An Encoder–Decoder Convolution Network With Fine-Grained Spatial Information for Hyperspectral Images Classification

ZHONGWEI LI¹, FANGMING GUO¹, QI LI¹, GUANGBO REN², AND LEIQUAN WANG³

¹College of Oceanography and Space Informatics, China University of Petroleum (East China), Qingdao 266580, China
²First Institute of Oceanography, Ministry of Natural Resources, Qingdao 266061, China
³College of Computer Science and Technology, China University of Petroleum (East China), Qingdao 266580, China

Corresponding author: Leiquan Wang (richiewlq@gmail.com)

This work was supported in part by the Joint Funds of the National Natural Science Foundation of China under Grant U1906217, in part by the National Key Research and Development Program of China under Grant 2018YFC1406204, and in part by the Fundamental Research Funds for the Central Universities under Grant 19CX05003A-3.

ABSTRACT Convolutional Neural Network (CNN) is widely used in Hyperspectral Images (HSIs) classification. However, the fine-grained spatial (FGS) details are discarded during a sequence of convolution and pooling operations for most of CNN-based HSIs classification methods. To address this issue, a unified encoder-decoder framework is proposed to integrate high-level semantics and FGS details for HSIs classification, denoted by FGSCNN. The encoder, including a series of convolution and pooling layers, captures the high-level semantic information with low resolution feature maps. The decoder fuses the high-level low-resolution semantic and the fine-grained high-resolution spatial information, namely, to get the FGS features with high-level semantics. The deconvolution layers and skip connection are used in the decoder to retain the FGS details, while, convolution layers are also used to combine the FGS features with high-level semantics. Based on the encoder-decoder framework, a unified loss function is exploited to integrate the high-level semantic information and FGS details with an end-to-end manner for HSIs classification. Experiments conducted on the three public datasets, i.e. the Indian Pines, Pavia University and Salinas, demonstrate the effectiveness of the proposed method on HSIs classification.

INDEX TERMS Convolutional neural networks (CNNs), encoder-decoder, hyperspectral image (HSI) classification, information fusion.

I. INTRODUCTION Hyperspectral Images (HSIs) contain a great deal of spatial geometric information and spectral information reflecting various characteristics of ground objects. The goal of HSIs classification is to assign each pixel to a set of land-use/land-over classes with a classifier. HSIs classification has attracted extensive attention due to the active effect in atmospheric environment research [1], ocean remote sensing [2], environmental monitoring [3], military reconnaissance, urban planning [4], [5] and so on. HSIs offer great potentials for finer classification [6]. Meanwhile, HSIs classification present numerous challenges, particularly in utilizing high-dimensional spectral data and high resolution spatial information.

In the past decades, many traditional machine learning methods are applied on HSIs classification, such as K-Nearest Neighbor (KNN), Support Vector Machine (SVM) [7], Low Rank [8] and Sparse Representation Classifier (SRC) [9]. These methods utilized spectral data of a pixel to determine its class label with the advantages of conceptual simplicity and easy implementation. However, due to the phenomenon that the same objects may present different spectral discrepancy and different material may share the same spectrum signatures, it often results in an unsatisfactory effect in distinguishing different classes with spectral information only. The abundant spatial information in HSIs were neglected. By observing the fact that the neighboring pixels usually carry correlated information with a smooth spatial domain, a large amount of researchers paid attention to integrate spectral and spatial information for HSIs classification. Nevertheless, the early spectral-spatial methods make use of
handcrafted features, which requires massive experience for specific application scenarios.

With the development of Convolutional Neural Network (CNN), various CNN architectures [10]–[13] are performed on HSIs to extract high-level spectral, spatial and spectral-spatial features [14]–[16], such as Google Inception [17], VGG, ResNet [18] and DenseNet [19]. These CNN-based methods [20]–[22] made an end-to-end training process with the supervision of high-level class labels [23]–[26]. Only the outputs of the last convolution layer with low resolution are sent to the fully connected layer for HSIs classification. The low resolution feature maps are certified to be effective to capture the high-level semantic labels. However, it leads to a major limitation that the fine-grained spatial (FGS) details are discarded during a sequence of convolution and pooling operations. Nevertheless, the high resolution feature maps lack of high-level semantic information. Therefore, extracting effective features to carry the FGS details with high-level semantics is a challenging task for HSIs classification.

By considering the issues mentioned above, we aimed at building a unified encoder-decoder framework to integrate high-level semantics and FGS details for HSIs classification, namely FGS based CNN (FGSCNN). In this method, the encoder is used to capture the high-level semantic information with low resolution feature maps. The decoder is exploited to retain the high resolution spatial details, which is a new branch compared to the traditional CNN-based HSIs classification methods. The high-level semantic information and FGS details are fused in a unified loss function for HSIs classification with an end-to-end manner. To summarize, the main contributions of this paper are as follows:

1) The encoder-decoder network with skip connection is proposed to capture the FGS features with high-level semantics. The FGS features with high-level semantics is achieved by the decoder part to retain the FGS distribution for HSIs classification.
2) The high-level low-resolution semantic information and the FGS features with high-level semantics are aggregated for end-to-end HSIs classification. To alleviate the loss of high-level semantic information during the decoder process, the high-level semantics from the encoder output are used as an auxiliary loss to assist the final classification.
3) Experimental results conducted on the public datasets demonstrate the effectiveness of the proposed FGSCNN, which achieves competitive performances with existing methods.

II. RELATED WORK

A. HAND-CRAFTED HSIS CLASSIFICATION

Over the past two decades, a great deal of effective methods have been developed for HSIs classification. Before the surge of deep learning, HSIs classification relies on handcrafted spatial-spectral features such as Markov Random Fields (MRFs), SVM and SRC. Tarabalka et al. [27] used the MRFs to combine spectral features with spatial features. Authors of [28], [29] enhance the traditional SVM by considering the spatial information as well as the spectrum. However, the hand-crafted methods are still unable to extract abundant semantic information and require massive experience for specific application scenarios.

B. CNN-BASED HSIS CLASSIFICATION

With the success of deep learning in machine learning and computer vision fields, deep learning based methods have shown great potential by learning discriminative spectral and spatial features adaptively. Lin et al. [30] used Stack AutoEncoders (SAEs) to extract the nonlinear features of HSIs followed by a SVM for classification. Li et al. [31] combined Deep Belief Network (DBN) with Logistic Regression (LR) classifier, and the performance is convincing. Nevertheless, the spatial information is still underutilized in the early deep learning methods. In order to solve this problem, CNN-based methods have been applied to HSIs classification. Wei et al. [32] developed a 1-D CNN with the Pixel-Pair Features (PPF) strategy to learn spectral features. It is noteworthy that the 1-D CNN also cannot make full use of spatial information. Therefore, the 2-D CNN based methods was taken turn in the spotlight. In [33], Lee et al. presented a 2-D CNN, which is similar to AlexNet, to extract the spectral features by using multi-scale windows sensing field. In [34], Zhang et al. exploit diverse region-based inputs to learn contextual interactional features named Diverse Region-Based CNN (DRCNN). Pan et al. [35] proposed a multi-grained network not only to extract the joint spectral-spatial information, but also to combine different grains spectral and spatial relationship for HSIs classification. However, the FGS details are discarded during a sequence of convolution and pooling operations for most of CNN-based HSIs classification. Jiao et al. [36] used the pre-trained Fully Convolutional Networks (FCN) [37] for spatial distribution prediction. However, the pre-trained FCN model on natural image dataset cannot retain the high-resolution spectrum information in HSIs. In this paper, a unified encoder-decoder network is builded to extract the FGS features with high-level semantics for HSIs classification.

III. METHODS

In this section, we explain the proposed FGSCNN for HSIs classification in details, which contains three blocks. The first block is applied to extract high-level semantics with low-resolution feature maps. The second block is implemented to retain the FGS distribution with high-resolution feature maps. The third one utilizes a balance parameter to unify the previous two parts into a loss function for HSIs classification.

A. OVERVIEW OF THE FGSCNN

In general, the traditional HSIs classification methods make use of the high-level semantics. However, these methods pay little attention to retain the FGS information. To address this issue, we propose an encoder-decoder convolution network with FGS information for hyperspectral images.
The overall frameworks of FGSCNN is illustrated in Figure 1. The encoder block is used to extract high-level semantics, while the decoder part is utilized to fuse the high-level low-resolution semantics and the fine-grained high-resolution spatial information, namely, to get the FGS features with high-level semantics. The FGS features with high-level semantics are fed into the fully-connected network. There are two fully-connected layers applied: the first one is used for dimensionality reduction of features to increase computational efficiency, and another one is served to generate the final classification results by a softmax classifier.

Different from AutoEncoder, which takes the output of the encoder as the feature, FGSCNN utilizes the FGS features with high-level semantics which is controlled with the supervision of high-level semantics to classify the pixel. The proposed FGSCNN is also similar to the U-Net. U-Net is designed for segmentation, while, FGSCNN is implemented for classification. And above all, an auxiliary item is performed on the output of encoder to retain the high-level semantic information as much as possible, which is one of the most important difference with U-Net.

**B. THE ENCODER ARCHITECTURE OF FGSCNN**

The encoder block via CNN are designed to exploit the spatial and spectral correlation across neighboring pixels. Let us assume $X = \{x_1, x_2, \ldots, x_n\} \in \mathbb{R}^{1 \times 1 \times B}$, where $B$ is the number of spectral bands of HSIs and $n$ is the number of pixels which have a ground truth label, namely $Y = \{y_1, y_2, \ldots, y_n\}$. For each pixel of $X$ set, the image patch $x_i^*$ with the size of $w \times w \times B$ ($w$ is the window size) are extracted, where $x_i$ is its centered pixel and the ground truth of the image patch is $y_i$. Therefore, the pre-processed $X$ is represented as $X^* = \{x_1^*, x_2^*, \ldots, x_n^*\} \in \mathbb{R}^{w \times w \times B}$. Then, each patch $x_i^*$ is fed into the encoder block. The encoder block can be formulated as:

$$\hat{x}_i^* = f(x_i^*)$$  \hspace{1cm} (1)

where $\hat{x}_i^*$ is the extracted high-level semantics and $f(\cdot)$ denotes the encoder model by CNN.

The general process of $f(\cdot)$ includes convolution, pooling and batch normalization steps. The detailed information of this encoder network is illustrated in Figure 2. Specifically, the encoder network regards $x_i^*$ as input, and employ convolution layers to extract semantics. The pooling layers reduce the number of parameters in networks, and enhance the generalization ability of networks. The hyperspectral pixels in a small neighborhood around the central pixel are jointly represented by the CNN model for spectral-spatial feature extraction. In general, after a series of operations, the low-resolution feature map is obtained, carrying the high-level semantics for HSIs classification. However, the obtained low-resolution feature map loses the correlations information across neighboring pixels. For example, the FGS details $\{x_1^*, x_2^*, x_3^*\}$ captured from early convolutional layers are abandoned.

**C. THE DECODER ARCHITECTURE OF FGSCNN**

In the decoder block, the FGS details of the encoder network are retained to generate the FGS features with high-level semantics. In the traditional CNN model, early convolutional layers with high spatial resolutions capture local details while the later convolutional layers with low spatial resolutions extract high-level semantics. To fuse the advantages of the two stages, the decoder network consists of deconvolution layers and skip connection layers which combines the FGS details generating from the encoder. Most of the HSIs classifier take advantage of the high-level semantics to assign the labels for pixels, discarding the FGS details. It is insufficient for high-resolution spatial and spectral HSIs classification by only exploiting high-level semantic features. To tackle this issue, a decoder network is designed, which translates the learned high-level low-resolution features to the FGS features with high-level semantics. The skip connection operations are exploited to retain the FGS details for the high-level semantics. The details of the decoder network are illustrated in Figure 4.

The inputs of the decoder block are the learned high-level semantic features $\hat{x}_i^*$ and the FGS details $\{x_{11}^*, x_{12}^*, x_{13}^*\}$,
which are extracted from the encoder block. Exploiting the FGS details \( \{x^*_1, x^*_2, x^*_3\} \) for HSIs classification directly is ineffective, lacking the guidance of semantics. In order to fuse the fine-grained details and high-level semantics, the deconvolution and the skip connection are implemented to generate joint features. It is inadequate to combine the two types of information directly for effective HSIs classification. Therefore, the convolution layers are also adopted to capture the FGS features with high-level semantics. In this way, the FGS features with high-level semantics is achieved with the supervision of semantic labels. The decoder block can be formulated as:

\[
s_i = g(x^*_1, x^*_2, x^*_3) \tag{2}
\]

where \( g(\cdot) \) denotes the decoder block, which includes deconvolution, convolution, skip connection and batch normalization steps, \( s_i \) is the FGS features with high-level semantics.

**D. LOSS FUNCTION**

Two kinds of information can be used for classification: the high-level semantics from the encoder output \( (\hat{x}^*_i) \) and the FGS features with high-level semantics from the decoder output \( (s_i) \). In the proposed FGSCNN, the FGS features with high-level semantics are used to calculate the main loss. In order to alleviate the loss of high-level semantics during the decoder process, the high-level semantics from the encoder output are used as an auxiliary to assist the final classification.

The main loss function is implemented on the FGS features with high-level semantics from the decoder output \( (s_i = g(\hat{x}^*_i, x^*_1, x^*_2, x^*_3)) \). The cross entropy is adopted for loss function calculation:

\[
L_m = -\sum_{k=1}^{m} \hat{y}_k \cdot \log p_k \tag{3}
\]

where \( p = \{p_1, p_2, \ldots, p_m\} \) is a probability distribution captured by the classification of decoder block, \( p_k \) is the probability of class \( k \), \( m \) is the number of kinds of pixel labels, and \( \hat{y} = [\hat{y}_1, \hat{y}_2, \ldots, \hat{y}_m] \) is the one-hot representation of the ground truth of pixel.

The auxiliary loss helps to optimize the learning process, which is performed on the high-level semantics from the encoder output:

\[
L_a = -\sum_{k=1}^{m} \hat{y}_k \cdot \log \hat{p}_k \tag{4}
\]

where \( \hat{p} = \{\hat{p}_1, \hat{p}_2, \ldots, \hat{p}_m\} \) is a probability distribution captured by the classification of encoder block, \( \hat{p}_k \) is the probability of class \( k \).

Hence, we add a parameter to balance the two loss functions. The unified loss function is then formulated as:

\[
L = (1 - \omega)L_m + \omega \cdot L_a \tag{5}
\]

where \( L_m, L_a \) denotes the main loss and auxiliary loss respectively, and \( \omega \) is a balance parameter.

**IV. EXPERIMENTS AND DISCUSSION**

**A. DATASETS**

The performance of the proposed FGSCNN is evaluated on the three publicly available datasets: Indian Pines, Pavia University and Salinas. For each dataset, we randomly selected 200 labeled pixels per class for training and the other pixels in the ground-truth map for testing. The details of the three datasets are as follows:
1) The Indian Pines dataset was collected by AVIRIS sensor over Northwestern Indiana, which contains $145 \times 145$ pixels. This dataset contains 200 spectral bands with a wavelength range from 0.4 to 2.5$\mu$m. There are 16 land-cover classes in ground-truth of this dataset. We removed the classes with fewer samples and selected 8 classes with more samples [34], [38], [39]. The numbers of training and testing samples are listed in Table 1.

2) The Pavia University dataset was captured by the ROSIS sensor over Pavia, northern Italy. This image contains $610 \times 340$ pixels and 103 spectral bands with the spatial resolution of 1.3m. There are 9 classes and 42,776 samples in total. Table 2 indicates the numbers of training and testing samples.

3) The last dataset is the Salinas dataset, and it is also the most sampled dataset. The Salinas dataset gathered with the AVIRIS sensor over Salinas Valley, which contains $512 \times 217$ pixels, 204 spectral bands. It has a total of 16 classes and 54,129 samples. The numbers of training and testing samples are listed in Table 3.

### B. DATA AUGMENTATION

A deep network usually requires sufficient training data to learn the model with a great deal of parameters. However, in HSIs classification tasks, only a few labeled samples may be available in practice. To address this issue, we take advantage of a data augmentation method to train the proposed FGSCNN in a small sample environment. The first step is flip, for which the process is applied by flipping the original samples horizontally or vertically. The second step is to add the Gaussian noise to the flipped data. The last step is to assign the newly generated data with the same class labels as the original data.

Table 4 investigates the effects of data augmentation. The performance of FGSCNN has an improvement of 0.88%, 0.44% and 0.42% on Indian Pines, Pavia University and Salinas respectively. On the whole, the augmentation operation for training data plays a positive role for HSIs classification.
### TABLE 5. On the Trade-off between $L_m$ and $L_o$ with varied $\omega$.

| $\omega$ | Indian Pines (OA%) | Pavia University (OA%) | Salinas (OA%) |
|----------|---------------------|------------------------|---------------|
| 0        | 98.18               | 99.58                  | 97.24         |
| 0.1      | 98.31               | 99.6                   | 97.57         |
| 0.2      | 98.65               | 99.65                  | 97.71         |
| 0.3      | 99.06               | 99.76                  | 97.79         |
| 0.4      | 98.79               | 99.75                  | 97.76         |
| 0.5      | 98.73               | 99.69                  | 97.58         |
| 0.6      | 98.65               | 99.71                  | 97.41         |
| 0.7      | 98.58               | 99.62                  | 97.44         |
| 0.8      | 98.42               | 99.65                  | 97.38         |
| 0.9      | 98.23               | 99.54                  | 97.33         |
| 1.0      | 98.08               | 99.47                  | 97.11         |

### TABLE 6. Effects of feature fusion on classification.

|                      | Indian Pines (OA%) | Pavia University (OA%) | Salinas (OA%) |
|----------------------|--------------------|------------------------|---------------|
| The output of encoder| 98.91              | 99.68                  | 97.66         |
| The output of decoder| 99.06              | 99.76                  | 97.79         |

### C. ON THE TRADE-OFF BETWEEN $L_m$ AND $L_o$

Parameter $w$ is used to balance $L_m$ and $L_o$. $L_m$ uses the FGS features with high-level semantics and $L_o$ utilizes the high-level semantics. When $\omega = 1$, the decoder block (auxiliary loss) is completely abandoned. With the decrease of $w$, the decoder block plays a more and more important role. When $\omega = 0$, the architecture of FGSCNN is similar to U-Net [40]. U-Net is proposed for image segmentation, while FGSCNN is exploited for HSIs classification. Table 5 reports the results with the varied $\omega$ on the three datasets. When $\omega = 0.3$, FGSCNN achieves the best classification performance on all the three datasets. The high-level semantics and the FGS features complement with each other for the best classification results. The output of the encoder (high-level semantics $\hat{x}^i$) is not used in the testing phase, however, it can provide the enhanced semantic information for $s^i$ (the FGS features with high-level semantics). Compared with no auxiliary loss, FGSCNN ($\omega = 0.3$) improves 0.88% on Indian Pines, 0.18% on Pavia University and 0.55% on Salinas.

Besides, we also investigate the classification performance by using the output from the encoder ($x^i$) and the decoder ($s_i$). The results are shown in Table 6. The output of decoder, carrying with FGS details by skip connection operations, contains more abundant information compared with that of encoder.

### D. THE EFFECT OF DIFFERENT WINDOW SIZES

The proposed FGSCNN utilizes the spatial information across neighboring pixels for HSIs classification. It is necessary to investigate the influence of the window size for HSIs classification. Figure 5 illustrates the classification performance with different window sizes. The window sizes are varied from $3 \times 3$ to $15 \times 15$. When the window size is $11 \times 11$, the classification performance of FGSCNN tends to be satisfied. Although, better performance can be found by using the window size of $13 \times 13$ and $15 \times 15$, considering the computational cost, the window size of $11 \times 11$ is applied for all the experiments except in this subsection.

### E. CLASSIFICATION PERFORMANCE

To validate the effectiveness of the proposed FGSCNN, the method is compared with some related HSI classification approaches, such as CNN [41], CNN-PPF [32], SS-CNN [42], DRCNN [34]. The overall accuracy (OA), average accuracy (AA) and kappa statistic (K) are used to evaluate the classification efficacy of each model. The configurations of the competing methods are as follows:

1) CNN: A simple CNN-based method for HSIs classification.
2) CNN-PPF: A CNN-based method, which utilizes Pixel-Pair Features, for HSIs classification.
3) SS-CNN: A CNN-based method, which uses spatial context and spatial-spectral feature, for HSIs classification.
4) DRCNN: A CNN-based method contains ‘multi-scale summation’ module and the central region feature extractor.
5) FGSCNN: A unified encoder-decoder framework to integrate high-level semantics and FGS details for HSIs classification.

All the methods listed above are based on CNN. The results of the experiments with 200 labeled pixels per class for training are given in Tables 7-9. In general, the classification performance of the proposed FGSCNN is superior to the other methods, especially on Indian Pines and Pavia University dataset. For example, in Table 8, the proposed FGSCNN yields OA 99.76%, nearly 7% higher than that of the CNN (i.e., 92.27%), 3% higher than that of the CNN-PPF.
FIGURE 7. The false color composite, the ground truth map and classification maps of Pavia University: (a) The false color composite, (b) the ground truth map, (c) classification maps of CNN, (d) classification maps of FGSCNN.

FIGURE 8. The false color composite, the ground truth map and classification maps of Salinas: (a) The false color composite, (b) the ground truth map, (c) classification maps of CNN, (d) classification maps of FGSCNN.

TABLE 7. Comparison of the overall classification accuracy (%) among the proposed method and the baselines using the Indian Pines dataset.

| Class  | CNN | CNN-PPF | SS-CNN | DRCNN | FGSCNN |
|--------|-----|---------|--------|-------|--------|
| 1      | 88.38 | 97.42 | 97.40 | 98.43 | 99.13 |
| 2      | 91.27 | 95.76 | 99.40 | 99.45 | 99.84 |
| 3      | 95.88 | 94.05 | 94.84 | 99.14 | 100.0 |
| 4      | 97.24 | 97.52 | 99.16 | 99.50 | 99.86 |
| 5      | 99.91 | 100.00 | 100.00 | 100.00 | 100.00 |
| 6      | 96.41 | 99.13 | 98.70 | 100.00 | 99.92 |
| 7      | 93.62 | 96.19 | 90.00 | 99.70 | 100.00 |
| 8      | 87.45 | 93.62 | 94.57 | 99.55 | 99.89 |
| 9      | 99.57 | 99.60 | 99.87 | 100.00 | 100.00 |
| AA(%)  | 93.36 | 97.03 | 98.22 | 99.53 | 99.85 |
| OA(%)  | 92.27 | 96.48 | 98.41 | 99.56 | 99.76 |
| Kappa  | -    | -      | -     | 99.41 | 99.68 |

TABLE 8. Comparison of the overall classification accuracy (%) among the proposed method and the baselines using the Pavia University dataset.

| Class  | CNN | CNN-PPF | SS-CNN | DRCNN | FGSCNN |
|--------|-----|---------|--------|-------|--------|
| 1      | 97.34 | 100.00 | 100.00 | 100.00 | 100.00 |
| 2      | 99.29 | 99.88 | 99.89 | 99.99 | 99.99 |
| 3      | 96.51 | 99.60 | 99.89 | 99.98 | 100.00 |
| 4      | 99.66 | 99.60 | 99.89 | 99.98 | 100.00 |
| 5      | 96.97 | 98.34 | 99.39 | 99.83 | 99.56 |
| 6      | 99.60 | 99.97 | 100.00 | 100.00 | 100.00 |
| 7      | 99.60 | 99.97 | 100.00 | 100.00 | 100.00 |
| 8      | 97.25 | 88.68 | 91.45 | 94.14 | 92.54 |
| 9      | 97.53 | 98.33 | 99.95 | 99.99 | 99.99 |
| 10     | 97.58 | 99.54 | 99.31 | 99.99 | 100.00 |
| 11     | 100.00 | 100.00 | 100.00 | 100.00 | 100.00 |
| 12     | 100.00 | 100.00 | 100.00 | 100.00 | 100.00 |
| 13     | 99.52 | 99.44 | 99.72 | 100.00 | 100.00 |
| 14     | 95.05 | 98.96 | 100.00 | 100.00 | 99.66 |
| 15     | 76.83 | 83.53 | 96.24 | 95.52 | 96.36 |
| 16     | 98.94 | 99.31 | 99.63 | 99.72 | 99.94 |
| AA(%)  | 94.84 | 97.73 | 98.94 | 99.26 | 99.21 |
| OA(%)  | 89.28 | 94.8 | 97.42 | 98.33 | 97.79 |
| Kappa  | -    | -     | -     | 98.14 | 97.53 |

TABLE 9. Comparison of the overall classification accuracy (%) among the proposed method and the baselines using the Salinas dataset.

| Class  | CNN | CNN-PPF | SS-CNN | DRCNN | FGSCNN |
|--------|-----|---------|--------|-------|--------|
| 1      | 97.34 | 100.00 | 100.00 | 100.00 | 100.00 |
| 2      | 99.29 | 99.88 | 99.89 | 99.99 | 99.99 |
| 3      | 96.51 | 99.60 | 99.89 | 99.98 | 100.00 |
| 4      | 99.66 | 99.60 | 99.89 | 99.98 | 100.00 |
| 5      | 96.97 | 98.34 | 99.39 | 99.83 | 99.56 |
| 6      | 99.60 | 99.97 | 100.00 | 100.00 | 100.00 |
| 7      | 99.60 | 99.97 | 100.00 | 100.00 | 100.00 |
| 8      | 97.25 | 88.68 | 91.45 | 94.14 | 92.54 |
| 9      | 97.53 | 98.33 | 99.95 | 99.99 | 99.99 |
| 10     | 97.58 | 99.54 | 99.31 | 99.99 | 100.00 |
| 11     | 100.00 | 100.00 | 100.00 | 100.00 | 100.00 |
| 12     | 100.00 | 100.00 | 100.00 | 100.00 | 100.00 |
| 13     | 99.52 | 99.44 | 99.72 | 100.00 | 100.00 |
| 14     | 95.05 | 98.96 | 100.00 | 100.00 | 99.66 |
| 15     | 76.83 | 83.53 | 96.24 | 95.52 | 96.36 |
| 16     | 98.94 | 99.31 | 99.63 | 99.72 | 99.94 |
| AA(%)  | 94.84 | 97.73 | 98.94 | 99.26 | 99.21 |
| OA(%)  | 89.28 | 94.8 | 97.42 | 98.33 | 97.79 |
| Kappa  | -    | -     | -     | 98.14 | 97.53 |

It is worth noting that the performance of DRCNN is better than our FGSCNN in Salinas dataset. We argue that FGSCNN can achieve a good performance by using a small amount of training samples. To further demonstrate the advantage, experiments are conducted on the three datasets with varied numbers of training samples per class. The results are shown in Table 10. From the Table 10, we can see that FGSCNN achieves the best performance among all the methods when the numbers of training samples are 100 and 150. Specially, the proposed FGSCNN yields OA 96.95%, nearly 3% higher than that of the CNN-PPF (i.e., 93.88%), 1% higher than that of the DRCNN (i.e., 95.54%). Both of the CNN-PPF and DRCNN are devoted to solve the problem of few label samples. It demonstrates that the proposed FGSCNN is also suitable for HSIs classification with inadequate training samples. FGSCNN depends on not only the traditional high-level semantics but also the proposed FGS details. However, the performance of FGSCNN is worse than CNN-PPF and DRCNN when the numbers of training samples is 50. There is a sharp decline in the performance of FGSCNN when the numbers of training samples vary from 100 to 50. The number
of training samples is too inadequate to extract effective high-level semantics for classification, even with the support of the FGS details.

All the experiments were conducted on a computer that has a NVIDIA GeForce RTX-2070 SUPER GPU (8GB GDDR5). Based on the stand back-propagation algorithms, the stochastic gradient descent algorithm is adopted to learn the network parameters, where the batch size is set to 160 and the learning rate is 0.001. Table 11 shows the computational complexity of training and test process using the proposed FGSCNN, CNN, CNN-PPF and DRCNN. The proposed FGSCNN has the faster convergence speed compared to other methods.

V. CONCLUSION

In recent years, deep learning methods have attracted much attention on HSIs. These methods are able to automatically capture spatial and spectral information. However, the FGS details are discarded during a sequence of convolution and pooling operations for HSIs classification. To tackle this issue, a new branch compared to the traditional CNN-based HSIs classification has been presented to build an encoder-decoder convolution network with FGS information for HSIs classification, denoted by FGSCNN. This model contains two main blocks: (i) the encoder block, which is used to extract high-level semantics from the original images, and (ii) the decoder block, which is utilized to retain the FGS details with high-level semantics by using a series of deconvolution, skip connection and convolution operations. In this study, the performance of the proposed FGSCNN has been assessed in three public datasets. From the results, it can be concluded that the proposed FGSCNN achieves convincing results compared with the state-of-the-art.

REFERENCES

[1] A. W. Nolin and J. Dozier, “A hyperspectral method for remotely sensing the grain size of snow,” Remote Sens. Environ., vol. 74, no. 2, pp. 207–216, Nov. 2000.
[2] E. Torrecilla, D. Stramski, R. A. Reynolds, E. Millán-Núñez, and J. Piera, “Cluster analysis of hyperspectral optical data for discriminating phytoplankton pigment assemblages in the open ocean,” Remote Sens. Environ., vol. 115, no. 10, pp. 2578–2593, Oct. 2011.
[3] R. J. Ellis and P. W. Scott, “Evaluation of hyperspectral remote sensing as a means of environmental monitoring in the St. Austell China clay (kaolin) region, Cornwall, U.K.,” Remote Sens. Environ., vol. 53, nos. 1–2, pp. 118–130, Oct. 2004.
[4] J. Franke, D. A. Roberts, K. Halligan, and G. Menz, “Hierarchical multiple endmember spectral mixture analysis (MESMA) of hyperspectral imagery for urban environments,” Remote Sens. Environ., vol. 113, no. 8, pp. 1712–1723, Aug. 2009.
[5] S. Roessner, K. Segl, U. Heiden, and H. Kaufmann, “Automated differentiation of urban surfaces based on airborne hyperspectral imagery,” IEEE Trans. Geosci. Remote Sens., vol. 39, no. 7, pp. 1525–1532, Jul. 2001.
[6] C. Chen, W. Li, H. Su, and K. Liu, “Spectral-spatial classification of hyperspectral image based on kernel extreme learning machine,” Remote Sens., vol. 6, no. 6, pp. 5795–5814, Jun. 2014.
[7] B. Yazi and F. Melgan, “Toward an optimal SVM classification system for hyperspectral remote sensing images,” IEEE Trans. Geosci. Remote Sens., vol. 54, no. 11, pp. 3367–3378, Nov. 2016.
[8] Z. He, L. Liu, R. Deng, and Y. Shen, “Low-rank group inspired dictionary learning for hyperspectral image classification,” Signal Process., vol. 120, pp. 209–221, Mar. 2016.
[9] Y. Chen, M. N. Nasrabad, and T. D. Tran, “Hyperspectral image classification using dictionary-based sparse representation,” IEEE Trans. Geosci. Remote Sens., vol. 49, no. 10, pp. 3973–3985, Oct. 2011.
[10] 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. 755–769, Feb. 2019.
[11] M. E. Paoletti, J. M. Haut, R. Fernandez-Beltran, J. Plaza, A. J. Plaza, and F. Pla, “Deep pyramidal residual networks for spectral–spatial hyperspectral image classification,” IEEE Trans. Geosci. Remote Sens., vol. 57, no. 2, pp. 740–754, Feb. 2019.
[12] J. M. Haut, M. E. Paoletti, J. Plaza, J. Li, and A. Plaza, “Active learning with convolutional neural networks for hyperspectral image classification using a new Bayesian approach,” IEEE Trans. Geosci. Remote Sens., vol. 56, no. 11, pp. 6440–6461, Nov. 2018.
[13] Y. Chen, K. Zhu, L. Zhu, X. He, P. Ghamisi, and J. A. Benediktsson, “Automatic design of convolutional neural network for hyperspectral image classification,” IEEE Trans. Geosci. Remote Sens., vol. 57, no. 9, pp. 7048–7066, Apr. 2019.
[14] W. Wang, S. Dou, Z. Jiang, and L. Sun, “A fast dense spectral–spatial convolution network framework for hyperspectral images classification,” Remote Sens., vol. 10, no. 7, p. 1068, Jul. 2018.
[15] X. Liu, Q. Sun, Y. Meng, M. Fu, and S. Bourouennah, “Hyperspectral image classification based on parameter-optimized 3D-CNNs combined with transfer learning and virtual samples,” Remote Sens., vol. 10, no. 9, p. 1425, Sep. 2018.
[16] C. Chen, Y. Ma, and G. Ren, “A convolutional neural network with fletcher–reeves algorithm for hyperspectral image classification,” Remote Sens., vol. 11, no. 11, p. 1325, Jun. 2019.
[17] S. Szegedy, W. Liu, Y. Jia, S. Reusser, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, “Going deeper with convolutions,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2015, pp. 1–9.
[18] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2016, pp. 770–778.
[19] G. Huang, Z. Liu, L. V. D. Maaten, and K. Q. Weinberger, “Densely connected convolutional networks,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jul. 2017, pp. 4700–4708.
[20] Q. Gao, S. Lim, and X. Jia, “Hyperspectral image classification using convolutional neural networks and multiple feature learning,” Remote Sens., vol. 10, no. 2, p. 299, Feb. 2018.
[21] M. Paolletti, J. Haut, J. Plaza, and A. Plaza, “Deep&dense convolutional neural network for hyperspectral image classification,” *Remote Sens.*, vol. 10, no. 9, p. 1454, 2018.

[22] K. Zhu, Y. Chen, P. Ghamisi, X. Jia, and J. A. Benediktsson, “Deep convolutional capsule network for hyperspectral image spectral and spectral-spatial classification,” *Remote Sens.*, vol. 11, no. 3, p. 223, Jan. 2019.

[23] Z. Gong, P. Zhong, Y. Yu, W. Hu, and S. Li, “A CNN with multiscale convolution and diversified metric for hyperspectral image classification,” *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 6, pp. 3599–3618, Jun. 2019.

[24] H. Zhang, Y. Li, Y. Jiang, P. Wang, Q. Shen, and C. Shen, “Hyperspectral classification based on lightweight 3-D CNN with transfer learning,” *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 8, pp. 5813–5828, Aug. 2019.

[25] S. Mei, J. Ji, Y. Geng, Z. Zhang, X. Li, and Q. Du, “Unsupervised spatial-spectral feature learning by 3D convolutional autoencoder for hyperspectral classification,” *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 9, pp. 6808–6820, Apr. 2019.

[26] Y. Jiang, Y. Li, and H. Zhang, “Hyperspectral image classification based on 3-D separable ResNet and transfer learning,” *IEEE Geosci. Remote Sens. Lett.*, vol. 16, no. 12, pp. 1949–1953, Dec. 2019.

[27] Y. Tarabalka, M. Fauvel, J. Chanussot, and J. A. Benediktsson, “SVM-and MRF-based method for accurate classification of hyperspectral images,” *IEEE Geosci. Remote Sens. Lett.*, vol. 7, no. 4, pp. 736–740, May 2010.

[28] G. Camps-Valls, L. Gomez-Chova, J. Munoz-Mari, J. Vila-Font, and J. Calpe-Maravilla, “Composite kernels for hyperspectral image classification,” *IEEE Geosci. Remote Sens. Lett.*, vol. 3, no. 1, pp. 93–97, Jan. 2006.

[29] M. Fauvel, J. Chanussot, and J. A. Benediktsson, “A spatial-spectral kernel-based approach for the classification of remote-sensing images,” *Pattern Recognit.*, vol. 45, no. 1, pp. 381–392, 2012.

[30] Z. Lin, Y. Chen, X. Zhao, and G. Wang, “Spectral-spatial classification of hyperspectral image using autoencoders,” in *Proc. 9th Int. Conf. Inf., Commun. Signal Process.*, Tainan, China, Dec. 2013, pp. 1–5.

[31] T. Li, J. Zhang, and Y. Zhang, “Classification of hyperspectral image based on deep belief networks,” in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Paris, France, Oct. 2014, pp. 5132–5136.

[32] W. Li, W. 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.

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

[34] 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.

[35] B. Pan, Z. Shi, and X. Xu, “MugNet: Deep learning for hyperspectral image classification using limited samples,” *ISPRS J. Photogram. Remote Sens.*, vol. 145, pp. 108–119, Nov. 2018.

[36] L. Jiao, M. Liang, H. Chen, S. Yang, X. Liu, and X. Cao, “Deep fully convolutional network-based spatial distribution prediction for hyperspectral image classification,” *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 10, pp. 5585–5599, Oct. 2017.

[37] J. Long, E. Shelhamer, and T. Darrell, “Fully convolutional networks for semantic segmentation,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Boston, MA, USA, Jun. 2015, pp. 3431–3440.

[38] F. Melgani and L. Bruzzone, “Classification of hyperspectral remote sensing images with support vector machines,” *IEEE Trans. Geosci. Remote Sens.*, vol. 42, no. 8, pp. 1778–1790, Aug. 2004.

[39] W. Li, E. W. Tramel, S. Prasad, and J. E. Fowler, “Nearest regularized subspace for hyperspectral classification,” *IEEE Trans. Geosci. Remote Sens.*, vol. 52, no. 1, pp. 477–489, Jan. 2014.

[40] O. Ronneberger, P. Fischer, and T. Brox, “U-Net: Convolutional networks for biomedical image segmentation,” in *Proc. Int. Conf. Med. Image Comput. Comput. Assisted Intervent.*, Munich, Germany, Oct. 2015, pp. 234–241.

[41] W. Hu, Y. Huang, L. Wei, F. Zhang, and H. Li, “Deep convolutional neural networks for hyperspectral image classification,” *J. Sens.*, vol. 2015, Jul. 2015, Art. no. 258619.

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

---

Z. Li et al.: Encoder–Decoder Convolution Network With FGS Information

---

ZHONGWEI LI received the Ph.D. degree from the China University of Petroleum. He is currently a Professor with the College of Oceanography and Space Informatics, China University of Petroleum, China. His current research interests include hyperspectral images classification.

FANGMING GUO is currently pursuing the master’s degree with the College of Oceanography and Space Informatics, China University of Petroleum, China. His current research interest includes hyperspectral images classification.

QI LI is currently pursuing the master’s degree with the College of Oceanography and Space Informatics, China University of Petroleum, China. His current research interest includes hyperspectral images classification.

GUANGBO REN received the Ph.D. degree from the Ocean University of China. He is currently an Associate Professor with the First Institute of Oceanography, Ministry of Nature Resources, China. His current research interests include ocean ecosystem hyperspectral remote sensing and coastal dynamics remote sensing.

LEIQUAN WANG received the Ph.D. degree in communication and electrical systems from the Beijing University of Posts and Telecommunications. He is currently a Lecturer with the College of Computer and Communication Engineering, China University of Petroleum, China. His current research interests include multimodal fusion, cross modal retrieval, and image/video caption.