An Effective Classification Method for Hyperspectral Image With Very High Resolution Based on Encoder–Decoder Architecture

Zhen Zhang, Tao Jiang, Chenxi Liu, and Linjing Zhang

***Abstract*—Hyperspectral images with very high resolution (VHR-HSI) have become considerably valuable due to their abundant spectral and spatial details. Classification of hyperspectral images (HSIs) is a basic and important procedure for diverse applications. However, low interclass spectral variability and high intraclass spectral variability in VHR-HSI, shadows, pedestrians, and low signal-to-noise ratio increase the fuzziness of different categories. To address the known challenges of VHR-HSI classification, an effective classification method based on encoder–decoder architecture is proposed. The proposed algorithm is an object-level contextual convolution neural network based on an improved residual network backbone with 3-D convolution, which fully considers the spatial–spectral and contextual features of HSIs. Two different spatial resolution aerial HSIs are used as experimental data. The results show that the overall accuracy of the proposed method is improved by 7.42% and 18.82%, respectively, compared to the pixelwise convolution neural network and DeepLabv3 algorithm, which is extraordinarily suitable for HSI classification with very high spatial resolution.**

***Index Terms*—Encoder–decoder, hyperspectral image (HSI) with high spatial resolution, image classification, 3-D convolution residual network.**

I. Introduction

HYPERSPECTRAL images (HSIs) consist of abundant spectral information covering the visible, near-infrared, and shortwave infrared bands and detailed ground spatial information. With the increasing demand for spatial resolution and the continuous innovation of unmanned aerial vehicle and airborne remote sensing technology, hyperspectral images with very high resolution (VHR-HSI) became considerably valuable due to their abundant spectral and spatial details, and are widely used in geological surveys [1], urban planning and management [2], crop growth monitoring [3], object detection [4], and environment monitoring [5], [6].

Classification of hyperspectral images is the basic procedure for these applications. At present, hyperspectral classification methods are mainly divided into two categories. One is based on spectral information, while the other is based on spectral–spatial information jointly. Traditional machine learning classification algorithms for hyperspectral images, such as k-nearest neighbor [7], spectral angle mapping [8], and multinomial logistic regression [9] are basically the former one. In order to improve classification performance, later researches used support vector machine [10], principal component analysis (PCA) [11], and independent component analysis [12] to decrease the redundant informative features. However, these methods do not fully consider spatial information, resulting in unsatisfactory classification accuracy. Subsequently, some spatial–spectral joint classification algorithms have emerged, such as morphological profiles [13], Markov random fields [14], [15], loop belief propagation [16], and sparse representation [17]. These methods promote the accuracy of classification to a certain extent, but establishing complex function expressions by above methods is extremely difficult due to the limitation of their structures. Besides, the poor generalization ability is another problem that makes the shallow machine learning methods unsuitable for VHR-HSI classification. Recently, deep learning has achieved remarkable success in image classification [18], [19], object detection [20], and other computer vision [21]. Some excellent classification algorithms have been developed for HSI data. Chen Y [22] took the lead in using stacked autoencoder for HSI classification. Subsequently, segmented stacked autoencoder [23], [24] and sparse autoencoder [25] were proposed to reduce complexity and improve efficacy. Deep belief network model is another typical deep learning network architecture similar to SAE, and it uses a multiple stacked restricted boltzmann machine to extract hierarchical features of HSI data [26]. The similarity between the two deep learning networks of SAE and DBN is that only 1-D vector data can be input, and the spatial features are flattened into a vector without considering the spatial adjacent relationship, which limits the classification performance. Meanwhile, the full connection between different layers adopted in SAE and DBN requires training a large number of weight parameters, leading to a huge calculation cost. The convolutional neural network (CNN) proposed by Krizhevsky [27] became the most

Manuscript received August 31, 2020; revised November 12, 2020; accepted December 14, 2020. Date of publication December 21, 2020; date of current version January 8, 2021. This work was supported in part by the National Natural Science Foundation of China under Grant 41801385, in part by the Shandong Provincial Natural Science Foundation under Grant ZR2018BD004 and Grant ZR2019QD010, and in part by the Shandong Province Key Research and Development Program of China under Grant 2019GGX101049. *(Corresponding author: Linjing Zhang.)*

The authors are with the College of Geodesy and Geomatics, Shandong University of Science and Technology, Qingdao 266590, China (e-mail: wfzhangzhen08@163.com; jiangtao@sdust.edu.cn; liuchenxirs@163.com; zhanglinjing@sdust.edu.cn).

Digital Object Identifier 10.1109/JSTARS.2020.3046245

widely used in HSI classification. Different from the above two networks, the CNN-based model can read 2-D or even 3-D HSI data directly. In addition, weight sharing strategy is used in CNN networks, greatly reducing the number of parameters and the cost of computation. CNN-based models are developed into a variety of network forms for HSI classification based on spectral–spatial features. Li *et al*. [28] developed a pixel-pair method to expand the number of training samples for the CNN classifier, and determined the final classification result by a voting strategy. Zhang *et al*. [29] exploited diverse region-based inputs to learn contextual interactional features on the basis of the pixel-pair CNN method. Zhang *et al*. [30] used a dual-channel CNN including a 1-D CNN channel to extract hierarchical spectral features and a 2-D CNN channel to extract the hierarchical space-related features for HSI classification. PCA and 2-D CNN are often combined to accomplish HSI classification, because PCA can effectively alleviate disastrous dimensions and maintain effective spatial characteristics [11], [31]. To reduce the complexity of the network and learn discriminative and abundant spectral–spatial features, 3-D convolutional layers are adopted to extract 3-D cubes image features directly from raw HSIs [32], [33]. Subsequently, different variants of 3-D CNN are generated by adding attention mechanism module [34], using a multiscale dense connection [35], [36], and integrating multiple networks [37] to optimize the classification results of traditional CNN.

However, these methods are still dominant in the processing of HSI with low and medium spatial resolution. With the continuous improvement of HSI spatial resolution, the noise of classification results gets increasingly serious, and classification performance becomes increasingly unsatisfactory. The reason for these consequences is that in a centimeter-level resolution image, the feature unit of remote sensing objects is composed of several neighboring pixels that may have different spectral attributes, and the spatial details of VHR-HSI are considerably complex and changeable. For example, the building in orthophoto VHR-HSI may consist of roll tiles, chimney, glasses, and other built components with different spectral properties. The classification should extract the building as a single category rather than the components of the building such as roll tiles, chimney, and glasses. By contrast, low interclass spectral variability and high intraclass spectral variability introduce difficulties in discriminating different classes. In general, the traditional CNN classification algorithm based on pixelwise method does not seem to be effective in terms of VHR-HSI, and the classification results usually show serious noise and blurring. Semantic segmentation is a powerful image segmentation method based on the CNN model. These algorithms can combine context information based on learning target features at different scales and levels via end-to-end training. Such an algorithm can effectively reduce or even eliminate noise and provide new possibilities for VHR-HSI image segmentation. At present, many excellent semantic segmentation algorithms have been successfully applied in ordinary natural digital photos [38]–[42] and remote sensing images [43]–[45]. On this basis, researches use the combination of geographic object-based image analysis [46] or super-pixel [47] with a semantic segmentation algorithm and achieve good classification results for very high-resolution remote sensing images. These algorithms aimed at the classification of ordinary natural digital photos and very high-resolution remote sensing images, but no attempt has been made to the processing of hyperspectral remote sensing images, not to mention VHR-HSI.

In this article, to alleviate the serious salt-and-pepper noise of the final results when classifying VHR-HSI, we design an effective classification network to fully capture the spatial–spectral and contextual features of VHR-HSI. Our main contributions are listed as follows:

1) The 3-D-ResNet is constructed by incorporating 3-D convolution and residual network.
2) 3-D-ResNet encoder–decoder architecture for VHR-HSI classification is proposed by end-to-end training, which is different from pixel-level feature learning of traditional CNN.
3) Atrous spatial pyramid pooling (ASPP) is integrated into the architecture to extract contextual features at multiple scales.
4) The proposed architecture has excellent performance in the classification result of hyperspectral images with very high spatial resolution, and is suitable for this kind of data.

The remainder of this article is organized as follows. In Section II, we describe the proposed classification method for VHR-HSI in detail. In Section III, we present experimental results and discussion. Section IV concludes the study.

II. Methods

Inspired by the effective performance of semantic segmentation algorithm in high spatial resolution, we first use the design of semantic segmentation to deal with the classification of VHR-HSI. Different from the traditional pixelwise CNN classification, the proposed algorithm is a contextual guided object-based CNN that uses the end-to-end training. Encoder–decoder structure and 3-D residual blocks are the core of the proposed algorithm. In this section, we give a detailed analysis of 3-D-ResNet and the overall architecture of encoder–decoder.

A. 3-D-ResNet

A skip connection is used in residual block to alleviate the phenomenon of gradient vanishing caused by increasing the depth in neural network. Residual network is a network model composed of a series of residual blocks [48], [52], [53]. The residual network can effectively alleviate the vanishing gradients problem caused by increasing the number of network layers in CNN, which has a wide range of applications in different network structure skeletons. In this article, the 3-D convolution is embedded into residual network for learning abundant spatial structure and spectral information.

The batch normalization (BN) can effectively alleviate the phenomenon of slow convergence or gradient explosion in the training process. Following by previous study [33], we employ this method to normalize the feature cubes after 3-D convolutional layers. If the kth 3-D convolutional layer has

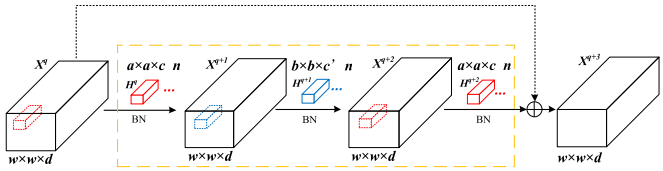

Fig. 1. Structure of the residual block used in this article.

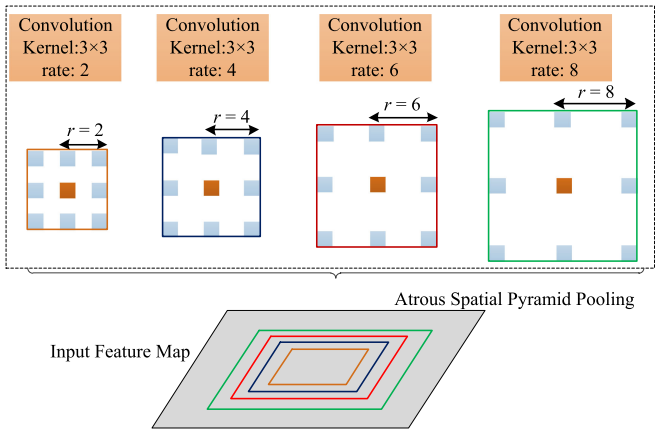

Fig. 2. ASPP exploits multiscale features by using multiple parallel filters with different rates. Effective field-of-views are shown in different colors.

m input feature cubes of size $w^k \times w^k \times d^k$, the ith output of $(k+1)$th 3-D convolutional layer with BN can be formulated as

$$X_i^{k+1} = F\left(\sum_{j=1}^{m} \hat{X}_j^k * H_i^{k+1} + b_i^{k+1}\right) \tag{1}$$

$$\hat{X}^k = \left(X^k - E\left(X^k\right)\right) / \mathrm{Var}\left(X^k\right) \tag{2}$$

where H_i^{k+1} and b_i^{k+1} are the parameters and bias of the ith convolutional filter feature in the $(k+1)$th layer, $*$ is a 3-D convolutional operation. $X_j^k \in \mathbb{R}^{\omega\times\omega\times d}$ is the jth input feature tensor of the $(k+1)$th layer. $F(\cdot)$ is the rectified linear unit activation function. $\hat{X}^k$ is the normalization result of batch feature cubes X^k in the kth layer. $E(\cdot)$ and $\mathrm{Var}(\cdot)$ represent the expectation and variance function of the input data, respectively.

The residual blocks in this article are composed of three 3-D convolution layers with different kernels and one shortcut. The structure of the residual block is as follows in Fig. 1. Successive 3-D convolutions with different scale convolutional kernels are used to extract features from VHR-HSI and the successive convolution kernels can be represented by H^q, H^{q+1}, and H^{q+2}, respectively, as shown in Fig. 1. The residual architecture for input 3-D feature cubes X^q with the size of $w \times w \times d$ can be formulated as follows:

$$X^{q+3} = X^q + F\left(X^q; \varepsilon\right) \tag{3}$$

$$F\left(X^q; \varepsilon\right) = k\left(\hat{X}^{q+2}\right) * H^{q+3} + b^{q+3} \tag{4}$$

$$X^{q+2} = k\left(\hat{X}^{q+1}\right) * H^{q+2} + b^{q+2} \tag{5}$$

$$X^{q+1} = k\left(\hat{X}^q\right) * H^{q+1} = k\left(\hat{X}^q\right) * H^{q+1} \tag{6}$$

where $\varepsilon = \left\{H^{q+1}, H^{q+2}, H^{q+3}; b^{q+1}, b^{q+2}, b^{q+3}\right\}$, H^{q+1}, and b^{q+1} represent the n 3-D convolutional kernels in the $(q+1)$th layer, respectively. Through the shortcut and successive convolution, the output feature X^{q+3} contains hierarchical 3-D feature cubes, including the low-level input feature X^q and a high-level 3-D convolution feature.

B. 3-D-ResNet Encoder–Decoder Architecture

3-D-ResNet encoder–decoder architecture is an end-to-end training structure for extracting context information of feature maps. The encoder can capture the spatial dimension of the feature maps by using continuous convolution and the longer range feature is more easily captured in the deeper encoder output. The decoder is based on different strategies, such as using deconvolution upsampling, reusing the pooling indices, and adding skip connections from encoder features to recover image details and spatial dimensions.

The 3-D-ResNet is composed of several consecutive residual blocks in series. A 3-D convolution with a kernel size of $1 \times 1 \times m$ is used after each residual block to alleviate the spectral dimension disaster of VHR-HSI. After the 3-D residual network operation, the spectral axis of the input 3-D feature cubes is compressed to 1, while the spatial scale remains the same as the input feature. At the same time, the label data is consistent with the scale of the output features, which can also be used for end-to-end training.

In the encoder structure, as shown in Fig. 3, first, the designed network uses $7 \times 7 \times$ 13-D convolution and $2 \times 2 \times$ 13-D max pooling to extract and compress the features of VHR-HSI. This is to consider that too high spatial resolution (e.g., <5 cm) may negatively impact the classification accuracy for land cover mapping. In fact this operation is to carry out feature extraction and down sampling on spatial scale to alleviate the damage. Subsequently, the four consecutive combinations of residual blocks and 3-D convolutions of spectral axis are used to compress the 3-D features into 2-D spatial features with small scale. Each residual block contains multiple units composed of three layers of 3-D convolution. From Block1 to Block4, there are 3, 4, 6, and 4 units, respectively. And each unit contains three 3-D convolution layers of $1 \times 1 \times 2$, $3 \times 3 \times 1$, and $1 \times 1 \times 2$. In the process, the residual block can reduce the image to half of the original in spatial scale, but does not change the number of features on the spectral axis. This is to capture features in a larger spatial scale and semantic features. And these blocks are followed by $1 \times 1 \times$ n 3-D spectral convolutions to extract and abstract spectral details. Subsequently, an ASPP including 3×3 atrous convolutions with rate = 2, 3, 4 and 1×1 conventional convolution is used in this encoder structure to handle multiscale semantic information. The ASPP structure as shown in Fig. 2 has an excellent classification effect on

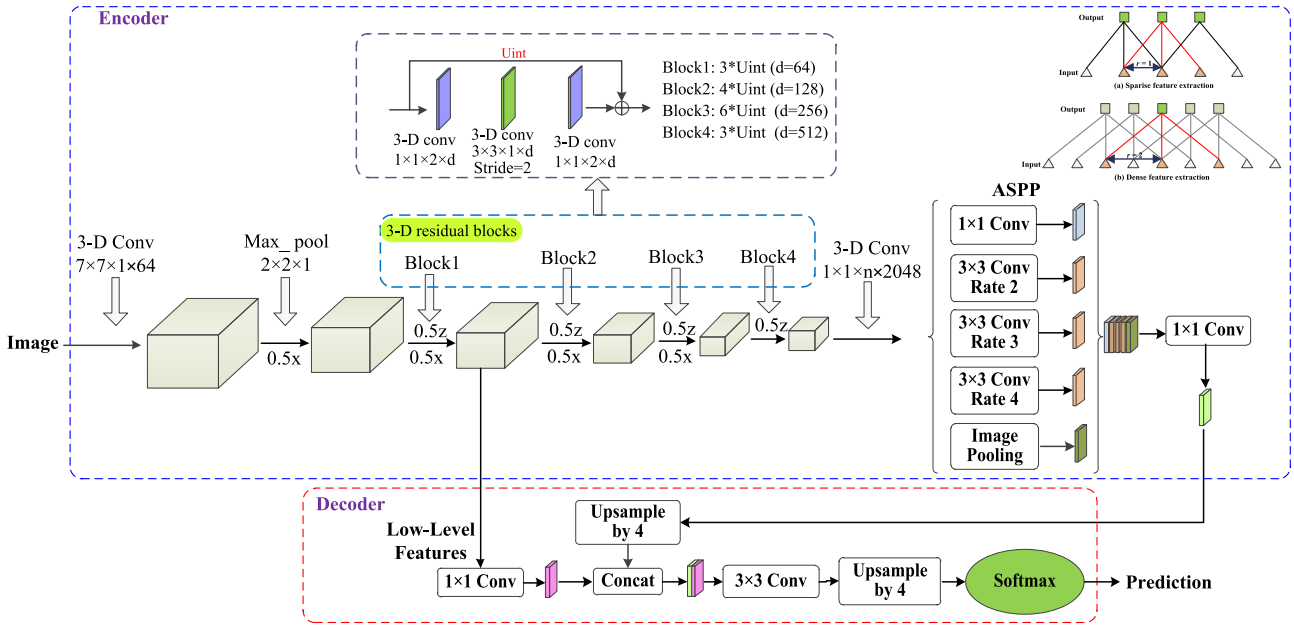

Fig. 3. 3-D-ResNet encoder–decoder architecture.

high spatial resolution remote sensing images [45], [51], [52] by using a convolutional feature layer with filters at multiple sampling rates to capture context information. In addition, the image pooling feature in ASPP is the mean value of the output of the previous layer. After concatenation, another 512 channels with 1×1 conventional convolution are used to reduce the number of channels and enhance the trainability of the network.

The encoder output feature contains rich semantic information. In the decoder procedure, the input features are first bilinearly upsampled by quadruple and then concatenated with the corresponding low-level features from the encoder structure with the same spatial resolution. After concatenation, 3×3 convolutions and another iteration of bilinear upsampling by quadruple are used to refine the features and amplify spatial scale to the same size as the input image. In other words, the size of the output image is the same as that of the label data, which can participate in end-to-end training. Then, the classification results are calculated by softmax function. After the architecture of deep learning models is built and the hyperparameters for training are configured, the models are trained by the training dataset of VHR-HSI and their corresponding ground-truth label vector set. In this process, the model parameters are updated according to the gradient of the cross-entropy objective function through backward propagation algorithm. During the training stage, the validation dataset is used for monitoring the training process by measuring the classification performance of interim models to select the optimal parameter model of the network with the highest classification accuracy. Finally, the testing dataset of VHR-HSI is used for assessing the generalizability and stability of the trained proposed model by calculating classification metrics and visualizing thematic maps. However, to avoid the overflow of computing memory caused by high-dimensional data, the original VHR-HSI is cut into the appropriate size for processing.

III. Result and Discussion

A. Dataset Description and Experimental Setup

Two different spatial resolutions of VHR-HSI are used in this article. The first data (we named it HySpex data) was obtained by the HySpex Mjolnir V-1240 hyperspectral imaging system developed by NEO in Norway. The system provides a unique combination of small form factor and low mass, which is highly suited for a wide range of unmanned aerial vehicle platforms for data imaging. This VHR-HSI used in this article was acquired by flight on October 26, 2016, in Huailai County, Hebei Province, China, with 0.04 m spatial resolution. Background targets such as pedestrians, cars, and shadows can be seen clearly in such centimeter spatial resolution HSI, which is a serious interference factor for hyperspectral image classification. The HySpex data has 200 bands covering the visible and near-infrared spectral range, from 400 to 1000 nm. The experimental image and corresponding ground truth are shown as Fig. 4(a) and (b). According to the on-the-spot investigation, we divided the ground objects into nine, namely, fallen withered grass, standing withered grass, impervious surface, trees, buildings, nudation, cars, flowers in clusters, and clutter/background.

The second dataset is grss_dfc_2018, which was collected by the National Center for Airborne Laser Mapping at the University of Houston on February 16, 2017, covering the University of Houston campus and its surrounding areas [53]. The hyperspectral data grss_dfc_2018 covers a 380–1050 nm spectral range with 48 bands at a 1-m ground sampling distance (GSD), as shown in Fig. 5(a). The sponsor provided the corresponding VHR RGB imagery at a 5-cm GSD, as shown in Fig. 5(b). Moreover, they also provide a ground truth map at 0.5-m GSD corresponding to 20 urban land use and land cover classes. We fuse the low spatial resolution hyperspectral (hyperLR) image

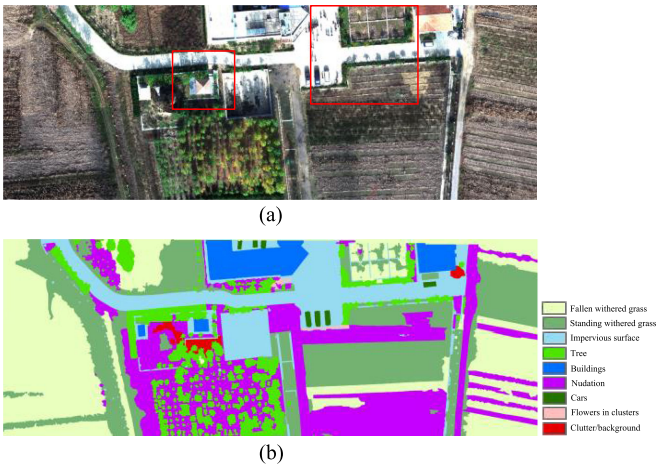

Fig. 4. HySpex data. (a) True composite image (R: 641 nm, G: 548 nm, B: 460 nm), the red rectangle is the data for testing, the rest is for training. (b) Ground truth map.

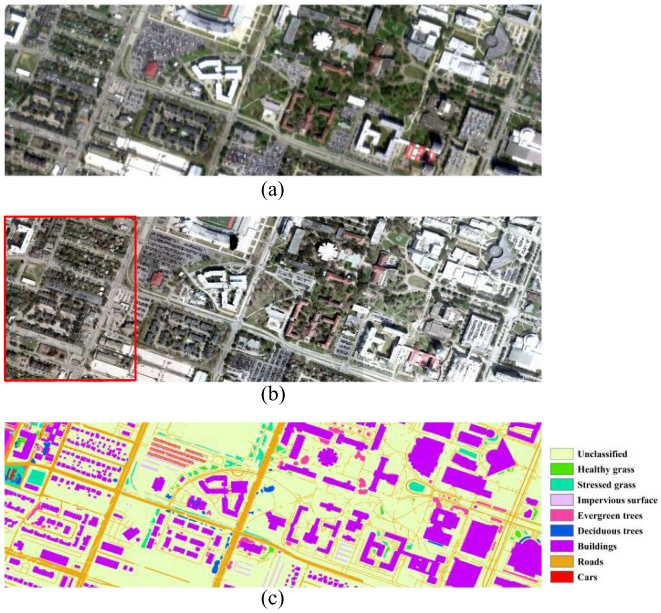

Fig. 5. grss_dfc_2018 data. (a) True composite image of HSI (R: 632 nm, G: 531 nm, B: 460 nm). (b) True composite image of VHR image, the red rectangle is the data for testing, the rest is for training. (c) Ground truth map.

and the high spatial resolution multispectral (mutilHR) image using Gram–Schmidt algorithm to obtain the VHR-HSI at 5-cm GSD. Meanwhile, we remake the labels according to the VHR RGB imagery, and merge some categories as shown in Fig. 5(c).

Owing to GPU memory limitations, the proposed algorithm cannot easily process large-scale VHR-HSI images directly. Therefore, the large-scale spatial image is sliced into multiple 224×224 patchwise images for processing.

The algorithm is based on Tensorflow framework and runs on CentOS 7.4 system with a Xeon Gold 5118 CPU @ 2.30 GHz and a Tesla P100 GPU with 16 GB graphics memory. The number of training iterations is 50 000, batch size is 2, and learning rate is 0.0001. The accuracy performance is evaluated with overall accuracy (OA), average accuracy (AA), and kappa coefficient (k).

B. Classification Results of HySpex data

We compare the proposed method with the pixelwise deep learning models, such as 2-D CNN, 3-D CNN, PCA combined with 2-D CNN (PCA-CNN), and 3-D-ResNet. Compared with the proposed method, the 3-D-ResNet still adopts pixelwise training instead of end-to-end training, and this network has no ASPP structure. Moreover, the semantic segmentation algorithm DeepLabv3 is also used for comparative analysis to demonstrate the effectiveness of the proposed framework. We construct three layers of convolution, three layers of pooling, and two layers of full connection for all these CNN algorithms. The Adam optimization method and cross-entropy loss function are used in these networks. The number of training is 50 000, the number of batch training is 20, the number of samples is 200, and the learning rate is 0.001. The number of training iterations of the DeepLabv3 algorithm is 50 000; batch size is 4; atrous rate $r = 6, 12, 18$; and learning rate is 0.0001.

The image with 0.04 m spatial resolution has many artificial objects showing complex spectral variances. Shadows and pedestrians wearing different colors will also increase the complexity in the spectrum, which is one of the major interference factors for the classification results. Figs. 6 and 7 present the classification results of different methods for visual interpretation. The traditional CNN networks, including 2-D CNN, 3-D CNN, PCA-CNN, and 3D-ResNet, have a good effect on spectral feature extraction, which is mainly reflected in the boundary of regular objects. However, the VHR-HSI with strong granular pixels and poor signal-to-noise ratio constrains the efficiency and effectiveness of classification. This is mainly manifested in the areas such as pedestrian, shadow, and ground object composed of different spectral characteristics. Pixelwise CNN classification methods show more serious salt-and-pepper noise in classification results, making them unsuitable for the classification of VHR-HSI data. However, they are also outstanding in some details. For example, in the red ellipse of Fig. 6(d), the boundary contour lines of standing withered grass using pixelwise CNN methods are extraordinarily clear and accurate. Fig. 6(e) is the classification result of 3-D-ResNet method without encoder–decoder architecture and ASPP. The results show that the method is slightly better than 2-D CNN, but worse than 3-D CNN and PCA-CNN. The anti-interference performance of this method for shadow and pedestrian is not satisfactory. The 200 bands of VHR-HSI are directly input into the DeepLabv3 algorithm, showing poor classification performance. As shown in Fig. 6(f), although the salt-and-pepper noise phenomenon in the classification results is effectively alleviated, the boundary of some ground targets is coarse and even some objects on the ground such as buildings, cannot be distinguished at all. A possible reason for this result is that the DeepLabv3 algorithm performs a convolution operation for each band of VHR-HSI separately, and it lacks the feature learning of pixel on spectral profile. By contrast, the proposed method can compromise between these two contradictions, which can effectively eliminate

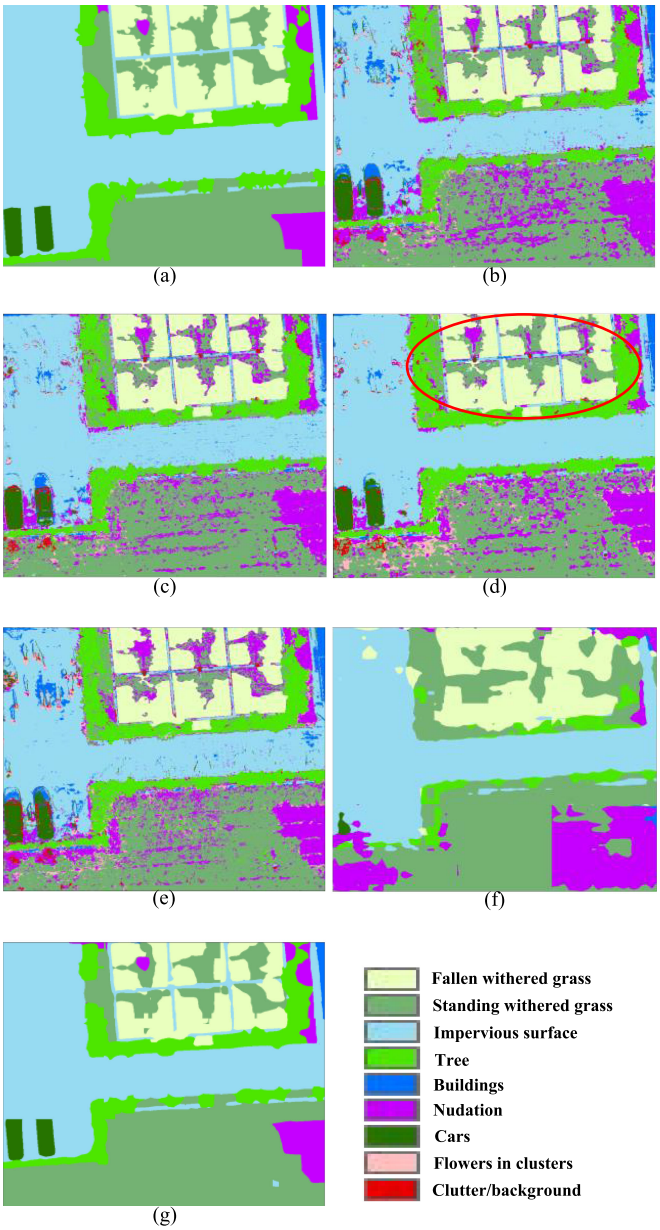

Fig. 6. Subarea1 classification results with different neural network algorithms. (a) Ground truth map. (b) 2-D CNN. (c) 3-D CNN. (d) PCA-CNN. (e) 3D-ResNet. (f) DeepLabv3. (g) Proposed method.

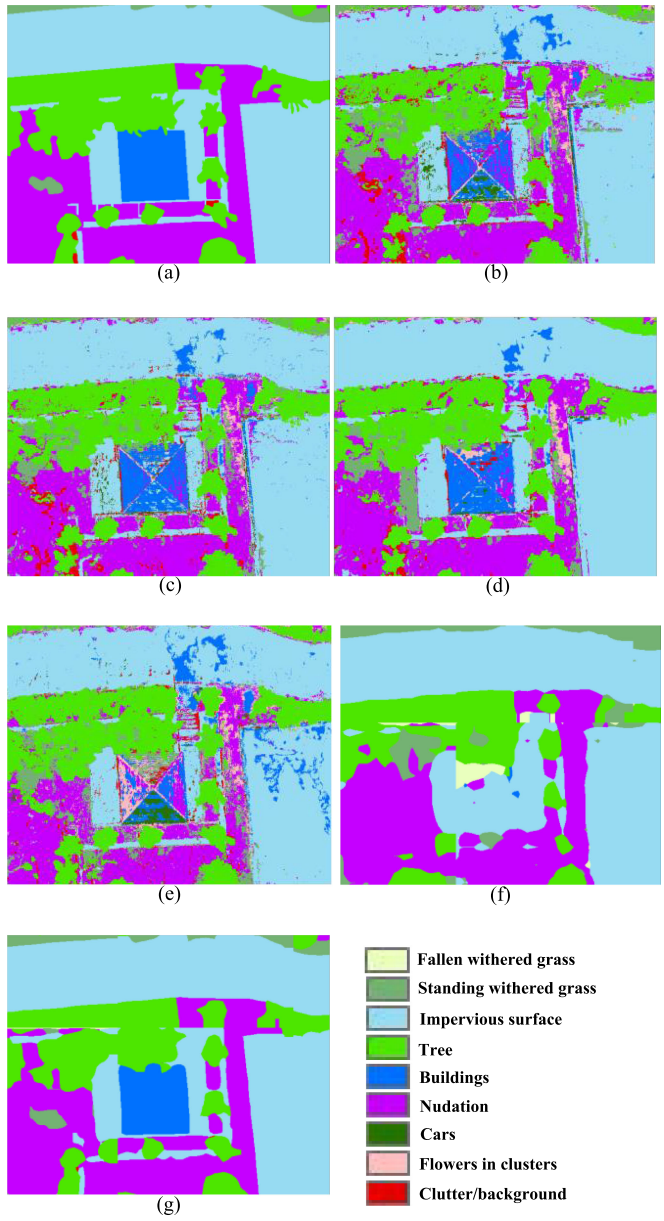

Fig. 7. Subarea2 classification results with different neural network algorithms. (a) Ground truth map. (b) 2-D CNN. (c) 3-D CNN. (d) PCA-CNN. (e) 3-D-ResNet. (f) DeepLabv3. (g) Proposed method.

the classification noise obtained by the pixelwise method and maintain fine boundary features. The proposed method achieves a predominant result with high internal compactness and fine boundary delineation. In particular, it outperforms the other three benchmark methods when classifying those same land cover classes with different spectra, such as buildings.

Fig. 7 shows the classification results with different neural network algorithms in the second subarea. Pixelwise CNN can extract the edge information that is located in the interior of the homogeneous objects and the boundary of a heterogeneous region, such as the building with complex texture information. Actually, the boundary between categories is needed to extract in classification applications rather than the edge information within homogeneous objects. The proposed algorithm can effectively eliminate the edge feature information of homogeneous objects in the classification and greatly improve the classification effect.

Table I shows the quantitative assessment of the classification performance. The proposed method achieves the highest OA 88.90% with k 0.861, compared with 2-D CNN (OA 63.17% and k 0.538), 3-D CNN (OA 79.50% and k 0.734), PCA-CNN (OA 81.48% and k 0.756), and 3-D-ResNet (OA 72.07% and k 0.640). The advantages of our algorithm can be observed obviously among conventional pixelwise classification methods. The classification accuracy of the DeepLabv3 algorithm is highly unsatisfactory under the condition of small samples, such as

TABLE I
CLASSIFICATION PERFORMANCE COMPARISON OF HYSPEX DATA

	2D-CNN	3D-CNN	PCA-CNN	3D-ResNet	DeepLabv3	Proposed method
Fallen withered grass	0.6038	0.7220	0.8413	0.7428	0.8827	0.9639
Standing withered grass	0.6398	0.8719	0.8169	0.7153	0.8032	0.9462
Impervious surface	0.7993	0.8906	0.8683	0.8143	0.7408	0.8604
Trees	0.6226	0.8075	0.8833	0.7991	0.6548	0.9740
Buildings	0.5755	0.7590	0.6538	0.5546	0.7519	0.9506
Nudation	0.6539	0.8995	0.8329	0.7105	0.4973	0.6885
Cars	0.7169	0.9561	0.9566	0.8024	0.5883	0.9116
Flowers in clusters	0.5727	0.9199	0.8730	0.6757	0.022	0.9158
Clutter/background	0.6502	0.8542	0.8426	0.7368	0.056	0.8562
OA	0.6317	0.7950	0.8148	0.7207	0.5528	**0.8890**
AA	0.6483	0.8533	0.8409	0.7279	0.7443	**0.8963**
Kappa	0.538	0.734	0.756	0.640	0.6763	**0.861**

Fig. 8. Fusion result by Gram–Schmidt algorithm. (a) MutilHR image. (b) HyperLR image (R: 760 nm, G: 632 nm, B: 531 nm). (b) VHR-HSI (R: 760 nm, G: 632 nm, B: 531 nm).

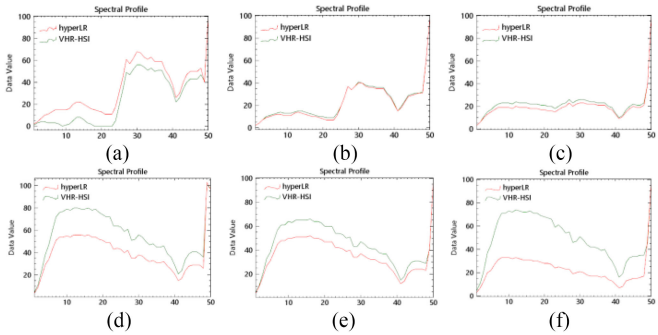

Fig. 9. Spectral changes of different urban land use and land cover classes before and after fusion.

flowers in clusters and clutter/background. Their classification accuracies are less than 0.1 because the classification performance of the DeepLabv3 needs to depend on a large number of training samples for support. The proposed method gains an OA improvement of 7.42% and a kappa coefficient improvement of 0.105. It has good performance in the classification results of each ground object, even under the condition of a small number of training samples.

C. Classification Results of grss_dfc_2018 Data

The VHR-HSI obtained by the fusion of hyperLR and mutilHR images is shown in Fig. 8(c). The spatial resolution has been effectively improved, and the texture details have been ameliorated and abundant, which is helpful to the image interpretation of urban land use and land cover.

In addition, we compare the spectral changes of several typical features before and after fusion. The spectral changes of several ground features before and after fusion are compared as shown in Fig. 9.

For those land cover classes with fixed spectral characteristics, such as healthy grass, evergreen trees, and deciduous trees, their spectrum of VHR-HSI after fusion has high fidelity compared with hyperLR image. But for those land cover classes with variable spectra, such as roads, cars, and buildings, their spectrum may be changed to a great extent before and after spectral fusion.

We also compare the proposed method with 2-D CNN, 3-D CNN, PCA-CNN, 3-D-ResNet, and DeepLabv3 algorithm. The setting of training parameters is the same as that of HySpex data. Fig. 10 shows the classification results with different methods.

The traditional pixelwise CNN networks, including 2-D CNN, 3-D CNN, PCA-CNN, and 3-D-ResNet, still have unsatisfactory classification results, showing a serious salt-and-pepper noise. Especially for the man-made objects with different spectra but the same urban land use and land cover feature class, such as cars and buildings, the performance of classification is more serious. The classification results of DeepLabv3 algorithm can better resist these noises. However, the stability of this model is not outstanding, and the training time is about two times more than that of the proposed method. When classifying the images, we can see that there are obvious stitching lines, which is caused by the instability of the model, as shown in Fig. 11(f). The proposed method is a context guided network, which can consider spectral, spatial, and contextual features. Fig. 11(g) shows the classification result of the proposed method, which can eliminate salt-and-pepper noise and stitching lines in VHR-HSI.

Table II shows the quantitative assessment of the classification performance. The proposed method achieves the highest OA 85.63% with k 0.830, compared with 2-D CNN (OA 45.52% and k 0.361), 3-D CNN (OA 60.82% and k 0.547),

Fig. 10. Classification results of grss_dfc_2018 data with different neural network algorithms. (a) Ground truth map. (b) 2-D CNN. (c) 3-D CNN. (d) PCA-CNN. (e) 3-D-ResNet. (f) DeepLabv3. (g) Proposed method.

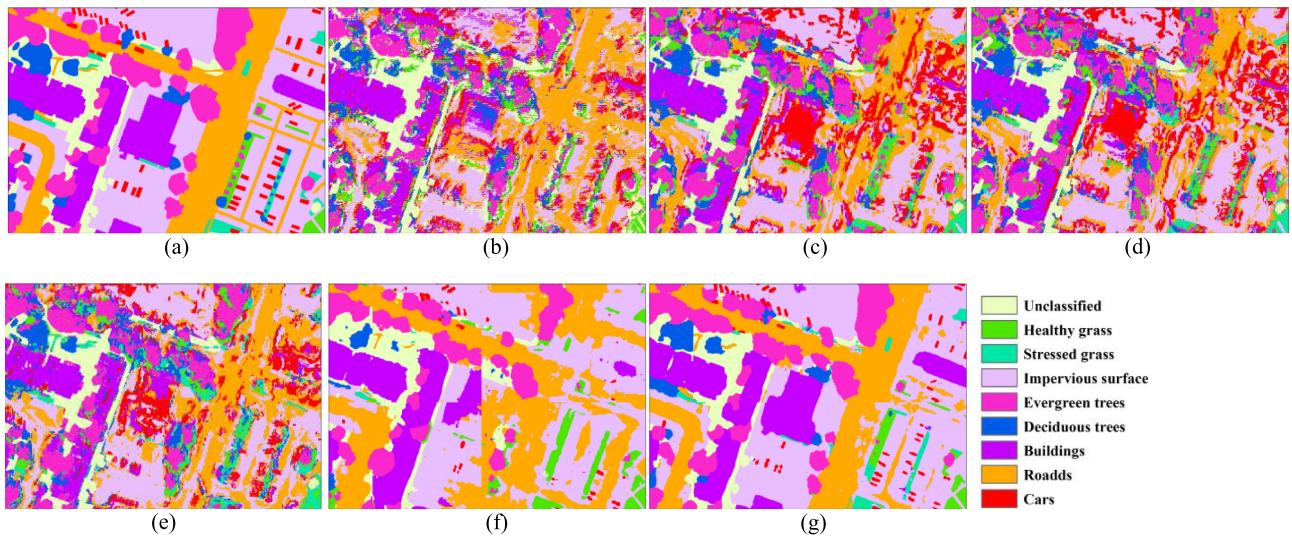

Fig. 11. Classification results of subarea with different neural network algorithms. (a) Ground truth map. (b) 2-D CNN. (c) 3-D CNN. (d) PCA-CNN. (e) 3-D-ResNet. (f) DeepLabv3. (g) Proposed method.

PCA-CNN (OA 59.45% and k 0.532), 3-D-ResNet (OA 58.13% and k 0.505), and DeepLabv3 (OA 66.81% and k 0.605). The classification result of this dataset using DeepLabv3 algorithm is better than that of HySpex data, this because the training sample number of grss_dfc_2018 data is much more sufficient. The proposed method gains an OA improvement of 18.82% and a kappa coefficient improvement of 0.225. Pixelwise CNN has poor classification performance in the man-made objects with great differences of spectrum, such as cars. On the contrary, the proposed method has excellent classification performance for both man-made and natural feature classes.

TABLE II
CLASSIFICATION PERFORMANCE COMPARISON OF GRSS_DFC_2018 DATA

	2D-CNN	3D-CNN	PCA-CNN	3D-ResNet	DeepLabv3	Proposed method
Unclassified	0.3703	1.000	1.000	0.5454	0.6139	0.8079
Healthy grass	0.4502	0.4720	0.4495	0.5669	0.4145	0.7511
Stressed grass	0.000	0.5612	0.5512	0.5465	0.8236	0.8448
Impervious surface	0.4425	0.6679	0.6524	0.6590	0.6568	0.8751
Evergreen trees	0.5378	0.5812	0.5600	0.6316	0.8379	0.8833
Deciduous trees	0.4641	0.3975	0.3695	0.6549	0.8049	0.9211
Buildings	0.5059	0.7880	0.7886	0.5513	0.7735	0.8844
Roads	0.4406	0.6312	0.6195	0.4870	0.6544	0.8484
Cars	0.3671	0.1716	0.1619	0.5602	0.9083	0.8894
OA	0.4552	0.6082	0.5945	0.5813	0.6681	**0.8563**
AA	0.4473	0.5856	0.5725	0.5781	0.6488	**0.7705**
Kappa	0.361	0.547	0.532	0.505	0.605	**0.830**

IV. CONCLUSION

To address the known challenges of VHR-HSI classification, an effective end-to-end encoder–decoder architecture is proposed. The proposed architecture is an object-level contextual guided CNN based on an improved residual network backbone with 3-D convolution, which completes classification effectively by fusing spectral, spatial, and contextual information. Compared with four pixelwise CNN methods and DeepLabv3, it is proved that the proposed method has a remarkable performance of suppressing salt-and-pepper noise in VHR-HSI classification. The proposed algorithm has the following attractive properties:

1) 3-D-ResNet encoder–decoder architecture is constructed for hyperspectral image classification by end-to-end training.
2) With multiscale contextual information used for classification, the method shows excellent superiority in withstanding predominant noise, especially in areas with shadows, pedestrians, and man-made features.
3) Compared with conventional pixel-wise classification methods and DeepLabv3 algorithm, the proposed method achieves high classification accuracy which is more suitable for centimeter level spatial resolution HSI.

Though the classification results are satisfactory, further experiments focusing on the boundary of small objects need to be carried out by the proposed method. In other words, the contradiction between multiscale semantic information and small object details in classification is still unsolved completely.

ACKNOWLEDGMENT

The authors would like to acknowledge NEO Company in Norway for providing us with the experimental data. They would also like to thank the National Center for Airborne Laser Mapping and the Hyperspectral Image Analysis Laboratory at the University of Houston for acquiring and providing the data used in this study, and the IEEE GRSS Image Analysis and Data Fusion Technical Committee.

REFERENCES

[1] C. Gomez *et al.*, "Sensitivity of clay content prediction to spectral configuration of VNIR/SWIR imaging data, from multispectral to hyperspectral scenarios," *Remote Sens. Environ.*, vol. 204, pp. 18–30, 2018.
[2] R. Sigrid, S. Karl, H. Uta, and K. Hermann, "Automated differentiation of urban surfaces based on airborne hyperspectral imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 39, no. 7, pp. 1525–1532, Jul. 2001.
[3] D. Haboudane, "Hyperspectral vegetation indices and novel algorithms for predicting green LAI of crop canopies: Modeling and validation in the context of precision agriculture," *Remote Sens. Environ.*, vol. 90, no. 3, pp. 337–352, 2004.
[4] T. Caras, J. Hedley, and A. Karnieli, "Implications of sensor design for coral reef detection: Upscaling ground hyperspectral imagery in spatial and spectral scales," *Int. J. Appl. Earth Observ. Geoinf.*, vol. 63, pp. 68–77, 2017.
[5] N. J. P. Anne, A. H. Abd-Elrahman, D. B. Lewis, and N. A. Hewitt, "Modeling soil parameters using hyperspectral image reflectance in subtropical coastal wetlands," *Int. J. Appl. Earth Observ. Geoinf.*, vol. 33, pp. 47–56, 2014.
[6] R. Pelta, N. Carmon, and E. Ben-Dor, "A machine learning approach to detect crude oil contamination in a real scenario using hyperspectral remote sensing," *Int. J. Appl. Earth Observ. Geoinf.*, vol. 82, 2019, Art. no. 101901.
[7] B. Tu, J. Wang, X. Kang, G. Zhang, X. Ou, and L. Guo, "KNN-based representation of superpixels for hyperspectral image classification," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 11, no. 11, pp. 4032–4047, Nov. 2018.
[8] M. Khodadadzadeh, J. Li, A. Plaza, and J. M. Bioucas-Dias, "A subspace-based multinomial logistic regression for hyperspectral image classification," *IEEE Geosci. Remote Sens. Lett.*, vol. 11, no. 12, pp. 2105–2109, Dec. 2014.
[9] B. Krishnapuram, L. Carin, M. A. Figueiredo, and A. J. Hartemink, "Sparse multinomial logistic regression: Fast algorithms and generalization bounds," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 27, no. 6, pp. 957–968, Jun. 2005.
[10] D. Vyas, N. S. R. Krishnayya, K. R. Manjunath, S. S. Ray, and S. Panigrahy, "Evaluation of classifiers for processing hyperion (EO-1) data of tropical vegetation," *Int. J. Appl. Earth Observ. Geoinf.*, vol. 13, no. 2, pp. 228–235, 2011.
[11] G. Licciardi, P. R. Marpu, J. Chanussot, and J. A. Benediktsson, "Linear versus nonlinear PCA for the classification of hyperspectral data based on the extended morphological profiles," *IEEE Geosci. Remote Sens. Lett.*, vol. 9, no. 3, pp. 447–451, May 2012.
[12] J. Xia, M. Dalla Mura, J. Chanussot, P. Du, and X. He, "Random subspace ensembles for hyperspectral image classification with extended morphological attribute profiles," *IEEE Trans. Geosci. Remote Sens.*, vol. 53, no. 9, pp. 4768–4786, Sep. 2015.
[13] P. Quesada-Barriuso, F. Arguello, and D. B. Heras, "Spectral–spatial classification of hyperspectral images using wavelets and extended morphological profiles," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 7, no. 4, pp. 1177–1185, Apr. 2014.

[14] X. Cao, L. Xu, D. Meng, Q. Zhao, and Z. Xu, "Integration of 3-dimensional discrete wavelet transform and Markov random field for hyperspectral image classification," *Neurocomputing*, vol. 226, pp. 90–100, 2017.
[15] J. Li, J. M. Bioucas-Dias, and A. Plaza, "Spectral–spatial hyperspectral image segmentation using subspace multinomial logistic regression and Markov random fields," *IEEE Trans. Geosci. Remote Sens.*, vol. 50, no. 3, pp. 809–823, Mar. 2012.
[16] J. Li, J. M. Bioucas-Dias, and A. Plaza, "Spectral–spatial classification of hyperspectral data using loopy belief propagation and active learning," *IEEE Trans. Geosci. Remote Sens.*, vol. 51, no. 2, pp. 844–856, Feb. 2013.
[17] B. Wu, Y. Zhu, X. Huang, and J. Y. Li, "Generalization of spectral fidelity with flexible measures for the sparse representation classification of hyperspectral images," *Int. J. Appl. Earth Observ. Geoinf.*, vol. 52, pp. 275–283, Oct. 2016.
[18] B. Pan, Z. W. Shi, and X. Xu, "MugNet: Deep learning for hyperspectral image classification using limited samples," *ISPRS J. Photogramm. Remote Sens.*, vol. 145, pp. 108–119, Nov. 2018.
[19] C. Shi and C.-M. Pun, "Multi-scale hierarchical recurrent neural networks for hyperspectral image classification," *Neurocomputing*, vol. 294, pp. 82–93, 2018.
[20] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 6, pp. 1137–1149, Jun. 2017.
[21] S. Lawrence, C. L. Giles, A. C. Tsoi, and A. D. Back, "Face recognition: A convolutional neural-network approach," *IEEE Trans. Neural Netw.*, vol. 8, no. 1, pp. 98–113, Jan. 1997.
[22] 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.
[23] S. Paul and D. Nagesh Kumar, "Spectral-spatial classification of hyperspectral data with mutual information based segmented stacked autoencoder approach," *ISPRS J. Photogramm. Remote Sens.*, vol. 138, pp. 265–280, 2018.
[24] J. Zabalza *et al.*, "Novel segmented stacked autoencoder for effective dimensionality reduction and feature extraction in hyperspectral imaging," *Neurocomputing*, vol. 185, pp. 1–10, 2016.
[25] T. Chao, P. Hongbo, L. Yansheng, and Z. Zhengrou, "Unsupervised spectral–spatial feature learning with stacked sparse autoencoder for hyperspectral imagery classification," *IEEE Geosci. Remote Sens. Lett.*, vol. 12, no. 12, pp. 2438–2442, Dec. 2015.
[26] 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.
[27] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," *Commun. ACM*, vol. 60, no. 6, pp. 84–90, 2017.
[28] W. Li, G. Wu, F. Zhang, and Q. Du, "Hyperspectral image classification using deep pixel-pair features," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 2, pp. 844–853, Feb. 2017.
[29] M. Zhang, W. Li, and Q. Du, "Diverse region-based CNN for hyperspectral image classification," *IEEE Trans. Image Process.*, vol. 27, no. 6, pp. 2623–2634, Jun. 2018.
[30] H. Zhang, Y. Li, Y. Zhang, and Q. Shen, "Spectral-spatial classification of hyperspectral imagery using a dual-channel convolutional neural network," *Remote Sens. Lett.*, vol. 8, no. 5, pp. 438–447, 2017.
[31] A. Appice and D. Malerba, "Segmentation-aided classification of hyperspectral data using spatial dependency of spectral bands," *ISPRS J. Photogramm. Remote Sens.*, vol. 147, pp. 215–231, 2019.
[32] C. Shi and C.-M. Pun, "Superpixel-based 3D deep neural networks for hyperspectral image classification," *Pattern Recognit.*, vol. 74, pp. 600–616, 2018.
[33] 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.
[34] Z. Y. Lu, B. Xu, L. Sun, T. M. Zhan, and S. Z. Tang, "3D Channel and spatial attention based multi-scale spatial spectral residual network for hyperspectral image classification," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 13, no. 7, pp. 4311–4324, Jul. 2020.
[35] Y. Xiao, Q. Xu, D. Y. Wang, J. Tang, and B. Luo, "Multi-scale densely 3D CNN for hyperspectral image classification," in *Proc. Chin. Conf. Pattern Recognit. Comput. Vis.*, 2019, pp. 596–606.
[36] Q. Xu, Y. Xiao, D. Y. Wang, and B. Luo, "CSA-MSO3DCNN: Multiscale octave 3D CNN with channel and spatial attention for hyperspectral image classification," *Remote Sens.*, vol. 12, no. 1, Jan. 2020, Art. no. 188.
[37] C. Y. Yu, R. Han, M. P. Song, C. Y. Liu, and C. I. Chang, "A simplified 2D-3D CNN architecture for hyperspectral image classification based on spatial-spectral fusion," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 13, no. 4, pp. 2485–2501, Apr. 2020.
[38] V. Badrinarayanan, A. Kendall, and R. Cipolla, "SegNet: A deep convolutional encoder-decoder architecture for image segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 12, pp. 2481–2495, Dec. 2017.
[39] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-Decoder with atrous separable convolution for semantic image segmentation," in *Proc. Comput. Vis.*, 2018, pp. 833–851.
[40] G. Lin, A. Milan, C. Shen, and I. Reid, "RefineNet: Multi-path refinement networks for high-resolution semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops*, 2017, pp. 5168–5177.
[41] E. Shelhamer, J. Long, and T. Darrell, "Fully convolutional networks for semantic segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 4, pp. 640–651, Apr. 2017.
[42] H. S. Zhao, J. P. Shi, X. J. Qi, X. G. Wang, and J. Y. Jia, "Pyramid scene parsing network," in *Proc. 30th IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 6230–6239.
[43] N. Audebert, B. L. Saux, and S. Lefèvre, "Beyond RGB: Very high resolution urban remote sensing with multimodal deep networks," *ISPRS J. Photogramm. Remote Sens.*, vol. 140, pp. 20–32, 2018.
[44] H. Yang, B. Yu, J. Luo, and F. Chen, "Semantic segmentation of high spatial resolution images with deep neural networks," *GISci. Remote Sens.*, vol. 56, no. 5, pp. 749–768, 2019.
[45] Z. Zhang, J. Huang, T. Jiang, B. Sui, and X. Pan, "Semantic segmentation of very high-resolution remote sensing image based on multiple band combinations and patchwise scene analysis," *J. Appl. Remote Sens.*, vol. 14, no. 01, 2020, Art. no. 016502.
[46] T. Liu and A. Abd-Elrahman, "Deep convolutional neural network training enrichment using multi-view object-based analysis of unmanned aerial systems imagery for wetlands classification," *ISPRS J. Photogramm. Remote Sens.*, vol. 139, pp. 154–170, 2018.
[47] W. Feng, H. Sui, W. Huang, C. Xu, and K. An, "Water body extraction from very high-resolution remote sensing imagery using deep U-Net and a superpixel-based conditional random field model," *IEEE Geosci. Remote Sens. Lett.*, vol. 16, no. 4, pp. 618–622, Apr. 2019.
[48] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. 2016 IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 770–778.
[49] Q. Bi, K. Qin, H. Zhang, Z. Li, and K. Xu, "RADC-Net: A residual attention based convolution network for aerial scene classification," *Neurocomputing*, vol. 377, pp. 345–359, 2020.
[50] 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, Oct. 2016.
[51] L. C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 4, pp. 834–848, Apr. 2018.
[52] B. Pan, X. Xu, Z. Shi, N. Zhang, H. Luo, and X. Lan, "DSSNet: A simple dilated semantic segmentation network for hyperspectral imagery classification," *IEEE Geosci. Remote Sens. Lett.*, vol. 17, no. 11, pp. 1968–1972, Nov. 2020.
[53] Y. Xu *et al.*, "Advanced multi-sensor optical remote sensing for urban land use and land cover classification: Outcome of the 2018 IEEE GRSS data fusion contest," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 12, no. 6, pp. 1709–1724, Jun. 2019.

Zhen Zhang was born in Weifang, Shandong province, China, in 1989. He received the B.S. degree in remote sensing science and technology, and M.S. degree in photogrammetry and remote sensing, Shandong University of Science and Technology, Qingdao, China, in 2012 and 2016, respectively. He is currently working toward the Ph.D. degree at Shandong University of Science and Technology.

His research interests include digital image processing and machine learning, including image segmentation and classification, and object recognition.

Tao Jiang was born in Jinan, Shandong province, China, in 1961. He received the Ph.D. degree in geodesy and survey engineering from Shandong University of Science and Technology, Qingdao, China, in 2005.

He is currently a Professor with the Remote Sensing Science and Technology Department, Shandong University of Science and Technology. His research interests include remote sensing information processing and analysis, resources and environment remote sensing, and application of remote sensing.

Chenxi Liu born in Zhengzhou, Henan province, China, in 1999. He is an undergraduate student with Remote Sensing Science and Technology Department from Shandong University of Science and Technology, Qingdao, China.

His research interests include machine learning, image classification, and algorithm programming.

Linjing Zhang received the B.S. degree in the resource environment and urban-rural planning management from Shandong University of Science and Technology, Qingdao, China, in 2011, the M.S. degree with Remote Sensing Science and Technology Department from Shandong University of Science and Technology, Qingdao, China, in 2014, and the Ph.D. degree in photogrammetry and remote sensing from the State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing, Wuhan University, Wuhan, China, in 2018.

Her research interests include remote sensing information processing and analysis, resources and environment remote sensing.