IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 24 February 2023, accepted 22 March 2023, date of publication 29 March 2023, date of current version 5 April 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3262992

RESEARCH ARTICLE

Improved Three-Dimensional Inception Networks for Hyperspectral Remote Sensing Image Classification

XIAOXIA ZHANG
School of Software Engineering, Chengdu University of Information Technology, Chengdu 610225, China
e-mail: zhangxiaoxia158@163.com

ABSTRACT Convolutional neural networks (CNNs) have been applied to hyperspectral image classification. Typically, fixed convolution kernel sizes are used in CNNs. This may be complicated by the high-dimensional features and the spatial-spectral features variability of hyperspectral data. In such cases, the fixed kernel size inhibits the ability to deal with different types of datasets flexibly. Therefore, we adopted a parallel architecture, an improved three-dimensional (3D) inception network, for hyperspectral image classification. Additionally, the dimensionality reduction based on interactive information entropy is used in the adaptive band selection of hyperspectral remote sensing data, then the low-redundant bands with more abundant and discriminative information can be selected. Experimental results based on the publicly available hyperspectral datasets demonstrated that the improved 3D inception network can perform well with limited training samples. Moreover, the combinations of different dimensional convolutions in the 3D-2D and 3D-2D-1D inception networks achieved comparable classification results. The experimental results indicate that the proposed networks are generalized classification models with high classification rates.

INDEX TERMS Hyperspectral image classification, remote sensing, deep learning, three-dimensional inception network, adaptive band selection.

I. INTRODUCTION

Hyperspectral images (HSIs) are widely used in the fields of remote sensing (e.g., urban planning, land use analysis, and environmental monitoring). HSIs are composed of several hundred spectral channels and simultaneously contain spatial and spectral information. They form three-dimensional data, improving the performance of accuracy-differentiating materials. However, increasing the spectral domain dimensionality of HSIs results in the curse of dimensionality, that is, the classification performance attenuates rather than improves as the dimension increases. To address this problem, accurate and efficient ways of extracting features are vital.

Several HSIs classification approaches using hand-designed feature descriptions have been developed. Fang et al. [1] used a local covariance matrix to encode the relationship between spectral bands. They used these matrices for HSIs training and support vector machine (SVM) as the classifier. Samiappan et al. [2] proposed a nonuniform random feature selection method within a multi-classifier system framework using an SVM as the base classifier. Fauvel et al. [3] introduced a method based on fusing morphological operators and SVMs. Their method yielded high classification accuracy.

Other hand-crafted methods include sparse representation [4], Boltzmann entropy-based band selection [5], fusing correlation coefficient, and sparse self-representation [6], [7] among others. Although these approaches were considered for classification, they are insufficient when coping with the abundant content of HSIs

Deep learning has become a state-of-art approach to extract features for HSIs classification. Instead of depending on shallow hand-designed features, deep learning can automatically obtain hierarchical features from data. Chen et al. [8]

The associate editor coordinating the review of this manuscript and approving it for publication was Geng-Ming Jiang.

pioneered the use of deep learning for HSIs classification. Meanwhile, other deep learning methods are also improved for HSIs classification. For example, Chen et al. [9] applied a stacked autoencoder (SAE) for feature extraction and classification. Other networks, like the deep belief network (DBN) [10], have been proposed for HSIs classification. DBNs, based on multivariate optical sensors, are constructed by stacking restricted Boltzmann machines. Although these methods have greatly improved HSIs classification, they require the training of numerous parameters owing to the full connection of various layers. Moreover, because SAEs and DBNs must represent spatial information as flattened vectors prior to training, they have limited ability to effectively extract spatial information.

Owing to the limitations of SAEs and DBNs, convolutional neural networks (CNNs) are often considered as alternatives. CNN-based models are capable of detecting local features and have achieved improved classification performance relative to fully connected SAEs and DBNs. However, directly applying a 2D-CNN to HSIs classification requires the convolution of each band. Considering that hyperspectral data usually have hundreds of channels, a large number of trained convolution kernels is required to handle the input data. This results in increased computational costs and overfitting.

To address this problem, researchers added dimensional reduction methods to reduce the spectral dimensions [11], [12]. Zhao and Du [13] obtained the spatial features of HSIs using a 2D-CNN model and principal component analysis (PCA). Makantasis et al. [14] used randomised principal component analysis (R-PCA) to compress the spectral dimension of the original HSIs data. Following the data compression, a 2D-CNN extracted features from the dimensionally reduced data. However, this prevents the comprehensive utilization of joint spectral-spatial features, thus affecting classification performance.

A three-dimensional CNN (3D-CNN) was introduced to extract joint HSIs features [15], [16]. Applying the 3D kernel to HSIs classification allowed the convolutions to extract spectral and spatial information simultaneously, and make full use of HSIs structural information. For example, a 3D and 2D data extraction method that combines spectral and spatial information was proposed by Zhang et al. [17]. Zhong et al. [18] proposed a spectral-spatial residual network (SSRN). SSRN uses identity maps to connect other 3D convolution layers. 3D-CNNs greatly increased accuracies.

However, the commonly used linear network architectures adopt a fixed convolution kernel size, inhibiting the ability to deal with different types of datasets flexibly. This limitation also leads to overfitting. Therefore, we developed an improved 3D CNN model for HSIs classification based on previous Inception architecture [19]. The model uses parallel architecture containing different convolution kernels to generate more flexible feature maps. The proposed method differs from the most advanced 3D CNN models in that it uses fewer but parallel 3D convolution operations for the spectral-spatial feature extraction stage and can produce richer feature maps. In addition, aiming at the problem of spectral band redundancy, the dimensionality reduction based on interactive information entropy is applied to the band selection.

The remainder of this paper is organized as follows. The proposed 3D inception frameworks and related information are discussed in Section II. The details of the experiment and its results are outlined in Section III. Finally, Section IV presents the conclusions.

II. RELATED METHODS

A. CONVOLUTION ACROSS DIMENSIONS

1) ONE-DIMENSIONAL CONVOLUTION

Neuroscientists have found that the human visual system is capable of classifying objects efficiently. Based on this finding, numerous data-processing methods for object classification have been developed. One such method is the CNNs, which can process data effectively with few computational parameters, primarily due to the locally connected and weight-sharing mechanisms [20]. Certain connections between the neurons in CNNs are replicated across a layer, and they share the same weights and biases. By adopting specific architecture (e.g., shared weights and local connections), CNNs can provide more optimized generalizations for computer vision problems.

The complete CNN architecture contains convolution, pooling, and fully connected layers. Among these, the convolution layer is the most widely used. Supposing the jth feature map in the ith convolutional layer has neuron at location z. Equation 1 is used to determine its value. ReLu is the most frequently used activation function (Eq. 2) [21].

$$v_{ij}^{z} = f\left(b_{ij} + \sum_{m}\sum_{p=0}^{P_i-1} w_{ijm}^{p} v_{(i-1)m}^{z+p}\right) \tag{1}$$

$$f(x) = \max(0, x) \tag{2}$$

Among them, m is the feature map from the previous layer that is connected to the current feature map. The kernel width, toward the spectral dimension, is P_i. w_{ijm}^{p} is the weight related to the mth feature map at position p. The bias of the jth feature map in the ith layer is b_{ij}.

2) TWO-DIMENSIONAL CONVOLUTION

A lot has been accomplished in image classification and computer vision via 2D-CNNs. The 2D convolution layer is denoted by extending Eq. 1. The value of the neuron v_{ij}^{xy} at position (x, y) of the jth feature map in the ith layer is expressed by Eq. 3.

$$v_{ij}^{xy} = f\left(b_{ij} + \sum_{m}\sum_{p=0}^{P_i-1}\sum_{q=0}^{Q_i-1} w_{ijm}^{pq} v_{(i-1)m}^{(x+p)(y+q)}\right) \tag{3}$$

where, b_{ij} denotes the bias. The feature map in the previous layer that is connected to the current feature map is denoted by m. The value of the convolution kernel at position (p, q)

is w_{ijm}^{pq}. P_i and Q_i represent the height and width of the corresponding convolution kernel in the ith layer, respectively. Finally, the activation function is $f(\cdot)$, which adopted in the proposed method is ReLU.

3) THREE-DIMENSIONAL CONVOLUTION

1D-CNNs extract only the spectral features and 2D-CNNs extract the spatial features. Simultaneously capturing spatial and spectral information and retaining their relationships are imperative when processing the 3D data of HSIs. Therefore, we designed a 3D CNN that obtains the spectral and spatial features of HSIs.

Meanwhile, 3D convolution is implemented by convolving a 3D kernel with 3D data. When processing HSIs data, feature maps are generated by the 3D kernel over multiple contiguous bands in the convolutional layer. In 3D convolution, the activation value v_{ij}^{xyz} of the position (x, y, z) in the jth feature map of the ith layer is calculated in accordance with Eq. 4.

$$v_{ij}^{xyz} = f\left(\sum_{m}\sum_{p=0}^{P_i-1}\sum_{q=0}^{Q_i-1}\sum_{r=0}^{R_i-1} w_{ijm}^{pqr} v_{(i-1)m}^{(x+p)(y+q)(z+r)} + b_{ij}\right) \tag{4}$$

where, m is the feature map in the (i-1)th layer that is connected to the current feature map. P_i and Q_i represent the height and width of the spatial convolution kernel in layer i, respectively. The size of the kernel along the spectral dimension is represented by R_i. w_{ijm}^{pqr} is the value of the weight connected to the jth feature map at position (p, q, r). Finally, $f(\cdot)$ is the activation function. It represents ReLU in the proposed method.

B. ADAPTIVE BAND SELECTION

Due to the high-dimensional characteristics of hyperspectral remote sensing images, their spectral bands inevitably have redundancy, thus the dimensionality reduction method based on mutual information entropy [22] is used for the analysis of hyperspectral remote sensing data. This adaptive band selection achieves effective dimensionality reduction by selecting spectral bands with strong discrimination.

Assuming that the original hyperspectral remote sensing data is represented as $X \in R^{I\times J\times K}$, where X is the original data, I stands for the width of the spectral bands, J is the height, and K is the number of spectral bands. A dimensionality reduction method based on interactive information entropy is used to perform adaptive band selection for hyperspectral remote sensing data, reducing the bands from K to $k(K \ll k)$. Therefore, there are k spectral bands after dimension reduction. At the same time, in order to process hyperspectral remote sensing data more efficiently and apply CNN for classification, the dimensionality-reduced neighborhood data is divided into $N \times N \times k$. Where $N \times N$ is the window size of neighborhood segmentation, that is, the neighborhood size of a pixel in the spatial dimension.

For band selection based on mutual information entropy, it is necessary to use labeled data to select k bands, and maximize the interaction information between samples and labels, which is expressed as follows:

$$\max_{S\in\mathbf{R}} \frac{1}{k}\sum_{j\in S} I\left(X_j;C\right) \tag{5}$$

where, $\mathbf{S}$ contains all the sub-bands of the k bands, and $I(X_j;C)$ represents the interaction information between the sample X_j and the label C, which can be defined by the following formula:

$$I\left(X_j;C\right) = H\left(X_j\right) + H\left(C\right) - H\left(X_j, C\right) \tag{6}$$

Among them, $H(X_j)$ represents the entropy of the sample X_j, and $H(X_j, C)$ represents the joint entropy of the sample X_j and label C. The purpose of $I(X_j;C)$ is to seek common information for the samples X_j and labels C.

The band selection with the largest mutual information does not necessarily preserve the original band information of the image to the greatest extent, because these bands may be adjacent bands with great similarities. While some bands are not rich in information, they can provide complementary information for other bands. In order to further reduce redundancy and minimize the interaction information between different bands, the implementation process is as follows:

$$\min_{S\in\mathbf{R}} \frac{2}{\left(k^2-k\right)} \sum_{j,m\in S;j<m} I\left(X_j;X_m;C\right) \tag{7}$$

where,

$$\begin{aligned} I\left(X_j;X_m;C\right) &= H\left(X_j\right) + H\left(X_m\right) + H\left(C\right) - H\left(X_j, X_m\right) \\ &\quad - H\left(X_j, C\right) - H\left(X_m, C\right) + H\left(X_j, X_m, C\right) \end{aligned} \tag{8}$$

By satisfying equations (5) and (7) at the same time, k bands under the condition of low redundancy can be selected, that is, the following equation can be realized:

$$\max_{S\in\mathbf{R}} \left[\left(\frac{1}{k}\sum_{j\in S} I\left(X_j;C\right) - \frac{2}{\left(k^2-k\right)}\sum_{j,m\in S;j<m} I\left(X_j;X_m;C\right)\right)\right] \tag{9}$$

In the specific implementation process, first, calculating the band with the largest interactive information entropy between the sample and the label in the original band as the initial band, and then selecting the band with the smallest interactive information with the initial band and the largest interactive information entropy between the sample and the label as the second band, forming a band subset. Then selecting the band with the smallest interaction information with all the bands in the sub-band and the band with the largest interactive information entropy between its own sample and the label as the selected sub-band, and so on, selecting the sub-band containing k bands as a result.

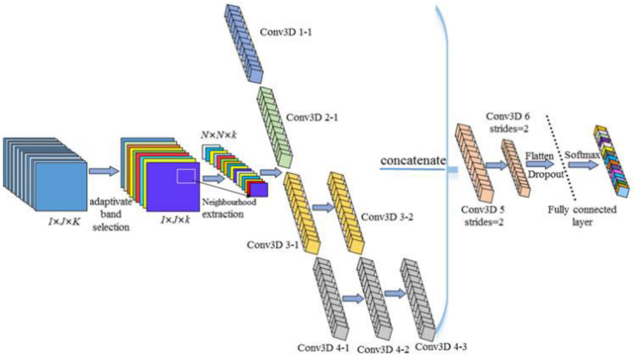


FIGURE 1. Architectural depiction of our proposed 3D inception network.

C. NETWORK ARCHITECTURE AND PARAMETER SETTING

The Inception network architecture is popular in CNNs. Instead of following the typical sequential process, this architecture stacks convolutional modules. These modules are organised into parallel branches and can be regarded as small independent networks. An inception network can use multiple kernel sizes for different branches, allowing for the generation of more flexible feature maps [23]. The addition of a 1×1 convolution operation in the inception architecture also allows for dimensional reduction. That significantly reduces the number of parameters and operations, saving computational costs and memory resources.

Both spectral and spatial factors influence a pixel's class label prediction. The label of a pixel is reflective of its spectral value in different spectra. Because pixels that are geographically close tend to belong to the same class, the labels of surrounding pixels should be considered when predicting the class label of a given pixel [24]. Therefore, hyperspectral image classification methods should consider the spectral and spatial factors simultaneously. 3D convolution does just that.

In this study, we present a 3D inception network for classifying hyperspectral images. Figure 1 is an illustration of the proposed network. The network has four branches and follows the general architecture of the Inception. The first branch has a 3D convolutional layer of $1 \times 1\times 1$. The second branch has a $3 \times 3 \times 3$ convolutional layer. The third branch has a $1 \times 1 \times 1$ and a $5 \times 5 \times 5$ convolutional layer. The third branch has a $1 \times 1 \times 1$ and a $5 \times 5 \times 5$ convolutional layer. It does not have a pooling layer, to retain more information. The fourth branch adds a $3 \times 3 \times 3$ convolutional layer to the general Inception architecture. Therefore, altogether, the fourth branch has a $1 \times 1 \times 1$ convolutional layer and two $3 \times 3 \times 3$ convolutional layers. Different branches with different numbers of convolutional layers are used to extract the features at different levels. To ensure the input of the deeper layers are filled with sufficient information and for the concatenation operation, convolutions in the four branches use the padding strategy. The outputs of each branch are concatenated and inputted into the next layer. After concatenation, the convolution kernel size is $3\times 3 \times 3$. The stride is set to 2 for the last two convolutional layers prior to flattening. In summary, we used convolution kernels of three different sizes in the proposed network (i.e., $1 \times 1 \times 1$, $3 \times 3 \times 3$, and $5 \times 5 \times 5$) to allow for the generation of more flexible feature maps. The convolution kernels were set to smaller sizes rather than larger ones. Because these kernels have relatively small receptive fields, the receptive fields can be enlarged by deepening the convolutional layers. Additionally, varying numbers of layers in the branches contribute to different feature extraction levels. Finally, the $1 \times 1 \times 1$ convolution operation significantly reduces the computational cost by reducing the number of parameters.

TABLE 1. Details of the proposed 3D inception network architecture.

Layer	Kernel size	Parameters
Conv3D 1-1	(1, 1, 1, 16)	32
Conv3D 2-1	(3, 3, 3, 16)	6928
Conv3D 3-1	(1, 1, 1, 16)	32
Conv3D 3-2	(5, 5, 5, 16)	32
Conv3D 4-1	(1, 1, 1, 16)	32,016
Conv3D 4-2	(3, 3, 3, 16)	448
Conv3D 4-3	(3, 3, 3, 16)	6928
Concatenate	—	0
Conv3D 5	(3, 3, 3, 64)	110,656
Conv3D 6	(3, 3, 3, 64)	110,656
Flatten	—	0
Dropout(0.5)	—	0
Dense	12	9228
Total trainable parameters	—	276,956

The parameters of our 3D inception network are trained via gradient descent (i.e., backpropagation). Moreover, we added dropout to the fully connected layer. Table 1 shows the details of the proposed 3D inception network.

Because the original CNN was developed for 2D image classification, the usefulness of 2D convolution should be tested on the HSIs classification. We used the 3D inception network as the feature extractor and replaced the last two 3D convolution layers with two 2D convolution layers. Table 2 lists this modified network's architecture. Additionally, 3D and 2D convolution are designed for hyperspectral image classification. To maintain integrity, the effectiveness of 1D convolution for HSIs classification should also be investigated. We also tested additional modifications. We replaced the last two 3D convolution layers with a 2D convolution layer and a 1D convolution layer. This architecture contains three different dimensional convolutions. This modified network's architecture is shown in Table 3.

III. EXPERIMENTS AND RESULTS

The experiments sought to verify the ability of the proposed inception networks to classify hyperspectral data. The model was trained with Keras, a deep learning framework, on Tensorflow. The experiments were carried out on a desktop equipped with an Intel I5-9400F CPU and an Nvidia GeForce

TABLE 2. Details of the proposed 3D-2D inception network architecture.

Layer	Kernel size	Parameters
Conv3D 1-1	(1, 1, 1, 16)	32
Conv3D 1-2	(5, 5, 5, 16)	32
Conv3D 2-1	(3, 3, 3, 16)	6928
Conv3D 3-1	(1, 1, 1, 16)	32,016
Conv3D 3-2	(3, 3, 3, 16)	448
Conv3D 3-3	(3, 3, 3, 16)	6928
Conv3D 4-1	(1, 1, 1, 16)	32
Concatenate	—	0
Reshape	—	0
Conv2D 1	(3, 3, 64)	36,928
Conv2D 2	(3, 3, 64)	36,928
Flatten	—	0
Dropout(0.5)	—	0
Dense	12	94,476
Total trainable parameters	—	214,748

TABLE 3. Details of the proposed 3D-2D-1D inception network architecture.

Layer	Kernel size	Parameters
Conv3D 1-1	(1, 1, 1, 16)	32
Conv3D 1-2	(5, 5, 5, 16)	32
Conv3D 2-1	(3, 3, 3, 16)	6928
Conv3D 3-1	(1, 1, 1, 16)	32,016
Conv3D 3-2	(3, 3, 3, 16)	448
Conv3D 3-3	(3, 3, 3, 16)	6928
Conv3D 4-1	(1, 1, 1, 16)	32
Concatenate	—	0
Reshape	—	0
Conv2D	(3, 3, 64)	36,928
Reshape	—	0
Conv1D	(3, 3, 3, 64)	86,080
Flatten	—	0
Dropout(0.5)	—	0
Dense	12	31,500
Total trainable parameters	—	200,924

TABLE 4. A summary of the number of classes on the KSC dataset.

Class	Training	Validation	Test
Scrub	76	76	609
willow swamp	24	24	194
wabbage palm hammock	26	26	205
cabbage palm/oak	25	25	202
slash pine	16	16	129
oak/broadleaf hammock	23	23	183
hardwood swamp	11	11	84
graminoid marsh	43	43	345
spartina marsh	52	52	416
cattail marsh	40	40	323
salt marsh	42	42	335
mud flats	50	50	402
Water	93	93	742
Total	521	521	4169

TABLE 5. A summary of the number of classes on the PC dataset.

Class	Training	Validation	Test
water	21	21	782
trees	21	21	778
asphalt	20	20	776
self-blocking	20	20	776
bitumen	20	20	768
titles	32	32	1196
shadows	12	12	452
meadows	21	21	782
bare soil	21	21	778
Total	188	188	7080

GTX 1660 GPU. The GPU was supported by the compute unified device architecture v10.0.

A. DATASET DESCRIPTION AND EXPERIMENTAL DESIGN

The first dataset was obtained from the Kennedy Space Center (KSC). It was an image of Florida, taken by the AVIRIS sensor in 1996. The 512 × 614-pixel image had 176 spectral bands that had spatial resolutions of 20 m per pixel. The bands' wavelengths ranged from 400 to 2500 nm. The KSC dataset contained 13 classes. The training, validation, and test samples were random 10%, 10%, and 80%, respectively. Table 4 lists the number of samples per class.

The second dataset was from the Pavia Centre (PC) and was taken by the ROSIS sensor during flights over Pavia in northern Italy. The study area included 9 classes. The 1096 × 1096-pixel image had 102 spectral bands with a geometric resolution of 1.3 m. This image has relatively extensive data. The training, validation, and test samples were random 2.5, 2.5, and 95%, respectively. Table 5 lists the number of samples per class.

The third dataset was collected by the Hyperion sensor on board EO-1 as it passed over the Okavango Delta in Botswana. The original data had 242 spectral bands covering wavelengths from 400 to 2500 nm. After removing uncalibrated and noisy bands that covered water absorption features, the remaining 145 bands are included as candidate features. The Botswana dataset contained 14 labelled classes, and the image had a resolution of 1476 × 256 pixels. The training, validation, and test samples were random 10, 10, and 80%, respectively. Table 6 lists the number of samples per class.

Several hyperparameters of the convolutional network were set according to experimental performance and experience. In our experiments, we chose Adam with the learning rate of 1e-3 and a decay term of 1e-6 for model training. We used trial and error to determine the optimal parameters of the networks. A mini-batch-based back-propagation method was used to optimize the parameters. The training epoch was set as 400 with mini-batches of 16.

TABLE 6. A summary of the number of classes on the botswana dataset.

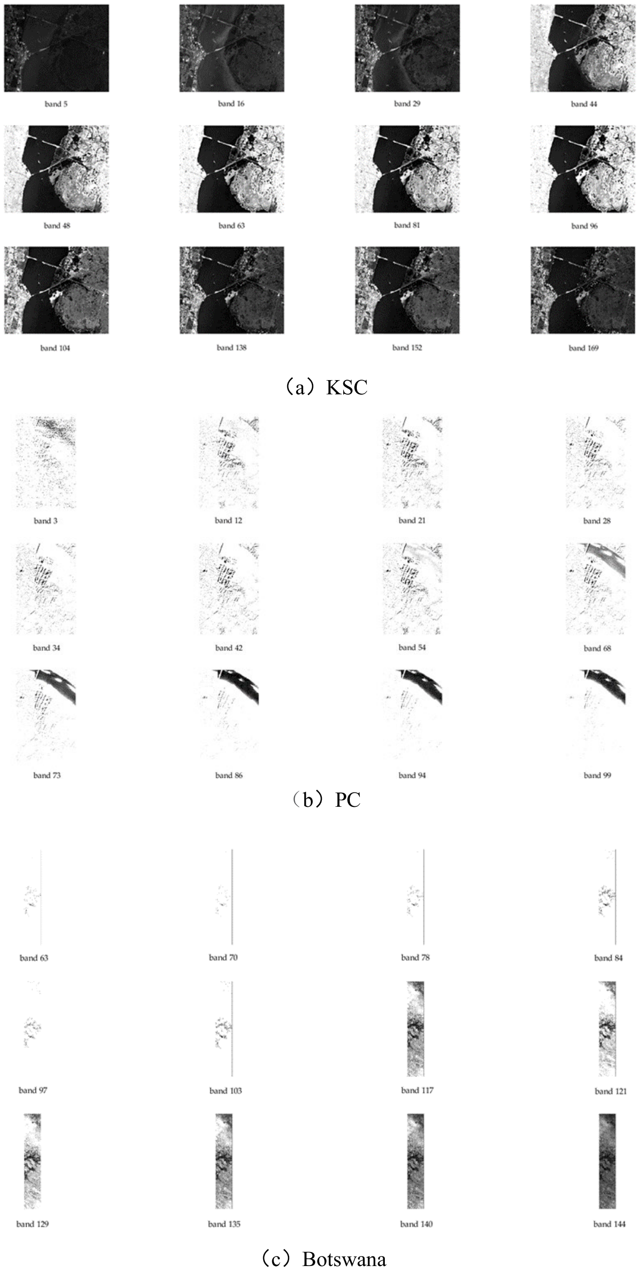
Class	Training	Validation	Test
Water	27	27	216
hippo grass	10	10	81
floodplain grasses 1	25	25	201
floodplain grasses 2	22	22	172
Reeds	27	27	215
riparian	27	27	215
firescar	26	26	207
island interior	20	20	162
acacia woodlands	31	31	251
acacia shrublands	25	25	199
acacia grasslands	30	30	244
short mopane	18	18	145
mixed mopane	27	27	215
exposed soils	10	10	76
Total	325	325	2598

The Kappa coefficient ($K \times 100$), overall accuracy (OA), average accuracy (AA), and execution time were used to measure the performance of the tested models.

B. ADAPTIVE BAND SELECTION RESULTS AND ANALYSIS

Because of the high dimensional characteristics of hyperspectral image, when the number of bands is too large, it does not produce a better classification effect, and in fact the classification effect decreases, especially when the number of training samples per class is small compared to the number of bands, the curse of dimensionality problem will be more pronounced. The curse of dimensionality restricts the classification ability of CNNs and easily leads to overfitting problems. The curse of dimensionality of hyperspectral remote sensing data can be fixed by adaptive band selection. Through the principle of adaptive band selection based on mutual information entropy, we can see that the adaptive band selection method based on mutual information entropy can filter out the bands with high discrimination from each another, and effectively reduce the band redundancy. Figure 2 shows the visualization of the selected bands on the three datasets by the adaptive band selection method based on mutual information entropy.

For the three datasets, it can be seen that the adaptive band selection method based on mutual information entropy selects the representative bands, which almost cover the range of each band of the dataset. On the Botswana dataset, although the first few bands selected contain less information, they can provide complementary information for other bands. For the KSC dataset, the filtered bands are [5, 16, 29, 44, 48, 63, 81, 96, 104, 138, 152, 169], respectively. For the PC dataset, the filtered bands are [3, 12, 21, 28, 34, 42, 54, 68, 73, 86, 94, 99], and for the Botswana dataset, the filtered bands are [63, 70, 78, 84, 97, 103, 117, 121, 129, 135, 140, 144], respectively.



（a）KSC

（b）PC

（c）Botswana

FIGURE 2. Selected spectral bands with adaptive band selection method.

C. EFFECTS OF THE TRAINING RATIO

We also investigated how the proposed 3D inception model's classification performance might be impacted by the ratio of training samples. The classification results with different numbers of training samples are shown in Table 7, where the random training ratio varied from 5 to 15 for the KSC and Botswana datasets, and the PC dataset from 0.5 to 10.

From Table 7, it is evident that the performance of the 3D inception network improves with the increasing training ratios. The experimental finding also indicates that deep convolutional networks-based hyperspectral image classifiers

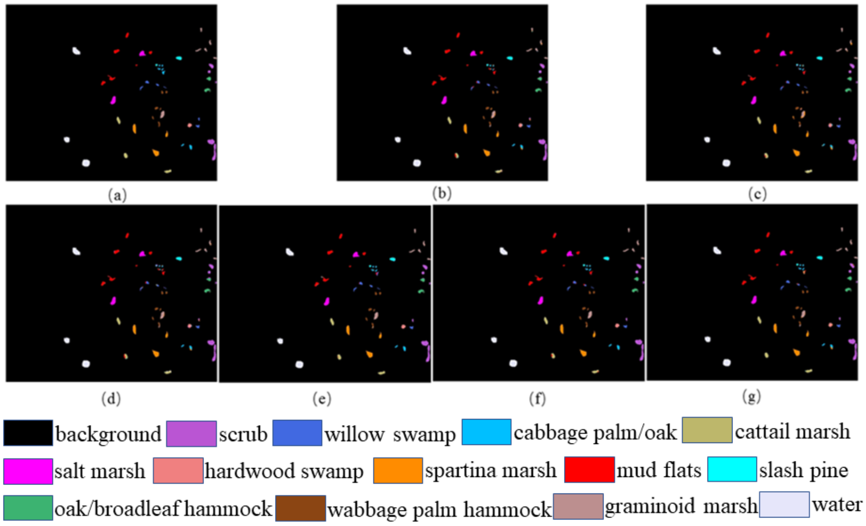


FIGURE 3. Classification maps for the methods of Table 7 on the KSC dataset. (a) ground truth; (b) SyCNN-S; (c) 3D-CNN; (d) 3-D DL; (e) 3D inception; (f) 3D-2D inception; and (g) 3D-2D-1D inception.

TABLE 7. Overall accuracies(%) obtained by the 3D inception network for different training ratios.

Dataset	Training ratio (%)				
	5	7.5	10	12.5	15
KSC	93.14	95.45	96.43	96.84	97.07
Botswana	96.39	98.44	99.46	99.57	99.63
PC	0.5	2.5	5	7.5	10
	98.30	99.73	99.81	99.86	99.88

typically require more training data to improve classification performance. Based on the experimental results, it is obvious that the proposed method can perform well in classification tasks even when only a small training ratio is available.

D. DISCUSSING AND COMPARING THE RESULTS

1) EXPERIMENTS ON THE KSC DATASET

On the KSC dataset, 19 × 19 ×12 image cubes were extracted from the original image by the dimensional-reduction. Table 8 provides a detailed comparison of the classification results of the three proposed networks (i.e., 3D inception network, 3D-2D inception network, 3D-2D-1D inception network), and three other typical classification methods (i.e., SyCNN-S [25], 3D-CNN [26], and 3-D DL [27]). The classification accuracies were also depicted visually. The classification maps of the compared methods for the KSC dataset are displayed in Fig. 3.

The KSC dataset only contained 521 training samples. From Table 8, based on the $K\times100$, OA, and AA metrics, the proposed 3D inception network showed the best classification results. Relative to the 3D-2D inception network, the proposed 3D inception network yielded accuracy increments of 0.73, 0.65 and 0.81, in the $K\times100$, OA, and AA metrics, respectively. The proposed 3D inception network yielded accuracy increments of 1.24, 1.11 and 1.49 in the $K\times100$, OA, and AA performance metrics, respectively, relative to the 3D-2D-1D inception network. The overall accuracies of the 3D-2D and 3D-2D-1D inception networks were slightly lower than that of the 3D inception network, probably due to the information loss inherent to 1D and 2D convolution compared to 3D convolution. Without any post-processing or pre-processing, 3D-CNN views the HSI cube data in its entirety to extract the deep spectral-spatial combination characteristics, but it has poor classification performance due to the problem of segmentation scale, notably on the ground objects of hardwood swamp and willow swamp.

The accurate distinction of cabbage palm/oak proved difficult owing to their unrecognisable appearance and limited training samples. Additionally, classification accuracies for certain classes were degraded by the lack of training samples (e.g., willow swamps). Graminoid marsh and cattail marsh are very similar in appearance, but the three proposed inception networks distinguished between the two classes well.

The classification results of the classification maps in Fig. 3 do not qualitatively show much difference between the methods. However, Table 8 quantitatively shows that the 3D inception network yielded optimal performance. Compared with 3D-CNN, inception networks were able to reduce the classification noise to some extent, but they cannot remove them entirely. An examination of Fig. 3 carefully reveals that the inception networks performed better than the other three tested methods, and the 3D inception network's classification map was almost identical to the ground truth map.

2) EXPERIMENTS ON THE PC DATASET

On the PC dataset, 17 × 17 ×12 image cubes were extracted from the original image by the dimensional-reduction.

TABLE 8. Comparisons of classification accuracies among different methods on the ksc dataset.

Class	SyCNN-S	3D-CNN	3-D DL	3D inception	3D-2D inception	3D-2D-1D inception
scrub	99.34	95.89	99.34	100	100	96.72
willow swamp	74.23	77.84	68.56	82.47	75.77	81.96
wabbage palm hammock	86.34	94.63	85.37	92.68	94.15	89.27
cabbage palm/oak hammock	47.52	50.50	40.59	75.25	71.78	64.36
slash pine	97.67	92.25	82.17	89.92	92.25	98.45
oak/broadleaf hammock	95.08	91.80	92.35	97.27	97.81	100
hardwood swamp	89.29	71.43	79.76	100	100	88.10
graminoid marsh	86.09	80.00	92.46	97.68	98.55	94.78
spartina marsh	100	100	99.76	100	100	97.84
cattail marsh	77.71	86.07	80.50	98.14	91.02	99.38
salt marsh	99.70	99.10	99.70	100	100	99.70
mud flats	99.25	91.04	96.02	95.52	97.01	99.00
water	100	100	100	100	100	100
K×100	91.08	89.80	89.92	96.02	95.29	94.78
OA (%)	92.01	90.86	90.98	96.43	95.78	95.32
AA (%)	88.63	86.97	85.89	94.53	93.72	93.04
Execution time (s)	447.69	105.25	691.57	251.76	157.15	106.82

TABLE 9. Comparisons of classification accuracies among different methods on the PC dataset.

Class	SyCNN-S	3D-CNN	3-D DL	3D inception	3D-2D inception	3D-2D-1D inception
water	99.99	99.98	99.86	100	100	100
trees	98.43	91.66	95.44	99.28	99.17	97.67
asphalt	69.81	79.90	79.15	97.51	96.59	99.25
self-blocking bricks	95.18	45.51	92.00	99.92	99.73	99.88
bitumen	92.18	88.17	66.83	99.02	99.54	99.18
titles	96.06	87.09	96.05	99.72	99.67	99.85
shadows	88.96	84.73	87.10	99.38	99.15	98.41
meadows	99.11	99.25	99.33	99.88	99.87	99.85
bare soil	97.43	69.56	98.93	97.21	95.18	95.15
K×100	96.82	93.28	95.10	99.62	99.53	99.44
OA (%)	97.75	95.27	96.55	99.73	99.67	99.60
AA (%)	93.02	82.87	90.52	99.10	98.77	98.81
Execution time (s)	917.55	248.72	1337.42	533.93	412.66	381.68

Similar to the KSC dataset, Table 9 compares, in detail, the classification results of the tested methods on the PC dataset.

The PC dataset only had 2.5% of the samples set aside for training. From Table 9, based on the $K \times 100$, OA, and AA metrics, the proposed 3D inception network performed optimally. It yielded scores of 99.62, 99.73, and 99.10% for $K \times 100$, OA, and AA, respectively. The overall accuracies of the 3D-2D and 3D-2D-1D inception networks were slightly lower, probably owing to the information loss inherent to 1D and 2D convolution relative to 3D convolution. The classification accuracy obtained by 3D-CNN is relatively low. 3-D DL combines the conventional CNN network with a unique twist by using 3-D convolution operations, while its average classification performance is poor.

More specifically, because of its small size and unrecognisable appearance, self-blocking bricks can hardly be distinguished. Additionally, the classification accuracies were also degraded by the lack of training samples for certain class (e.g., asphalt). Despite those challenges, the three proposed inception networks performed well for those classes. Finally, the classification accuracies were depicted visually. Figure 4 shows the classification maps of the tested methods for the PC dataset.

Figure 4 shows that except for the class of shadows, there does not seem to be much difference between the classification results of these methods from naked eye observation. The outstanding performance of the proposed inception networks is evident. The classification maps of the suggested inception networks and the ground truth map are almost exact replicas. The morphology of the ground objects is well preserved, and their internal smoothness is significantly higher. This again highlights the usefulness of the extracted features of the inception networks for HSI classification.

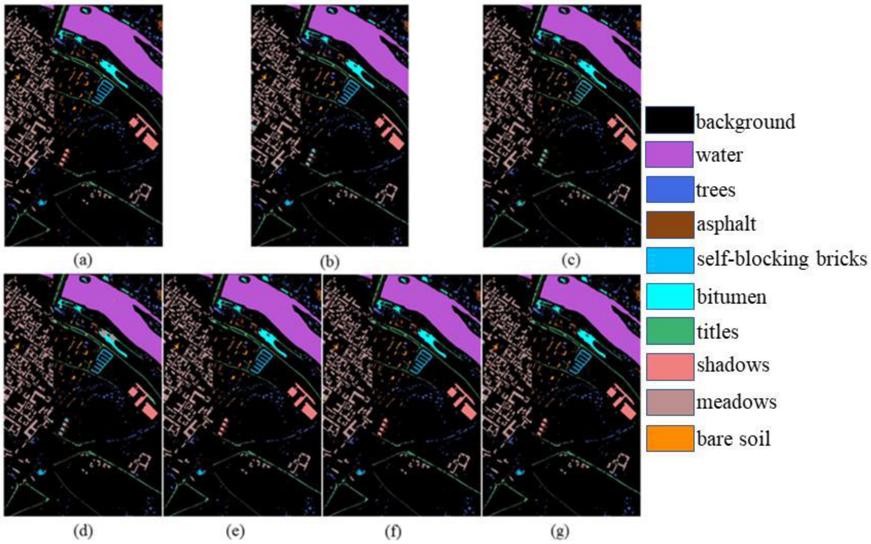


FIGURE 4. Classification maps for the methods of Table 8 on the PC dataset. (a) ground truth; (b) SyCNN-S; (c) 3D-CNN; (d) 3-D DL; (e) 3D inception; (f) 3D-2D inception; and (g) 3D-2D-1D inception.

TABLE 10. Comparisons of classification accuracies among different methods on the botswana dataset.

Class	SyCNN-S	3D-CNN	3-D DL	3D inception	3D-2D inception	3D-2D-1D inception
Water	100	100	100	100	100	100
hippo grass	96.30	100	91.21	100	100	100
floodplain grasses 1	97.51	95.02	97.35	100	100	100
floodplain grasses 2	100	100	99.48	100	100	100
Reeds	91.63	88.84	93.39	99.07	96.74	91.63
Riparian	92.56	92.56	84.71	97.21	99.07	99.53
Firescar	100	100	97.85	98.55	100	96.28
island interior	95.68	100	100	100	99.38	100
acacia woodlands	98.80	99.60	100	99.60	99.20	99.60
acacia shrublands	88.44	89.95	87.44	100	100	100
acacia grasslands	96.31	94.67	97.45	99.18	98.77	100
short mopane	99.31	90.34	91.41	100	98.62	100
mixed mopane	98.14	99.07	99.59	100	100	100
exposed soils	85.53	82.89	100	100	97.37	98.68
K×100	95.83	95.29	95.44	99.42	99.21	99.12
OA (%)	96.15	95.65	95.79	99.46	99.27	99.19
AA (%)	95.73	95.21	95.71	99.54	99.23	99.25
Execution time (s)	58.62	19.03	84.53	29.28	29.16	28.11

3) EXPERIMENTS ON THE BOTSWANA DATASET

On the Botswana dataset, 19 × 19 ×12 image cubes were extracted from the original image by the dimensional-reduction. Similar to the KSC and PC datasets, Table 9 provides a detailed comparison of the classification results of the tested methods on the Botswana dataset.

The Botswana dataset only had 325 samples for training. Based on the $K\times100$, OA, and AA metrics shown in Table 10, the proposed 3D inception network yielded optimal classification results. It yielded 99.42, 99.46%, and 99.54% for $K\times100$, OA, and AA, respectively. The 2D and 1D convolutions in the proposed, modified inception networks do not seem to significantly decrease classification accuracies, probably because the total number of trainable parameters vary slightly between the three inception networks. Relative to two 3D convolutional layers, the information loss of a 1D and 2D convolutional layer is marginal. Our approach can simultaneously identify the more distinctive spectral bands and reduce the spectral dimension, and the inception networks can extract the diversified features to enhance the ability to identify ground objects, resulting in superior classification results.

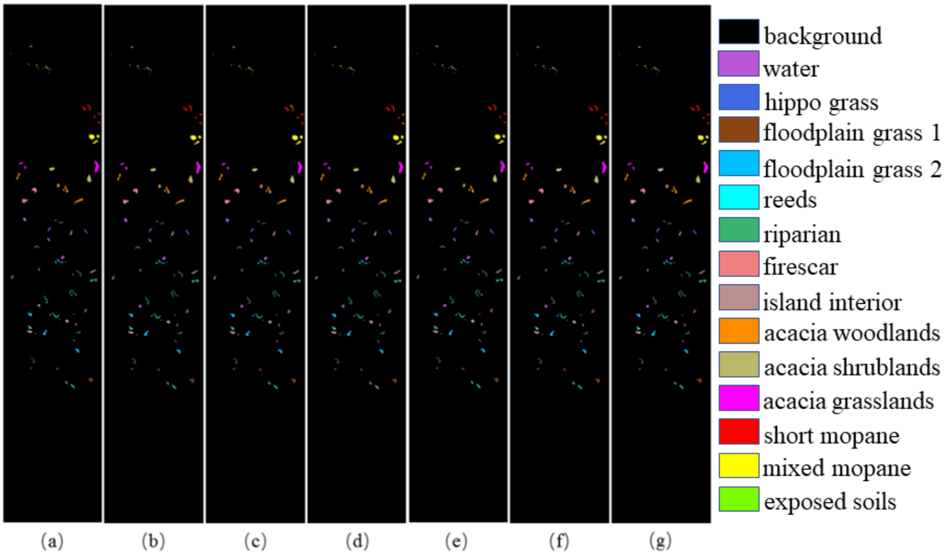


FIGURE 5. Classification maps for the methods of Table 8 on the Botswana dataset. (a) ground truth; (b) SyCNN-S; (c) 3D-CNN; (d) 3-D DL; (e) 3D inception; (f) 3D-2D inception; and (g) 3D-2D-1D inception.

More specifically, the distinction of the exposed soils class is difficult because it occupies a small area in the study area. This phenomenon is especially obvious for the 3D-CNN and SyCNN-S methods. In addition, acacia shrublands and acacia grasslands are very similar in appearance, it is hard to distinguish them, but the three proposed inception networks categorized them remarkably.

Tables 8 to 10 also present the execution time of different methods. Execution time includes training and testing time. It can be seen that even though the proposed models have a number of parameters, this does not significantly increase the complexity of the models. The execution time is not as high as 3-D DL and SyCNN-S's. The possible reason is that the models have more operations such as convolution and data interaction. 3D-CNN only has two 3D convolution layers, and there are few 3D kernels, so its execution time is the shortest. However, its classification accuracy is inferior compared with the proposed methods. The classification accuracies were visually depicted for the Botswana dataset. Figure 5 shows the classification maps of the methods.

The suggested inception networks exhibit remarkable performance in Fig. 5, and the classification maps do not seem to qualitatively differ much between them. However, Table 10 quantitatively shows that the 3D inception network yields optimal results. Additionally, it is evident from the classification details that 3D inception performs superbly when classifying small, irregularly shaped ground objects.

IV. CONCLUSION

In this study, we presented a 3D inception network based on adaptive band selection for spectral-spatial joint feature extraction and classification to improve HSIs classification accuracy. Dimensionality reduction based on interactive information entropy seeks to find more distinctive and informative spectral bands. The proposed 3D inception network provides excellent classification accuracy with limited training samples. The improved 3D-2D and 3D-2D-1D inception networks also yield satisfactory classification rates. The experimental results also demonstrated that the proposed approaches have potential to help boost the classification performance.

REFERENCES

[1] L. Fang, N. He, S. Li, A. J. Plaza, and J. Plaza, "A new spatial–spectral feature extraction method for hyperspectral images using local covariance matrix representation," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 6, pp. 3534–3546, Jun. 2018.

[2] S. Samiappan, S. Prasad, and L.M. Bruce, "Non-uniform random feature selection and kernel density scoring with SVM based ensemble classification for hyperspectral image analysis," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 6, no. 2, pp. 792–800, Apr. 2013.

[3] M. Fauvel, J. A. Benediktsson, J. Chanussot, and J. R. Sveinsson, "Spectral and spatial classification of hyperspectral data using SVMs and morphological profiles," *IEEE Trans. Geosci. Remote Sens.*, vol. 46, no. 11, pp. 3804–3814, Nov. 2008.

[4] B. Tu, X. Zhang, X. Kang, G. Zhang, J. Wang, and J. Wu, "Hyperspectral image classification via fusing correlation coefficient and joint sparse representation," *IEEE Geosci. Remote Sens. Lett.*, vol. 15, no. 3, pp. 340–344, Mar. 2018.

[5] P. Gao, J. Wang, H. Zhang, and Z. Li, "Boltzmann entropy-based unsupervised band selection for hyperspectral image classification," *IEEE Geosci. Remote Sens. Lett.*, vol. 16, no. 3, pp. 462–466, Mar. 2019.

[6] P. Hu, X. Liu, Y. Cai, and Z. Cai, "Band selection of hyperspectral images using multiobjective optimization-based sparse self-representation," *IEEE Geosci. Remote Sens. Lett.*, vol. 16, no. 3, pp. 452–456, Mar. 2019.

[7] B. Zhou, B. Li, X. He, H. Liu, and F. Wang, "Classification of camouflages using hyperspectral images combined with fusing adaptive sparse representation and correlation co-efficient," *Spectrosc. Spectral Anal.*, vol. 41, no. 12, pp. 3851–3856, 2022.

[8] Y. Chen, Z. Lin, X. Zhao, G. Wang, and Y. Gu, "Deep learning-based classification of hyperspectral data," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 7, no. 6, pp. 2094–2107, Jun. 2014.

[9] Y. Chen, X. Zhao, and X. Jia, "Spectral–spatial classification of hyperspectral data based on deep belief network," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 8, no. 6, pp. 2381–2392, Jun. 2015.

[10] S. Zhao and H. Lang, "Improving deep subdomain adaptation by dual-branch network embedding attention module for SAR ship classification," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 15, pp. 8038–8048, 2022.

[11] J. Feng, J. Chen, L. Liu, X. Cao, X. Zhang, L. Jiao, and T. Yu, "CNN-based multilayer spatial–spectral feature fusion and sample augmentation with local and nonlocal constraints for hyperspectral image classification," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 12, no. 4, pp. 1299–1313, Apr. 2019.

[12] N. He, M. E. Paoletti, J. M. Haut, L. Fang, S. Li, A. Plaza, and J. Plaza, "Feature extraction with multiscale covariance maps for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 2, pp. 1–15, Aug. 2018.

[13] W. Zhao and S. Du, "Spectral–spatial feature extraction for hyperspectral image classification: A dimension reduction and deep learning approach," *IEEE Trans. Geosci. Remote Sens.*, vol. 54, no. 8, pp. 4544–4554, Apr. 2016.

[14] 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)*, Milan, Italy, Jul. 2015, pp. 4959–4962.

[15] H. Firat, M. E. Asker, M. I. Bayindir, and D. Hanbay, "3D residual spatial–spectral convolution network for hyperspectral remote sensing image classification," *Neural Comput. Appl.*, vol. 35, no. 6, pp. 4479–4497, Oct. 2022.

[16] W. Pi, J. Du, Y. Bi, X. Gao, and X. Zhu, "3D-CNN based UAV hyperspectral imagery for grassland degradation indicator ground object classification research," *Ecol. Informat.*, vol. 62, May 2021, Art. no. 101278.

[17] H. Zhang, M. Wang, F. Wang, G. Yang, Y. Zhang, J. Jia, and S. Wang, "A novel squeeze-and-excitation W-Net for 2D and 3D building change detection with multi-source and multi-feature remote sensing data," *Remote Sens.*, vol. 13, no. 3, p. 440, Jan. 2021.

[18] Z. Zhong, J. Li, Z. Luo, and M. Chapman, "Spectral–spatial residual network for hyperspectral image classification: A 3-D deep learning framework," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 2, pp. 847–858, Feb. 2018.

[19] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Boston, MA, USA, Jun. 2015, pp. 1–9.

[20] S. Bera and V.K. Shrivastava, "Analysis of various optimizers on deep convolutional neural network model in the application of hyperspectral remote sensing image classification," *Int. J. Remote Sens.*, vol. 41, no. 7, pp. 2664–2683, Dec. 2019.

[21] Y. Chen, H. Jiang, C. Li, X. Jia, and P. Ghamisi, "Deep feature extraction and classification of hyperspectral images based on convolutional neural networks," *IEEE Trans. Geosci. Remote Sens.*, vol. 54, no. 10, pp. 6232–6251, Jul. 2016.

[22] H. Matsuda, "Physical nature of higher-order mutual information: Intrinsic correlations and frustration," *Phys. Rev. E, Stat. Phys. Plasmas Fluids Relat. Interdiscip. Top.*, vol. 62, no. 3, pp. 3096–3102, Sep. 2000.

[23] D. Ruiz Hidalgo, B. Bacca Cortés, and E. Caicedo Bravo, "Data classification of hyperspectral images based on inception networks and extended attribute profiles," *Int. J. Remote Sens.*, vol. 41, no. 22, pp. 8717–8738, Sep. 2020.

[24] X. Yang, Y. Ye, X. Li, R. Y. K. Lau, X. Zhang, and X. Huang, "Hyperspectral image classification with deep learning models," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 9, pp. 5408–5423, Sep. 2018.

[25] X. Yang, X. Zhang, Y. Ye, R. Y. K. Lau, S. Lu, X. Li, and X. Huang, "Synergistic 2D/3D convolutional neural network for hyperspectral image classification," *Remote Sens.*, vol. 12, no. 12, p. 2033, Jun. 2020.

[26] Y. Li, H. Zhang, and Q. Shen, "Spectral–spatial classification of hyperspectral imagery with 3D convolutional neural network," *Remote Sens.*, vol. 9, no. 1, p. 67, Jan. 2017.

[27] A. B. Hamida, A. Benoit, P. Lambert, and C. B. Amar, "3-D deep learning approach for remote sensing image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 8, pp. 4420–4434, Aug. 2018.

XIAOXIA ZHANG received the Ph.D. degree from the College of Geophysics, Chengdu University of Technology. She is currently a Lecturer with the School of Software Engineering, Chengdu University of Information Technology, Chengdu, China. Her research interests include hyperspectral image classification, high-resolution image processing, and computer vision.