Signal Processing 212 (2023) 109151

Contents lists available at ScienceDirect

Signal Processing

journal homepage: www.elsevier.com/locate/sigpro

Three-dimensional Softmax Mechanism Guided Bidirectional GRU Networks for Hyperspectral Remote Sensing Image Classification

Guoqiang Wu[a,b], Xin Ning[c,*], Luyang Hou[d,*], Feng He[e], Hengmin Zhang[f], Achyut Shankar[g,h,i]

[a] School of Mechatronical Engineering, Beijing Institute of Technology, Beijing, China
[b] Intelligent Unmanned System Overall Technology Research and Development Center, China Aerospace Science and Technology Corporation, Beijing, China
[c] High-speed Circuits and Artificial Neural Networks Laboratory, Institute of Semiconductors, Chinese Academy of Sciences, Beijing 100083, China
[d] School of Computer Science (National Pilot Software Engineering School), Beijing University of Posts and Telecommunications, Beijing 100876, China
[e] University of Science and Technology of China, Hefei, 230026, China
[f] School of Electrical and Electronic Engineering, Nanyang Technological University (NTU), Singapore
[g] WMG, University of Warwick, Coventry, CV74AL, United Kingdom
[h] Department of Computer Science and Engineering, Graphic Era Deemed to be University, Dehradun, 248002, India
[i] School of Computer Science Engineering, Lovely Professional University, Phagwara - 144411, Punjab, India

ARTICLE INFO

Article history:
Received 27 February 2023
Revised 15 April 2023
Accepted 28 April 2023
Available online 14 June 2023

Keywords:
Three-dimensional Softmax Mechanism
Neural Network
Hyperspectral Image Classification
Remote Sensing Image

ABSTRACT

Hyperspectral data is a valuable source of both spectral and spatial information. However, to enhance the classification accuracy of hyperspectral image features, it is crucial to capture the spatial spectral features of image elements. The recent years have witnessed the potentials of deep learning methods have shown great promise in the hyperspectral image classification due to their ability to model complex structures and extract multiple features in an end-to-end fashion. Since hyperspectral images can be viewed as sequential data, we propose a novel three-dimensional Softmax mechanism-guided bidirectional GRU network (TDS-BiGRU) for HSI classification. By utilizing a bidirectional GRU to process the sequence data, our method can significantly reduce the processing time. Furthermore, the proposed three-dimensional Softmax mechanism leverages three branches to capture cross-latitude interactions and calculate Softmax weights, which enables us to obtain deeper features with greater discriminative power. The experimental results demonstrate that the proposed method outperforms several prevalent algorithms on four hyperspectral remote sensing datasets. Additionally, we conduct thorough comparisons and ablation tests, which further confirm the effectiveness of our approach.

1. Introduction

In recent years, remote sensing technology has gained significant attention due to the launch of high-resolution remote sensing satellites. This technology provides a new technical solution for disaster risk warning [1–2], geological monitoring [3], and meteorological monitoring [4]. To distinguish features through remote sensing, it is necessary to ensure a spectrospatial spacing between the centroids of two reflected peaks greater than the half-wave width of each reflected wave. Traditional remote sensing can be considered as discrete sampling in spectral space and is limited to distinguishing features with obvious differences in spectral space, such as water bodies, vegetation, sand, and other features with distinct optical behaviors. In contrast, high spectral resolution remote sensing can distinguish different classes of the same feature by satisfying the requirements of continuity and spectral differentiability.

Hyperspectral image classification [28–31] is a crucial aspect of remote sensing applications, and numerous classification algorithms based on spectral information have emerged in the early stages [5–7]. Since different features possess unique spectral profiles, many semi-supervised band selection algorithms based on affinity propagation and band conversion algorithms based on linear [36] or nonlinear transformation of feature statistics have been developed to achieve feature extraction, such as PCA [8] and LDA [9]. However, these classification algorithms are susceptible to noise and mainly serve as a means of dimensionality reduction in conjunction with other methods. Kernel methods, represented by support vector machines, have become a popular option for hyperspectral classification due to their excellent performance with small samples and elegant mathematical expressions. As a generalized linear classifier with spatial mapping and exploration of deci-

* Corresponding authors.
E-mail addresses: ningxin@semi.ac.cn (X. Ning), luyang.hou@bupt.edu.cn (L. Hou), ashankar2711@gmail.com (A. Shankar).

https://doi.org/10.1016/j.sigpro.2023.109151
0165-1684/

sion boundaries through kernel functions, it has gradually become a benchmark for measuring hyperspectral classification.

The spatial distribution of hyperspectral images is complex and spectrally heterogeneous, making accurate classification difficult if only spectral features are utilized. Therefore, the importance of spatial characteristics in hyperspectral image classification [35] research has grown as image processing [34] and pattern recognition techniques [32–33] have advanced. In the existing literature, researchers have applied many spatial feature extraction techniques, including GLCM and Gabor filter. Mathematical morphological features have also gained significant attention, with researchers such as Plaza using extended morphological contours to mine spatial information via multiple morphological operations [10]. However, the above methods rely solely on spectral or spatial features, which may result in the loss of some useful information and poor classification results. To address this issue, HSI classification algorithms are proposed based on spatial spectral features, adding spatial contextual information into the pixel classifier. Nonetheless, most of these methods extract features manually, which are shallow and less robust in more complex environments. In addition, different a priori information and feature extraction methods may lead to poor generalization ability and difficulty adapting to different datasets. In terms of these, it is crucial to extract more discriminative features to distinguish subtle differences between different classes and large differences between the same classes. Therefore, automated feature extraction methods have been proposed to extract more complex and discriminative features, which can learn high-level features from raw data without relying on handcrafted features.

Recurrent neural networks (RNNs) are a powerful tool for processing continuous inputs with dynamic temporal behavior, thanks to their periodic hidden state that depends on the previous step's activation. The hyperspectral imagery (HSI) data can be viewed as a collection of ordered and continuous spectral sequences in the spectral dimension. While RNNs are typically used for processing sequential data, researchers have explored their potential in HSI classification. For instance, Mou *et al*. proposed a RNN with a parametrically corrected tanh activation function for analyzing hyperspectral sequence data [11]. They used a single pixel as input to fully exploit the correlation between spectra, but their method utilized only spectral information and not spatial information, which is not conducive to improving classification accuracy. Moreover, the authors assumed that all pixels in the same local region belong to the same ground label class [12], which is not strict and the network's classification accuracy could be affected by the presence of interfering pixels in the same local region with inconsistent labels.

Deep neural networks are one of the most prevalent classification techniques for hyperspectral images. Nonetheless, the spatial variability of spectral features and the absence of labelled samples in hyperspectral images pose significant obstacles. These obstacles cause a decline in accuracy and enormous computational difficulties for deep classification networks. In addition, local classification networks generate unsatisfactory classification outcomes for edge pixels. This paper proposes a novel three-dimensional Softmax mechanism-guided bidirectional GRU model for the classification of HSI to address these issues. The proposed model uses a bidirectional gated recurrent unit (GRU) to reduce the processing time of sequence data. The proposed three-dimensional Softmax mechanism calculates Softmax weights by capturing cross-latitude interactions via three branches, thereby yielding more discriminatively potent and in-depth features. Extensive comparisons and ablation tests show that the proposed TDS-BiGRU model achieves exceptional results on four hyperspectral remote sensing datasets. In addition, the outcomes demonstrate that the proposed model is competitive with respect to other prevalent algorithms.

The contributions of this paper are summarized as follows:

(1) A new HSI classification model is proposed based on a bidirectional GRU network guided by a three-dimensional Softmax mechanism. This addresses a limitation of existing methods, which do not consider the correlation between feature maps in spatial feature extraction and the contextual information of spectral sequences in spectral feature extraction.
(2) The proposed model leverages the fact that hyperspectral images can be treated as a type of sequence data by introducing bidirectional GRU to process sequences. This reduces the time cost of the classification process and captures cross-latitude interactions to calculate Softmax weights through a three-dimensional Softmax mechanism.
(3) The model is evaluated on four widely used hyperspectral datasets, and extensive experiments demonstrate that it is robust and capable of extracting more discriminative joint features to improve classification accuracy of hyperspectral images.

The rest of the paper is organized as follows: Section 2 presents motivation, Section 3 elaborates our approach, and Section 4 provides sufficient experiments and analysis. Section 5 presents the conclusions.

2. Motivation

Convolutional neural networks (CNNs) have emerged as a popular approach for hyperspectral remote sensing due to their ability to minimize model parameters while preserving spatial information through local connections and parameter sharing techniques. To improve the discriminative power of CNN-based hyperspectral image classification, researchers have explored various techniques. For instance, Song et al. [13] developed a deep feature fusion network that leverages residual learning to overcome the gradient disappearance issue associated with deep networks. Roy et al. [14] proposed a hybrid 2D and 3D CNN classification algorithm that learns joint spectral-spatial feature representations using 3D CNN and then applies 2D CNN to obtain a deeper spatial representation. Xie et al. [15] designed the DMC-CNN algorithm, which crops pixels into patches of different sizes and feeds them into a densely connected CNN to obtain multiscale features for final classification. Zhu et al. [16] introduced deformable convolution to adaptively adjust the convolution kernel size and shape for objects of different scales and sizes, leading to better classification performance. To further enhance the features extracted by CNNs, attention mechanisms have been introduced to adaptively distinguish higher-value spatial feature regions and optical feature spectral channels. Additionally, the CNN-based spatial-spectral dual-branching framework, which learns separate features for the spectral and spatial information of HSI and fuses and classifies the features at the end of the network, has received extensive attention in the literature.

The CNN-based hyperspectral image classification method operates at the image-level by selecting a pixel-centered neighborhood block as the input to the network, rather than individual pixels. The resulting labels are then assigned to the center pixel corresponding to the input neighborhood block to generate the final classification map. However, recent research has focused on achieving true pixel-level classification and leveraging pixel space context information through a segmentation approach. For example, Xu et al [17] proposed a spectral space fully convolutional hyperspectral classification network that directly predicts the entire hyperspectral image without expanding each pixel into a neighborhood block. Similarly, Zheng et al. [18] suggested an encoder-decoder network with lateral connections to merge spatial data and semantic information, while Jiang et al. [19] integrated a 3D

complete convolutional network with a convolutional spatial propagation network to use the spatial consistency of hyperspectral images to obtain more discriminative features with lower model complexity. Furthermore, Wang et al. [20] combined multiscale features with a self-attentive mechanism to propose a completely contextual hyperspectral scene parsing network, while Shen et al. [21] proposed an efficient deep learning network with non-local features for hyperspectral classification, which calculates the similarity between each pixel and its neighboring pixels in the same row and column and uses a round-robin operation to obtain the similarity of each pixel for the global. By reducing the model's parameters, these techniques can produce a tighter classification result map with improved accuracy.

The research described above highlights the effectiveness of deep learning-based methods for hyperspectral image classification, which enable the extraction of more robust and discriminatory features by leveraging the complex spectral information through hierarchical learning. However, while most current studies focus on the basic integration of spectral and spatial information, recent research has shown that further enhancing the classifier's ability can be achieved by fully exploiting the contextual linkages between spectral dimensions and capturing all contextual spatial correlations. Future research in this field should aim to fully leverage the contextual information to improve hyperspectral image classification accuracy and performance.

3. Methodology

Fig. 1 illustrates the proposed TDS-BiGRU method. First, we apply PCA to preprocess the HSI data, and then use the multi-scale convolution module to extract shallow features at different perceptual fields. Next, we splice and divide the multi-scale features into three groups and input them into Average Pooling softmax, GlobalAvgPool softmax, and MaxPooling softmax. We then multiply the softmax outputs separately to obtain three-dimensional Softmax weights. Global and local depth features are obtained by applying BiGRU and dense fully connected layers. Finally, we perform further semantic mining using Conv1D.

3.1. Multi-scale convolution

The use of multi-scale convolution kernels provides two key benefits, as discussed in this work. Firstly, these kernels allow for the extraction of hyperspectral image characteristics at different scales, enabling filters to learn more features. This results in a more comprehensive and accurate understanding of the data. Secondly, CNNs rely on the continuous learning of filter parameters to find the optimal values for achieving the closest match to the labels. The use of multiscale convolution kernels enhances this learning process by enabling a certain convolutional layer to learn more diverse weights and biases. This facilitates the effective and complete extraction of semantic features from hyperspectral images. Fig. 2 illustrates the multiscale convolution layer used in this work, which consists of three different scales of convolution.

X_1,X_2 and X_3 are the feature maps from different scales of convolution. To represent image features with fewer neurons, the output of the multiscale CNN can be written as:

$$Y = \phi\left\{\sum_{i=0}^{n1} w_{ij}^1 * x_i^1 + \sum_{i=0}^{n2} w_{ij}^2 * x_i^2 + \sum_{i=0}^{n3} w_{ij}^3 * x_i^3\right\} \tag{1}$$

where x_i^1,x_i^2,x_i^3,w_{ij}^1,w_{ij}^2,w_{ij}^3denote the neurons and weights from the multi-scale convolutional layer kernel, while denoting a filter in CONV_*i*. Y is fed into the BiGRU dense network as the output of the convolutional layer layers.

In computer vision models, multi-scale inference approaches are commonly employed to achieve superior results. The idea is to capture fine-grained details at larger scales, while detecting larger objects at smaller scales. This is because the network's receptive field can better comprehend the image at smaller scales. However, traditional multi-scale structures may not always be the best choice. In our approach, we use three convolutional kernel sizes (7×7, 5×5, and 3×3) to extract features.

$$\begin{aligned} Y_c{}^1 &= \phi\left\{\sum_{i=0}^{n1} w_{ij}^1 * x_i^1 + b_j^1\right\} \\ Y_d{}^1 &= \phi\left\{\sum_{l=0}^{n1}\sum_{m=0}^{n1} w_{l,m}^1 * x_{j+l,k+m}^1 + b_j^1\right\} \end{aligned} \tag{2}$$

$$\begin{aligned} Y_c{}^2 &= \phi\left\{\sum_{i=0}^{n2} w_{ij}^2 * x_i^2 + b_j^2\right\} \\ Y_d{}^2 &= \phi\left\{\sum_{l=0}^{n2}\sum_{m=0}^{n2} w_{ij}^2 * x_{j+1,k+m}^2 + b_j^2\right\} \end{aligned} \tag{3}$$

$$\begin{aligned} Y_c{}^3 &= \phi\left\{\sum_{i=0}^{n3} w_{ij}^3 * x_i^3 + b_j^3\right\} \\ Y_d{}^3 &= \phi\left\{\sum_{l=0}^{n3}\sum_{m=0}^{n3} w_{ij}^3 * x_{j+1,k+m}^3 + b_j^3\right\} \end{aligned} \tag{4}$$

where h_j denotes the content of the pixel feature vector's hidden state; k denotes the feature points;$j * k$ represents the size of the

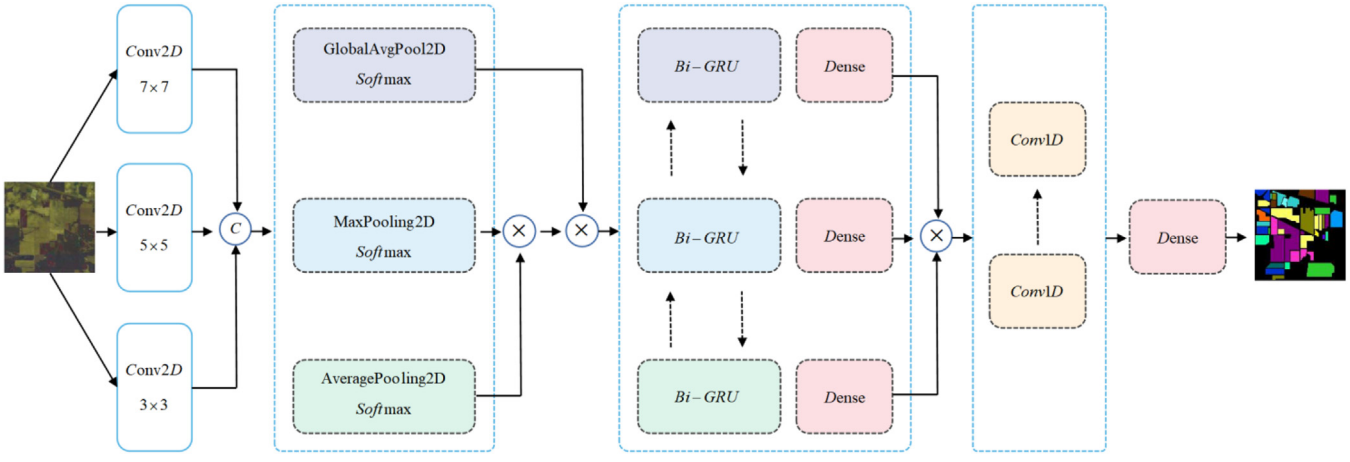

Fig. 1. Schematic of the detailed design of our proposed TDS-BiGRU. ⊗ represents the multiplication operation. Ⓒ represents the concatenation operation.

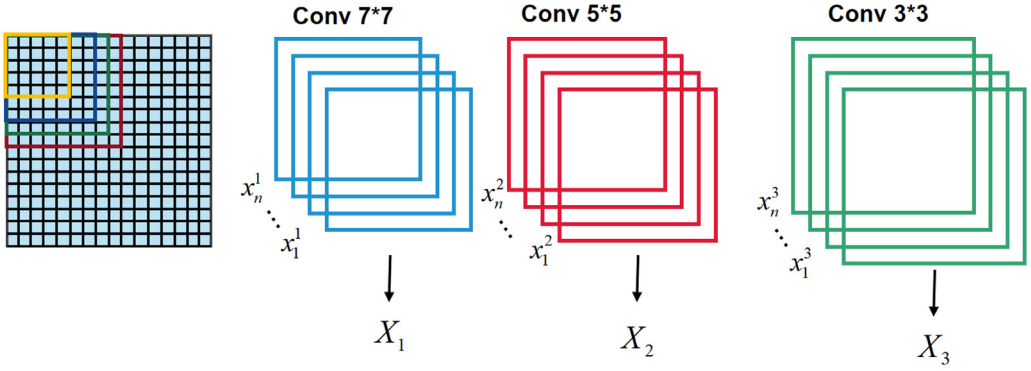

Fig. 2. Schematic diagram of the structure of multi-scale convolution.

feature map; and $l * m$ denotes the size of the local perceptual field of the null convolution. The above equations extract multi-scale semantic features of HSI data using conventional convolution and null convolution at three scales, respectively.

3.2. Three-dimensional Softmax Mechanism

3.2.1. Softmax mechanism

The softmax attention mechanism is designed to emulate human perception by taking into account the varying levels of attention that the human brain assigns to different stimuli. This mechanism can be conceptualized as a scoring mechanism that evaluates the relative importance of different features for classification. In computer vision, the attention mechanism is used to identify the most representative features while suppressing less significant ones. The softmax attention mechanism is defined by the following equation:

$$\zeta = \frac{e^{-e_{ij}}}{\sum_{k=1}^{n} e^{-e_{ik}}} \tag{5}$$

3.2.2. Three-dimensional Softmax Model

Fig. 3 shows the Three-dimensional Softmax Mechanism module. Classifying hyperspectral images requires dividing the original image into blocks of a certain size and using neighbourhood information to guide the classification process, which improves data differentiation. However, the null-spectral unity structure is only partially redundant and still contains useful data. In addition, the average spectral profile of the HSI dataset indicates that vegetation has a high degree of similarity in its spectral profiles, which can make it difficult to distinguish between different landscapes using the majority of hyperspectral image classification algorithms. These algorithms frequently fail to utilise the null-spectral structure effectively, resulting in confusion between landscapes with high spectral similarity. A softmax mechanism can be used to highlight important features while reducing the influence of useless features, making it suitable for the classification of HSI with redundant features.

In this study, we propose a three-part Three-dimensional Softmax mechanism for image classification. The first part is the global Softmax average, where each pixel's feature vector is fed into the global Softmax, and a weight coefficient is assigned to each pixel. We preserve 50% of the features with higher weights and feed them into the BiGRU layer as new feature vectors. Next, we employ a three-channel Softmax technique to extract global and local semantic information. Finally, we use a dense layer for integration. To begin with, we compute the global average attention mechanism using the following equation:

$$\mathrm{G} = GAP\left\{\frac{1}{H \times W}\sum_{i=1}^{W}\sum_{j=1}^{H} Y_C(i,j)\right\} \tag{6}$$

In Fig. 4, global average pooling is illustrated, which involves computing the average value of each channel to capture the feature map value distribution. This information is compressed and then fed into the excitation layer. The activation layer can be described by the following function:

$$\mathrm{s} = \zeta\left(g(G, W)\right) \tag{7}$$

where ζ denotes the softmax operation.

The calculation equations for average pooling and max pooling softmax mechanisms are shown as follows:

$$A = AP\left\{\frac{1}{H \times W}\sum_{i=1}^{W}\sum_{j=1}^{H} Y_C(i,j)\right\} \tag{8}$$

$$M = MP\left\{\frac{1}{H \times W}\sum_{i=1}^{W}\sum_{j=1}^{H} Y_C(i,j)\right\} \tag{9}$$

The output of the Three-dimensional Softmax mechanism model is obtained by fusing the output of above three softmax

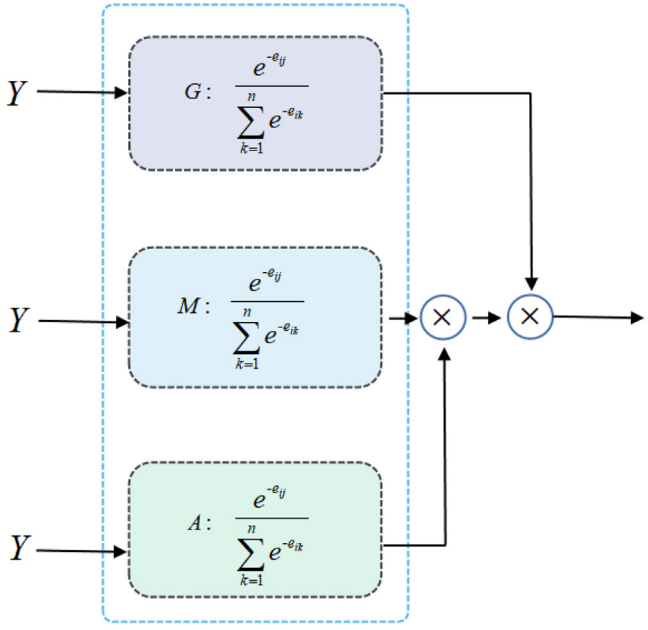

Fig. 3. Three-dimensional Softmax Mechanism module.

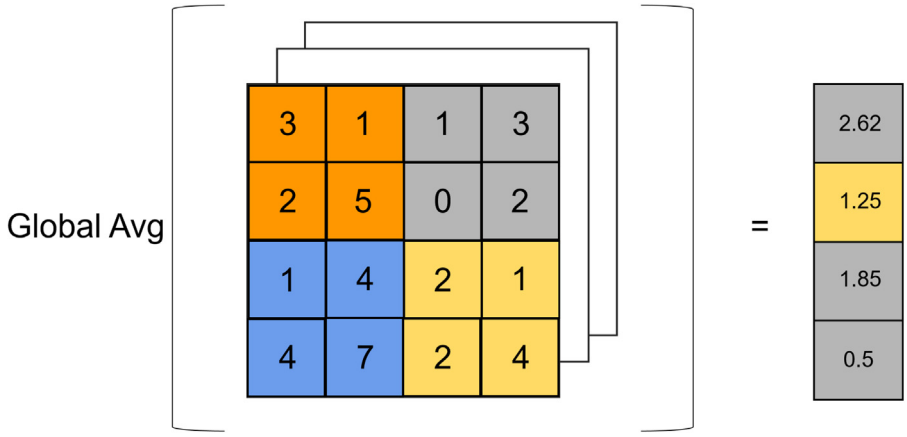

Fig. 4. Global average pooling.

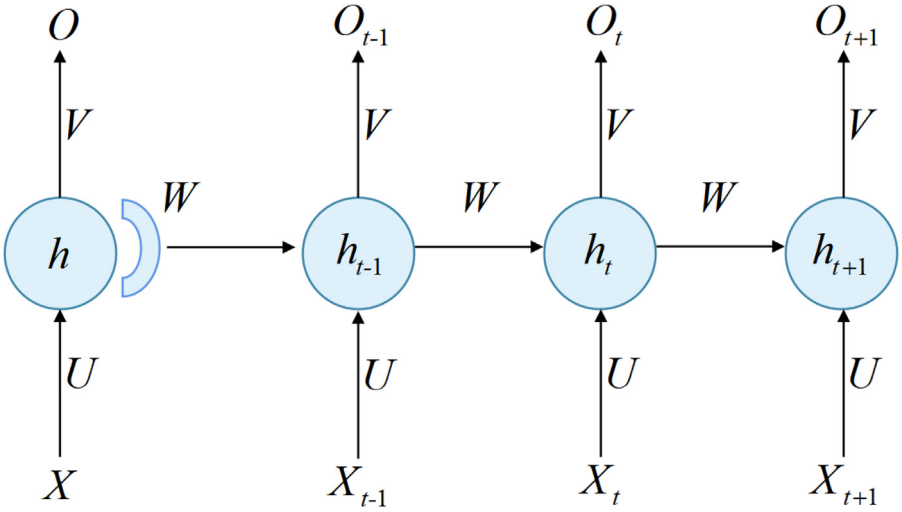

Fig. 5. Schematic diagram of the recurrent unit structure.

mechanisms . The formula for calculation is shown as follows:

$$TDS = cat\{(A \otimes M) \otimes G\} \tag{10}$$

After generating and assigning weights using the three-dimensional softmax mechanism equation, the three aforementioned softmax weighting strategies result in three features with updated weights. By fusing these features using the three-dimensional softmax technique, the BiGRU layer can benefit from improved semantic qualities.

3.3. Dense BiGRU Module

In the classification and discrimination of hyperspectral data, the spectral profile of each band is highly correlated with its previous band, making it difficult to process using traditional neural networks that cannot pass information between layers. RNNs are designed to handle sequential data with a recursive structure that includes feedback loops and self-repetition, as shown in Fig. 5. The nodes in an RNN are connected in a chain-like manner, allowing for sequential relationships and interdependencies among input data elements to be captured.

After unfolding by each time step, the sequence evolves using the previous time state h^{t-1} and the current input x^t to iteratively compute the current system state h^t: $h_t = \varphi(Ux_t + Wh_{t-1})$. U is the weight between the input and the hidden layer, W is the weight coefficient between the before and after moments of the hidden layer node itself, and φ is the excitation function.

Unlike CNNs that share weight parameter matrices across spatial dimensions, RNNs share these matrices across time steps. This sharing of parameters in the temporal dimension allows RNNs to fully exploit time-domain correlations and share statistical properties at different sequence lengths and temporal locations. RNNs aim to minimize the loss function L(U,W,V) and optimize model parameters using error backpropagation with multiple iterations of gradient descent, which is similar to fully connected neural networks (FCNNs). However, RNNs differ from FCNNs in that they use time-based backpropagation, which involves summing gradients over different time steps.

$$L = \sum_{t=1}^{n} L^{(t)} \tag{11}$$

$$\frac{\partial L}{\partial V} = \sum_{t=1}^{n} \frac{\partial L^{(t)}}{\partial O^{(t)}} \cdot \frac{\partial O^{(t)}}{\partial V} \tag{12}$$

$$\frac{\partial L^{(t)}}{\partial W} = \sum_{k=0}^{t} \frac{\partial L^{(t)}}{\partial O^{(t)}} \cdot \frac{\partial O^{(t)}}{\partial h^{(t)}} \left(\prod_{j=k+1}^{t} \frac{\partial h^{(j)}}{\partial h^{(j-1)}} \right) \frac{\partial h^{(k)}}{\partial W} \tag{13}$$

$$\frac{\partial L^{(t)}}{\partial U} = \sum_{k=0}^{t} \frac{\partial L^{(t)}}{\partial O^{(t)}} \cdot \frac{\partial O^{(t)}}{\partial h^{(t)}} \left(\prod_{j=k+1}^{t} \frac{\partial h^{(j)}}{\partial h^{(j-1)}} \right) \frac{\partial h^{(k)}}{\partial U} \tag{14}$$

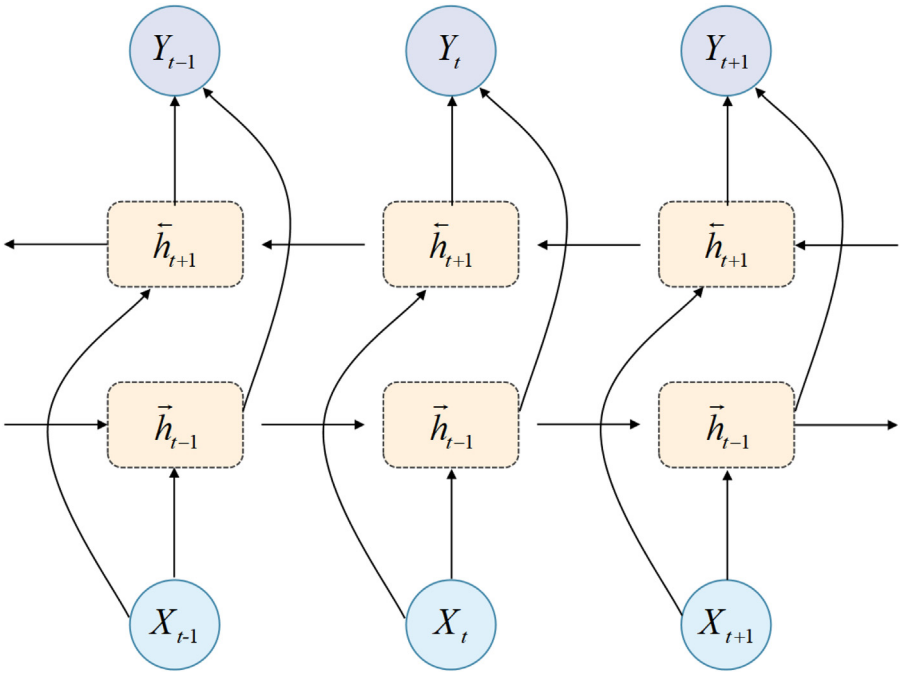

Fig. 6. A schematic diagram of BiGRU, which can be regarded as two unidirectional GRUs, so the hidden layer state of BiGRU at time t is obtained by the weighted sum of the forward hidden layer state and the backward hidden layer state.

Substituting the activation function f into the middle cumulative multiplication part of the above generalization leads to an error term of $\delta_k^T = \delta_t^T \prod_{i=k}^{t-1} W diag[f'(n_i)]$. As the cumulative multiplication of $f'(n_i)$ leads to smaller and smaller gradient updates until it is close to 0, making it difficult to perform effective learning for longer history time steps.

The original RNN suffers from limitations in long-range learning due to the cumulative multiplication of gradient updates. To address this issue, we propose the use of BiGRU to overcome the limitation of long-range dependence in the original RNN, thereby enhancing the effectiveness of learning for longer time steps.

Fig. 6 demonstrates the construction of BiGRU. The current hidden layer state of the BiGRU is determined by the three components of the current input x_t, the input $\vec{h}_{t-1}$ of the forward hidden layer state at the moment $(t-1)$ and the output $\overleftarrow{h}_{t-1}$ of the reverse hidden layer state. Since the BiGRU can be viewed as two one-way GRUs, the hidden layer state of the BiGRU at moment t is obtained by weighting and summing the forward hidden layer state $\vec{h}_{t-1}$ and the reverse hidden layer state $\overleftarrow{h}_{t-1}$ is shown as follows:

$$\vec{h}_t = GRU\left(x_t, \vec{h}_{t-1}\right) \tag{15}$$

$$\overleftarrow{h}_t = GRU\left(x_t, \overleftarrow{h}_{t-1}\right) \tag{16}$$

$$h_t = w_t \vec{h}_t + v_t \overleftarrow{h}_t + b_t \tag{17}$$

where the $GRU()$ function represents a nonlinear transformation of the input features, encoding them into the corresponding GRU hidden states. w_t, v_t denote the weights corresponding to the forward hidden layer state ht and the reverse hidden state ht corresponding to the bidirectional GRU at time t, respectively, and b_t denotes the bias corresponding to the hidden layer state at time t.

After modeling the three-dimensional softmax features with BiGRU, we further process them using dense layers and Conv1D. The dense layer applies nonlinear transformations to capture the associations between these features, and maps them to the output space using the following computational equation:

$$D = Dense(BiGRU(TDS)) \tag{18}$$

And the advantage of using Conv1D over only using Dense is that it can handle dynamic input sizes to better mine the deep semantic features of HSI, the computational equation Conv1D is shown as follows:

$$F = \sigma(\text{Conv1}D(D)) \tag{19}$$

where F denotes the final output of the TDS-BiGRU algorithm, and σ denotes the softmax function.

4. Experiments

In HSI classification tasks, several evaluation criteria are commonly used, including the Kappa coefficient, average accuracy (AA), and overall accuracy (OA). Higher values of these metrics indicate that the model has better classification performance. The computational equations for these metrics are as follows:

$$OA = \frac{1}{N} \sum_{j=1}^{N_c} m_{nn} \tag{20}$$

$$AA = \frac{\sum_{j=1}^{N_c} \frac{m_{nn}}{m+n}}{N_c} \tag{21}$$

$$K = \frac{N \cdot \sum_{i=1}^{N_c} X_{ii} - \sum_{i=1}^{N_c} (X_{i+} \cdot X_{+i})}{N^2 - \sum_{i=1}^{N_c} (X_{i+} \cdot X_{+i})} \tag{22}$$

where N denotes the total number of real feature samples, X_{ii} denotes the samples in the i-th row and i-th column, and the higher K coefficient represents the better model subclassification effect, i.e., the samples are less missed and misclassified.

4.1. DataSet

The **Indian Pines dataset (IP)** is a widely-used hyperspectral remote sensing image, captured in 1992 from the Indian Pines test region in northwestern Indiana by the AVIRIS sensor. The image contains 224 spectral reflection channels in the 400-2500 nm wavelength range and the 145 × 145 pixel image has a spatial resolution of 20 meters per pixel and comprises a total of 16 features, mainly local agriculture and woodlands. Due to its agroforestry nature, the agricultural area is geometrically regular, while the forest area is irregular. To ensure optimal results, four bands significantly impacted by water absorption were eliminated, leaving 220 spectral bands for experimentation.

The **Pavia University dataset (PU)** is another widely-used hyperspectral dataset, collected over the University of Pavia, Italy, in 2001 by the German Aerospace Agency's hyperspectral project team (DLR). The dataset contains a remote sensing image of 610 × 340 pixels, captured using Reflection Optical System Imaging Spectrometer (ROSIS) scanning in 103 spectral bands within the wavelength range of 430-860 nm. The image includes nine urban characteristics, such as asphalt, metal plates, and tarmac, with a spatial resolution of 1.3 meters per pixel.

The **Salinas dataset (SA)** is similar to the Indian Pines dataset, as it also contains 16 types of feature information and has a strong local spatial homogeneity. The images in the Salinas dataset primarily represent large feature areas, with fewer details and no subtle or trivial areas like those found in the Pavia University dataset.

The **Houston University dataset (Houston)** was funded by the University of Houston Airborne Laser Mapping Center and captured on June 23, 2012, using the ITRES-CASI 1500 sensor. The dataset covers the University of Houston campus and surrounding metropolitan areas and has a spatial resolution of 2.5 meters per pixel, with a resolution of 349 × 1905 pixels. The dataset contains spectral information in a total of 144 bands, ranging from 380 nm to 1050 nm, covering the visible to the near-infrared range. The ground scenes captured in this dataset include civil buildings, such as parking lots and commercial areas, as well as natural landscapes, such as trees and soils. The key information of all datasets and examples of ground truth are shown in Table 1 and Fig. 7, respectively.

Table 1
Key information about the datasets

DataSet	Size	Bands	Wavelength range
Indian Pines (IP)	145 × 145	200	400-2500 nm
Pavia University (PU)	610 × 340	103	430-860 nm
Salinas (SA)	512 × 217	204	360-2500 nm
Houston	349 × 1905	144	380-1050 nm

4.2. Experimental setup

In this paper, we conduct an experimental evaluation of hyperspectral classification algorithms using four widely-used public datasets, namely IP, PU, SA, and Houston. All experiments are performed on an Nvidia 1080Ti GPU, with Python 3.6 as the programming language, and Tensorflow 1.10 and Keras 2.1 as the software environment. For IP, (5%, 10%, 15%) of the samples are taken as the training set, while for PU, SA, and Houston, (1%, 5%, 10%) of the samples are taken as the training set, with the remaining samples used as the test set.

4.3. Comparison experiments

IP: Table 2 and Fig. 8 demonstrate that the classification results of SAGP and the proposed TDS-BiGRU method are more distinct and exhibit a smoother local area classification than those of PRAN and FSSFNet, which only consider single-pixel informa-

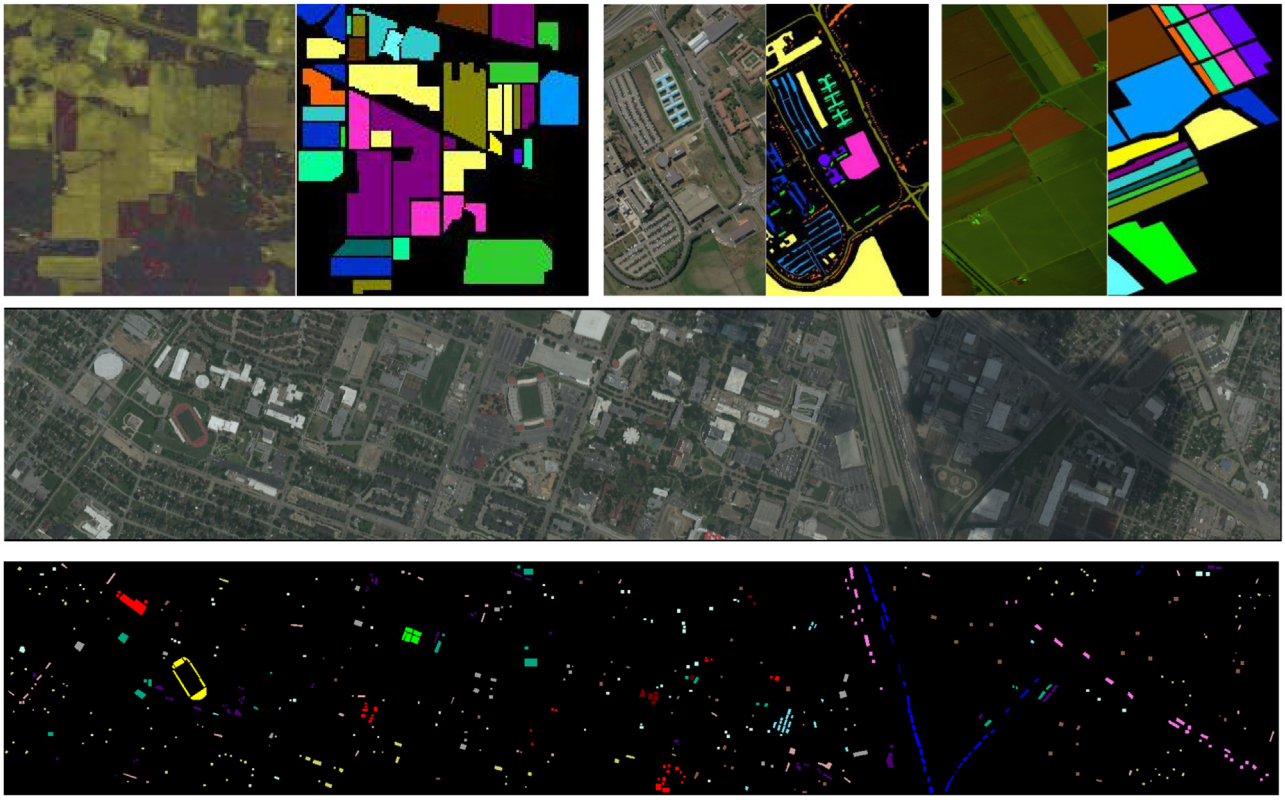

Fig. 7. Example of ground truth for four datasets.

Table 2
Comparative experimental results with other 6 methods on IP.

Models	5%			10%			15%		
	Kappa	OA	AA	Kappa	OA	AA	Kappa	OA	AA
ResNet [23]	66.59%	76.70%	69.69%	75.12%	78.29%	79.22%	86.90%	83.21%	85.90%
AlexNet [22]	64.23%	68.97%	56.29%	75.20%	74.11%	65.26%	79.12%	81.59%	79.62%
PRAN [25]	69.29%	72.67%	73.95%	74.61%	77.58%	73.51%	84.20%	82.79%	76.72%
DenseNet [24]	66.66%	71.41%	67.23%	75.81%	78.77%	75.62%	82.15%	84.32%	81.51%
SAGP [27]	73.61%	73.49%	76.58%	76.72%	78.36%	88.90%	82.89%	81.59%	86.50%
FSSFNet [26]	69.98%	73.75%	67.95%	75.47%	78.66%	71.26%	80.50%	82.61%	74.48%
TDS-BiGRU	**76.75%**	**78.67%**	**79.91%**	**81.15%**	**83.21%**	**81.99%**	**86.33%**	**87.22%**	**87.18%**

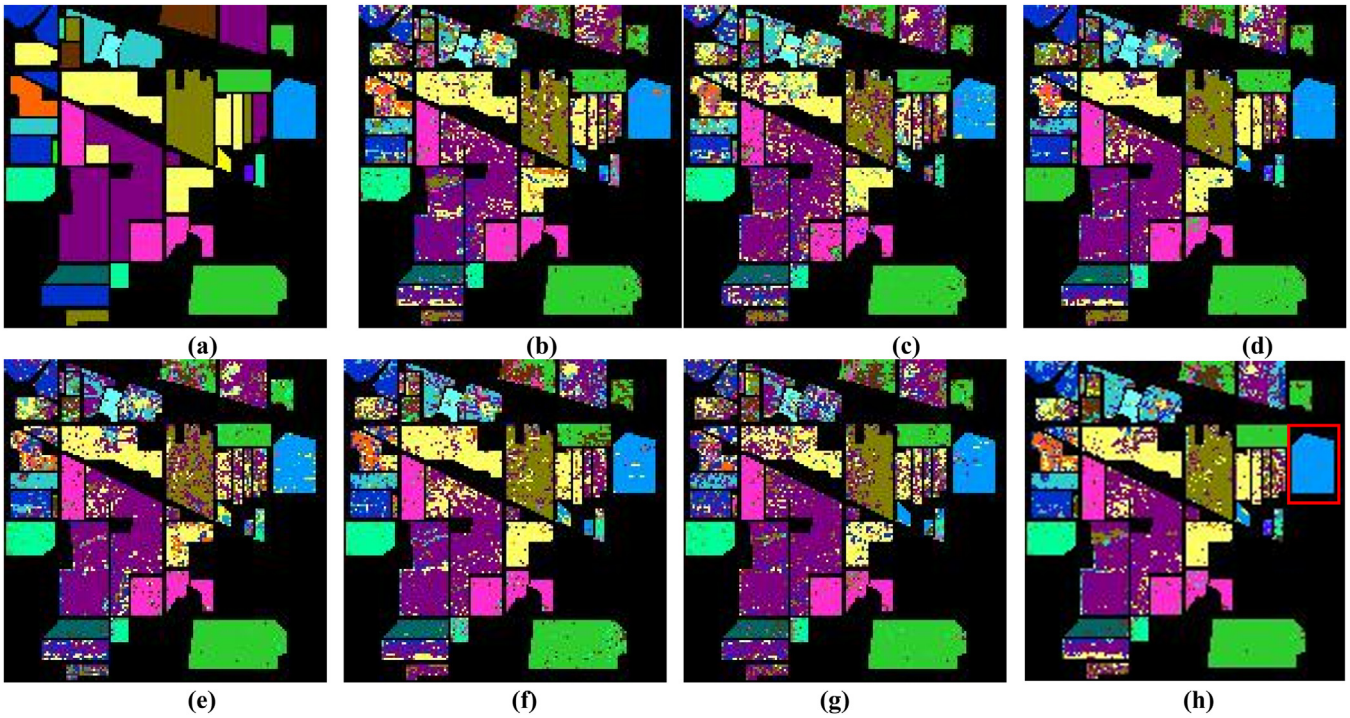

Fig. 8. Visualization of comparative classification results on IP. (a) Ground truth. (b) ResNet. (c) AlexNet. (d) PRAN. (e) DenseNet. (f) FSSFNet. (g) SAGP. (h) TDS-BiGRU.

Table 3
Comparative experimental results with other 6 methods on PU.

Models	1%			5%			10%		
	Kappa	OA	AA	Kappa	OA	AA	Kappa	OA	AA
ResNet [23]	77.58%	84.22%	82.39%	87.91%	90.88%	91.51%	93.38%	94.33%	94.78%
AlexNet [22]	82.69%	87.11%	84.41%	90.49%	92.89%	91.79%	91.53%	92.88%	93.66%
PRAN [25]	85.86%	89.42%	88.69%	91.65%	93.61%	92.53%	93.29%	94.31%	93.71%
DenseNet [24]	75.68%	82.53%	80.87%	87.23%	90.43%	89.48%	91.52%	92.46%	93.42%
SAGP [27]	78.83%	84.32%	83.84%	88.14%	91.09%	89.75%	91.65%	93.73%	92.97%
FSSFNet [26]	82.11%	86.63%	84.23%	91.65%	93.72%	92.35%	92.52%	94.37%	92.65%
TDS-BiGRU	**88.01%**	**89.81%**	**88.56%**	**94.12%**	**95.23%**	**93.18%**	**94.66%**	**95.88%**	**95.28%**

tion. The extraction of high-level semantic features using softmax attention for low-level features substantially improves the classification effect for both low-level and semantic features. This suggests that high-level semantic features have superior discriminative and representable properties and can effectively integrate local spatial data. The semantic features extracted by the three-dimensional softmax module of TDS-BiGRU are especially advantageous for the high spectral image classification task and can improve the classification effect in some detail. For instance, while the classification overall accuracy (OA) of AlexNet with only a single spectral feature is about 68.9%, SAGP with the same softmax attention reaches 73.4%, an improvement of 4.5 percentage points. This result demonstrates the advantage of the TDS-BiGRU method.

PU: In regards to PU, the results presented in Table 3 and Fig. 9 highlight the superior performance of the proposed TDS-BiGRU method in extracting high-level semantic features compared to conventional low-level features for hyperspectral image classification. The use of single-pixel low-level features by DenseNet leads to a significant amount of noise, while FSSFNet, which incorporates spatial information, exhibits relatively less noise. On the other hand, SAGP and TDS-BiGRU, which employ attention mechanisms, produce the clearest, smoothest, and most accurate classification maps.

SA: The results in Table 4 and Fig. 10 demonstrate that the superior performance of the proposed TDS-BiGRU method for hyperspectral image classification on the SA dataset. The OA of TDS-BiGRU, which utilized deep spatial semantic features extracted us-

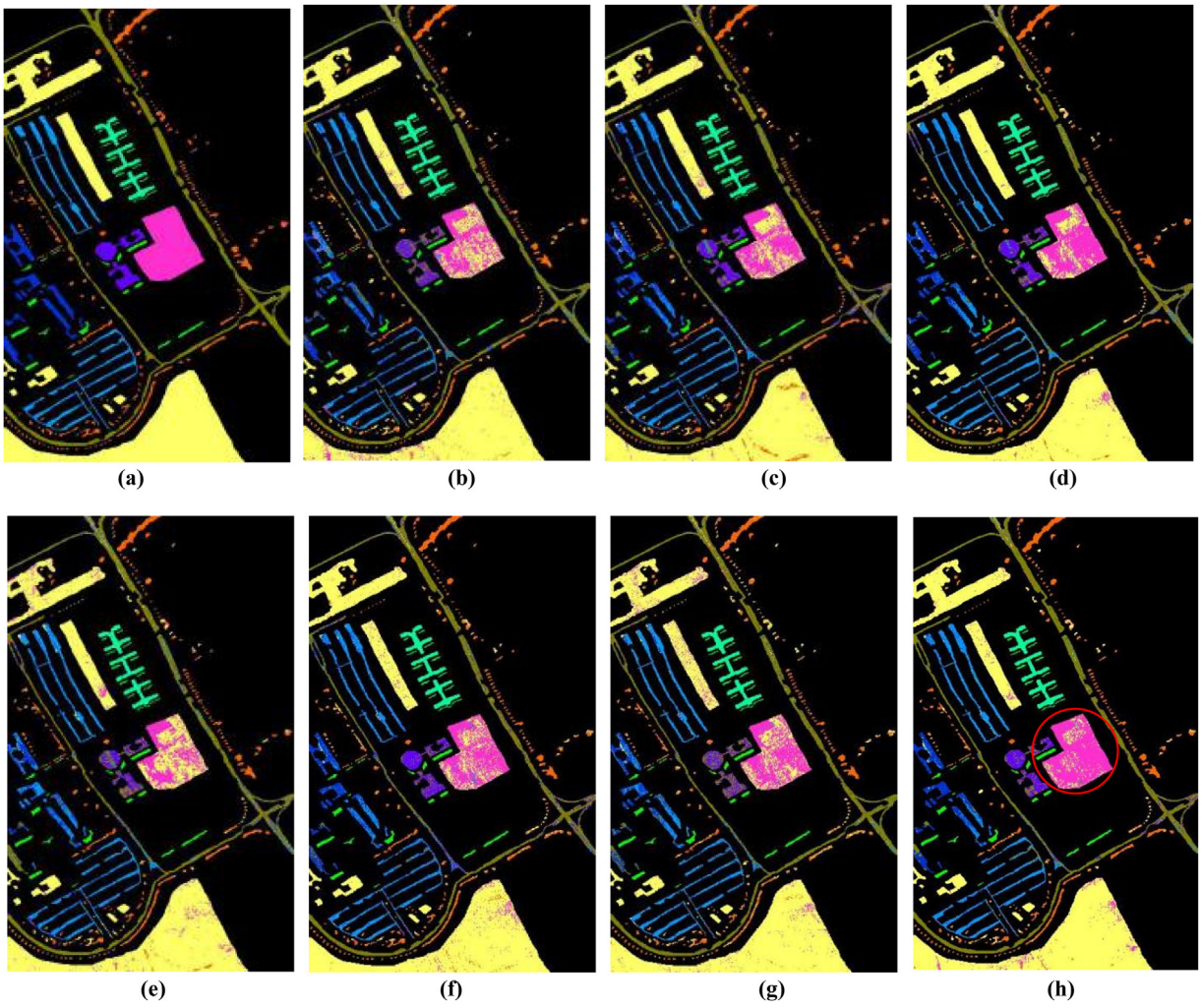

Fig. 9. Visualization of comparative classification results on PU. (a) Ground truth. (b) ResNet. (c) AlexNet. (d) PRAN. (e) DenseNet. (f) FSSFNet. (g) SAGP. (h) TDS-BiGRU.

Table 4
Comparative experimental results with other 6 methods on sa.

Models	1%			5%			10%		
	Kappa	OA	AA	Kappa	OA	AA	Kappa	OA	AA
ResNet [23]	85.87%	87.41%	92.34%	90.92%	91.79%	95.33%	93.17%	93.81%	96.37%
AlexNet [22]	89.91%	90.88%	93.79%	93.62%	94.21%	96.69%	95.42%	94.41%	95.81%
PRAN [25]	76.49%	78.89%	76.43%	89.36%	90.41%	88.38%	91.19%	91.87%	89.55%
DenseNet [24]	84.15%	85.72%	90.87%	90.31%	91.23%	94.54%	92.36%	93.12%	96.19%
SAGP [27]	89.82%	90.87%	94.68%	91.81%	92.64%	95.69%	94.16%	94.76%	97.45%
FSSFNet [26]	90.02%	91.04%	94.74%	92.82%	93.56%	96.69%	95.37%	95.85%	98.03%
TDS-BiGRU	**92.55%**	**92.11%**	**93.24%**	**96.24%**	**96.77%**	**97.13%**	**98.55%**	**97.14%**	**98.22%**

ing BiGRU with spectral features as the base features, reached about 92%, indicating its competitive experimental results compared to the traditional DenseNet and PRAN methods. Furthermore, the TDS-BiGRU method achieved significantly higher OA, AA, and Kappa coefficients, with values of 92.11%, 93.24%, and 92.55%, respectively, emphasizing its superiority in accurately classifying hyperspectral images.

Houston: Regarding the Houston dataset, the results presented in Table 5 and Fig. 11 demonstrate that the proposed TDS-BiGRU method is capable of extracting better spatial neighborhood information and reducing mislabeled regions. Traditional methods such as ResNet and DenseNet, which rely solely on spectral features, suffer from the problem of homospectral dissimilarity, leading to many misclassification points. In contrast, methods like SAGP that use spatial information have improved the classification results by inferring the feature attributes of the center point from the contextual information of the surrounding pixels with high probability. However, models that focus on spatial features may lose some spectral information, resulting in fine misclassification within the local area, which ultimately affects the overall classification effect. By using the softmax attention module, the proposed TDS-BiGRU method is able to focus on the most informative and useful features, enabling it to make full use of global contextual information for learning nonlinear feature interactions and modeling long-term dependencies, particularly at the edges of the feature coverage region.

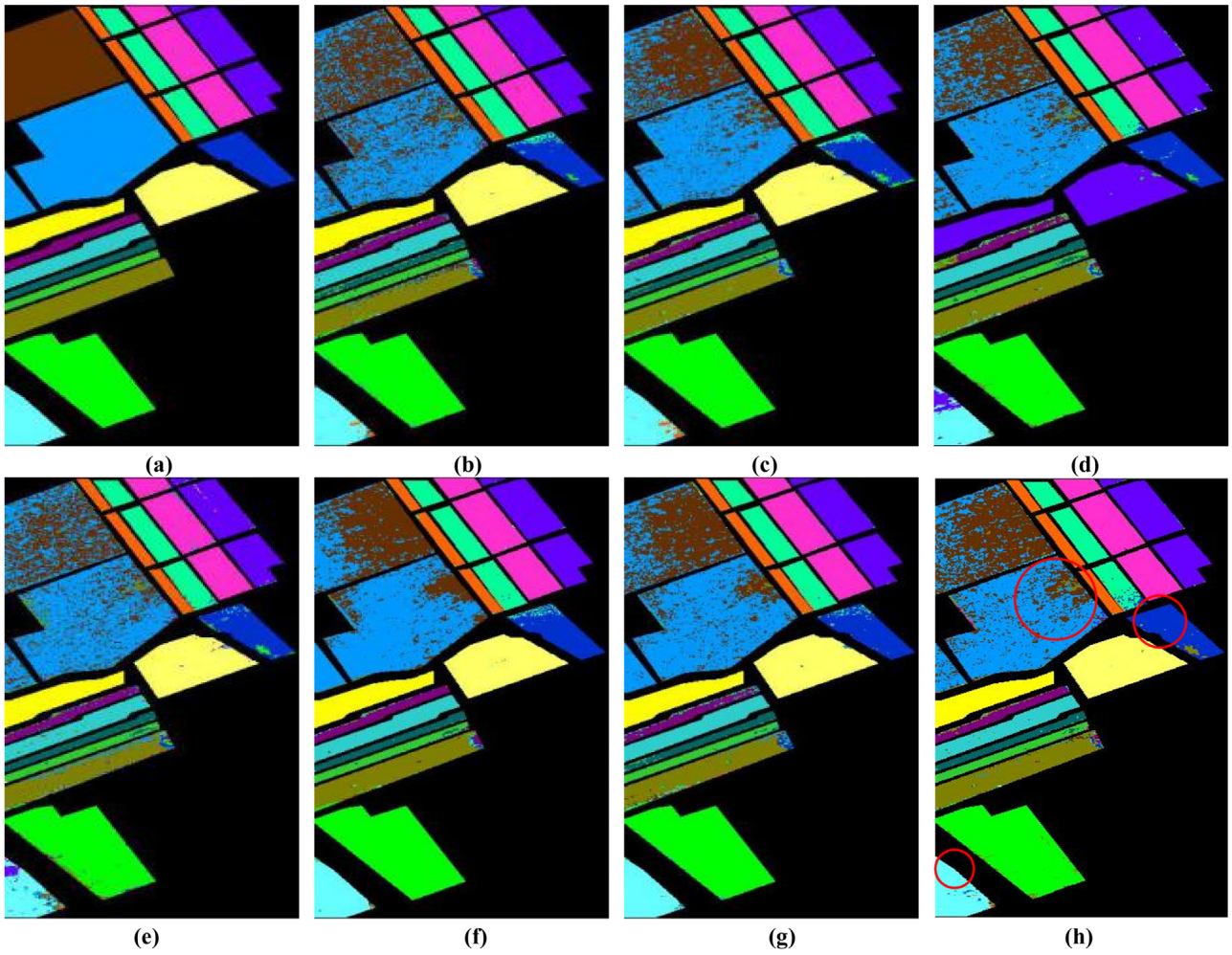

Fig. 10. Visualization of comparative classification results on SA. (a) Ground truth. (b) ResNet. (c) AlexNet. (d) PRAN. (e) DenseNet. (f) FSSFNet. (g) SAGP. (h) TDS-BiGRU.

Table 5
Comparative experimental results with other 6 methods on Houston.

Models	1%			5%			10%		
	Kappa	OA	AA	Kappa	OA	AA	Kappa	OA	AA
ResNet [23]	91.24%	91.01%	89.76%	93.87%	92.80%	93.29%	95.09%	95.79%	96.95%
AlexNet [22]	90.01%	91.03%	89.72%	92.53%	92.19%	91.76%	95.31%	93.37%	94.79%
PRAN [25]	93.25%	94.81%	92.49%	95.22%	95.33%	96.23%	97.02%	96.95%	97.44%
DenseNet [24]	91.43%	93.56%	91.52%	94.12%	94.14%	93.47%	97.29%	95.09%	96.15%
SAGP [27]	96.98%	97.19%	96.55%	97.81%	97.64%	98.69%	99.16%	98.76%	98.45%
FSSFNet [26]	96.29%	96.45%	95.55%	97.82%	96.56%	96.69%	97.37%	98.85%	98.03%
TDS-BiGRU	**99.25%**	**98.66%**	**99.11%**	**98.98%**	**99.01%**	**98.91%**	**99.60%**	**99.34%**	**99.11%**

4.4. Ablation experiment results of three-dimensional softmax mechanism

To further demonstrate the effectiveness of the three-dimensional softmax attention mechanism, we conducted ablation experiments in this section. Specifically, we evaluated the performance of using only GlobalAvgPool2D (G), MaxPooling2D (M), or AveragePooling2D (A) as the attention mechanism. We also evaluated the performance of combining two attention mechanisms, including G+M, G+A, and M+A.

Table 6 demonstrates the superiority of the three-dimensional softmax attention mechanism in achieving the best classification results. Moreover, the combined softmax attention outperforms the single attention methods. In addition, among the single attention methods, AVERAGEPooling2D outperforms both GlobalAvgPool2D and MaxPooling2D. These findings provide further evidence for the superiority of the three-dimensional softmax mechanism.

Table 6
Ablation experiment results of three-dimensional softmax mechanism.

Methods	IP (5%) OA	PU (1%) OA	SA (1%) OA	Houston (1%) OA
G	75.22%	86.18%	89.07%	96.69%
M	75.61%	85.07%	89.22%	97.01%
A	75.69%	86.77%	90.03%	97.53%
G+M	76.68%	88.66%	90.31%	97.89%
G+A	77.81%	89.15%	91.66%	98.11%
M+A	76.14%	87.95%	91.29%	98.34%
G+M+A (ours)	**78.67%**	**89.81%**	**92.11%**	**98.66%**

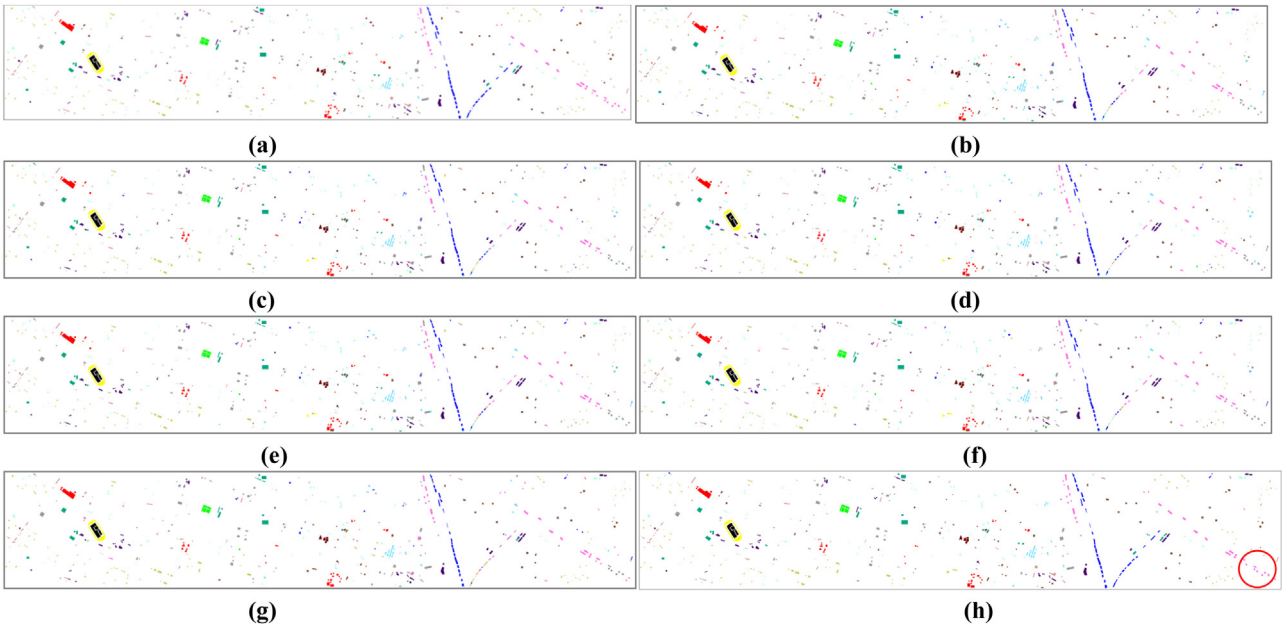

Fig. 11. Visualization of comparative classification results on Houston. (a) Ground truth. (b) ResNet. (c) AlexNet. (d) PRAN. (e) DenseNet. (f) FSSFNet. (g) SAGP. (h) TDS-BiGRU.

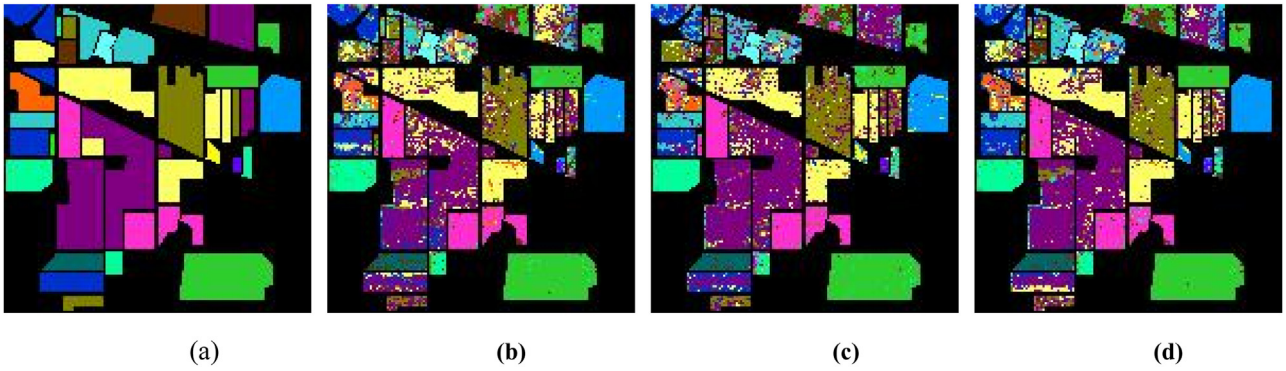

Fig. 12. Classification results of the IP using sub-modules. (a) Ground truth. (b) No-TGS. (c) No-BiGRU. (d) TDS-BiGRU.

4.5. Ablation experiment results of sub-modules

The results of the proposed TDS-BiGRU submodule ablation tests are presented in Table 7 and Fig. 12, revealing that the combination of TDS and BiGRU yields the best outcomes. Extracting additional discriminatory features is crucial in hyperspectral data processing. While the three-dimensional softmax attention mechanism effectively characterizes the semantic content of HSI data, BiGRU can extract the output characteristics of the three-dimensional softmax attention module more effectively, leading to a highly competitive performance of the presented algorithm.

Table 7 reveals the indispensable role of both the three-dimensional attention mechanism and the BiGRU network in achieving competitive performance, as their absence would greatly diminish the algorithm's overall performance. The experimental results demonstrate that the combined use of both modules yields superior outcomes compared to their individual use. Notably, the three-dimensional attention mechanism has the greatest impact on classification ability, underscoring the superiority of the three-dimensional dense attention mechanism in deep feature mining and selection.

4.6. Ablation experiment results of PCA

Table 8 shows the ablation experiment results of PCA. This section presents a comparison of the experimental results (as shown in Fig. 8) before and after dimension reduction using PCA. The findings demonstrate that PCA dimensionality reduction can effectively reduce redundant features, leading to improved computational efficiency without significant loss of classification accuracy. Therefore, the results indicate that PCA dimensionality reduction is a practical approach for feature selection in hyperspectral image classification.

Table 7
Ablation experiment results of sub-modules.

Methods	IP (5%) OA	PU (1%) OA	SA (1%) OA	Houston (1%) OA
No-TGS	74.28%	86.46%	90.58%	95.91%
No-BiGRU	76.55%	87.27%	91.03%	97.18%
TDS-BiGRU	**78.67%**	**89.81%**	**92.11%**	**98.66%**

Table 8
Ablation experiment results of PCA

Methods	IP (5%) OA	PU (1%) OA	SA (1%) OA	Houston (1%) OA
No-PCA	76.11%	88.43%	92.03%	98.27%
TDS-BiGRU	**78.67%**	**89.81%**	**92.11%**	**98.66%**

5. Conclusions

This paper presented a novel approach for hyperspectral remote sensing image classification, namely the three-dimensional Softmax mechanism-guided bi-directional GRU network (TDS-BiGRU). By incorporating a GRU to process the sequence data, our approach significantly reduced the computational time. Furthermore, we introduced a three-dimensional Softmax mechanism that captures cross-latitude interactions through three branches to obtain deeper features with stronger discriminative power. This mechanism effectively addressed the problem of insufficient utilization of spatial information in traditional recurrent neral network-based classification algorithms. The proposed framework combined the merits of different processing units and resulted in improved classification accuracy. Extensive comparison and ablation experiments demonstrated the effectiveness of our approach on four hyperspectral remote sensing datasets and its competitiveness with other mainstream algorithms. Additionally, we investigated the impact of PCA dimensionality reduction and show that it can reduce redundant features and enhance computational efficiency. In future work, we will further investigate the effectiveness of sequence models and their variants on our tasks.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

CRediT authorship contribution statement

Guoqiang Wu: Conceptualization, Methodology, Software, Validation. **Xin Ning:** Data curation, Conceptualization, Software. **Luyang Hou:** Methodology, Software. **Feng He:** Software, Validation. **Hengmin Zhang:** Writing – review & editing, Visualization. **Achyut Shankar:** Visualization, Investigation.

Data availability

Data will be made available on request.

References

[1] S. Ghaffarian, N. Kerle, T. Filatova, Remote sensing-based proxies for urban disaster risk management and resilience: A review, Remote sensing 10 (11) (2018) 176.

[2] A. Fekete, K. Tzavella, I. Armas, J. Binner, M. Garschagen, C. Giupponi, … D. Serre, Critical data source; tool or even infrastructure? Challenges of geographic information systems and remote sensing for disaster risk governance, ISPRS International Journal of Geo-Information 4 (4) (2015) 1848–1869.

[3] H. Ahmadi, E. Pekkan, Fault-based geological lineaments extraction using remote sensing and GIS—a review, Geosciences 11 (5) (2021) 183.

[4] T. Javed, N. Yao, X. Chen, S. Suon, Y. Li, Drought evolution indicated by meteorological and remote-sensing drought indices under different land cover types in, China. Environmental Science and Pollution Research 27 (2020) 4258–4274.

[5] W. Cai, X. Ning, G. Zhou, X. Bai, Y. Jiang, W. Li, P. Qian, A Novel Hyperspectral Image Classification Model Using Bole Convolution with Three-Directions Attention Mechanism: Small sample and Unbalanced Learning, IEEE Transactions on Geoscience and Remote Sensing (2022).

[6] Z. Zhang, Y. Ding, X. Zhao, S. Li, N. Yang, Y. Cai, Y. Zhan, MultiReceptive Field: An Adaptive Path Aggregation Graph Neural Framework for Hyperspectral Image Classification, Expert Systems with Applications (2023) early accepted.

[7] W. Cai, B. Liu, Z. Wei, M. Li, J. Kan, TARDB-Net: triple-attention guided residual dense and BiLSTM networks for hyperspectral image classification, Multimedia Tools and Applications 80 (2021) 11291–11312.

[8] M.P. Uddin, M.A. Mamun, M.A. Hossain, Effective feature extraction through segmentation-based folded-PCA for hyperspectral image classification, International Journal of Remote Sensing 40 (18) (2019) 7190–7722.

[9] S.D. Fabiyi, P. Murray, J. Zabalza, J. Ren, Folded LDA: extending the linear discriminant analysis algorithm for feature extraction and data reduction in hyperspectral remote sensing, IEEE Journal of selected topics in applied earth observations and remote sensing 14 (2021) 12312–12331.

[10] A. Plaza, J.A. Benediktsson, J.W. Boardman, J. Brazile, L. Bruzzone, G. Camps-Valls, … G. Trianni, Recent advances in techniques for hyperspectral image processing, Remote sensing of environment 113 (2009) S110–S122.

[11] L. Mou, P. Ghamisi, X.X. Zhu, Deep recurrent neural networks for hyperspectral image classification, IEEE Transactions on Geoscience and Remote Sensing 55 (7) (2017) 3639–3655.

[12] X. Zhang, Y. Sun, K. Jiang, C. Li, L. Jiao, H. Zhou, Spatial sequential recurrent neural network for hyperspectral image classification, IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing 11 (11) (2018) 4141–4155.

[13] W. Song, S. Li, L. Fang, T. Lu, Hyperspectral image classification with deep feature fusion network, IEEE Transactions on Geoscience and Remote Sensing 56 (6) (2018) 3173–3184.

[14] S.K. Roy, G. Krishna, S.R. Dubey, B.B. Chaudhuri, HybridSN: Exploring 3-D–2-D CNN feature hierarchy for hyperspectral image classification, IEEE Geoscience and Remote Sensing Letters 17 (2) (2019) 277–281.

[15] J. Xie, N. He, L. Fang, P. Ghamisi, Multiscale densely-connected fusion networks for hyperspectral images classification, IEEE Transactions on Circuits and Systems for Video Technology 31 (1) (2020) 246–259.

[16] J. Zhu, L. Fang, P. Ghamisi, Deformable convolutional neural networks for hyperspectral image classification, IEEE Geoscience and Remote Sensing Letters 15 (8) (2018) 1254–1258.

[17] Y. Xu, B. Du, L. Zhang, Beyond the patchwise classification: Spectral-spatial fully convolutional networks for hyperspectral image classification, IEEE Transactions on Big Data 6 (3) (2019) 492–506.

[18] Z. Zheng, Y. Zhong, A. Ma, L. Zhang, FPGA: Fast patch-free global learning framework for fully end-to-end hyperspectral image classification, IEEE Transactions on Geoscience and Remote Sensing 58 (8) (2020) 5612–5626.

[19] Y. Jiang, Y. Li, S. Zou, H. Zhang, Y. Bai, Hyperspectral image classification with spatial consistence using fully convolutional spatial propagation network, IEEE Transactions on Geoscience and Remote Sensing 59 (12) (2021) 10425–10437.

[20] D. Wang, B. Du, L. Zhang, Fully contextual network for hyperspectral scene parsing, IEEE Transactions on Geoscience and Remote Sensing 60 (2021) 1–16.

[21] Y. Shen, S. Zhu, C. Chen, Q. Du, L. Xiao, J. Chen, D. Pan, Efficient deep learning of nonlocal features for hyperspectral image classification, IEEE Transactions on Geoscience and Remote Sensing 59 (7) (2020) 6029–6043.

[22] M Salman, S E Yüksel, Fusion of hyperspectral image and LiDAR data and classification using deep convolutional neural networks, in: 2018 26th Signal Processing and Communications Applications Conference (SIU), IEEE, July 2018, pp. 1–4.

[23] M E Paoletti, J M Haut, R Fernandez-Beltran, et al., Deep pyramidal residual networks for spectral-spatial hyperspectral image classification, IEEE TranMSCtions on Geoscience and Remote Sensing 57 (2) (Feb. 2019) 740–754.

[24] G. Yang, U.B. Gewali, E. Ientilucci, M. Gartley, S.T. Monteiro, Dual-Channel Densenet for Hyperspectral Image Classification, in: IGARSS 2018-2018 IEEE International Geoscience and Remote Sensing Symposium, IEEE, July 2018, pp. 2595–2598.

[25] H Gao, Y Yang, D Yao, et al., Hyperspectral Image Classification With Pre-Activation Residual Attention Network, IEEE Access 7 (2019) 176587–176599.

[26] J Chen, S Chen, P Zhou, et al., Deep Neural Network Based Hyperspectral Pixel Classification With Factorized Spectral-Spatial Feature Representation, IEEE Access 7 (2019) 81407–81418.

[27] H You, S Tian, L Yu, Y Lv, Pixel-level remote sensing image recognition based on bidirectional word vectors, IEEE Transactions on Geoscience and Remote Sensing 58 (2) (2019) 1281–1293.

[28] Y. Zhang, W. Li, W. Sun, R. Tao, Q. Du, Single-Source Domain Expansion Network for Cross-Scene Hyperspectral Image Classification, IEEE Transactions on Image Processing 32 (2023) 1498–1512, doi:10.1109/TIP.2023.3243853.

[29] Y. Zhang, W. Li, M. Zhang, Y. Qu, R. Tao and H. Qi, "Topological Structure and Semantic Information Transfer Network for Cross-Scene Hyperspectral Image Classification," in IEEE Transactions on Neural Networks and Learning Systems, doi: 10.1109/TNNLS.2021.3109872.

[30] Y. Zhang, M. Zhang, W. Li, S. Wang, R. Tao, Language-Aware Domain Generalization Network for Cross-Scene Hyperspectral Image Classification, IEEE Transactions on Geoscience and Remote Sensing 61 (2023) 1–12 Art no. 5501312, doi:10.1109/TGRS.2022.3233885.

[31] Y. Zhang, W. Li, M. Zhang, S. Wang, R. Tao, Q. Du, Graph Information Aggregation Cross-Domain Few-Shot Learning for Hyperspectral Image Classification, IEEE Transactions on Neural Networks and Learning Systems (2022), doi:10.1109/TNNLS.2022.3185795.

[32] D. Hong, L. Gao, J. Yao, B. Zhang, A. Plaza, J. Chanussot, Graph Convolutional Networks for Hyperspectral Image Classification, IEEE Transactions on Geoscience and Remote Sensing 59 (7) (July 2021) 5966–5978, doi:10.1109/TGRS.2020.3015157.

[33] D. Hong, et al., More Diverse Means Better: Multimodal Deep Learning Meets Remote-Sensing Imagery Classification, IEEE Transactions on Geoscience and Remote Sensing 59 (5) (May 2021) 4340–4354, doi:10.1109/TGRS.2020.3016820.

[34] X. Wu, D. Hong, J. Chanussot, UIU-Net: U-Net in U-Net for Infrared Small Object Detection, IEEE Transactions on Image Processing 32 (2023) 364–376, doi:10.1109/TIP.2022.3228497.

[35] D. Hong, X. Wu, P. Ghamisi, J. Chanussot, N. Yokoya, X.X. Zhu, Invariant Attribute Profiles: A Spatial-Frequency Joint Feature Extractor for Hyperspectral Image Classification, IEEE Transactions on Geoscience and Remote Sensing 58 (6) (June 2020) 3791–3808, doi:10.1109/TGRS.2019.2957251.

[36] D. Hong, N. Yokoya, J. Chanussot, X.X. Zhu, An Augmented Linear Mixing Model to Address Spectral Variability for Hyperspectral Unmixing, IEEE Transactions on Image Processing 28 (4) (April 2019) 1923–1938, doi:10.1109/TIP.2018.2878958.