±4.20	73.84±1.24	57.32±3.46	72.54±5.16	65.42±3.36	90.06±2.56	83.19±3.70	90.07±3.11
Soybean-clean	50/543	72.54±11.97	58.14±5.73	45.16±4.21	65.30±3.98	87.11±4.05	74.00±4.46	81.54±6.65	92.94±6.47	96.76±2.30
Woods	50/1215	95.48±2.18	89.65±4.55	97.46±1.31	91.23±3.41	98.07±0.87	96.11±1.61	99.52±0.20	99.85±0.20	99.93±0.13
AA(%)	-	77.94±2.56	66.04±2.49	68.99±1.18	74.34±1.84	87.55±0.60	82.49±1.30	87.44±1.40	90.04±1.78	94.22±0.90
OA(%)	-	71.65±3.80	59.36±3.21	67.61±1.25	69.57±1.80	83.52±1.15	77.77±1.72	86.62±1.36	87.25±1.75	92.51±1.04
$\kappa \times 100$	-	66.59±4.13	52.15±3.52	61.80±1.35	64.05±2.07	80.37±1.30	73.67±1.97	84.00±1.61	84.70±2.14	91.02±1.23

TABLE III
COMPARISON OF THE CLASSIFICATION ACCURACIES AMONG THE PROPOSED METHOD AND THE BASELINES USING THE PAVIAU IMAGE

Classes	Train/Test	CNNL	DCNN	RNN	CRNN	**TCNN**	MRFCNN	PPFCNN	ANNC	**TCNNS**
Asphalt	50/6581	78.32±4.18	74.73±1.79	93.42±3.62	73.39±2.79	84.59±2.95	89.60±0.72	97.41±1.58	92.36±3.81	96.08±1.55
Meadows	50/18599	76.71±4.78	81.08±2.44	93.75±1.07	64.28±10.54	90.12±1.75	89.83±1.50	97.00±1.30	95.38±3.52	97.94±1.85
Gravel	50/2049	81.62±4.20	76.03±4.33	58.70±9.25	69.62±15.33	84.31±2.66	86.99±0.54	83.53±4.22	91.70±4.51	95.24±3.04
Trees	50/3014	91.55±4.70	91.38±3.96	73.69±9.14	92.48±1.65	93.12±1.30	95.63±0.94	80.34±7.58	94.49±0.99	89.83±2.50
Metal sheets	50/1295	99.50±0.24	99.85±0.12	94.72±2.31	99.54±0.19	99.80±0.14	99.61±0.63	99.76±0.44	100±0	99.99±0.02
Bare soil	50/4979	77.50±5.47	86.74±4.88	59.41±8.14	68.82±7.53	89.41±3.39	82.95±1.71	73.65±4.48	99.97±0.08	99.95±0.14
Bitumen	50/1280	90.40±6.03	88.73±3.07	53.67±7.63	90.34±2.68	93.34±1.15	91.70±0.77	86.25±7.61	97.50±2.12	99.19±0.45
Bricks	50/3632	81.68±4.20	74.49±1.88	77.10±5.71	78.06±7.83	80.41±3.58	80.31±2.53	86.51±6.21	89.45±6.44	96.87±2.22
Shadows	50/897	99.88±0.09	99.87±0.16	97.05±1.86	99.81±0.10	99.90±0.11	98.72±0.87	96.98±2.93	96.63±0.73	93.51±1.70
AA(%)	-	86.35±1.23	85.88±1.00	77.95±1.16	81.82±1.85	90.55±0.32	90.59±0.39	89.05±1.67	95.28±0.71	96.51±0.63
OA(%)	-	80.38±2.07	81.89±1.18	79.84±2.14	72.30±4.36	88.87±0.66	88.99±0.62	90.12±1.56	94.93±1.70	97.09±1.01
$\kappa \times 100$	-	74.86±2.43	76.69±1.41	74.16±2.58	65.30±4.60	85.43±0.82	85.56±0.77	87.14±1.99	93.33±2.19	96.15±1.33

TABLE IV
COMPARISON OF THE CLASSIFICATION ACCURACIES AMONG THE PROPOSED METHOD AND THE BASELINES USING THE SALINAS IMAGE

Classes	Train/Test	CNNL	DCNN	RNN	CRNN	**TCNN**	MRFCNN	PPFCNN	ANNC	**TCNNS**
Brocoli Weeds_1	50/1959	98.36±0.70	99.03±0.48	98.36±3.71	95.59±3.03	99.59±0.28	99.26±1.33	99.88±0.29	100±0	100±0
Brocoli Weeds_2	50/3676	98.88±0.47	99.02±0.76	99.11±0.38	97.20±4.43	99.74±0.13	97.65±1.12	99.44±0.31	100±0	100±0
Fallow	50/1926	94.70±6.05	97.04±1.69	93.11±2.18	93.83±6.69	99.76±0.18	98.90±1.38	95.49±3.42	100±0	100±0
Fallow_plow	50/1344	99.53±0.37	99.46±0.23	95.98±2.48	99.27±0.56	99.61±0.15	99.50±0.44	96.61±1.27	99.82±0.17	99.91±0.08
Fallow_smooth	50/2628	94.79±2.83	94.98±1.07	97.82±1.13	97.68±0.49	98.82±0.44	96.48±6.80	99.03±1.10	98.92±0.63	99.68±0.28
Stubble	50/3909	98.96±0.86	98.94±0.54	99.86±0.13	99.49±0.22	99.71±0.15	99.49±0.62	99.86±0.07	100±0	100±0
Celery	50/3529	99.35±0.21	99.47±0.12	97.65±1.36	99.23±0.40	99.61±0.09	98.98±0.81	99.32±1.41	100±0	100±0
Grapes_untrained	50/11221	67.9±16.59	64.02±8.96	74.94±3.24	59.80±11.91	73.57±4.59	74.69±4.95	84.82±2.06	87.35±3.76	88.68±2.70
Soil vinyard	50/6153	96.42±1.63	98.59±0.70	99.21±0.29	97.14±3.85	99.42±0.43	97.48±2.41	99.15±0.35	99.99±0.02	99.99±0.02
Corn weeds	50/3228	87.96±3.68	90.93±1.54	85.43±3.93	84.96±4.89	94.80±1.26	92.17±2.46	87.71±3.45	97.68±1.31	98.38±0.72
Lettuce_4wk	50/1018	95.97±5.38	95.76±1.90	86.77±5.58	95.19±2.30	99.47±0.41	98.36±1.20	88.55±7.96	100±0	100±0
Lettuce_5wk	50/1877	99.37±0.80	99.89±0.12	95.72±1.78	97.82±2.22	99.98±0.04	99.90±0.25	98.60±1.02	100±0	100±0
Lettuce_6wk	50/866	97.08±2.23	97.85±0.78	95.61±1.71	96.54±1.39	98.59±0.91	99.58±0.77	98.36±1.67	99.65±0.40	99.18±0.71
Lettuce_7wk	50/1020	94.99±2.81	94.53±1.50	89.14±9.10	92.68±1.34	96.99±2.07	96.68±1.47	90.98±7.67	99.95±0.12	99.63±0.58
Vinyard_untrained	50/7218	63.73±13.50	70.47±5.29	54.84±3.77	67.46±8.59	73.87±3.73	75.82±4.97	71.99±6.15	84.56±7.26	90.23±6.83
Vinyard_trellis	50/1757	98.04±1.36	98.35±0.48	93.08±6.13	96.15±1.47	98.93±0.31	97.31±1.52	98.55±0.95	99.74±0.25	100±0
AA(%)	-	92.88±1.04	93.65±0.57	91.04±0.94	91.88±0.98	95.78±0.38	95.14±0.65	94.27±1.19	97.98±0.51	98.48±0.48
OA(%)	-	86.17±2.17	86.86±1.63	85.22±1.08	84.61±1.71	90.24±0.87	89.92±0.87	90.97±1.42	95.04±1.20	96.16±1.06
$\kappa \times 100$	-	84.63±2.36	85.41±1.78	83.60±1.16	82.93±1.85	89.14±0.96	88.80±0.96	89.96±1.57	94.47±1.34	95.72±1.18

B. Comparisons with the State-of-the-art

We compare the classification performance of the proposed two-stream CNN model involving the proposed spatial post-processing (dubbed by TCNNS) with the following state-of-the-art deep learning-based methods. The reference methods are divided into three groups:

1) *CNN-based spectral classifiers:* 1D-CNN with a large reception field (CNNL) [23] and 1D-CNN with a deep network (DCNN) [24].
2) *RNN-based spectral classifiers:* recurrent neural network with a band grouping strategy (RNN) [26] and cascaded RNN (CRNN) [27];
3) *Deep classifiers with spatial post-processing:* pixel-pair CNN model with a majority voting (PPFCNN) [25], 2D-CNN combined with Markov random field prior posteriorly (MRFCNN) [46], and artificial neural network with an adaptive majority voting strategy (ANNC) [30].

The parameters of the reference methods are set to the default values indicated in their original works. We also compare the proposed method to its reduced version: TCNNS without involving the spatial post-processing (TCNN), to verify the effectiveness of its spectral feature extraction capability.

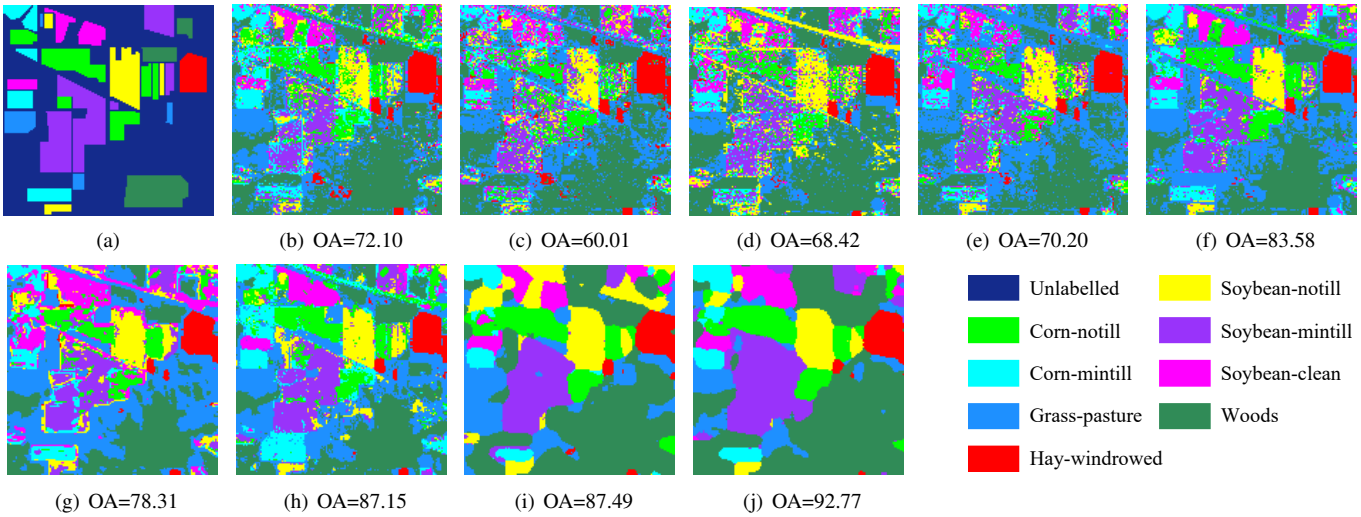

Fig. 6. Classification maps for the Indian Pines image. (a) Ground truth, (b) CNNL, (c) DCNN, (d) RNN, (e) CRNN, (f) TCNN, (g) MRFCNN, (h) PPFCNN, (i) ANNC, and (j) TCNNS.

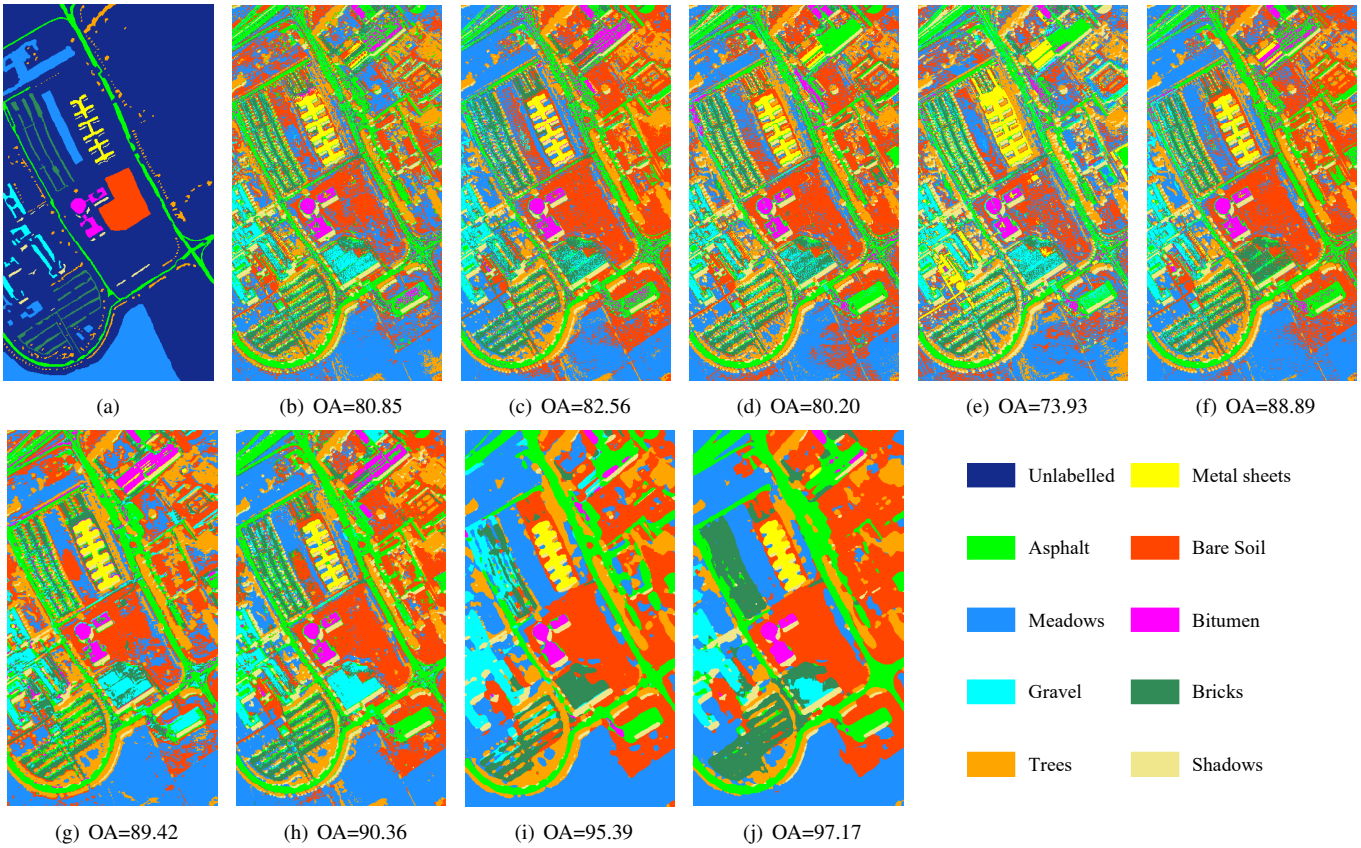

Fig. 7. Classification maps for the PaviaU image. (a) Ground truth, (b) CNNL, (c) DCNN, (d) RNN, (e) CRNN, (f) TCNN, (g) MRFCNN, (h) PPFCNN, (i) ANNC, and (j) TCNNS.

Tables II-IV report the class-specific accuracy, AA, OA, and κ of the tested methods on the three data sets. As can be observed, the proposed TCNNS consistently yields the best AA, OA, and κ with a significant improvement over the reference methods for all the data sets. For example, on Indian Pines (Table II), the improvement in OA compared to CNNL, DCNN, RNN, CRNN, MRFCNN, PPFCNN, and ANNC methods is about 20.8%, 33.1%, 24.9%, 22.9%, 14.7%, 5.9%, and 5.3%, respectively. The gains in OA compared to the best baseline are approximately 5.3%, 2.2%, and 1.1% for the Indian Pines, PaviaU, and Salinas images, respectively.

For the comparison of spectral classifiers, it is also evident that the proposed TCNN improves significantly the classification performance compared to both CNN-based spectral classifiers (CNNL and DCNN) and RNN-based spectral classifiers (RNN and CRNN) on the three data sets. The gains in OA

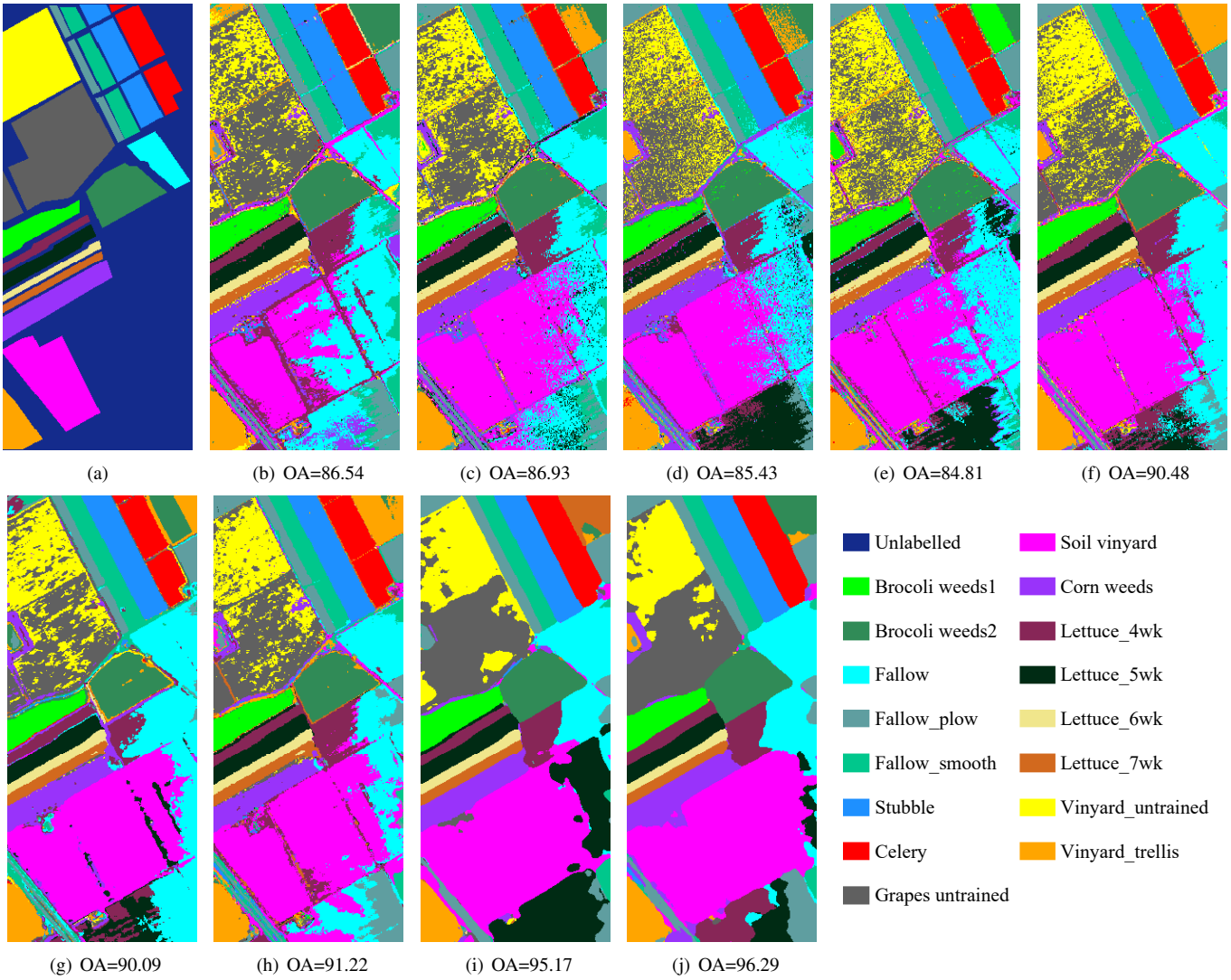

Fig. 8. Classification maps for the Salinas image. (a) Ground truth, (b) CNNL, (c) DCNN, (d) RNN, (e) CRNN, (f) TCNN, (g) MRFCNN, (h) PPFCNN, (i) ANNC, and (j) TCNNS.

compared to the best spectral classifiers are approximately 11.8%, 7%, and 3.4% for the Indian Pines, PaviaU, and Salinas images, respectively. This indicates that the proposed TCNN has powerful spectral feature extraction capability. For the comparison of spectral classifiers with spatial post-processing, the proposed TCNNS consistently yields again better classification performance than PPF-CNN, MFRCNN, and ANNC for all the three data sets. In addition, the proposed TCNNS improves significantly the classification performance compared to TCNN, with gains in OA of 9%, 8.2%, and 5.9% for the three data sets, which verifies the effectiveness of the proposed majority voting method.

It is also of interest to compare the proposed TCNNS with representative spectral-spatial feature extraction methods: CAE [47], PCNN [48], and a recent graph convolutional method (NLGCN) [49] under the same settings as in Tables II-IV. As indicated [47, 48], the input spatial size of CAE and PCNN is set to 16×16, which inevitably suffers from the pixel-sharing problem [13]. The results in Table V show that our TCNNS

TABLE V
COMPARISON BETWEEN THE PROPOSED METHOD AND SEVERAL SPECTRAL-SPATIAL FEATURE EXTRACTION METHODS.

Image	CAE	PCNN	NLGCN	**TCNNS**
Indian Pines	86.46±9.50	92.45±1.21	86.93±0.75	92.51±1.04
PaviaU	95.82±4.18	96.73±1.78	93.69±0.78	97.09±1.01
Salinas	93.74±3.12	93.52±1.17	92.96±0.47	96.16±1.06

without pixel-sharing still yields better overall accuracy than these reference methods for all the data sets.

Figs. 6-8 show the classification maps obtained by different methods on the three data sets. Obviously, the reference spectral classifiers (CNNL, DCNN, RNN, and CRNN) exhibit noisier estimations, mainly because of various noise, spectral variability, and mixture pixels. The proposed TCNN mitigates this phenomenon, benefiting from its discriminative spectral feature extraction. The methods with post-processing (MRFCNN, PPFCNN, ANNC, and TCNNS) also alleviate this phenomenon due to the use of spatial information posteriorly.

TABLE VI
OA OBTAINED BY DIFFERENT METHODS ON HOUSTON 2013 AND HOUSTON 2018.

Image	DCNN	CRNN	**TCNN**	PPFCNN	CAE (4×4)	PCNN (4×4)	CAE (16×16)	PCNN (16×16)	**TCNNS**
Houston 2013	77.69±0.68	75.59±0.35	81.33±0.21	81.26±0.48	79.18±2.29	84.71±0.51	79.17±0.92	85.20±0.53	81.66±0.33
Houston 2018	52.34±1.84	48.71±2.68	58.84±1.88	55.80±3.72	63.79±2.14	63.94±2.02	65.64±1.45	73.08±1.46	64.91±2.09

Among them, our TCNNS presents more similar results to the reference maps than all the reference methods. Nevertheless, our classification maps tend to be oversmoothed for small objects and are blurred in some borders. This is because our TCNNS cannot extract spatial features since it simply exploits the spatial information posteriorly to avoid pixel-sharing.

We further test the proposed method on two challenging hyperspectral images: Houston 2013[4] and Houston 2018[5]. Note that the training and testing sets of the two images are isolated from each other with region-shape as shown in [48]. The initial learning rate of the proposed method is set to 0.5 for these two images. The other hyperparameters of the proposed method for these two images are the same as for the other three images (Indian Pines, PaviaU, and Salinas). Four kinds of reference methods are used: (i) CNN-based spectral classifier: DCNN[24], (ii) RNN-based spectral classifier: CRNN [27], (iii) CNN-based classifier with spatial post-processing: PPFCNN [25], and (iv) spectral-spatial classifiers: CAE [47] and PCNN [48], which report two input spatial sizes, i.e., 4×4 and 16×16. As shown in Table VI, the proposed TCNN yields better OA than the reference spectral classifiers (DCNN and CRNN) on the two images. For the comparison of the use of spatial information, our TCNNS performs better than PPFCNN and performs comparable to CAE on the two images. PCNN yields the best OA when using a larger input spatial size (i.e., 16×16), especially on Houston 2018. The main reason is that a larger input spatial size provides richer spatial information and shares more pixels between the training and testing sets [13], improving the classification performance. It can be concluded that the proposed method without pixel-sharing yields favorable classification accuracy compared to the reference methods on the two challenging images.

C. Hyperparameter Analysis

1) Analysis of the number of training samples: We analyze the effect of the proposed method on the overall accuracy with respect to the number of training samples. The results in Fig. 9 show that the overall accuracy first drastically increases and then continues to rise gradually when the number of training samples increases. Particularly, our method yields about 80% overall accuracy with only 15 training samples per class in three data sets, which verifies the effectiveness of the proposed method under limited training data.

2) Analysis of the intergroup and groupwise classifiers: To validate the proposed two-stream architecture, we compare it with the networks that only contain the intergroup and groupwise classifiers. The results in Fig. 10 show that the proposed

[4]Available online: https://hyperspectral.ee.uh.edu/?page_id=459
[5]Available online: https://hyperspectral.ee.uh.edu/?page_id=1075

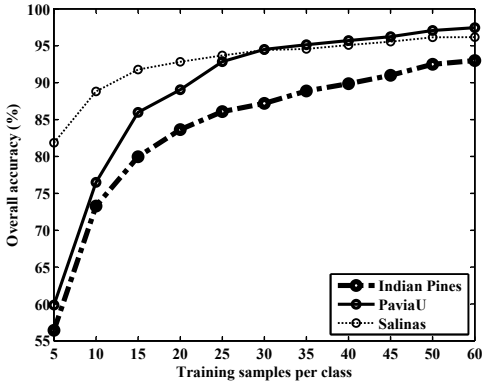

Fig. 9. The overall accuracy of the proposed method with different numbers of training samples per class in three data sets.

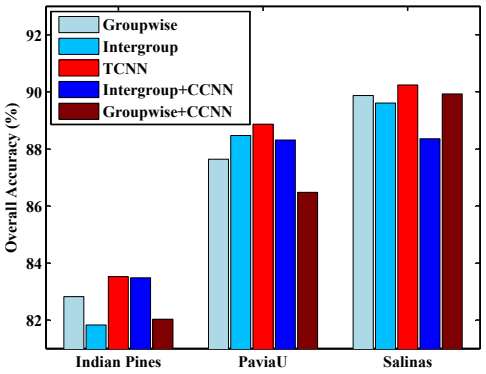

Fig. 10. The effect of the proposed decision fusion on the overall classification in three data sets. CCNN corresponds to the conventional 1D-CNN.

architecture (labelled by TCNN) consistently yields better accuracy than any of its two classifiers alone. The proposed two classifiers consistently perform better than the reference spectral classifiers (CNNL, DCNN, RNN and CRNN) on three data sets (see Tables II-IV), which verifies their effectiveness. It is also of interest to compare the proposed TCNN with its reduced versions: one stream of TCNN is replaced with its conventional 1D-CNN (labelled by CCNN). Also, our TCNN performs better than the two reduced versions, which demonstrates the usefulness of the proposed interlocal and discriminative local spectral feature extractors.

Furthermore, we analyze the number of non-adjacent channels S in the intergroup classifier. The results in Fig. 11 show that the overall accuracy generally increases and then declines as S increases. The main reason is that a smaller S underfits the interlocal spectral features and an excessive S tends to overfit them and requires more learning parameters. We choose $S = 4$ which yields nearly optimal performance on the tested data sets. In this case, the results show clearly the benefit

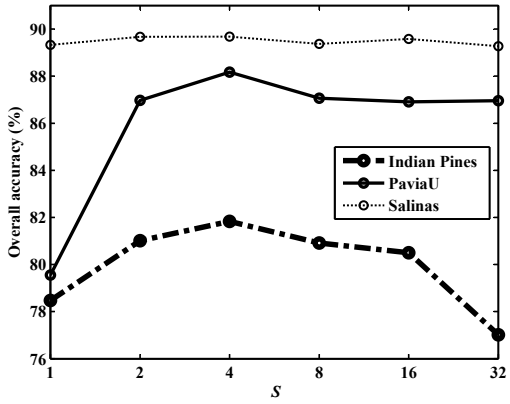

Fig. 11. The overall accuracy in function of the number of non-adjacent channels S in the proposed intergroup spectral classifier.

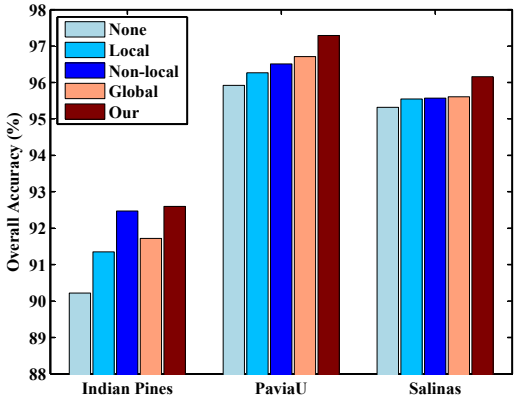

Fig. 13. The influence of the different channel attention methods on the overall accuracy in three data sets.

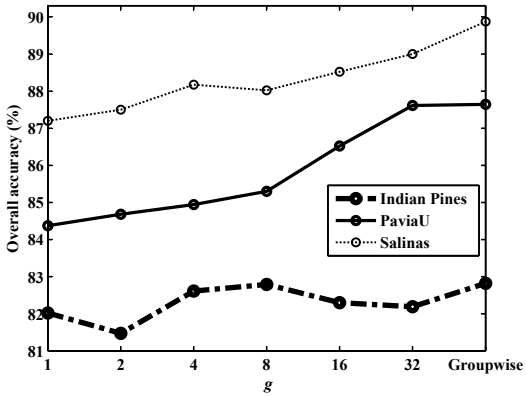

Fig. 12. The overall accuracy in function of the number of groups g in the proposed groupwise spectral classifier.

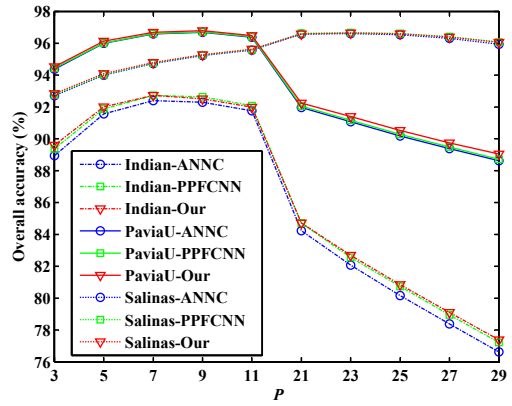

Fig. 14. The influence of the pixel-sharing with different window sizes P on the overall accuracy in three data sets. Best zoomed-in view.

of interlocal spectral feature extraction than the local spectral extraction (i.e., $S = 1$) in the Indian Pine and PaviaU images.

We also test the number of groups g in the groupwise classifier. Observe that the case $g = 1$ corresponds to the traditional local spectral feature extraction. The results in Fig. 12 show that the overall accuracy generally increases with increasing g in the PaviaU and Salinas images. For the Indian Pines image, the overall accuracy fluctuates with g values, but these fluctuations are within 1.5%. Clearly, our groupwise band convolution yields optimal performance on all the data sets due to its discriminative local spectral feature extraction.

3) Analysis of the dual-channel attention: We compare the proposed dual-channel attention with two state-of-the-art attention methods: the global channel attention (Global) [35] and the local channel attention (Local) [37], and with our reduced version: the non-local channel attention (Non-local) as well as the version without any attention (None). The results in Fig. 13 show that all the attention methods consistently perform better than the version without any attention in terms of overall accuracy for three data sets, which verify the effectiveness of attention mechanisms in spectral classification. Observe that our dual-channel attention provides the best overall accuracy for three data sets because it learns both the non-local and global inter-channel correlations.

4) Analysis of the majority voting: We first analyze the effect of the pixel-sharing with different window sizes in the majority voting on the overall accuracy. We compare our reduced majority voting that used one window size with two related methods: the majority voting that used the training set as did in PPFCNN [25] and the majority voting that excluded the training set as did in ANNC [30]. The results in Fig. 14 show that PPFCNN performs slightly better than ANNC, which verifies that the pixel-sharing indeed improves the classification accuracy in the tests. Also, our method performs slightly better than ANNC, since we exploit the predicted probabilities of unlabelled pixels to make a reliable voting. It is also evident that each data set has its own optimal window size.

To adapt to different data sets, the proposed method integrates both small and large window sizes. The results in Fig. 15 show the contributions of the weighting coefficients (i.e., λ_L and λ_G in (11)) in terms of overall accuracy. Note that λ_L and λ_G correspond to the large and small window sizes, respectively. Clearly, the performance is more sensitive to λ_G than λ_L. For the Indian Pines and PaviaU images, the OA first improves with increasing λ_G and then declines because they have more detailed regions. For the Salinas image, the OA continues to rise as λ_G increases since it has many large smooth regions. We choose $\lambda_L = 25 * 25$ and $\lambda_G = 7 * 7$ as a trade-off between the classification performance and the

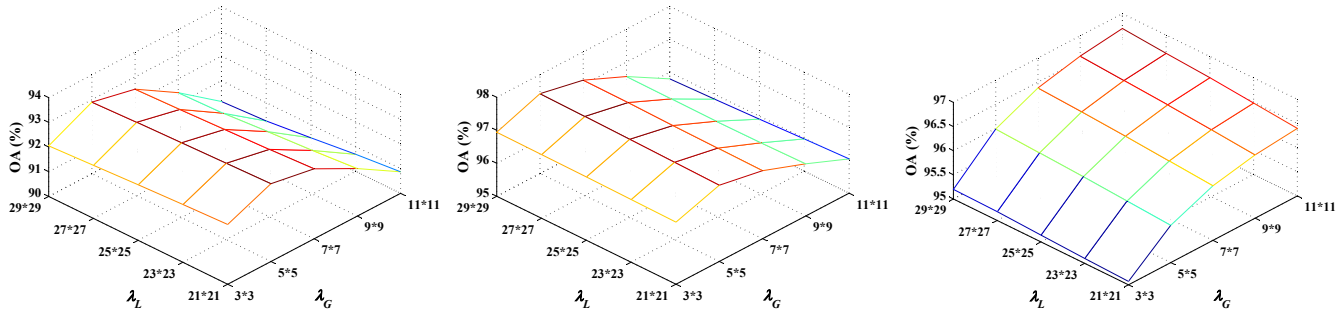

Fig. 15. The overall accuracy of the proposed method with the contributions of weighting coefficients λ_L and λ_G (a) Indian Pines, (b) PaviaU, (c) Salinas.

TABLE VII
THE EFFECT OF THE PIXEL-SHARING WITH DIFFERENT NUMBERS OF TRAINING SAMPLES PER CLASS ON OA FOR THE INDIAN PINES IMAGE

Training data	20	40	60	80	100
With pixel-sharing	83.53	90.65	93.21	94.94	95.42
Without pixel-sharing	83.15	90.24	92.78	94.45	94.82
Difference	0.38	0.41	0.43	0.49	0.60

TABLE VIII
COMPARISON OF THE NUMBER OF PARAMETERS ON DIFFERENT METHODS FOR THE PAVIAU IMAGE

Method	DCNN	CRNN	MRFCNN	CAE	TCNNS
#Params (×M)	0.05	0.35	0.88	0.37	6.90

TABLE IX
COMPARISON OF THE PROCESSING TIME ON DIFFERENT POST-PROCESSING METHODS FOR THE PAVIAU IMAGE

Method	MRFCNN	PPFCNN	ANNC	TCNNS
Training (s)	107.8	786.5	743.3	318.3
Testing (s)	41.2	26.9	165.6	20.2

running time for three data sets.

It is of interest to analyze the effect of the pixel-sharing with different numbers of training samples per class on the performance of the proposed method. Table VII reports the results for a particular test image (Indian Pines). Similar trends hold for other test images. It can be seen that the pixel-sharing indeed improves the classification accuracy, and the gains increases with more training data, benefiting from more shared pixels between the training and test sets.

5) Analysis of the computational efficiency: A comparative analysis of the number of parameters for different representative methods is summary in Table VIII. Four kinds of reference methods are used: (i) CNN-based spectral classifier: [24], (ii) RNN-based spectral classifier: CRNN [27], (iii) CNN-based classifier with spatial post-processing: MRFCNN [46], and (iv) spectral-spatial classifier: CAE [47]. The reported values correspond to one testing image (PaviaU) and are similar for the other two images. Obviously, the proposed method involves much more parameters than all the reference methods due to a wider two-stream network with attention mechanism.

The results in Table IX provide comparative analysis of the processing time of the proposed TCNNS and three related post-processing methods: PPFCNN [25], MRFCNN [46] and ANNC [30]. Compared to the reference methods, the proposed TCNNS is moderately fast in training and yields considerably faster testing. It can be concluded that the proposed TCNNS is competitive in the classification performance and is also computationally efficient compared to the reference methods.

V. CONCLUSION

In this paper, we proposed a unified two-stream spectral feature fusion approach based on 1D-CNN for hyperspectral image classification. In this approach, we devised a novel intergroup spectral classifier and an original groupwise spectral classifier, which simultaneously captures the interlocal and discriminative local spectral features. Moreover, we developed a novel dual-channel attention method to boost the spectral feature learning capability based on non-local and global inter-channel correlations. This approach improves significantly the spectral classification performance under limited training data. In addition, we introduced a decision fusion method and a joint loss to facilitate the training process. At the testing stage, we utilized the local and global spatial context to smooth the spectral classification results, without involving any training pixels. Experimental results on real data sets demonstrated the state-of-the-art classification performance.

VI. ACKNOWLEDGMENT

We would like to thank the Hyperspectral Image Analysis Laboratory at the University of Houston, and the IEEE GRSS Image Analysis for providing the Houston University 2013 and 2018 data sets used in this paper, and the Data Fusion Technical Committee for their preparation and pre-process for the two data sets. We also thank the Associate Editor and the anonymous Reviewers for their insightful comments and helpful suggestions which have greatly improved this paper.

REFERENCES

[1] X. X. Zhu, D. Tuia, L. Mou, G.-S. Xia, L. Zhang, F. Xu, and F. Fraundorfer, "Deep learning in remote sensing: A comprehensive review and list of resources," *IEEE Geosci. Remote Sens. Mag.*, vol. 5, no. 4, pp. 8–36, 2017.

[2] P. Ghamisi, N. Yokoya, J. Li, W. Liao, S. Liu, J. Plaza, B. Rasti, and A. Plaza, "Advances in hyperspectral image and signal

processing: A comprehensive overview of the state of the art," *IEEE Geosci. Remote Sens. Mag.*, vol. 5, no. 4, pp. 37–78, 2017.
[3] A. Signoroni, M. Savardi, A. Baronio, and S. Benini, "Deep learning meets hyperspectral image analysis: a multidisciplinary review," *J. Imaging*, vol. 5, no. 5, p. 52, 2019.
[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] K. Golhani, S. K. Balasundram, G. Vadamalai, and B. Pradhan, "A review of neural networks in plant disease detection using hyperspectral data," *Inf. Process. Agric.*, vol. 5, no. 3, pp. 354–371, 2018.
[6] J. Marcello, E. Ibarrola-Ulzurrun, C. Gonzalo-Martín, J. Chanussot, and G. Vivone, "Assessment of hyperspectral sharpening methods for the monitoring of natural areas using multiplatform remote sensing imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 10, pp. 8208–8222, 2019.
[7] P. Ghamisi, E. Maggiori, S. Li, R. Souza, Y. Tarablaka, 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 Geosci. Remote Sens. Mag.*, vol. 6, no. 3, pp. 10–43, 2018.
[8] 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, 2019.
[9] X. Li, M. Ding, and A. Pižurica, "Deep feature fusion via two-stream convolutional neural network for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 4, pp. 2615–2629, 2020.
[10] Y. Chen, K. Zhu, L. Zhu, X. He, P. Ghamisi, and J. A. Benediktsson, "Automatic design of convolutional neural network for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 9, pp. 7048–7066, 2019.
[11] 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, 2018.
[12] Q. Gao, S. Lim, and X. Jia, "Spectral–spatial hyperspectral image classification using a multiscale conservative smoothing scheme and adaptive sparse representation," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 10, pp. 7718–7730, 2019.
[13] J. Liang, J. Zhou, Y. Qian, L. Wen, X. Bai, and Y. Gao, "On the sampling strategy for evaluation of spectral-spatial methods in hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 2, pp. 862–880, 2017.
[14] Z. Gong, P. Zhong, Y. Yu, W. Hu, and S. Li, "A CNN with multiscale convolution and diversified metric for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 6, pp. 3599–3618, 2019.
[15] G. Hughes, "On the mean accuracy of statistical pattern recognizers," *IEEE Trans. Inf. Theory.*, vol. 14, no. 1, pp. 55–63, 1968.
[16] P. Ghamisi, J. Plaza, Y. Chen, J. Li, and A. J. Plaza, "Advanced spectral classifiers for hyperspectral images: A review," *IEEE Geosci. Remote Sens. Mag.*, vol. 5, no. 1, pp. 8–32, 2017.
[17] M. Pal and G. M. Foody, "Feature selection for classification of hyperspectral data by svm," *IEEE Trans. Geosci. Remote Sens.*, vol. 48, no. 5, pp. 2297–2307, 2010.
[18] M. Belgiu and L. Drăguţ, "Random forest in remote sensing: A review of applications and future directions," *ISPRS J. Photogramm. Remote Sens.*, vol. 114, pp. 24–31, 2016.
[19] M. Khodadadzadeh, J. Li, A. Plaza, H. Ghassemian, J. M. Bioucas-Dias, and X. Li, "Spectral–spatial classification of hyperspectral data using local and global probabilities for mixed pixel characterization," *IEEE Trans. Geosci. Remote Sens.*, vol. 52, no. 10, pp. 6298–6314, 2014.
[20] Y. Zhong and L. Zhang, "An adaptive artificial immune network for supervised classification of multi-/hyperspectral remote sensing imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 50, no. 3, pp. 894–909, 2011.
[21] P. Zhou, J. Han, G. Cheng, and B. Zhang, "Learning compact and discriminative stacked autoencoder for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 7, pp. 4823–4833, 2019.
[22] P. Zhong, Z. Gong, S. Li, and C.-B. Schönlieb, "Learning to diversify deep belief networks for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 6, pp. 3516–3530, 2017.
[23] W. Hu, Y. Huang, L. Wei, F. Zhang, and H. Li, "Deep convolutional neural networks for hyperspectral image classification," *J. Sens.*, vol. 2015, 2015.
[24] 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, 2016.
[25] W. Li, G. Wu, F. Zhang, and Q. Du, "Hyperspectral image classification using deep pixel-pair features," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 2, pp. 844–853, 2017.
[26] Y. Xu, L. Zhang, B. Du, and F. Zhang, "Spectral-spatial unified networks for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, no. 99, pp. 1–17, 2018.
[27] R. Hang, Q. Liu, D. Hong, and P. Ghamisi, "Cascaded recurrent neural networks for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 8, pp. 5384–5394, 2019.
[28] 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, 2017.
[29] Y. Tarabalka, J. A. Benediktsson, and J. Chanussot, "Spectral–spatial classification of hyperspectral imagery based on partitional clustering techniques," *IEEE Trans. Geosci. Remote Sens.*, vol. 47, no. 8, pp. 2973–2987, 2009.
[30] A. J. X. Guo and F. Zhu, "Spectral-spatial feature extraction and classification by ANN supervised with center loss in hyperspectral imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 3, pp. 1755–1767, 2019.
[31] S. Asadzadeh and C. R. de Souza Filho, "A review on spectral processing methods for geological remote sensing," *Int. J. Appl. Earth. Obs. Geoinf.*, vol. 47, pp. 69–90, 2016.
[32] Y. Zhou, J. Peng, and C. P. Chen, "Extreme learning machine with composite kernels for hyperspectral image classification," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 8, no. 6, pp. 2351–2360, 2014.
[33] Y. Chen, Z. Lin, X. Zhao, G. Wang, and Y. Gu, "Deep learning-based classification of hyperspectral data," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 7, no. 6, pp. 2094–2107, 2014.
[34] V. Mnih, N. Heess, A. Graves *et al.*, "Recurrent models of visual attention," in *Adv. Neural. Inf. Process. Syst.*, 2014, pp. 2204–2212.
[35] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proc. CVPR.*, 2018, pp. 7132–7141.
[36] J. Fu, J. Liu, H. Tian, Y. Li, Y. Bao, Z. Fang, and H. Lu, "Dual attention network for scene segmentation," in *Proc. CVPR.*, 2019, pp. 3146–3154.
[37] Q. Wang, B. Wu, P. Zhu, P. Li, W. Zuo, and Q. Hu, "ECA-Net: Efficient channel attention for deep convolutional neural networks," in *Proc. CVPR.*, 2020, pp. 11 534–11 542.
[38] L. Mou and X. X. Zhu, "Learning to pay attention on spectral domain: A spectral attention module-based convolutional network for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 1, pp. 110–122, 2019.
[39] 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, 2020.

[40] 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.*, pp. 1–14, 2020.
[41] T. Zhang, G.-J. Qi, B. Xiao, and J. Wang, "Interleaved group convolutions," in *Proc. CVPR.*, 2017, pp. 4373–4382.
[42] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," in *Proc. CVPR.*, 2017, pp. 1251–1258.
[43] H. Zhang, Y. Li, Y. Jiang, P. Wang, Q. Shen, and C. Shen, "Hyperspectral classification based on lightweight 3-D-CNN with transfer learning," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 8, pp. 5813–5828, 2019.
[44] M. D. Zeiler, "ADADELTA: An adaptive learning rate method," *CoRR.*, vol. abs/1212.5701, 2012. [Online]. Available: https://arxiv.org/abs/1212.5701.
[45] M. Zhang, W. Li, and Q. Du, "Diverse region-based CNN for hyperspectral image classification," *IEEE Trans. Image Process.*, vol. 27, no. 6, pp. 2623–2634, 2018.
[46] X. Cao, F. Zhou, L. Xu, D. Meng, Z. Xu, and J. Paisley, "Hyperspectral image classification with Markov random fields and a convolutional neural network," *IEEE Trans. Image Process.*, vol. 27, no. 5, pp. 2354–2367, 2018.
[47] R. Kemker and C. Kanan, "Self-taught feature learning for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 5, pp. 2693–2705, 2017.
[48] B. Rasti, D. Hong, R. Hang, P. Ghamisi, X. Kang, J. Chanussot, and J. A. Benediktsson, "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, 2020.
[49] L. Mou, X. Lu, X. Li, and X. X. Zhu, "Nonlocal graph convolutional networks for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 12, pp. 8246–8257, 2020.

Xian Li (S'19) received the M.S. degree from Harbin Institute of Technology, Harbin, China, in 2016, where he is currently pursuing the Ph.D. degree in instrument science and technology with the School of Instrumentation Science and Engineering. He is also a doctoral researcher with the Department of Telecommunications and Information Processing, UGent-GAIM, Ghent University, Belgium, supported by the China Scholarship Council. His research interests include deep learning, hyperspectral remote sensing image analysis.

Mingli Ding received the B.S. and the Ph.D. degrees from Harbin Institute of Technology, Harbin, China, in 2000 and 2005, respectively. From 2009 to 2010, he was a Visiting Scholar with the French National Center for Scientific Research, Toulouse, France. He is currently a Full Professor with the School of Instrumentation Science and Engineering, Harbin Institute of Technology, China.

His research interests include deep learning, image classification, object detection, automation test technology, and information processing.

Aleksandra Pižurica (SM'15) received the Diploma in electrical engineering from the University of Novi Sad, Serbia, in 1994, the Master of Science degree in telecommunications from the University of Belgrade, Serbia, in 1997, and the Ph.D. degree in engineering from Ghent University, Belgium, in 2002.

She is a Professor in statistical image modeling with Ghent University. Her research interests include the area of signal and image processing and machine learning, including multiresolution statistical image models, Markov Random Field models, sparse coding, representation learning, and image and video reconstruction, restoration, and analysis.

Prof. Pižurica served as an Associate Editor for the IEEE TRANSACTIONS ON IMAGE PROCESSING (2012 – 2016), Senior Area Editor for the IEEE TRANSACTIONS ON IMAGE PROCESSING (2016 – 2019) and currently an Associate Editor for the IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY. She was also the Lead Guest Editor for the EURASIP Journal on Advances in Signal Processing for the Special Issue "Advanced Statistical Tools for Enhanced Quality Digital Imaging with Realistic Capture Models" (2013). The work of her team has been awarded twice the Best Paper Award of the IEEE Geoscience and Remote Sensing Society Data Fusion contest, in 2013 and 2014. She received the scientific prize "de Boelpaepe" for 2013–2014, awarded by the Royal Academy of Science, Letters and Fine Arts of Belgium for her contributions to statistical image modeling and applications to digital painting analysis.