This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

Central Attention Network for Hyperspectral Imagery Classification

Huan Liu, Wei Li, *Senior Member, IEEE*, Xiang-Gen Xia, *Fellow, IEEE*, Mengmeng Zhang, Chen-Zhong Gao, and Ran Tao, *Senior Member, IEEE*

***Abstract*— In this article, the intrinsic properties of hyperspectral imagery (HSI) are analyzed, and two principles for spectral–spatial feature extraction of HSI are built, including the foundation of pixel-level HSI classification and the definition of spatial information. Based on the two principles, scaled dot-product central attention (SDPCA) tailored for HSI is designed to extract spectral–spatial information from a central pixel (i.e., a query pixel to be classified) and pixels that are similar to the central pixel on an HSI patch. Then, employed with the HSI-tailored SDPCA module, a central attention network (CAN) is proposed by combining HSI-tailored dense connections of the features of the hidden layers and the spectral information of the query pixel. MiniCAN as a simplified version of CAN is also investigated. Superior classification performance of CAN and miniCAN on three datasets of different scenarios demonstrates their effectiveness and benefits compared with state-of-the-art methods.**

***Index Terms*— Central attention, hyperspectral imagery (HSI), spectral–spatial feature extraction, transformer.**

I. Introduction

BY RECORDING reflectance spectral information of the ground on an aircraft or satellite platform, hyperspectral imagery (HSI), occupying dozens of or even hundreds of contiguous narrow bands, possesses abundant discriminative information for land use and land cover (LULC) classification [1]–[6]. However, reflectances in adjacent bands are often highly correlated, and spectral and spatial resolutions are unbalanced, which deteriorates the classification performance.

Many traditional feature extraction methods have been proposed. These methods are partitioned into three categories: spectral-based, spectral plus spatial feature extraction, and spatial–spectral feature extraction methods. For spectral-based feature extraction, feature extraction is performed only in the spectral domain. The principal component analysis (PCA) [7] and the linear discriminant analysis (LDA) [8] are the most classical unsupervised and supervised feature extraction methods, respectively. Under the framework of graph embedding [9], many methods have been proposed for spectral feature extraction, including the unsupervised sparsity preserving graph embedding (SPGE) [10], the sparsity graph-based discriminant analysis (SGDA) [11], and the collaboration-competition graph preserving embedding (CCPGE) [12]. They learn low-dimensional features from spectral information by particularly designed graphs. For spectral plus spatial feature extraction, extended morphological profiles (EMP) [13], morphological attribute profiles (AP) [14], and local binary pattern (LBP) [15] are commonly used methods for spatial feature extraction on HSI. In general, the extracted spatial features are stacked with the spectral feature, which incurs the curse of dimensionality, causing information redundancy and overfitting of classification models. For spectral–spatial feature extraction, spectral and spatial features are jointly extracted to keep the most discriminative information for HSI classification. The representative methods include the spatial coherence distance (SCD) [16], the image patches distance (IPD) [17], the spatial–spectral combined distance (SSCD) [18], the tensor sparse and low-rank graph-based discriminant analysis (TSLGDA) [19], and the orthogonal total variation component analysis (OTVCA) [20]. In general, for those methods, feature extraction is performed on a patch around a query pixel or on the whole HSI.

Compared with traditional feature extraction, deep-learning-based methods extract deeper and more discriminative features. For convolutional neural network (CNN), a strategy was proposed to classify HSI (CNNHSI) directly in the spectral domain [21], while spatial information was not utilized at all. CNNPPF [22] was fed with a pair of pixels to learn the pixel-pair features, and the final classification result of a query pixel was decided by the voting strategy on the results of all pairs of the query pixel and its neighboring pixels. The benefit of this network is that training samples are augmented tremendously, but spatial information is not utilized. In [23], the randomized principal component analysis (RPCA) was first performed on the original spectral–spatial features, and the transformed features were fed into a designed CNN-based network (RPCACNN) to extract deeper spectral–spatial features. Similarly, in SSCNN [24],

Manuscript received October 18, 2021; revised January 8, 2022; accepted February 24, 2022. This work was supported in part by the Beijing Natural Science Foundation under Grant JQ20021 and in part by the National Natural Science Foundation of China under Grant 61922013 and Grant U1833203. *(Corresponding author: Wei Li.)*

Huan Liu, Wei Li, Mengmeng Zhang, Chen-Zhong Gao, and Ran Tao are with the School of Information and Electronics, Beijing Institute of Technology, Beijing 100081, China (e-mail: huanliu233@gmail.com; liwei089@ieee.org; xianggen@udel.edu; 7520200002@bit.edu.cn; 3120205425@bit.edu.cn; rantao@bit.edu.cn).

Xiang-Gen Xia is with the School of Information and Electronics, Beijing Institute of Technology, Beijing 100081, China, and also with the Department of Electrical and Computer Engineering, University of Delaware, Newark, DE 19716 USA.

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TNNLS.2022.3155114.

Digital Object Identifier 10.1109/TNNLS.2022.3155114

a traditional spectral–spatial extraction method was exerted on the original features, followed by a designed CNN. Contextual deep CNN (CDCNN) performed spectral–spatial extraction on a patch by a multiscale convolutional filter bank [25]. A novel network called diverse region-based CNN (DRCNN) was fed with a patch and different parts of the patch around a query pixel simultaneously to learn contextual interactional features, and the learned spectral–spatial features from different parts of the patch were merged to obtain the final feature [26]. For tree species classification based on HSI, a 3-D CNN framework was proposed [27] to fully exploit the local spectral structure information. To tackle the small sample set problem, a lightweight convolutional neural network (LWCNN) was proposed [28], where the joint spectral–spatial information was extracted by spectral–spatial Schroedinger eigenmaps (SSSE) and then fed into a designed CNN network. To address the problem of a limited number of training samples, CNNs with multiscale convolution (MS-CNNs) were proposed by extracting deep multiscale features from HSI [29]. To reduce the computational cost of the patch-based learning framework, a fast patch-free global learning (FPGA) framework was proposed for HSI classification [30]. To maintain detailed information while preserving semantic information, a multiscale CNN-based module called information compensation was proposed by integrating the original input with more abstract hierarchical learning feature maps [31]. As a variant of CNN, the graph convolutional network (GCN) has combined the concept of convolution with graph [32] and attracted a lot of attention. In [33], by combining CNN and GCN, a framework called FusCNNGCN was proposed for HSI classification, where features were generated by CNN and GCN, and then fused by a full connection layer. Similarly, a deep feature aggregation framework driven by GCN (DFAGCN) was developed for scene classification in remote sensing by the combination of CNN and GCN [34], where features in different layers of a pretrained VGGNet-16 were concatenated and fed into a graph-level classification model based on GCN.

Recently, based on the self-attention mechanism, transformer [35] has been applied in many other fields of machine learning and become more and more successful. For HSI classification, bidirectional encoder representations from transformers (BERTHSI) [36] were introduced. In BERTHSI, a pixel sequence with a query pixel and its neighbors was first fed into an embedding module with features and positional embedding, and then, the output embedded pixels were fed into a BERT module to learn the final features. By combining attention mechanism with CNN, an attention mechanism-based method termed the multilevel feature network with spectral–spatial attention model (MFNSAM) was proposed [37]. This network consists of a multilevel CNN and a spectral–spatial attention module. For many other deep-learning-based methods, please refer to the literature [38].

Although the existing deep-learning-based methods achieve a certain level of HSI classification performance, some problems exist. For CNN-based or CNN-mixed methods, on the one hand, some methods extract features only in the spectral domain and do not utilize spatial information or utilize spatial information partly, such as CNNHSI [21] and CNNPPF [22]. On the other hand, spatial information is excessively utilized in some methods, and the spectral feature of a query pixel does not stand out sufficiently, such as RPCACNN [23], SSCNN [24], CDCNN [25], DRCNN [26], 3DCNN [27], and FusCNNGCN [33]. These methods treat all pixels of a patch around a query pixel equally without highlighting the query pixel. For attention-based methods, such as BERTHSI [36], they treat all relationships with equal right without highlighting the relationships between the query pixel and its surrounding pixels. This is unreasonable since it is the query pixel that is labeled, not the patch. Though the pixels around the query pixel provide discriminative information as spatial information, the query pixel should be more important than these pixels.

In this article, to highlight a query pixel and correctly extract the spatial information brought by the pixels around the query pixel simultaneously, the central attention network (CAN) is proposed. In CAN, shallower features of a query pixel and the surrounding pixels are fed into multilayer central attention modules, where deeper features of each pixel and the similarity weight of each pixel relative to the query pixel are obtained in the spectral domain or channel dimension. Then, by weighted averaging pooling, pixels with more weights have more contribution in the feature extraction of the next layer. CAN adopts a dense connection strategy, that is, all the former features and weights are reused in the current layer. Specifically, for the original HSI patch, the spectral feature of a query pixel is reused in the third last layer; for hidden nodes, both weight matrices and features are reused for all the deeper layers. In addition, a simplified version of CAN called miniCAN is investigated with lower computation complexity.

The main contributions can be summarized as follows.

1) To demonstrate the intrinsic properties of HSI, pixel-level HSI classification is revisited, and two principles for spectral–spatial feature extraction on HSI are built: one emphasizes the important role of the spectral information of a query pixel and the other defines the exploitable spatial information. They provide a brand-new perspective for designing excellent feature extraction methods for pixel-level HSI classification.
2) Based on the two principles, the proposed central attention module in CAN performs feature extraction on HSI and calculates the similar weights between a query pixel and its neighbors, simultaneously. By doing so, spectral–spatial feature extraction is obtained by collecting the information of a query pixel and its neighbors that are similar to the query pixel. The design is completely HSI-tailored.
3) The proposed dense connection strategy is also HSI-tailored. For the input original HSI patch, the reuse of the spectral information of a query pixel emphasizes the spectral information; for hidden nodes, both weight matrices and features are reused for all the deeper layers and help collect detailed and useful spatial information.

The remainder of this article is organized as follows. In Section II, the problem of pixel-level HSI classification is revisited, and two principles for HSI classification and their rationality are introduced. Section III explicitly explains

Version of record at: https://doi.org/10.1109/TNNLS.2022.3155114

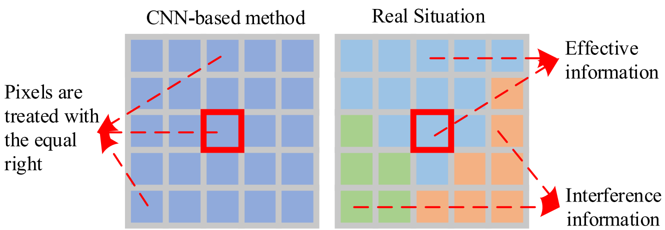

Fig. 1. Gap between CNN-based method and real situation of HSI classification (the pixel in the red rectangle is a query pixel).

the proposed CAN. Experimental results are displayed in Section IV to demonstrate the superior performance of CAN. In Section V, we conclude this article.

II. Preliminary

The earlier work of CNN for HSI classification only utilizes spectral information while neglecting spatial information [21]. To incorporate spatial information, a query pixel is replaced by a patch around the query pixel as input so that spatial information can be utilized as well. By spatial convolution, the local spatial structures of an image are obtained to discriminate images of different classes. However, for pixel-level HSI classification, the most important information mainly comes from the spectral domain with dozens or even hundreds of dimensions, and the spatial information is the relationships between a query pixel and its neighbors. As shown in Fig. 1, for the CNN-based methods, pixels in an HSI patch share the same right with a query pixel. However, in reality, pixels that are similar to a query pixel provide effective spatial information that helps for the correct classification of the query pixel, but those that are dissimilar to the query pixel provide interference spatial information, which hinders the correct classification of the query pixel. Therefore, the specific spatial information of HSI may not be well-extracted by CNN.

Inspired by the success of transformer and BERT [39], the vision transformer (ViT) [40] was proposed and achieved better classification performance than CNN. It is not surprising since the self-attention module in ViT encodes the relationships between different pixels, and thus, the global spatial structure is extracted and more discriminating than the local spatial structures extracted by CNN. However, it may be not well-suited for HSI classification since the relationships of a query pixel to its neighbors are more important than other relationships between the surrounding pixels, but all relationships are treated equally in ViT. BERTHSI [36] follows the same rule as ViT; therefore, the extracted features still have redundant and interference information from other relationships.

Based on the above analysis for pixel-level classification of HSI, two principles are presented for the spectral–spatial feature extraction of pixel-level HSI classification in the following.

Principle 1: The class of a query pixel in HSI pixel-level classification is mainly decided by its spectral feature.

Principle 1 points out that spectral information is the most discriminative information for LULC classification. One material is made up of a combination of molecules or atoms, and different materials have different molecular or atomic structures. Starting from the quantum hypothesis of energy proposed by Planck, the absorption and radiation of energy by atoms and molecules are not continuous but discrete, and thus, different atoms and molecules have different reflectance spectra. Furthermore, different materials are made of different combinations of atoms or molecules; therefore, different materials have different reflectance spectral features.

By combining spectroscopic and imaging technologies, hyperspectral imaging is an image-spectrum merging technology, and every pixel of an HSI is a reflectance spectral feature that expands a wide spectrum band. Taking different spectral features, pixels of categories, such as water and building, can be easily distinguished. Moreover, by occupying a sufficient number of spectrum bands, pixels of categories that cannot be differentiated by RGB images is well-differentiated, such as the fine classification of tree species [27]. Therefore, the discriminative nature of spectral information is the foundation of pixel-level HSI classification.

Principle 2: Besides the spectral feature of a query pixel, neighboring pixels that are similar to the query pixel also contain discriminative information for the classification of the query pixel, which is called effective spatial information.

Principle 2 points out that discriminative spatial information can be derived from neighboring pixels that are similar to the query pixel. Although the spectral feature of a query pixel provides primary discriminative information for classification, in reality, one class of LULC consists of different materials, and the material composition in the same LULC class varies from pixel to pixel. Therefore, in order to reduce the interference caused by the variation of material composition, it is necessary to use the information of the surrounding pixels to reduce the negative effect on classification from the variation. How can we reduce the variation?

By extracting the effective spatial information defined in Principle 2, we can reduce the variation. This definition of spatial information is different from most of the existing methods. Some of them define spatial information as texture information, such as LBP [15]. Indeed, this information can be used for object detection of different shapes. However, it is not well-suited for pixel-level HSI classification as it may sabotage the fine structure of the most discriminative information—spectral information. Other methods define spatial information as the query information and all its adjacent pixels without highlighting the query pixel and the more similar adjacent pixels. It is obviously unreasonable to use all the surrounding pixels without differentiation and treat them equally. This is because some of the surrounding pixels are in the same category as the query pixel, and some may not be. In this case, the information that is not in the same category becomes interference information and may prevent the correct feature extraction. Therefore, in order to extract effective spatial information, we need to give more attention to pixels that are similar to the query pixel and ignore those that are less similar to the query pixel since similar pixels are more likely to be the same class as the query pixel.

Accepted Manuscript
Version of record at: https://doi.org/10.1109/TNNLS.2022.3155114

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

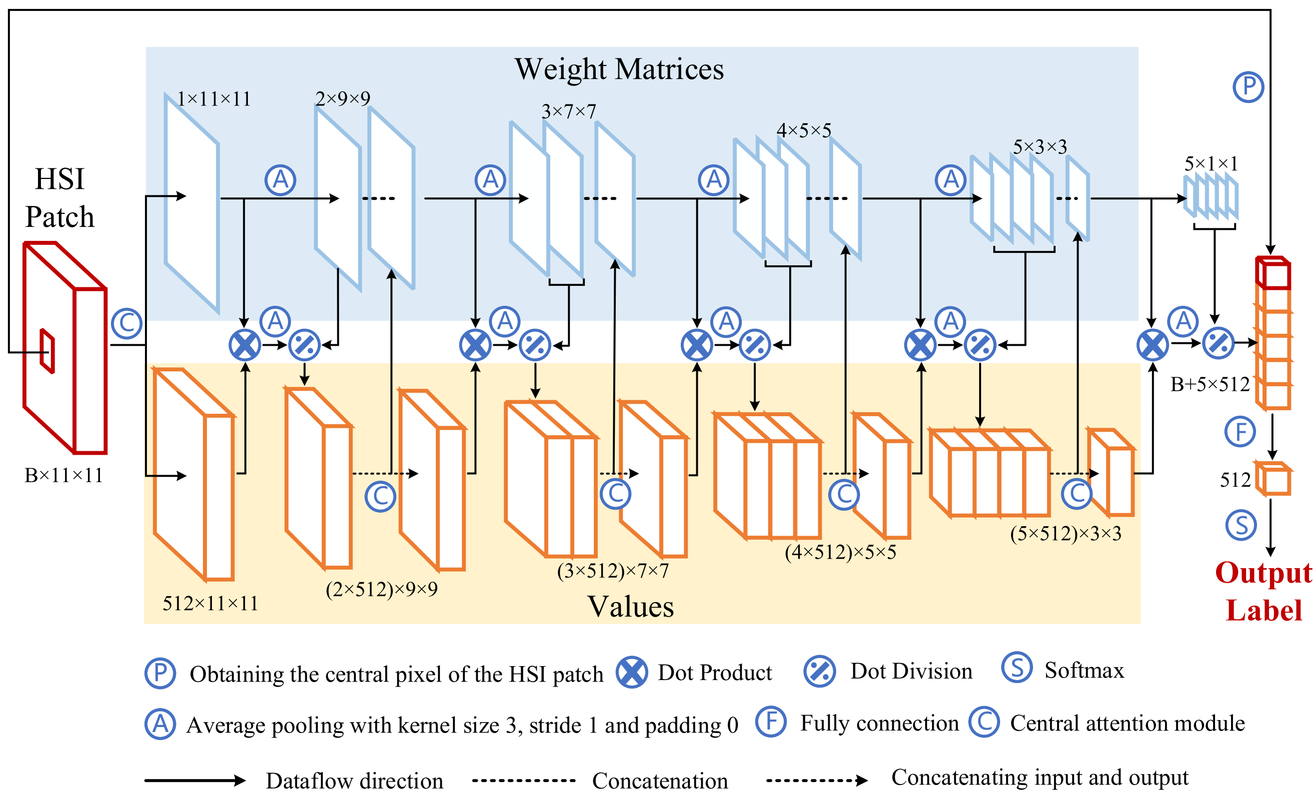

Fig. 2. Illustration of the proposed CAN for HSI classification, including central attention module and HSI-tailored dense connection.

III. Proposed Central Attention Network

Based on the two principles for pixel-level HSI classification, the proposed model architecture of CAN is illustrated in Fig. 2, including the central attention module and HSI-tailored dense connection.

A. Central Attention Module

In the central attention module, a 3-D tensor made of pixels as input is mapped along the channel dimension into two new 3-D tensors, called key and value tensors, respectively, made of the same number of pixels as the input. Every pixel of the key and value tensors are called the value and key of the corresponding pixel of the input tensor, respectively. By a compatibility function of the key tensor and its central pixel, a weight matrix is obtained. The output is the weighted averaging pooling of the value tensor with the weight matrix as the corresponding weights. In the following, we describe it in more detail.

1) Scaled Central Dot-Product Attention: In analogy to the self-attention proposed in [35], our particular central attention module is called "scaled central dot-product attention," as shown in Fig. 3.

Suppose that an input is a 3D tensor patch $\mathbf{X}_i \in \mathbb{R}^{C_i \times m \times n}$, where m and n are the height and width, respectively, and C_i is the number of channels of the input. Note that $m = n$ is an odd integer, and the central pixel of an HSI patch as an input tensor is a query pixel to be classified. In the process of calculation, we transform the 3D tensor into a pixelwise 2-D matrix $\mathbf{x}_i \in \mathbb{R}^{C_i \times \mathrm{mn}}$. The value matrix $\mathbf{y} \in \mathbb{R}^{C_o \times \mathrm{mn}}$ corresponding to a value tensor is obtained by

$$\mathbf{y} = \mathbf{ReLU}(\mathbf{W}_o \mathbf{x}_i + \mathbf{b}_o) \tag{1}$$

where $\mathbf{W}_o \in \mathbb{R}^{C_o \times C_i}$ and $\mathbf{b}_o \in \mathbb{R}^{C_o \times \mathrm{mn}}$ are learnable parameters, C_o is the number of channels of the value tensor, and **ReLU** is a rectified linear unit. Then, $\mathbf{y}$ is transformed into $\mathbf{Y} \in \mathbb{R}^{C_o \times m \times n}$ as the value tensor. The key matrix $\mathbf{z} \in \mathbb{R}^{C_o \times \mathrm{mn}}$ corresponding to a key tensor is obtained by

$$\mathbf{z} = \mathbf{W}_1 \mathbf{x}_i + \mathbf{b}_1 \tag{2}$$

where $\mathbf{W}_1 \in \mathbb{R}^{C_o \times C_i}$ and $\mathbf{b}_1 \in \mathbb{R}^{C_o \times \mathrm{mn}}$ are learnable parameters. Let $\mathbf{z}_o$ be the middle column vector of the key matrix $\mathbf{z}$, i.e., the key of the central pixel of the input $\mathbf{X}_i$. Then, the 1-D weight vector of size mn is

$$\mathbf{w} = \mathbf{SoftMax}\left(\frac{\mathbf{z}_o^T \mathbf{z}}{\sqrt{C_o}}\right) \tag{3}$$

where $\mathbf{SoftMax}(\cdot)$ is the softmax activation function. The scaling factor of $(1/(C_o)^{1/2})$ is analogical to that of the self-attention in [35]. Then, $\mathbf{w}$ is transformed into a weight matrix $\mathbf{W} \in \mathbb{R}^{m \times n}$. The output is $\mathbf{X}_o \in \mathbb{R}^{C_o \times (m-2) \times (n-2)}$ obtained by

$$\mathbf{X}_o = \mathbf{AvgPool}(\mathbf{Y} \otimes \mathbf{W}) / \mathbf{AvgPool}(\mathbf{W}) \tag{4}$$

where $\otimes$ and $/$ are pointwise multiplication and division, respectively, $\mathbf{W}$ is repeated along channel dimension C_0 times to adapt to the size of $\mathbf{Y}$, and **AvgPool** is the average pooling

Accepted Manuscript
Version of record at: https://doi.org/10.1109/TNNLS.2022.3155114

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

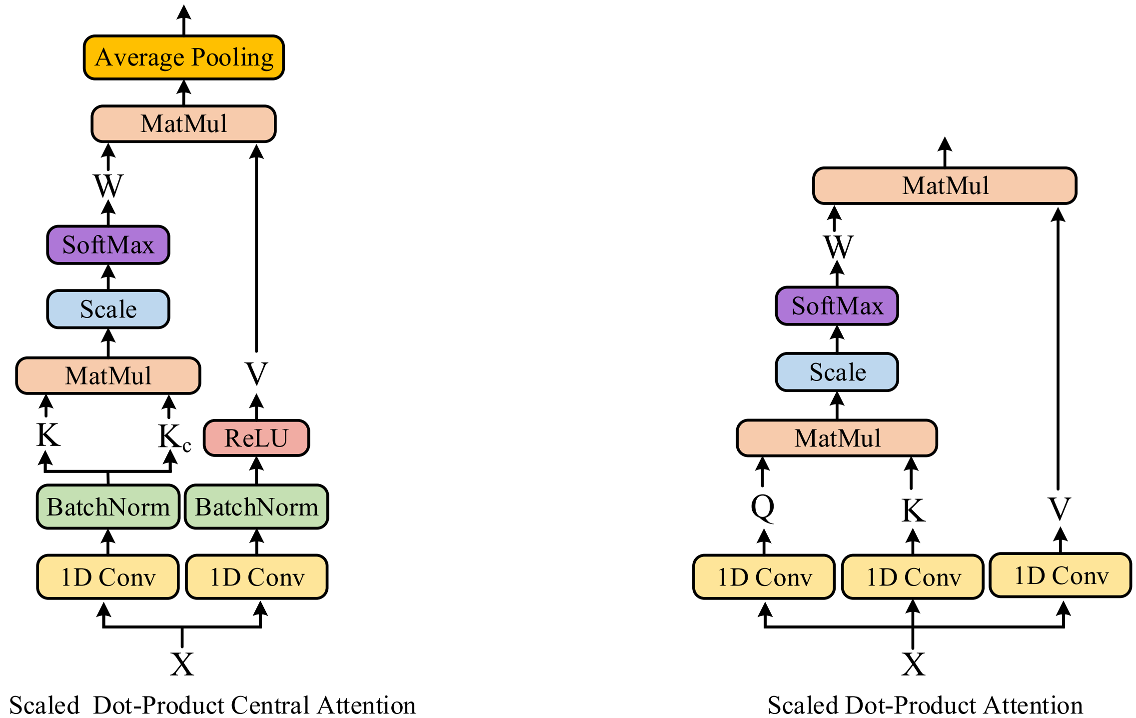

Fig. 3. Comparison between scaled dot-product central attention of the proposed CAN and scaled dot-product attention of transformer. Q, K, V, and K_c are the abbreviations of query, key, value, and the central pixel of the key (i.e., the key of the query pixel for an HSI patch), respectively.

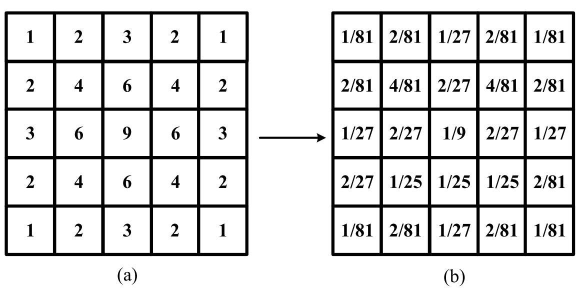

Fig. 4. (a) Participant times in the calculation of average pooling of kernel size 3. (b) Corresponding implicit spatial weights.

of kernel size 3, stride 1, and padding 0. By weighted average pooling, information of pixels with more similarity weights to the central pixel can flow with more ratio into the next layer, but that of dissimilar pixels to the central pixel has little influence on the next layer. In addition, since the stride of the weighted average pooling is 1, the pixels that are closer to the central pixel participate more times in the calculation of the average pooling. This indicates an implicit spatial weight for every pixel, and this implicit spatial weight is defined as the ratio between the participation times of the pixel and that of all pixels, as exemplified in Fig. 4. This is reasonable since closer pixels are more likely to be the same class as the central pixel (i.e., the query pixel of an HSI patch as an input tensor), and the central pixel stands out with the highest implicit spatial weight.

There are differences between the proposed central attention module and self-attention module in the transformer, as shown in Fig. 3. First, relationships between all pixels are calculated in the self-attention module, while only the relationships between the central pixel of an input tensor and its surrounding pixels are calculated. Second, the self-attention module contains value, query, and key, and the weight matrix is a function of query and key [35]; however, the central attention module contains value and key, and the weight matrix is a function of key and key's central pixel. Finally, the central attention module has used batch normalization and **ReLU** activation that are not used in the self-attention module. For language translation, the needed features are extracted from word sequences, and layer normalization is empirically better than batch normalization; therefore, in the self-attention based methods, such as transformer and ViT as its variant for RGB image classification, layer normalization and **ReLU** are used in a feed-forward network (FFN) after the self-attention to generate nonlinear features. However, for pixel-level HSI classification, spectral information possesses the most discriminative information; therefore, feature extraction can be directly conducted in the spectral domain, and the uses of batch normalization and **ReLU** are beneficial for better nonlinear feature extraction empirically.

2) Multihead Central Attention: As it is beneficial to use multihead attention in transformer, by analogy, multihead central attention is designed.

Let the number of heads be h. Then, the value tensor $Y \in \mathbb{R}^{C_o \times m \times n}$ is divided along the channel dimension into $Y_1, Y_2, \ldots, Y_h \in \mathbb{R}^{(C_o/h) \times m \times n}$, and the key matrix $z \in \mathbb{R}^{C_o \times \mathrm{mn}}$ is divided along the channel dimension into $z_1, z_2, \ldots, z_h \in \mathbb{R}^{(C_o/h) \times \mathrm{mn}}$. By Eqs. (2) and (3), X_{oj} corresponding to (Y_j, z_j) in (4) for each j is obtained. The output tensor is

$$X_o = \mathbf{Concat}(X_{o1}, \ldots, X_{oh}) \tag{5}$$

Accepted Manuscript
Version of record at: https://doi.org/10.1109/TNNLS.2022.3155114

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

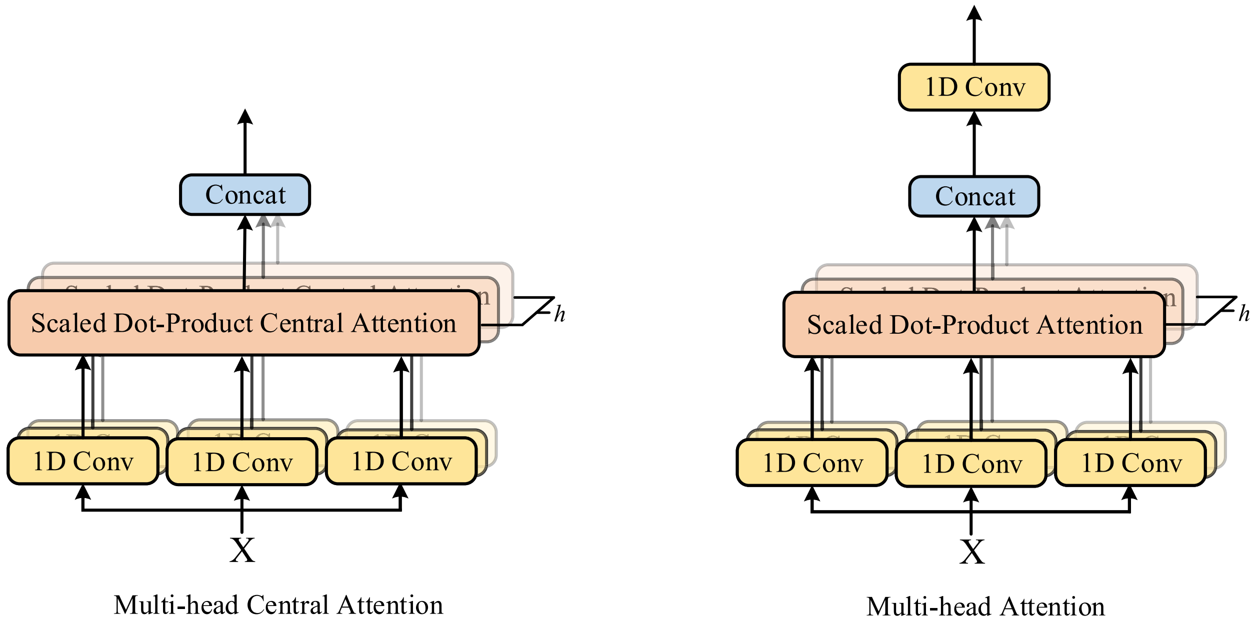

Fig. 5. Comparison between multihead central attention of the proposed CAN and multihead attention of transformer.

where **Concat** means that all the tensors are concatenated along the channel dimension. As shown in Fig. 5, compared with multihead attention in transformer, the proposed multihead central attention is generalized trivially from the single-head central attention module and has no extra 1-D convolution layer. The benefits of this are twofold. First, different from self-attention, central attention extracts favorable feature without any extra layer; therefore, the extra 1-D convolution is not needed for multihead attention. Second, an extra 1-D convolution means extra parameters to be trained. Accordingly, the complexity is reduced without the extra 1-D convolution.

B. HSI-Tailored Dense Connection

As pointed out in [41], based on dense connection, DenseNets alleviates the vanishing-gradient problem, strengthens feature propagation, encourages feature reuse, and substantially reduces the number of parameters. Inspired by the significant improvement of DenseNets over other state-of-the-art methods, two HSI-tailored dense connections are designed for CAN, i.e., dense connection for value and weight matrix, and dense connections for HSI patch.

1) Dense Connection for Value Tensor and Weight Matrix: In DenseNet, to reuse all the former features, the features of all former layers are concatenated as the input of the current layer, and the widths and heights of the output features of all layers are set equal for the convenience of concatenation. In the proposed CAN, values in different layers are supposed to be reused. However, unlike DenseNet, the widths and heights of the values of different layers are different. How can we concatenate all the former values when they have different patch sizes? We reuse not only all the former values but also all the former weight matrices. By weighted average pooling with different kernel sizes in different layers, i.e., using a smaller kernel size for value with a weight matrix in a shallower layer, and a larger kernel size for that in a deeper layer, the patch sizes of all the former layers adapt to the patch size of the latest former layer. Only by doing so, the features of all the former layers can be reused as the input for the next layer.

The weight matrix represents the relationships between the central pixel and its surrounding pixels. On the one hand, by weighted average pooling with a small kernel size in the central attention module, more detailed spatial information flows into the next layer. For example, if there is one pixel in the upper left corner that is the same category as the central pixel, then its spatial information is meaningful. By weighted average pooling with a small kernel size, spatial information is kept well. However, by weighted average pooling with a large kernel size, the information of this pixel is submerged by pixels that are more similar to the central pixel. Therefore, to keep more detailed spatial information, a small kernel size should be used. On the other hand, by weighted average pooling with a large kernel size in dense connection, more useful spatial information flows into the deeper layer. For example, if there is no pixel in the upper left corner that is the same category as the central pixel, then its spatial information is meaningless. By weighted average pooling with a small kernel size, this information flows into deeper layers. However, in deeper layers, the pixels are more similar to the central pixel. Therefore, to reuse more accurate and useful spatial information, the pixels in shallower layers that are more similar to the central pixel should be reused in deeper layers. By weighted average pooling with a large kernel size, more similar pixels stand out.

In practical implementation, for the convenience of calculation, the weight matrices and values in different former layers are reused by performing average pooling and weighted average pooling with the same kernel size (in practice, 3) different times, to adapt to the patch sizes of the value tensor and weight matrix of the latest former layer. As shown in Fig. 6, performing average pooling with a small kernel

Accepted Manuscript
Version of record at: https://doi.org/10.1109/TNNLS.2022.3155114

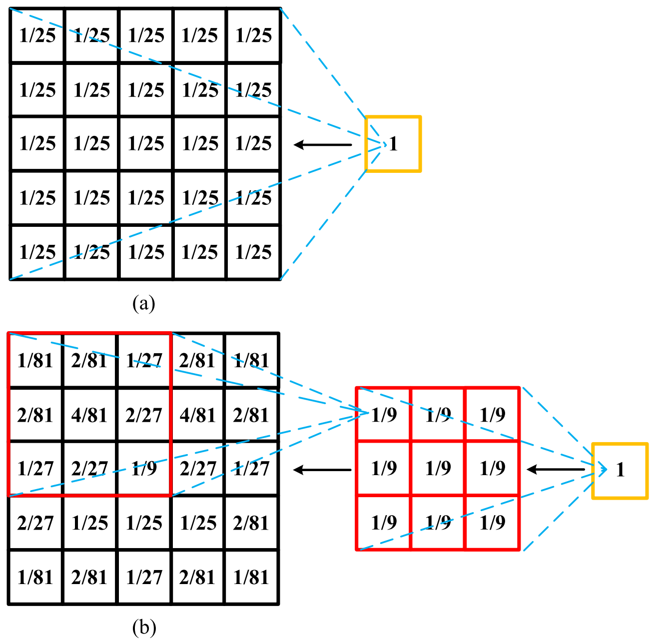

Fig. 6. Implicit spatial weights for different cases. (a) Average pooling with a kernel size 5. (b) Two average pooling with a kernel size 3.

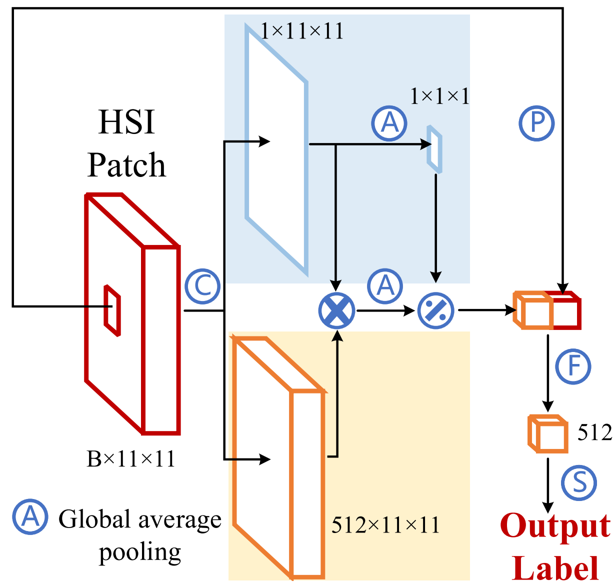

Fig. 7. Illustration of the proposed miniCAN for HSI classification.

size multiple times obtains a patch with the same size as performing average pooling with a large kernel size once; however, one benefits more from the former strategy, i.e., the central pixel and the pixels that are spatially closer to the central pixel have more implicit spatial weights in the former strategy. For example, in Fig. 6, the implicit spatial weight of the central pixel by average pooling of a kernel size 5 once is $1/25$, which is less than that by average pooling of a kernel size 3 twice, i.e., $1/9$. Therefore, the central pixel and the pixels closer to it are highlighted with more ratio by average pooling of a small kernel size multiple times, which conforms to the intuition of central attention.

2) Dense Connection for HSI Patch: The patch size of an HSI patch is different from that of the value tensor in deeper layers. However, since an HSI patch has no weight matrix, the strategy of dense connection for value tensor and weight matrix cannot be used for an HSI patch.

How can we reuse an HSI patch in deep layers? The answer is that the spectral information of the query pixel on an HSI patch is reused in the third last layer by concatenating it with one-dimensional features extracted by multilayer central attention modules. By multilayer central attention modules, information of the query pixel and its neighbors is similar to it is extracted. However, the potential drawback is that the information of the query pixel may be submerged by surrounding pixels. The benefit of this dense connection strategy is that the spectral features of the query pixel are highlighted, which conforms to the first principle, i.e., the class of a query pixel in HSI pixel-level classification is mainly decided by its spectral information.

C. MiniCAN

In the proposed CAN, the structure of CAN varies for input HSI patches of different sizes, i.e., the number of layers increases with the increase in the patch size, and more layers mean more trainable parameters, which increases training and testing times, and raises computational complexity. In light of this, a simplified version of CAN called miniCAN is further designed, as shown in Fig. 7. In miniCAN, only one layer of the central attention module is used, and adaptive global average pooling replaces the average pooling with a kernel size of 3×3 in (4) so that the number of layers is fixed no matter how large the patch size is. In addition, the computational complexity of miniCAN is much lower than that of CAN since the number of trainable parameters is much less than CAN.

IV. Experiments and Analysis

In this section, the proposed CAN and miniCAN are utilized on three HSI datasets to validate their effectiveness. First, three HSI datasets of different scenarios are introduced, including Gaofeng State Owned Forest Farm, Houston2013, and Yellow River Estuary that correspond to woodland, city, and wetland scenarios, respectively. Second, the classification performances of different patch sizes are analyzed for the three datasets. Then, ablation experiments are conducted to verify the effectiveness of the proposed miniCAN and CAN. Finally, experiments of CAN are carried out in comparison with several traditional and state-of-the-art algorithms, including OTVCA [20], CNNHSI [21], RPCACNN [23], SSCNN [24], CDCNN [25], DRCNN [26], BERTHSI [36], FusCNNGCN [33], and 3DCNN [27].

A. Datasets

1) Gaofeng State Owned Forest Farm: The first HSI dataset is the Gaofeng State Owned Forest Farm dataset, which was obtained by the AISA Eagle II diffraction grating push-broom hyperspectral imager carried by the airborne LiCHy (LiDAR, CCD, and Hyperspectral) system of the Chinese Academy of Forestry Sciences over Gaofeng State Owned Forest Farm in

Accepted Manuscript
Version of record at: https://doi.org/10.1109/TNNLS.2022.3155114

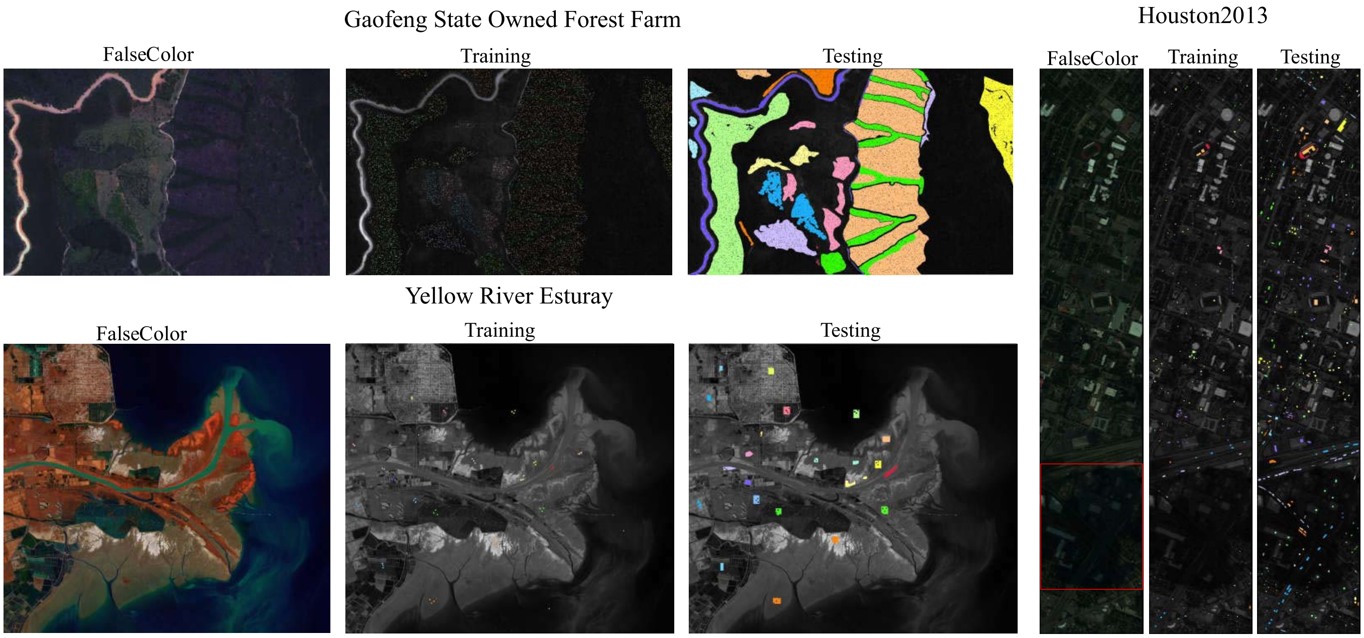

Fig. 8. False-color images and the distributions of training and testing samples for Gaofeng State Owned Forest Farm, Houston2013, and Yellow River Estuary datasets, respectively.

TABLE I

CLASS LABELS AND TRAIN–TEST DISTRIBUTION OF SAMPLES FOR GAOFENG STATE OWNED FOREST FARM

Class No.	Class Name	Train	Test
1	Cuninghamia lanceolata	4038	76705
2	Pinus massoniana	268	5072
3	Pius elliottii	208	3967
4	Eucalyptus grandis x urophylla	746	14185
5	Eucalyptus urophylla	2014	38261
6	Catanopsis hystrix	1274	24187
7	Mytilaria laosensis	150	2850
8	Camellia oleifera	416	7908
9	Other broadleaf forest	538	10203
10	Road	524	9948
11	Cutting blank	508	9662
12	Building land	4	64
Total		10688	203012

TABLE II

CLASS LABELS AND TRAIN–TEST DISTRIBUTION OF SAMPLES FOR HOUSTON2013

Class No.	Class Name	Train	Test
1	Grass healthy	198	1053
2	Grass stressed	190	1064
3	Grass synthetics	192	505
4	Tree	188	1056
5	Soil	186	1056
6	Water	182	143
7	Residential	196	1072
8	Commercial	191	1053
9	Road	193	1059
10	Highway	191	1036
11	Railway	181	1054
12	Parking lot 1	192	1041
13	Parking lot 2	184	285
14	Tennis court	181	247
15	Running track	187	473
Total		2832	12197

Guangxi province in south China in January 2018. It consists of 572 × 906 pixels with the spatial resolution of 1 m × 1 m and 125 spectral bands in the wavelength range of 0.4–1.0 μm with the spectral resolution of 3.3 nm. It contains nine different forest vegetation classes and three nonforest vegetation classes, and the numbers of training and testing samples of each class are listed in Table I. The false-color image of Yellow River Estuary and the distribution of those samples are shown in Fig. 8.

2) Houston2013: The second HSI dataset is the Houston2013 dataset, which was gathered by the CASI-1500 senor over the University of Houston and neighboring areas in June 2012 for the 2013 GRSS Data Fusion Contest. It consists of 349 × 1905 pixels with the spatial resolution of 2.5 m and 144 spectral bands in the wavelength range of 0.38–1.05 μm with the spectral resolution of 4.65 nm. It contains five natural objects and ten man-made objects, and the numbers of training and testing samples of each class are listed in Table II. The false-color image of Houston2013 and the distribution of those samples are shown in Fig. 8. Note that, in the red rectangle of the false-color image of Houston2013, there is an area affected by the cloud.

3) Yellow River Estuary: The third dataset is the Yellow River Estuary dataset, which was collected by GF-5 satellite over the Yellow River Delta in November 2018. It consists of 1185 × 1342 pixels with the spatial resolution of 30 m, 150 spectral bands in VNIR 0.40–1.00 μm with the spectral resolution of 5 nm, and 180 spectral bands in SWIR 1.00–2.50 μm with the spectral resolution of 10 nm, respectively. After moving the bands 1 and 2 in VNIR and 42–53, 96–115, 119–121, 172–173, and 175–180 in SWIR, 285 spectral bands remain. It contains 20 typical wetland

Accepted Manuscript
Version of record at: https://doi.org/10.1109/TNNLS.2022.3155114

TABLE III
CLASS LABELS AND TRAIN–TEST DISTRIBUTION OF SAMPLES FOR THE YELLOW RIVER ESTUARY DATASET

Class No.	Class Name	Train	Test
1	Spartina alterniflora	84	743
2	Low-tide mudflat	84	709
3	High-tide mudflat	91	645
4	Mixed area 1	16	123
5	Mixed area 2	95	774
6	Mixed area 3	75	731
7	Oil field	72	612
8	Sea	135	756
9	Intertidal phragmite	86	758
10	Ecological reservoir	75	563
11	Suaeda salsa	67	610
12	Salt fields	81	663
13	Arable land	68	681
14	Phragmite	83	756
15	Woodland	49	502
16	Lotus pond	75	461
17	Typha orientals presl	6	30
18	Robina pseudoacacia	65	516
19	Pond	86	763
20	Yellow river	88	771
Total		1481	12167

classes, and the numbers of training and testing samples of each class are listed in Table III. The false-color image of Yellow River Estuary and the distribution of those samples are shown in Fig. 8.

B. Analysis of Patch Size

In this section, the influence of HSI patch size on the classification performance is analyzed experimentally. The patch size is chosen from the set {7, 9, 11, 13}. For the training of the proposed miniCAN and CAN, we use the Adam optimizer with $\beta_1 = 0.9$, $\beta_2 = 0.99$, and the learning rate is decayed by multiplying $\gamma = 0.5$ every 20 epochs. The initial learning rate is set as 0.001 for both Gaofeng State Owned Forest Farm and Houston2013 datasets, and 0.1 for the Yellow River Estuary dataset empirically in the subsequent experiments. The overall accuracy (OA) is used to evaluate the classification performance.

As shown in Fig. 9 (GSOFF and YSE are the abbreviations of Gaofeng State Owned Farm and Yellow River Estuary, respectively), a general trend is that OA is increased with the increase in the patch size. This is because CAN and miniCAN gather more information from a larger HSI patch so that better classification performance is obtained. However, the number of training samples is limited, and more trainable parameters are needed with the increase in the patch size. As a result, CAN is overfitted with a large HSI patch. OAs for Gaofeng State Owned Forest Farm and Houston2013 datasets are slightly worse when the patch size is 13, as shown in Fig. 9. For the proposed CAN, the best patch sizes for Gaofeng State Owned Farm Forest, Houston2013, and Yellow River Estuary datasets are 11, 9, and 13, respectively. However, competitive classification performance for Yellow River Estuary is also achieved in patch size 11. To reduce computational complexity, the patch size for Yellow River Estuary is set as 11. Therefore, the patch sizes for the three datasets are set as 11, 9, and 11,

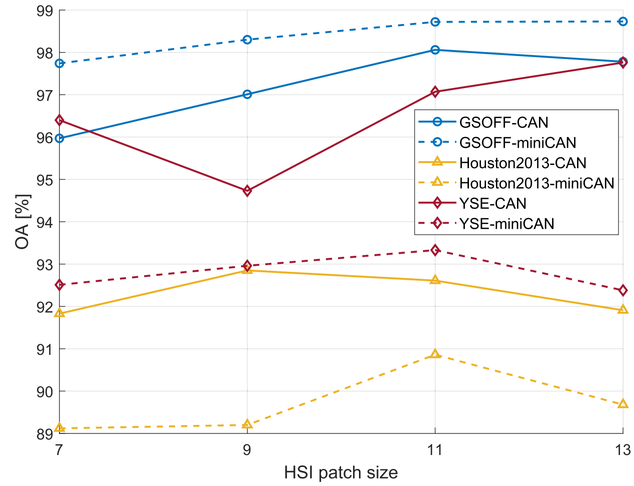

Fig. 9. OA [%] versus HSI patch size on Gaofeng State Owned Farm Forest, Houston 2013, and Yellow River Estuary datasets for the proposed CAN.

respectively, in the subsequent experiments. For the proposed miniCAN, the best patch sizes for the three datasets are all 11, so the patch size is set as 11 for miniCAN in the subsequent experiments. Note that 11 is empirically a good number for patch size that is recommended to other datasets.

C. Ablation Studies

The ablation experiment is conducted to verify the effectiveness of different parts in the proposed miniCAN and CAN, i.e., the scaled dot-product central attention (SDPCA), the dense connection of hidden layer (DCHL), and the dense connection of the central pixel of HSI patch (DCCP). If SDPCA is not used, then only 1-D convolution is used for feature extraction in every layer, and the number of layers is the same as CAN or miniCAN. OA is used to evaluate classification performance.

As listed in Table IV, after SDPCA is introduced into HSI feature extraction, the improvements in classification performance are significant. In miniCAN, the overall accuracy increases from 64.67% to 98.66%, 82.54% to 90.86%, and 83.39% to 92.24% for Gaofeng State Owned Forest Farm, Houston2013, and Yellow River Estuary datasets, respectively. In CAN, the overall accuracy increases from 63.78% to 97.53%, 80.88% to 90.94%, and 79.96% to 95.66% for the Gaofeng State Owned Forest Farm, Houston2013, and Yellow River Estuary datasets, respectively. Therefore, the effectiveness of SDPCA for HSI spectral–spatial feature extraction is confirmed. In addition, the performance of DCHL and DCCP is different in three experimental datasets for miniCAN and CAN. In miniCAN, the performance gaining from DCCP is nearly 0.9% and 0.24%, respectively, for the Houston2013 and Yellow River Estuary datasets. Nevertheless, the gaining is insignificant for the Gaofeng State Owned Forest Farm dataset, i.e., only 0.06%. In CAN, the performance gained from DCHL/DCCP is significant for both Houston2013 and Yellow River Estuary datasets, i.e., 0.8%/1% and 0.3%/0.5%, respectively. For the Gaofeng State Owned Forest Farm dataset, the performance gained from DCCP is insignificant, i.e., only 0.01%, but that from DCHL is significant, i.e., 0.4%. However,

Accepted Manuscript
Version of record at: https://doi.org/10.1109/TNNLS.2022.3155114

TABLE IV

ABLATION EXPERIMENTS FOR THE PROPOSED MINICAN AND WITH RESPECT TO THE CLASSIFICATION PERFORMANCE (OVERALL ACCURAY [%]) FOR GAOFENG STATE OWNED FOREST FARM, HOUSTON2013, AND YELLOW RIVER ESTUARY DATASETS

	SDPCA	DCHL	DCCP	Gaofeng State Owned Forest Farm	Houston2013	Yellow River Estuary
minCAN	✗	✗	✗	64.67	82.54	83.39
	✓	✗	✗	98.66	90.62	92.24
	✓	✗	✓	98.72	90.86	93.33
CAN	✗	✗	✗	63.78	80.88	79.96
	✓	✗	✗	97.53	90.94	95.66
	✓	✓	✗	97.98	91.71	96.08
	✓	✗	✓	97.54	92.05	96.23
	✓	✓	✓	98.06	92.85	97.07

TABLE V

CLASSIFICATION PERFORMANCE [%] OF DIFFERENT METHODS FOR THE GAOFENG STATE OWNED FOREST FARM DATASET

Class No.	OTVCA [20]	CNNHSI [21]	RPCACNN [23]	SSCNN [24]	CDCNN [25]	DRCNN [26]	BERTHSI [36]	FusCNNGCN [33]	3DCNN [27]	miniCAN	CAN
1	95.34	86.83	83.48	91.61	92.05	98.00	96.99	95.78	95.43	**98.91**	98.62
2	**99.70**	11.73	34.27	75.51	77.43	92.61	92.94	89.33	92.84	98.94	95.84
3	**99.24**	3.05	20.04	85.53	73.56	92.74	93.80	84.93	97.08	99.04	97.08
4	**100**	77.99	52.23	86.07	91.62	99.28	97.46	98.06	96.84	99.87	99.75
5	**99.93**	61.86	66.16	96.73	99.09	99.58	92.44	98.48	97.51	99.68	99.77
6	83.31	45.59	29.80	85.07	76.52	86.41	90.09	83.82	84.41	**95.35**	91.71
7	99.89	1.16	6.91	70.28	73.86	96.91	83.58	92.67	94.28	**99.93**	99.12
8	99.03	58.80	72.03	95.52	94.92	99.06	**99.97**	97.77	98.47	99.67	99.48
9	91.30	45.53	54.82	68.81	77.59	89.16	93.46	83.40	90.74	**97.89**	97.19
10	98.43	97.63	98.13	99.10	99.66	99.76	99.69	99.49	99.01	98.84	**99.52**
11	**99.98**	94.78	95.74	99.15	99.22	99.83	98.34	99.81	99.69	99.64	99.72
12	0.00	7.81	0.00	90.63	65.63	26.56	68.75	48.44	45.31	**93.75**	89.06
OA	95.63	69.59	67.48	90.32	90.60	96.50	95.12	94.42	94.80	**98.72**	98.06
AA	88.85	49.40	51.13	87.00	85.09	89.99	92.29	89.33	90.97	**98.46**	97.24
Kappa*100	94.48	60.91	58.69	87.79	88.11	95.57	93.84	92.95	93.45	**98.39**	97.54

TABLE VI

CLASSIFICATION PERFORMANCE [%] OF DIFFERENT METHODS FOR THE HOUSTON2013 DATASET

Class No.	OTVCA [20]	CNNHSI [21]	RPCACNN [23]	SSCNN [24]	CDCNN [25]	DRCNN [26]	BERTHSI [36]	FusCNNGCN [33]	3DCNN [27]	miniCAN	CAN
1	80.72	82.53	82.34	82.53	82.43	82.34	83.00	82.62	80.63	81.96	**89.27**
2	79.04	83.55	82.33	**84.96**	**84.96**	84.77	83.36	78.10	**84.96**	84.77	84.30
3	99.21	99.60	85.35	98.81	**100**	88.71	91.09	91.68	95.84	**100**	99.01
4	84.75	98.58	91.10	92.23	93.37	92.33	93.56	96.31	71.40	93.37	**100**
5	98.20	98.67	93.47	99.24	**100**	99.91	99.81	**100**	90.06	99.53	99.91
6	94.41	99.30	83.92	93.01	95.80	99.30	98.60	98.60	**100**	**100**	**100**
7	89.93	71.08	74.16	94.12	79.38	89.09	87.22	91.23	**93.56**	87.13	87.13
8	49.76	62.58	63.34	68.95	71.23	67.90	68.76	85.85	85.57	**92.69**	90.60
9	85.08	59.87	76.02	80.64	78.75	78.00	77.15	72.71	72.71	87.54	**91.03**
10	67.37	96.24	46.04	43.24	49.71	63.51	48.46	61.49	42.57	96.04	**97.10**
11	74.10	81.69	66.03	59.49	79.22	72.68	82.73	79.51	65.65	82.16	**85.29**
12	80.02	61.19	60.42	90.11	78.39	91.64	86.17	92.03	78.48	92.70	**95.58**
13	58.25	59.30	78.95	85.61	**92.63**	89.47	86.32	90.53	67.02	91.58	**92.63**
14	**100**	**100**	98.38	95.95	**100**	**100**	99.60	99.60	99.19	**100**	**100**
15	97.25	98.10	93.23	94.50	**100**	83.09	98.94	90.70	84.36	**100**	99.58
OA	80.62	81.30	75.60	81.63	82.32	83.28	82.90	85.23	78.25	90.96	**92.85**
AA	82.54	83.49	78.34	84.23	85.73	85.52	85.65	87.40	80.80	92.63	**94.10**
Kappa*100	79.13	79.69	73.66	80.05	80.95	81.94	81.51	83.96	76.49	90.18	**92.24**

the performance gained from the combination of DCHL and DCCP is consistent for all the three datasets, i.e., 0.5%, 1.9%, and 1.4%, respectively, which confirms the effectiveness of the designed DCHL and DCCP.

D. Classification Performance

To validate the effectiveness of the proposed miniCAN and CAN, experiments are conducted on the three datasets in comparison with the aforementioned state-of-the-art methods, i.e., OTVCA [20], CNNHSI [21], RPCACNN [23], SSCNN [24], CDCNN [25], DRCNN [26], BERT-HSI [36], FusCNNGCN [33], and 3DCNN [27]. The hyperparameters of those methods are set by their recommendations. As shown in our previous work [42], length normalization (LN) can reduce the intraclass difference. However, this normalization may not suit some comparison methods. To be fair, normalization

Accepted Manuscript
Version of record at: https://doi.org/10.1109/TNNLS.2022.3155114

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

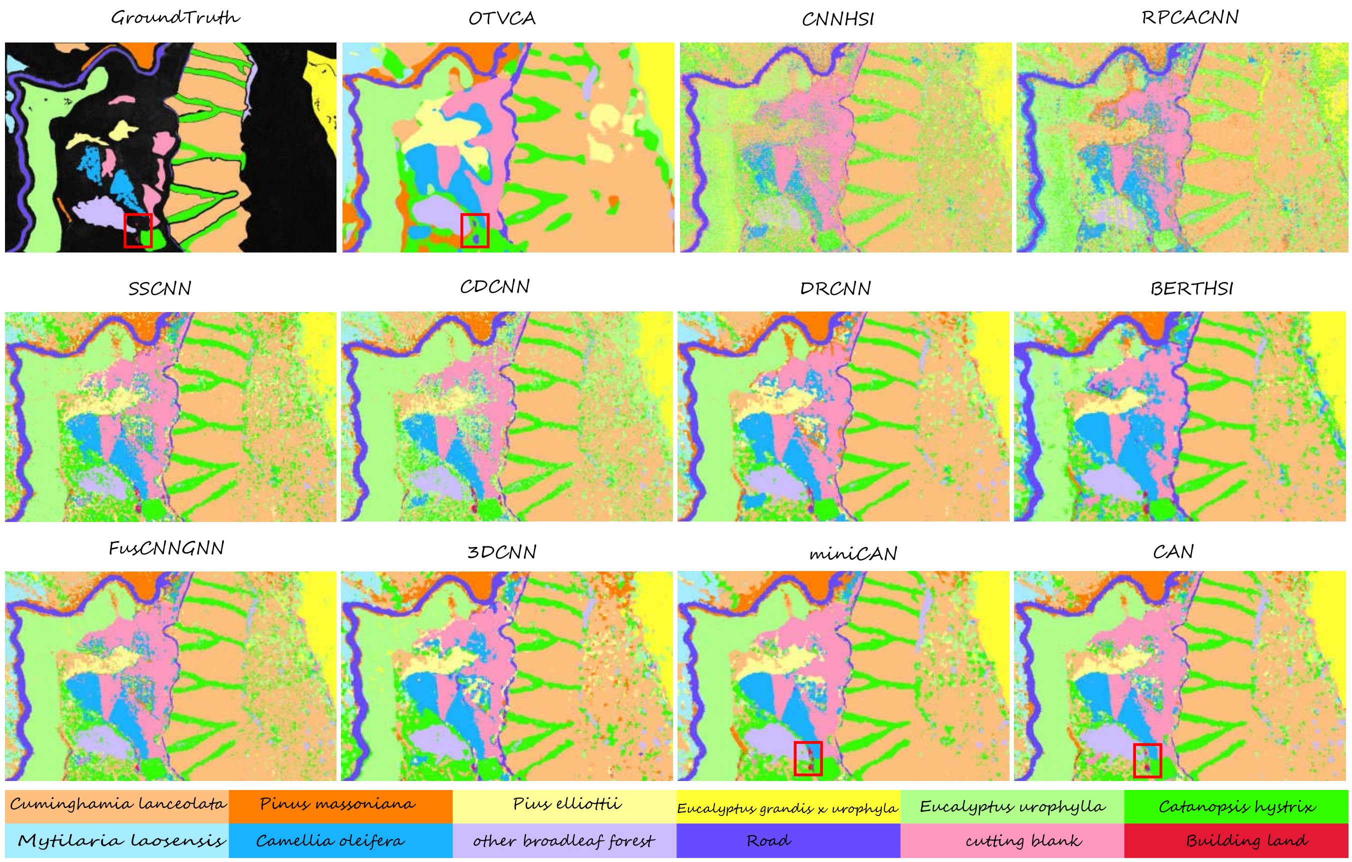

Fig. 10. Ground truth and classification maps of the Gaofeng State Owned Forest Farm dataset produced by different methods.

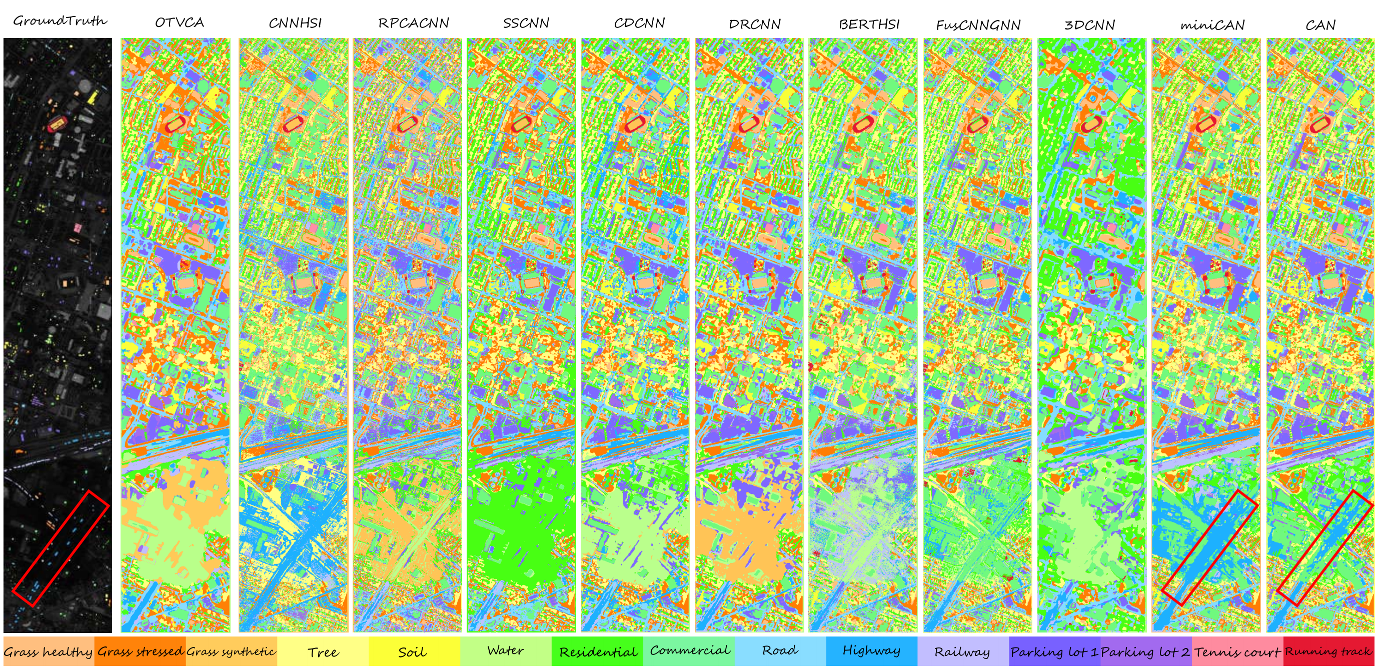

Fig. 11. Ground truth and classification maps of the Houston2013 dataset produced by different methods.

methods, including standardization, 0-1 normalization, and LN, are all tried to preprocess the three datasets for all the comparison methods, and the best classification results based on the preprocessed dataset by the three normalization methods are chosen. The class-specific accuracy (CA), OA, average accuracy (AA), and Kappa coefficient are employed

Accepted Manuscript
Version of record at: https://doi.org/10.1109/TNNLS.2022.3155114

TABLE VII
CLASSIFICATION PERFORMANCE [%] OF DIFFERENT METHODS FOR THE YELLOW RIVER ESTUARY DATASET

Class No.	OTVCA [20]	CNNHSI [21]	RPCACNN [23]	SSCNN [24]	CDCNN [25]	DRCNN [26]	BERTHSI [36]	FusCNNGCN [33]	3DCNN [27]	miniCAN	CAN
1	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	89.91	**100**	99.33
2	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**
3	**100**	97.52	**100**	95.97	99.22	**100**	**100**	**100**	**100**	**100**	**100**
4	**100**	82.93	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**
5	44.32	75.19	66.93	85.66	20.41	48.84	43.15	25.58	23.00	43.15	**95.35**
6	**100**	73.60	64.71	81.81	86.05	90.70	**100**	84.27	74.83	**100**	84.40
7	**100**	54.58	46.90	69.61	**100**	97.71	**100**	98.37	**100**	**100**	**100**
8	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**
9	**100**	99.21	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**
10	**100**	86.50	78.69	89.52	**100**	**100**	**100**	97.87	**100**	**100**	**100**
11	92.79	96.39	73.11	51.80	91.80	97.38	45.57	42.79	89.02	97.21	**100**
12	90.95	63.95	65.91	84.16	70.44	73.30	70.29	62.90	**97.74**	62.75	88.99
13	48.90	60.35	47.28	46.99	98.97	91.04	77.97	99.41	89.57	**100**	96.77
14	**100**	96.69	91.80	89.81	**100**	**100**	**100**	**100**	**100**	**100**	**100**
15	**100**	79.08	79.88	73.90	86.25	99.60	92.03	82.67	98.80	97.81	94.22
16	**95.44**	60.95	30.37	77.22	87.64	86.55	64.64	78.09	84.60	79.18	84.82
17	**100**	93.33	10.00	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**
18	99.81	81.98	30.43	33.91	0.00	22.09	70.74	62.79	92.05	**100**	99.42
19	**100**	81.65	85.45	97.64	97.90	**100**	**100**	**100**	**100**	**100**	99.48
20	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**
OA	92.56	84.73	77.90	84.00	86.57	90.15	87.89	86.09	90.75	93.33	**97.07**
AA	93.61	84.20	73.57	83.90	86.93	90.36	88.22	86.74	91.98	94.00	**97.14**
Kappa*100	92.13	83.83	76.59	83.06	85.78	89.58	87.19	85.27	90.21	92.95	**96.90**

to evaluate the classification performance. Classification performances of different methods on the Gaofeng State Owned Forest Farm, Houston2013, and Yellow River Estuary datasets are listed in the Tables V–VII, respectively.

In Tables V–VII, for all experiments on the three datasets, the classification performance of spectral-only feature extraction method CNNHSI is generally worse than those of spectral–spatial feature extraction methods except for RPCACNN with the worst performance where randomized PCA is performed on HSI patch beforehand and may reduce the discriminative information from the complete HSI patch. For spectral–spatial feature extraction methods, 3DCNN performs better in Gaofeng State Owned Forest Farm and Yellow River Estuary datasets but worse in the Houston2013 dataset since Houston2013 as an urban HSI dataset is more complex than the other two datasets. However, the classification performances of OTVCA, DRCNN, BERTHSI, and FusCNNGCN are consistently good in all three datasets. The performance of the proposed miniCAN/CAN is superior and better than all the other aforementioned state-of-the-art methods by approximately 2.2%/1.6% in Gaofeng State Owned Forest Farm, 5.7%/7.6% in Houston2013, and 0.8%/4.5% in Yellow River Estuary, respectively, over the second-highest other methods in terms of OA. In particular, the result of the proposed miniCAN is a litter better than that of the proposed CAN in the Gaofeng State Owned Forest Farm dataset. This may result from that the training samples are randomly sampled from the ground-truth map and evenly distributed, and the testing samples are neighboring pixels of these training samples, as shown in Fig. 9. It may reveal that this classification task is not complex compared with the other two tasks. As a result, as a less complex spectral–spatial feature extraction model, miniCAN may achieve a better classification result, indicating that miniCAN with lower computational complexity can perform as good as CAN in noncomplex scenarios.

To visually demonstrate the effectiveness of miniCAN and CAN, classification maps of all the aforementioned methods are illustrated in Figs. 10–12. The produced maps by CNNHSI contain many salt-and-pepper pixels, which confirms the poor classification results listed in Tables V–VII. This is because spatial information is not utilized in CNNHSI. By comparison, although the classification performance of OTVCA, as listed in Tables V–VII is acceptable, the produced maps are not satisfactory since they are oversmoothed, and many small objects are submerged. For example, for Gaofeng State Owned Forest Farm dataset, the building land in the red rectangle has been covered by other classes, as shown in Fig. 10. However, the proposed miniCAN and CAN produce more accurate and spatially smoother classification maps with fewer mislabeled pixels than other methods, which are consistent with the results listed in Tables V–VII. In particular, the building land is well-preserved in the maps of Gaofeng State Owned Forest Farm produced by miniCAN and CAN. For the sea area marked by the red rectangle in Fig. 12, it is easily mislabeled with the Yellow River or ecological reservoir by other methods. However, that area is correctly labeled as sea by miniCAN and CAN. For the Houston2013 dataset, the challenge mainly comes from the shadow area covered by the cloud where pixels of the same class in the cloud-covered area and other areas have different spectral characteristics. What is worse, most of the training data come from the areas that are not covered by the cloud. Not surprisingly, in the cloud area, the highway is mislabeled by all the other state-of-the-art methods except for CNNHSI (see the red rectangle area in Fig. 11). However, it is well-classified by miniCAN and CAN. This is because CAN and miniCAN extract the most effective spatial information around a patch and reduce the interference

Accepted Manuscript
Version of record at: https://doi.org/10.1109/TNNLS.2022.3155114

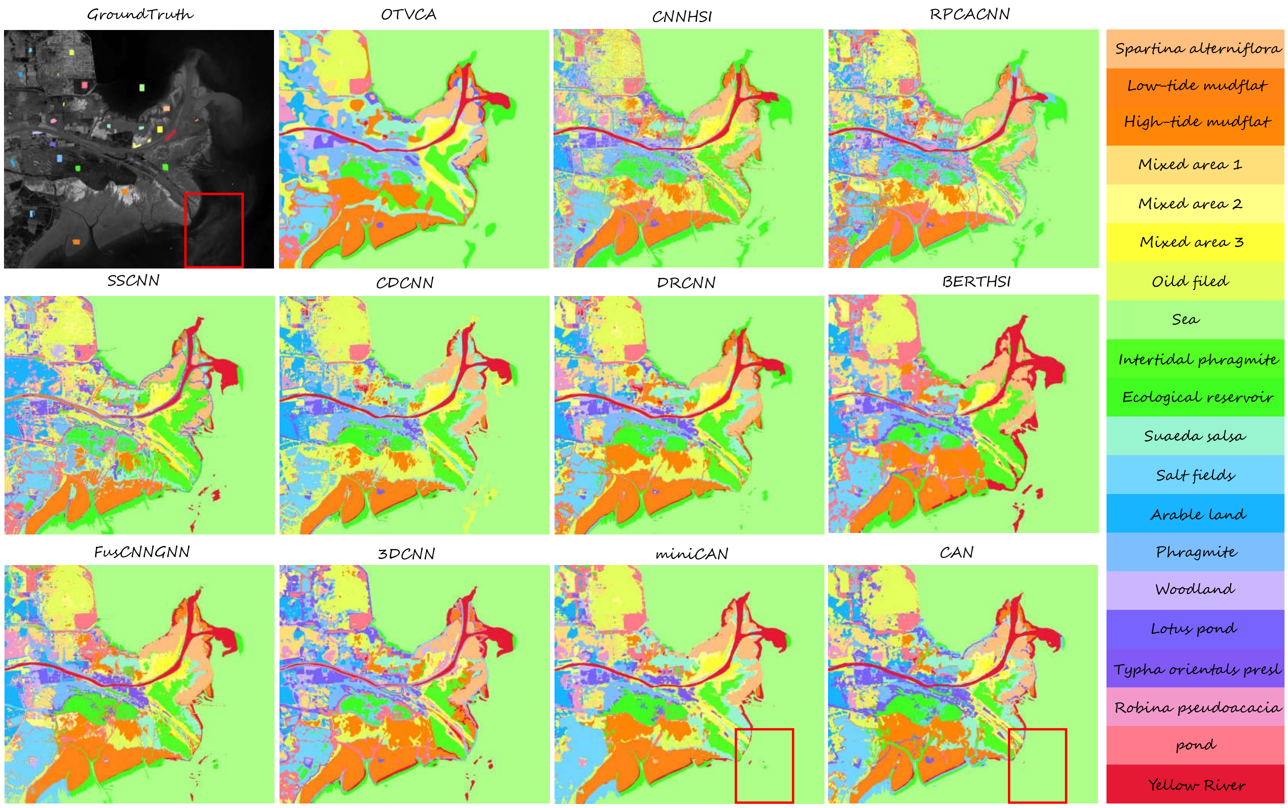

Fig. 12. Ground truth and classification maps of the Yellow River Estuary dataset produced by different methods.

TABLE VIII
COMPUTATION TIME [SECOND] OF DIFFERENT METHODS FOR THE GAOFENG STATE OWNED FOREST FARM, HOUSTON2013, AND YELLOW RIVER DATASETS

		Gaofeng State Owned Forest Farm	Houston2013	Yellow River Estuary
CNNHSI	Training time (s)	116.93	33.07	29.86
	Testing time (s)	18.20	25.30	92.23
BERTHSI	Training time (s)	374.21	103.83	89.76
	Testing time (s)	35.96	51.21	215.88
DRCNN	Training time (s)	1461.72	443.96	373.73
	Testing time (s)	36.76	53.77	247.92
CAN	Training time (s)	961.11	143.73	142.7
	Testing time (s)	57.08	73.25	220.67
miniCAN	Training time (s)	316.77	89.79	73.39
	Testing time (s)	35.80	50.44	207.79

information that may prevent from correct classification. As a result, most essential properties of a pixel in cloud-covered or other areas are extracted, and by these properties, the pixels in the cloud-covered area can be discriminated to some extent by the spectral and label information from other areas.

To illustrate the computational complexities of the proposed CAN and miniCAN compared to other methods, Table VIII provides the computation time of training and testing of several methods on the aforementioned three datasets by using Pytorch in Python on AMD Ryzen 7 5800X eight-core Processor with 32-GB RAM and NVIDIA Geforce RTX 3070 with 8-GB RAM. For all three datasets, the training and testing costs of CNNHSI are the lowest since the network of CNNHSI is the simplest. By comparison, the training costs of DRCNN are the highest since DRCNN contains two stages, i.e., the pretraining and fine-tuning stages. Followed by DRCNN are CAN, BERTHSI, and miniCAN. As the simplified version of CAN, miniCAN is nearly half of that of CAN in Houston2013 and Yellow River Estuary datasets, and a third in Gaofeng State Owned Forest Farm dataset in terms of training time.

V. CONCLUSION

In this article, an efficient spectral–spatial feature extraction network called CAN has been proposed based on two principles for HSI classification. In CAN, the HSI-tailored scaled dot-product central attention and the multihead central attention are utilized to extract spectral–spatial information from a query pixel and the pixels that are similar to the query pixel on an HSI patch. Also, to reuse the former features and the spectral information of the query pixel, HSI-tailored dense connections are utilized. To reduce the complexity of CAN, miniCAN as a simplified version of CAN has been investigated with lower computational complexity. Experiments on three HSI datasets of different scenarios have demonstrated that the proposed CAN and miniCAN outperform the existing state-of-the-art HSI feature extraction methods, which has confirmed their effectiveness.

REFERENCES

[1] J. Peng, L. Li, and Y. Y. Tang, "Maximum likelihood estimation-based joint sparse representation for the classification of hyperspectral remote sensing images," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 30, no. 6, pp. 1790–1802, Jun. 2019.

[2] C. Liu, J. Li, L. He, A. J. Plaza, S. Li, and B. Li, "Naive Gabor networks for hyperspectral image classification," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 32, no. 1, pp. 376–390, Jan. 2021.

Accepted Manuscript
Version of record at: https://doi.org/10.1109/TNNLS.2022.3155114

[3] Y. Duan, H. Huang, Z. Li, and Y. Tang, "Local manifold-based sparse discriminant learning for feature extraction of hyperspectral image," *IEEE Trans. Cybern.*, vol. 51, no. 8, pp. 4021–4034, Aug. 2021.

[4] Y. Duan, H. Huang, and Y. Tang, "Local constraint-based sparse manifold hypergraph learning for dimensionality reduction of hyperspectral image," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 1, pp. 613–628, Jan. 2021.

[5] J. Yue, L. Fang, H. Rahmani, and P. Ghamisi, "Self-supervised learning with adaptive distillation for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–13, 2022, doi: 10.1109/TGRS.2021.3057768.

[6] J. Xie, L. Fang, B. Zhang, J. Chanussot, and S. Li, "Super resolution guided deep network for land cover classification from remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–12, 2022, doi: 10.1109/TGRS.2021.3120891.

[7] B. Schölkopf, A. Smola, and K.-R. Müller, "Nonlinear component analysis as a kernel eigenvalue problem," *Neural Comput.*, vol. 10, no. 5, pp. 1299–1319, Jul. 1998.

[8] K. Fukunaga, *Introduction to Statistical Pattern Recognition*. Amsterdam, The Netherlands: Elsevier, 2013.

[9] S. Yan, D. Xu, B. Zhang, H.-J. Zhang, Q. Yang, and S. Lin, "Graph embedding and extensions: A general framework for dimensionality reduction," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 29, no. 1, pp. 40–51, Jan. 2007.

[10] B. Cheng, J. Yang, S. Yan, Y. Fu, and T. S. Huang, "Learning with ℓ^1-graph for image analysis," *IEEE Trans. Image Process.*, vol. 19, no. 4, pp. 858–866, Apr. 2010.

[11] N. H. Ly, Q. Du, and J. E. Fowler, "Sparse graph-based discriminant analysis for hyperspectral imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 52, no. 7, pp. 3872–3884, Jul. 2014.

[12] N. Liu, W. Li, and Q. Du, "Unsupervised feature extraction for hyperspectral imagery using collaboration-competition graph," *IEEE J. Sel. Topics Signal Process.*, vol. 12, no. 6, pp. 1491–1503, Dec. 2018.

[13] J. A. Benediktsson, J. A. Palmason, and J. R. Sveinsson, "Classification of hyperspectral data from urban areas based on extended morphological profiles," *IEEE Trans. Geosci. Remote Sens.*, vol. 43, no. 3, pp. 480–491, Mar. 2005.

[14] M. Dalla Mura, J. A. Benediktsson, B. Waske, and L. Bruzzone, "Morphological attribute profiles for the analysis of very high resolution images," *IEEE Trans. Geosci. Remote Sens.*, vol. 48, no. 10, pp. 3747–3762, Oct. 2010.

[15] W. Li, C. Chen, H. Su, and Q. Du, "Local binary patterns and extreme learning machine for hyperspectral imagery classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 53, no. 7, pp. 3681–3693, Jul. 2015.

[16] A. Mohan, G. Sapiro, and E. Bosch, "Spatially coherent nonlinear dimensionality reduction and segmentation of hyperspectral images," *IEEE Geosci. Remote Sens. Lett.*, vol. 4, no. 2, pp. 206–210, Apr. 2007.

[17] H. Pu, Z. Chen, G.-M. Jiang, and B. Wang, "A novel spatial–spectral similarity measure for dimensionality reduction and classification of hyperspectral imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 52, no. 11, pp. 7008–7022, Nov. 2014.

[18] H. Huang, G. Shi, H. He, Y. Duan, and F. Luo, "Dimensionality reduction of hyperspectral imagery based on spatial–spectral manifold learning," *IEEE Trans. Cybern.*, vol. 50, no. 6, pp. 2604–2616, Jun. 2020.

[19] L. Pan, H.-C. Li, Y.-J. Deng, F. Zhang, X.-D. Chen, and Q. Du, "Hyperspectral dimensionality reduction by tensor sparse and low-rank graph-based discriminant analysis," *Remote Sens.*, vol. 9, no. 5, p. 452, May 2017.

[20] B. Rasti, M. O. Ulfarsson, and J. R. Sveinsson, "Hyperspectral feature extraction using total variation component analysis," *IEEE Trans. Geosci. Remote Sens.*, vol. 54, no. 12, pp. 6976–6985, Dec. 2016.

[21] W. Hu, Y. Huang, L. Wei, F. Zhang, and H. Li, "Deep convolutional neural networks for hyperspectral image classification," *J. Sensors*, vol. 2015, pp. 1–12, Jan. 2015.

[22] W. Li, G. Wu, F. Zhang, and Q. Du, "Hyperspectral image classification using deep pixel-pair features," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 2, pp. 844–853, Feb. 2017.

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

[24] S. Mei, J. Ji, J. Hou, X. Li, and Q. Du, "Learning sensor-specific spatial–spectral features of hyperspectral images via convolutional neural networks," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 8, pp. 4520–4533, Aug. 2017.

[25] H. Lee and H. Kwon, "Going deeper with contextual CNN for hyperspectral image classification," *IEEE Trans. Image Process.*, vol. 26, no. 10, pp. 4843–4855, Oct. 2017.

[26] 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, Jun. 2018.

[27] B. Zhang, L. Zhao, and X. Zhang, "Three-dimensional convolutional neural network model for tree species classification using airborne hyperspectral images," *Remote Sens. Environ.*, vol. 247, Sep. 2020, Art. no. 111938.

[28] S. Jia *et al.*, "A lightweight convolutional neural network for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 5, pp. 4150–4163, May 2021.

[29] 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, Jun. 2019.

[30] Z. Zheng, Y. Zhong, A. Ma, and L. Zhang, "FPGA: Fast patch-free global learning framework for fully end-to-end hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 8, pp. 5612–5626, Aug. 2020.

[31] D. Wang, B. Du, L. Zhang, and S. Chu, "Hyperspectral image classification based on multi-scale information compensation," *Remote Sens. Lett.*, vol. 11, no. 3, pp. 293–302, Mar. 2020.

[32] T. N. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," in *Proc. Int. Conf. Learn. Represent.*, 2017, pp. 1–14.

[33] D. Hong, L. Gao, J. Yao, B. Zhang, A. Plaza, and J. Chanussot, "Graph convolutional networks for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 7, pp. 5966–5978, Jul. 2021.

[34] K. Xu, H. Huang, P. Deng, and Y. Li, "Deep feature aggregation framework driven by graph convolutional network for scene classification in remote sensing," *IEEE Trans. Neural Netw. Learn. Syst.*, early access, Apr. 15, 2021, doi: 10.1109/TNNLS.2021.3071369.

[35] A. Vaswani *et al.*, "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 5998–6008.

[36] J. He, L. Zhao, H. Yang, M. Zhang, and W. Li, "HSI-BERT: Hyperspectral image classification using the bidirectional encoder representation from transformers," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 1, pp. 165–178, Jan. 2020.

[37] C. Pu, H. Huang, and L. Yang, "An attention-driven convolutional neural network-based multi-level spectral–spatial feature learning for hyperspectral image classification," *Expert Syst. Appl.*, vol. 185, Dec. 2021, Art. no. 115663.

[38] S. Li, W. Song, L. Fang, Y. Chen, P. Ghamisi, and J. A. Benediktsson, "Deep learning for hyperspectral image classification: An overview," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 9, pp. 6690–6709, Sep. 2019.

[39] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," 2018, *arXiv:1810.04805*.

[40] A. Dosovitskiy *et al.*, "An image is worth 16×16 words: Transformers for image recognition at scale," in *Proc. Int. Conf. Learn. Represent.*, 2020.

[41] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 4700–4708.

[42] H. Liu, W. Li, X.-G. Xia, M. Zhang, C.-Z. Gao, and R. Tao, "Spectral shift mitigation for cross-scene hyperspectral imagery classification," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 6624–6638, 2021.

Huan Liu received the B.S. degree from the China University of Mining and Technology, Xuzhou, China, in 2013, and the M.S. degree from the Beijing Institute of Technology, Beijing, China, in 2016, where he is currently pursuing the Ph.D. degree with the School of Information and Electronics.

His research interests include image processing, image classification, and transfer learning.

Accepted Manuscript
Version of record at: https://doi.org/10.1109/TNNLS.2022.3155114

Wei Li (Senior Member, IEEE) received the B.E. degree in telecommunications engineering from Xidian University, Xi'an, China, in 2007, the M.S. degree in information science and technology from Sun Yat-Sen University, Guangzhou, China, in 2009, and the Ph.D. degree in electrical and computer engineering from Mississippi State University, Starkville, MS, USA, in 2012.

He spent one year as a Post-Doctoral Researcher with the University of California, Davis, CA, USA. He is currently a Professor with the School of Information and Electronics, Beijing Institute of Technology, Beijing, China. He has published more than 160 peer-reviewed articles and 100 conference papers. His research interests include hyperspectral image analysis, pattern recognition, and target detection.

Dr. Li is currently serving as an Associate Editor for the IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING (TGRS). He has served as an Associate Editor for the IEEE JOURNAL OF SELECTED TOPICS IN APPLIED EARTH OBSERVATIONS AND REMOTE SENSING (JSTARS) and the IEEE SIGNAL PROCESSING LETTERS (SPL). He received the JSTARS Best Reviewer in 2016 and TGRS Best Reviewer Award in 2020 from the IEEE Geoscience and Remote Sensing Society (GRSS) and the Outstanding Paper Award at IEEE International Workshop on Hyperspectral Image and Signal Processing: Evolution in Remote Sensing (Whispers) in 2019.

Xiang-Gen Xia (Fellow, IEEE) received the B.S. degree in mathematics from Nanjing Normal University, Nanjing, China, in 1983, and the M.S. degree in mathematics from Nankai University, Tianjin, China, in 1986, and the Ph.D. degree in electrical engineering from the University of Southern California, Los Angeles, in 1992.

He was a Senior/Research Staff Member with Hughes Research Laboratories, Malibu, CA, USA, from 1995 to 1996. In September 1996, he joined the Department of Electrical and Computer Engineering, University of Delaware, Newark, Delaware, where he is the Charles Black Evans Professor. He is the author of the book *Modulated Coding for Intersymbol Interference Channels* (New York, Marcel Dekker, 2000). His current research interests include space-time coding, MIMO and OFDM systems, digital signal processing, and SAR and ISAR imagings.

Dr. Xia received the National Science Foundation (NSF) Faculty Early Career Development (CAREER) Program Award in 1997, the Office of Naval Research (ONR) Young Investigator Award in 1998, and the Outstanding Overseas Young Investigator Award from the National Nature Science Foundation of China in 2001. He received the 2019 Information Theory Outstanding Overseas Chinese Scientist Award from The Information Theory Society of Chinese Institute of Electronics. He has served as an Associate Editor for numerous international journals including IEEE TRANSACTIONS ON SIGNAL PROCESSING, the IEEE TRANSACTIONS ON WIRELESS COMMUNICATIONS, the IEEE TRANSACTIONS ON MOBILE COMPUTING, and the IEEE TRANSACTIONS ON VEHICULAR TECHNOLOGY. He is a Technical Program Chair of the Signal Processing Symp., Globecom 2007 in Washington, DC, USA, and the General Co-Chair of ICASSP 2005 in Philadelphia.

Mengmeng Zhang received the B.S. degree in computer science and technology from the Qingdao University of Science and Technology, Qingdao, China, in 2014, and the Ph.D. degree in control science and engineering from the Beijing University of Chemical Technology, Beijing, China, in 2019.

She is currently a Post-Doctoral Researcher with the School of Information and Electronics, Beijing Institute of Technology, Beijing, China. Her research interests include remote sensing image processing and pattern recognition. Her current research interests include remote sensing, image process, and pattern recognition.

Chen-Zhong Gao received the B.S. degree in information and communication engineering from the Beijing Institute of Technology, Beijing, China, where he is currently pursuing the Ph.D. degree with the School of Information and Electronics. His main research direction is remote sensing and image processing.

Ran Tao (Senior Member, IEEE) was born in 1964. He received the B.S. degree from the Electronic Engineering Institute of PLA, Hefei, China, in 1985, and the M.S. and Ph.D. degrees from the Harbin Institute of Technology, Harbin, China, in 1990 and 1993, respectively.

In 2001, he was a Senior Visiting Scholar with the University of Michigan, Ann Arbor, MI, USA. He is currently a Professor with the School of Information and Electronics, Beijing Institute of Technology, Beijing, China. He has three books and over 100 peer-reviewed journal articles. His research interests include fractional Fourier transform and its applications, theory, and technology for radar and communication systems.

Dr. Tao is a member of the Wireless Communication and Signal Processing Commission of the International Union of Radio Science (URSI). He was a recipient of the National Science Foundation of China for Distinguished Young Scholars in 2006, and the First Prize of Science and Technology Progress in 2006 and 2007, and the First Prize of Natural Science in 2013, both awarded by the Ministry of Education. He was a Distinguished Professor of the Changjiang Scholars Program in 2009. He was a Chief Professor of the Program for Changjiang Scholars and Innovative Research Team in University from 2010 to 2012. He has been a Chief Professor of the Creative Research Groups of the National Natural Science Foundation of China since 2014. He is the Vice-Chair of the IEEE China Council and the URSI China Council. He is an Associate Editor of the IEEE SIGNAL PROCESSING LETTERS (SPL).

Accepted Manuscript
Version of record at: https://doi.org/10.1109/TNNLS.2022.3155114