A NOVEL CLASSIFICATION FRAMEWORK FOR HYPERSPECTRAL IMAGE CLASSIFICATION BASED ON MULTISCALE SPECTRAL-SPATIAL CONVOLUTIONAL NETWORK

Zhen Xu[1]*, Haoyang Yu*[1]*, Ke Zheng*[2]*, Lianru Gao*[2] *and Meiping Song*[1]

[1]Center of Hyperspectral Imaging in Remote Sensing (CHIRS), Information Science and Technology College, Dalian Maritime University, Dalian, 116026, China
[2]Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing, 100094, China

ABSTRACT

Multiscale spectral-spatial classification has been widely applied to hyperspectral image (HSI). Convolution neural networks (CNN) with multiscale spectral-spatial features have been introduced for hyperspectral image classification (HSIC) in recent years. However, most of current methods mainly use patches as input, which may cause a lot of redundancy in the testing phase and reduce processing efficiency. In this paper, we design a multiscale spectral-spatial CNN for HSIs (HyMSCN) based on a novel image-based classification framework. This network integrates multiple receptive fields fused features with multiscale spatial features at different levels. Experimental results from two real hyperspectral images demonstrate the efficiency of the proposed method.

Index Terms— Hyperspectral image, Classification, Convolution neural network, Multiscale spectral-spatial feature.

1. INTRODUCTION

The rise of hyperspectral remote sensing is of great significance in the field of modern remote sensing [1]. Compare with ordinary images, Hyperspectral image (HSI) has more information, wider bands and higher spectral resolution, which helps HSI to better detect the properties of ground features and makes HSI used in many tasks, such as classification. Among current classification framework, Supervised classification has been widely used in HSI [2]. A large number of studies have shown that the spectral-based classification method achieves some valid results. However, spectral-based classification methods do not show a better effect for more complex ground features. Therefore, we could consider both spectral information and spatial information to improve HSIC.

In recent years, CNN can efficiently extract spectral-spatial features in HSI, which makes CNN have gradually become hotspots of HSIC. Meanwhile, in order to retain more relevant information, HSIC introduced a variety of convolutional neural networks with multi-scale spectral-spatial features [3-5]. Fang et al. [3] proposed a 3-D dense convolutional network with a spectral attention network. Liang et al. [4] proposed a cooperative sparse autoencoder method to fuse deep spatial features and spectral information. In this method, a pretrained VGG-16 is also introduced to extract multiscale spatial structures. Pan et. al [5] proposed rolling guidance filter and vertex component analysis network to utilize spectral-spatial information. However, these methods have certain limitations in HSI, because they only use image patches as model inputs, which may cause a lot of redundancy and reduces the processing efficiency of the testing phase.

In order to solve the inefficiency of existing methods, we propose a multiscale spectral-spatial CNN for HSIs (HyMSCN) based on a novel image-based classification framework. The proposed method considers residual learning and feature pyramids. Specifically, the main contributions of the proposed method are divided into the following three points: 1) A new image-based classification framework is introduced to replace the traditional patch-based classification framework, which makes training and testing processes more efficient and convenient. 2) A residual multiple receptive field fusion block (ReMRFF) can extract local neighbor spatial information to make feature extraction lightweight and efficient. 3) HyMSCN method is proposed using multiscale spectral-spatial features. The feature pyramid structure is introduced into this method and multiple receptive field fused features are considered. Experimental results on the two sets of public hyperspectral datasets show that our method is significantly improved in time efficiency and accuracy than existing methods.

2. PROPOSED METHOD

In Section 2.1, a novel image-based classification framework is proposed. In Section 2.2, a HyMSCN is designed based on this framework. This network is mainly divided into two parts, including residual multiple receptive field fusion block

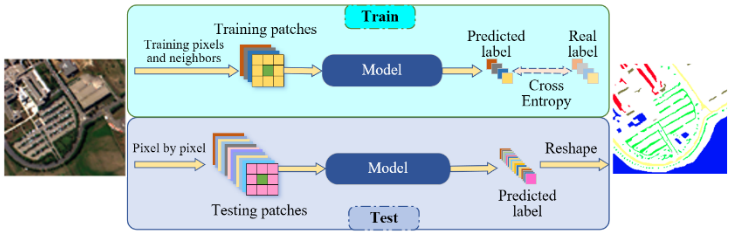

Figure 1. An illustration of patch-based classification for HSI.

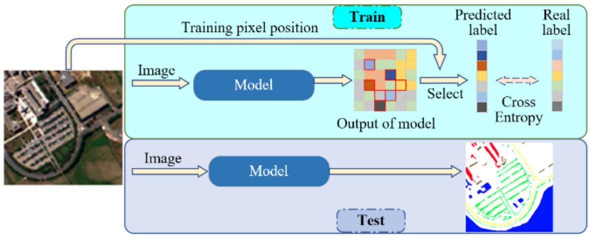

Figure 2. An illustration of image-based classification for HSI.

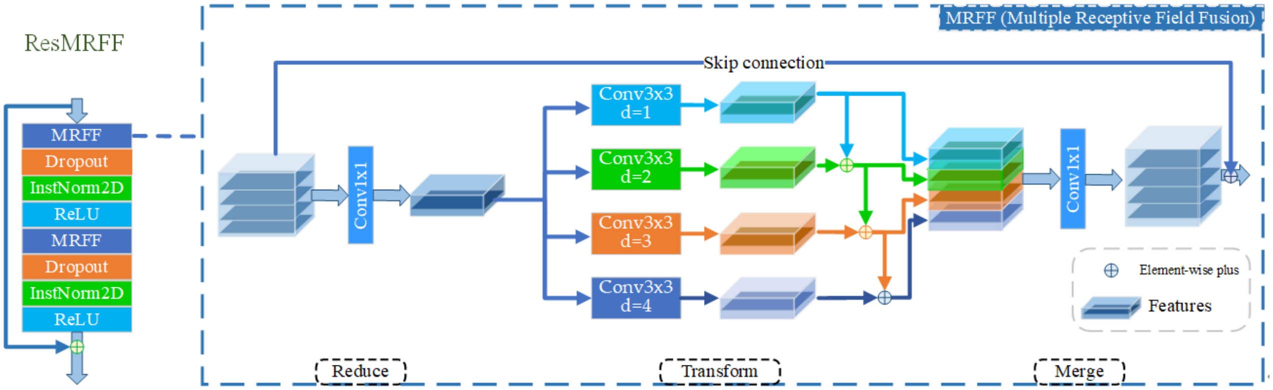

Figure 3. A schematic of the residual multiple receptive field fusion block (ResMRFF). The basic strategy of the multiple receptive field fusion block is represented as Reduce-Transform-Merge.

(ResMRFF) and feature pyramid. ResMRFF optimizes the network structure of the model to make it lightweight, while also avoids overfitting. In Section 2.3, a feature pyramid-based network structure is developed and multiple features at different scales are extracted. Then the multiscale features are fused for the final classification.

2.1. Patch-based and image-based for HSIC

The traditional deep learning HSIC framework is a patch-based framework (Figure 1). The training patch is composed of training pixels and its neighboring pixels, and then the patch is input into the model and the label of the center pixel of the patch is predicted. This method generates patches pixel by pixel during testing. Then, the patches are imported into the model one by one. Finally, the predicted labels are reshaped into an image of the same size as the original image.

Obviously, the patch-based classification method has some disadvantages. Firstly, the receptive field of the model is limited by the patch size. Secondly, the model needs to be redesigned if the patch size changes. The optimal patch size is related to the image spatial resolution, making it difficult to design a general classification model. Finally, during the testing phase, the pixel-by-pixel patch generation method results in information redundancy, consumes a lot of computing resources, and has low time efficiency.

In this paper, we proposed an image-based classification framework (Figure 2), which can overcome the above issues associated with patch-based classification. Arbitrary semantic segmentation model can be applied to proposed framework for HSIC. During the training phase, input is the image containing training samples, and the predicted labels for all corresponding pixels are used as the output. We use the position of labeled samples as a mask, it covers the output to select the corresponding pixels. During the testing phase, we bring the image into the model and then predict the corresponding labels to all pixels. Compared to patch-based classification frameworks, image-based classification frameworks have a faster inference process in testing since the information is non-redundant. In addition, the test result map can be directly output during each inference process, which can save a lot of computing resources.

2.2. Residual multiple receptive field fusion block

In recent years, many deep convolutional neural networks are introduced into HSIC. However, the deepening of the network layer brings the problem of network degradation. The residual network solves this problem well. The residual network is composed of many residual blocks. The core content of the residual block is the use of a skip connection, and the signal can be directly propagated from one residual

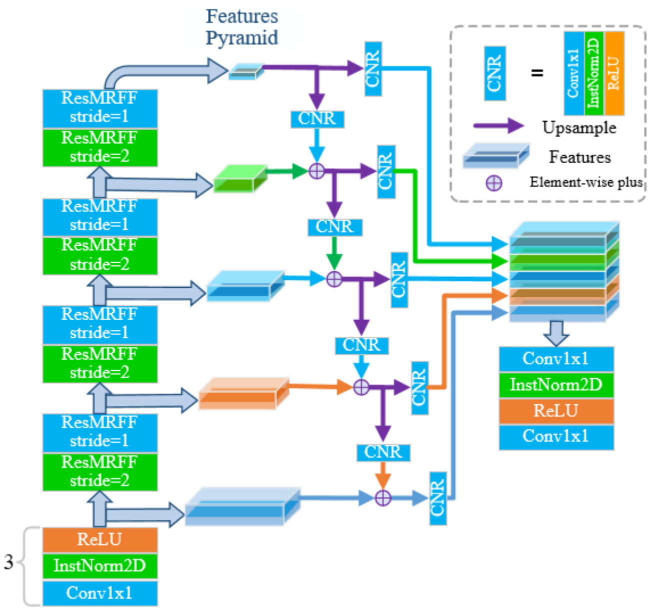

Figure 4. An illustration of the proposed HyMSCN network.

block to another block. This method can ease the gradients disappearance in deep networks.

We combined residual learning and multiple receptive field fusion block (MRFF) to design a new ResMRFF block (Figure 3). MRFF is designed based on dilated convolution to obtain a larger reception field with fewer parameters. The reduce-transform-merge strategy is the core of this structure. Firstly, dimensionality reduction through a 1×1 convolutional layer. Secondly, feature extraction through dilated convolutional layers with various dilation factors. Finally, multiple features fusion. The combination of receiving field features of different sizes can improve representation ability. This strategy can reduce the number of parameters and improve network computing efficiency. We define C and k as channels and kernel sizes, respectively. For a standard convolution layer, it contains $C_{in} \times k \times k \times C_{out}$ parameters. This MRFF module only has $(C_{in})^2/4 + (kC_{in})^2/4 + C_{in}C_{out}$ parameters. Thus, the module is more efficient. In addition, multiple dilated convolutions are introduced into MRFF to obtain multiple receptive field features and enlarge the diversity of features.

In ResMRFF, we add a dropout layer, instance normalization layer and a ReLU nonlinear activation function to prevent overfitting and enhance the adaptability of the network. In addition, the stride is set to 1 or 2. When the stride is set to 1, the output feature size is consistent with the input. Conversely, ResMRFF produce hierarchical features at a low spatial scale when stride=2.

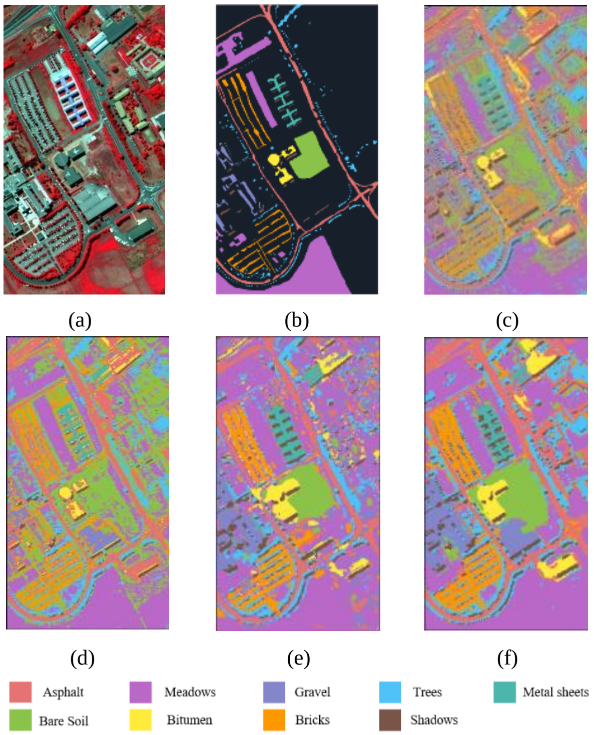

Figure 5. Classification map for ROSIS Pavia University data with 50 samples per class: (a) False color composite image, (b) Ground truth, (c) SVM (84.73%), (d) 3DCNN (90.08%), (e) UNet (95.32%), (f) HyMSCN (99.40%)

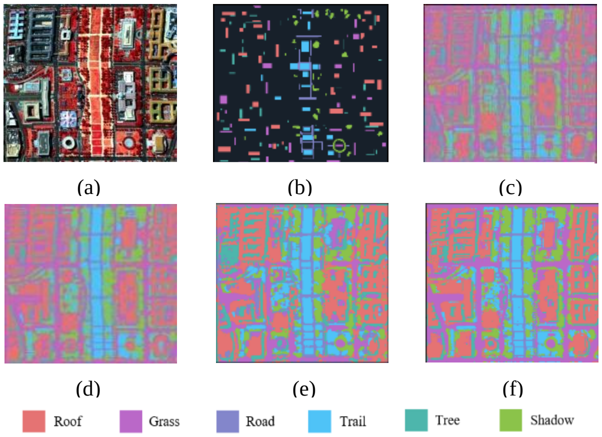

Figure 6. Classification map for HYDICE Washington DC Mall data with 50 samples per class: (a) False color composite image, (b) Ground truth, (c) SVM (89.91%), (d) 3DCNN (90.59%), (e) UNet (93.77%), (f) HyMSCN (98.61%)

2.3. Multiscale spectral-spatial convolutional network

The HyMSCN network are designed based on the ResMRFF module, and this network uses the bottom-up pathway to compute hierarchical features at different feature levels.

Table 1. Overall and individual class accuracy for ROSIS Pavia University data with 50 training samples per class. (Best results in bold)

Class	SVM	3DCNN	UNet	HyMSCN
Asphalt	78.41%	92.73%	95.69%	**99.93%**
Meadows	83.53%	93.66%	98.10%	**99.88%**
Gravel	81.81%	80.24%	76.60%	**98.95%**
Trees	93.97%	93.66%	99.12%	**99.93%**
Metal sheets	99.25%	99.37%	99.48%	**99.92%**
Bare Soil	86.39%	90.82%	**100.00%**	**100.00%**
Bitumen	92.20%	85.30%	81.94%	**97.43%**
Bricks	82.48%	88.00%	89.58%	**95.80%**
Shadows	**99.88%**	94.23%	99.16%	99.16%
Overall accuracy	84.73%	90.08%	95.32%	**99.40%**

Table 2. Overall and individual class accuracy for HYDICE Washington DC Mall data with 50 training samples per class. (Best results in bold)

Class	SVM	3DCNN	UNet	HyMSCN
Roof	79.24%	89.05%	93.49%	**98.87%**
Grass	87.47%	78.84%	92.69%	**95.81%**
Road	97.24%	95.73%	86.72%	**98.98%**
Trail	95.75%	97.25%	98.99%	**99.77%**
Tree	98.25%	93.30%	91.30%	**99.11%**
Shadow	95.33%	97.73%	98.96%	**99.68%**
Overall accuracy	89.91%	90.59%	93.77%	**98.61%**

Table 3. The overall accuracies produced by various classification methods for Pavia University data using a different number of training samples. (Best results in bold)

Sample	Classification Method			
(Per Class)	SVM	3DCNN	UNet	HyMSCN
90(10)	69.31%	73.24%	85.65%	**88.08%**
180(20)	75.43%	78.75%	88.01%	**96.82%**
270(30)	79.35%	83.87%	90.31%	**97.75%**
360(40)	82.35%	85.88%	92.05%	**98.26%**
450(50)	84.73%	90.08%	95.32%	**99.40%**

Table 4. The overall accuracies produced by various classification methods for Washington DC Mall data using a different number of training samples. (Best results in bold)

Sample	Classification Method			
(Per Class)	SVM	3DCNN	UNet	HyMSCN
60(10)	81.04%	82.33%	83.65%	**89.09%**
120(20)	85.03%	85.72%	86.75%	**95.26%**
180(30)	86.45%	87.93%	90.71%	**96.37%**
240(40)	87.10%	89.89%	92.26%	**97.96%**
300(50)	89.91%	90.59%	93.77%	**98.61%**

Table 5. The training and testing times for patch-based and image-based classification.

	Pavia University				Washington DC Mall			
	3DCNN	3DCNN	3DCNN	HyMSCN	3DCNN	3DCNN	3DCNN	HyMSCN
Patch size	9	15	21	-	9	15	21	-
Train Time of One Epoch (s)	0.18	0.39	0.94	0.54	0.19	0.47	0.96	0.39
Test Time of One Epoch (s)	16.12	47.64	100.62	0.25	11.24	34.65	72.97	0.20

In HyMSCN, Feature extraction module contains a 1 × 1 convolution layer, instance normalization layer and non-linearity activation function. In order to extract abundant and complete spectral features, we execute this module three times. In addition, feature pyramid structure is introduced into the network as shown in Figure 4. This structure produces hierarchical features with different spatial scales.

Firstly, different features are fused by a top-down approach. Because the features of the upper pyramid have strong semantics, they can be merged with the features of the lower pyramid by upsampling to achieve a highest spatial resolution. Secondly, the output feature maps are generated by integrating two 1×1 convolution layers, a normalized layer and an activation function. Finally, the output feature is sliced into a vector according to the position of training samples, and the loss is calculated between the vector labels and the corresponding true labels.

3. EXPERIMENT RESULTS

3.1. Experiment Setup

Two hyperspectral datasets are used to evaluate the performance of the proposed model. The first one is ROSIS Pavia University data. It contains 610×340 pixels and 103 spectral bands. The reference of this image includes nine mutually exclusive classes, with a total of 42776 labeled samples. The second one is the HYDICE Washington DC Mall data, it contains 280×307 pixels and 191 spectral bands. The reference of this image includes six classes, with a total of 10190 labeled samples. In both sets of experiments, 50 samples from each class are randomly selected. We compare the proposed model with support vector machines (SVM) [6], 3DCNN [7]and UNet [8]. SVM represent a typical spectral-domain classifier. 3DCNN, UNet are used as means for comparing patch-based classification and image-based classification, respectively. The computer configuration for this experiment is as follows: i7-7820X CPU, 32 GB of RAM, and a GTX 2080TI 11GB GPU.

3.2. Experiments with the Pavia University dataset and the Washington DC mall dataset.

In this section, Pavia University dataset and Washington DC mall dataset are used to compare proposed method with other methods, respectively. Figure 5 and Figure 6 displays the classification maps. According to the results reported in Table 1 and Table 2, we can draw several conclusions.

1) Compared with SVM, 3DCNN shows better classification performance, which proves the effectiveness of combining the spatial information.
2) Compared with 3DCNN, UNet produces higher accuracies, which indicates that image-based classification framework is more effective than patch-based classification framework.
3) Compared with UNet, the proposed HyMSCN yields better results, which demonstrating that a well-designed network combining multiscale and multi-level features is suitable for HSIC.

Moreover, two training sets are generated by randomly selecting 10, 20, 30, 40, and 50 samples per class in two different datasets. Table 3 and Table 4 display the overall accuracy for each group. The results reveal similar conclusions to the first experiment presented. HyMSCN still show the best performance in fewer training samples, which proved the reliability and robustness of the proposed method.

3.3. Comparing Time Consumed by Patch-Based Classification and Image-Based Classification

Improving the time efficiency of CNN for HSI has always been a research hotspot. In order to compare the time consumption of two different frameworks, we chose 3DCNN and HyMSCN to represent patch-based frameworks and image-based frameworks, respectively. The batch size for 3DCNN is set to 100, and the batch size for HyMSCN is set to 1 due to the image-based classification processes an entire image at one time. 50 samples per class are randomly selected as training samples for each test. In addition, the patch size is set to three different sizes, including 9, 15 and 21. Table 5 shows the experimental results.

1) The training times of 3DCNN are shorter than HyMSCN when patch size is smaller.
2) As the patch size increases, 3DCNN spends more training time than HyMSCN in one epoch.
3) The results reveal 3DCNN lasted 60 to 400 times longer than HyMSCN on test time.

It can be imagined that the patch-based classification method will consume more time due to adjacent patches contain a lot of redundant information when processing a larger image. Conversely, the image-based classification method is faster than the patch-based classification method because there is no redundant information in the test phase.

4. CONCLUSIONS

This paper proposes a multiscale spectral-spatial CNN for HSIs (HyMSCN) based on a novel image-based classification framework. The main contributions of the proposed method first an image-based classification framework for HSI to solve the inefficiency of patch-based classification，followed by a HyMSCN is designed based on the proposed classification framework to integrate multiple local neighbor information and multiscale spatial features. Experiments performed on two hyperspectral real datasets demonstrate that the proposed HyMSCN network can achieve a high classification accuracy and robust performance.

FUNDING

This work was supported by the Chinese Postdoctoral Science Foundation (No. 2020M680925), the Open Research Fund of Key Laboratory of Digital Earth Science, Institute of Remote Sensing and Digital Earth, Chinese Academy of Sciences (No. 2020LDE001), Fundamental Research Funds for the Central Universities, grant numbers 3132019341.

REFERENCES

[1] H. Yu, X. Shang, X. Zhang, L. Gao, M. Song and J. Hu, “Hyperspectral image classification based on adjacent constraint representation,” *IEEE Geoscience and Remote Sensing Letters*, pp. 1-5, Apr. 2020.

[2] H. Yu, L. Gao, W. Liao, P. Gamba, B. Zhang, “Global spatial and local spectral similarity-based manifold learning group sparse representation for hyperspectral imagery classification,” *IEEE Transactions on Geoscience and Remote Sensing*, pp. 3579-3582, May. 2020.

[3] B. Fang, Y. Li, H. Zhang and J. Chan, “Hyperspectral Images Classification Based on Dense Convolutional Networks with Spectral-Wise Attention Mechanism,” *Remote Sensing*, vol. 11, no. 2, Jan. 2019.

[4] M. Liang, L. Jiao, S. Yang, F. Liu, B. Hou and H. Chen, “Deep Multiscale Spectral-Spatial Feature Fusion for Hyperspectral Images Classification,” *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, pp. 1-14, Jun. 2018.

[5] 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 Sensing*, Jan. 2019.

[6] H. Yu, L. Gao, J. Li, S. Li, B. Zhang and J.A. Benediktsson, "Spectral-spatial hyperspectral image classification using subspace-based support vector machines and adaptive Markov random fields," *Remote Sensing*, vol. 11, no. 3, Apr. 2016.

[7] A. B. Hamida, A. Benoit, P. Lambert and C. B. Amar, “3-D Deep Learning Approach for Remote Sensing Image Classification,” *IEEE Transactions on Geoscience and Remote Sensing*, pp. 1-15, Apr. 2018.

[8] O. Ronneberger, P. Fischer and T. Brox, *U-Net: Convolutional Networks for Biomedical Image Segmentation*, Nov. 2015.