



# A Multi-scale Densely Connected and Feature Aggregation Network for Hyperspectral Image Classification

Jian Zhu<sup>1</sup>, Yi Liu<sup>1</sup>(✉), Jiajie Feng<sup>1</sup>, and Caihong Mu<sup>2</sup>

<sup>1</sup> School of Electronic Engineering, Xidian University, Xi'an 710071, China  
yiliuxd@foxmail.com

<sup>2</sup> School of Artificial Intelligence, Xidian University, Xi'an 710071, China

**Abstract.** Convolutional neural networks have been widely used in the field of hyperspectral image (HSI) classification due to their excellent ability to model local regions, and have achieved good classification performance. However, HSI classification still faces challenges such as insufficient representation of spectral-spatial features and inadequate fusion of multi-level features. To address these issues, we propose a Multi-scale Densely Connected and Feature Aggregation Network (MSDC-FAN) for HSI classification. The network mainly consists of a Spectral-Spatial Feature Extraction (SSFE) module, three Multi-scale Feature Extraction (MSFE) modules, and a Multilevel Feature Aggregation Module (MFAM). Firstly, the SSFE module is carried out to extract more comprehensive spectral-spatial features. Secondly, three MSFE modules are used in sequence to extract multi-scale features and highlight significant features, thus further improving the model's performance. Finally, the MFAM is designed to aggregate features at different levels, enhancing the model's feature representation ability. Experimental results on two commonly used hyperspectral datasets demonstrate the superiority of the proposed method.

**Keywords:** Hyperspectral image · Multi-scale feature · Dense connection

## 1 Introduction

Hyperspectral images (HSIs), containing abundant spectral-spatial information, have been widely used in environmental monitoring [1], mineral exploration [2], precision agriculture [3] and other fields. Early HSI classification methods only considered spectral features, such as support vector machine [4],  $k$ -nearest neighbor [5], and random forest [6], which had certain limitations. To simultaneously consider spectral-spatial information, methods such as sparse representation [7] and Markov random field [8] were proposed. However, spectral-spatial feature extraction and fusion still face challenges.

Convolutional neural networks (CNNs) have been widely used in HSI classification tasks due to their excellent ability to model local regions [9, 10]. Zhong et al. [11] designed spectral and spatial residual blocks to learn discriminative features and

alleviate the problem of accuracy degradation in deep networks. Song et al. [12] constructed a very deep network to extract more discriminative features and fused low-level, mid-level, and high-level features by addition to improve performance. Yu et al. [13] introduced an image-based global learning framework into HSI classification to fully utilize global information and achieved good classification results. Shi et al. [14] designed spectral feature extraction branch and spatial feature extraction branch to fully exploit the spectral-spatial information of HSIs and further improve the classification performance.

In addition, redundant information and noise in HSIs can interfere with model classification decisions, so it is important to extract effective information. To address this issue, attention mechanisms have been introduced into HSI classification [15] to help the model focus on more important features and regions. Zhu et al. [16] designed a spectral attention module and a spatial attention module to emphasize useful bands and pixels, achieving good classification results. In addition, the Transformer has been introduced into HSI classification tasks due to its excellent global feature modeling ability [17]. Sun et al. [18] introduced the Transformer for global features modeling, achieved good classification results and improved computational efficiency. Mei et al. [19] significantly improved the HSI classification accuracy by introducing a grouped pixel embedding module and constructing the Transformer in a hierarchical manner.

These networks have improved the HSI classification performance to some extent. However, limited by the fixed size of the CNN's convolutional kernel, the local features it extracts are limited, resulting in insufficient spectral-spatial feature representation. Additionally, multi-level features that are complementary and correlated have not been fully fused and utilized, and the classification performance needs to be further improved. To address the above issues, we propose a Multi-scale Densely Connected and Feature Aggregation Network (MSDC-FAN) for HSI classification. The main contributions of this paper are summarized as follows.

- (1) A spectral-spatial feature extraction (SSFE) module is devised to capture the spectral-spatial features of HSIs more comprehensively. Firstly, features of different scales are extracted by dilated convolution, and then concatenated and fused by skip connections.
- (2) A multi-scale feature extraction (MSFE) module is designed to fully extract HSI features. The multi-scale branch is adopted to extract multi-scale features, and the residual branch is carried out to make the information flow between the shallow layer and the deep layer. Then the cross-attention module is employed to enhance the feature fusion of the two branches, thus improving the model's performance.
- (3) A multi-level feature aggregation module (MFAM) is proposed to enhance the model's feature representation ability. Three MSFE modules are used in sequence to extract multi-scale features at different levels, which are then aggregated through the top-down channel to enhance the feature representation.

The rest of this paper is organized as follows. Section 2 provides a detailed introduction to the MSDC-FAN method. Section 3 presents the experimental results and analysis. Section 4 concludes the paper.

## 2 Proposed Method

In this section, we will introduce in detail our proposed MSDC-FAN for HSI classification, the overall framework of which is shown in Fig. 1. Firstly, the SSFE module is adopted to extract features from the dimension-reduced HSI after principal component analysis (PCA), in order to acquire the spectral-spatial information effectively. Then, the two-dimensional (2D) convolution is used to unify the number of channels. Secondly, multi-scale and multi-level features are extracted by stacking three MSFE modules in sequence, and features of different levels are aggregated by the MFAM module to obtain the final feature representation for classification. Finally, the aggregated features are fed into the linear layer for classification.

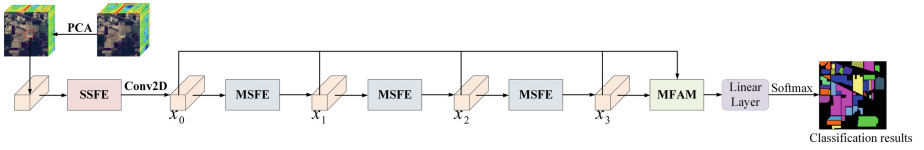


Fig. 1. Overall framework of the proposed MSDC-FAN

### 2.1 Spectral-Spatial Feature Extraction Module

The three-dimensional (3D) convolution can move simultaneously in both spectral and spatial dimensions, making it more effective to extract the spectral-spatial features of HSIs. Therefore, we construct the SSFE module to capture the features in HSIs more comprehensively, and the specific structure of the SSFE module is shown in Fig. 2.

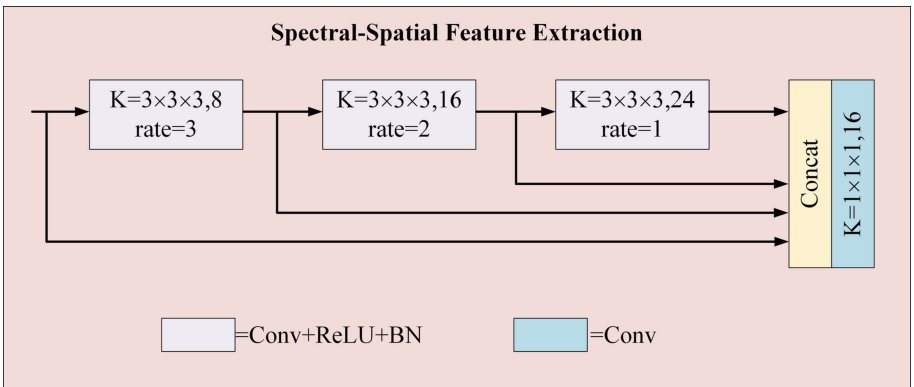
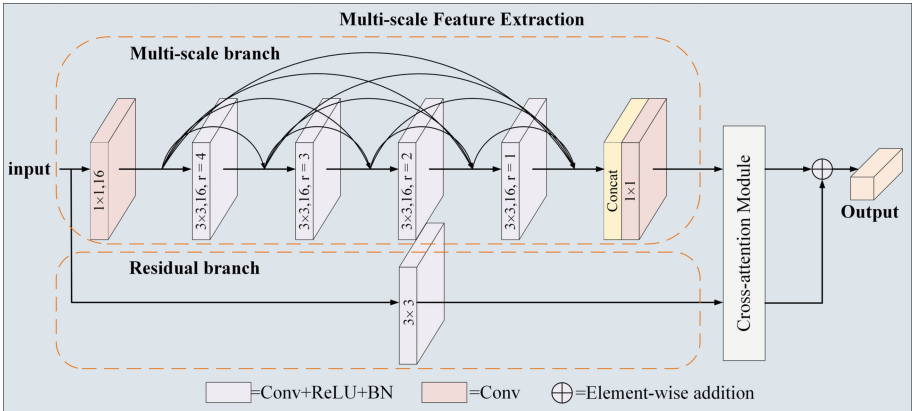


Fig. 2. Specific structure of the SSFE module. Conv, ReLU and BN represent convolution operation, ReLU activation function and batch normalization, respectively.  $K = 3 \times 3 \times 3$ , 8 represents eight convolution kernels of size  $3 \times 3 \times 3$ . Rate and Concat represent the dilation rate and the concatenation operation, respectively

The SSFE module mainly consists of dilated convolutions and skip connections. By increasing the dilation rate (i.e., rate in Fig. 2), the receptive field of the convolution kernel can be expanded while keeping the kernel size fixed, thereby improving the network performance. Skip connections can fully utilize the multi-level features extracted by the module, enhancing the feature representation capability of the network. Firstly, three convolution kernels are cascaded for multi-scale feature extraction. Then, the features extracted by each kernel are concatenated along the channel dimension through skip connections. Finally, the channel number is unified to 16 by a convolution kernel of size  $1 \times 1 \times 1$ .

## 2.2 Multi-scale Feature Extraction Module

**Overall Structure of the MSFE Module.** Using fixed receptive field, CNNs cannot capture all details and features in the HSI. To enhance the feature representation capability of the model, multi-scale convolutions are used to extract features of different scales. In addition, redundant and noisy information in HSI may lead to decreased classification accuracy. The introduction of attention mechanisms can help the model focus on pixels and spectral bands with important information, thereby improving classification accuracy and performance. Therefore, we propose the MSFE module to extract multi-scale features and highlight significant features, and the specific structure of the MSFE module is shown in Fig. 3.



**Fig. 3.** Specific structure of the MSFE module.  $3 \times 3, 16$  represents 16 convolution kernels of size  $3 \times 3$ , and  $r$  represents the size of dilation rate

The MSFE module consists of the multi-scale branch, the residual branch and the cross-attention module. The multi-scale branch is constructed by dilated convolutions and dense connections. The dilated convolutions extract features of different scales by cascading convolution kernels with different dilation rates. The dense connections connect the features of all previous layers to the input of the current layer, increasing the reusability of features and avoiding information loss. The residual branch uses skip

connections to allow shallow and deep features to complement each other, improving the ability of the model to represent the features. The cross-attention module aims to highlight the significant features and enhance the feature fusion of the two branches, thereby improving the performance of the model.

As shown in Fig. 3, MSFE first uses  $1 \times 1$  convolution kernel to unify the channel number of the input feature  $x_i \in R^{s \times s \times c}$  to 16, denoted as  $x_{i,0} \in R^{s \times s \times 16}$ . Secondly, four convolution kernels are cascaded and dense connections are used to extract multi-scale features, where the size and number of the four convolution kernels are  $3 \times 3$  and 16, and the dilation rates (i.e.,  $r$  in Fig. 3) are 4, 3, 2 and 1, respectively. In dense connections, for the  $l$ th layer, it receives  $x_{i,0}$  and all the previously extracted features, denoted as  $x_{i,0}, x_{i,1}, x_{i,2}, \dots, x_{i,l-1}$ , and uses them as input to calculate the output, as shown in Eq. (1):

$$x_{i,l} = BN(\delta(\text{Conv}([x_{i,0}, x_{i,1}, x_{i,2}, \dots, x_{i,l-1}]))) , l = 1, 2, 3, 4 \quad (1)$$

where  $BN$  and  $\delta$  represent the batch normalization and ReLU activation functions, respectively.  $[\cdot]$  represents concatenation operation,  $\text{Conv}$  represents  $3 \times 3$  convolution operation. Then, the input feature  $x_{i,0}$  and the output feature  $x_{i,l}$  of each convolutional layer are concatenated along the channel dimension, and the output is unified to have the same channel number as  $x_i$  by a  $1 \times 1$  convolution. In addition, in the residual branch, a  $3 \times 3$  convolution kernel is used to convolve the input feature  $x_i$ .

**Cross-Attention Module.** The cross-attention module aims to combine the complementary information from two different-level features to better highlight the effective information and enhance the feature fusion of the two branches, and the specific structure is shown in Fig. 4. It mainly consists of the spectral attention block and the spatial attention block. The former aims to generate band weights to recalibrate the importance of each band and adjust the correlation of each band, while the latter aims to enhance spatial information of pixels that have the same class as the center pixel and suppress pixels from other classes. The specific implementation of the cross-attention module is described as follows.

The features  $M$  and  $N$  extracted by the residual and the multi-scale branches have the shape of  $H \times W \times C$ . The feature  $F_{spe} \in R^{H \times W \times C}$  is obtained by adding and fusing the two features element-by-element and sent to the spectral attention.

In spectral attention, firstly, the input features are processed by the global average pooling and max pooling operations in two branches, respectively, to obtain  $F_{spe,avg}$  and  $F_{spe,max}$ , whose shapes are  $1 \times 1 \times C$ . Then, they are passed through the multilayer perceptron (MLP) to obtain features  $F'_{spe,avg}$  and  $F'_{spe,max}$ . Finally, the two obtained features are added and normalized using the Softmax function to obtain the band weight  $W_{spe} \in R^{1 \times 1 \times C}$ . The operations above are described as shown in Eqs. (2)–(4):

$$F'_{spe,avg} = \delta(W_1 * \delta(W_0 * F_{spe,avg})) \quad (2)$$

$$F'_{spe,max} = \delta(W_1 * \delta(W_0 * F_{spe,max})) \quad (3)$$

$$W_{spe} = \sigma\left(\text{add}\left[F'_{spe,avg}, F'_{spe,max}\right]\right) \quad (4)$$

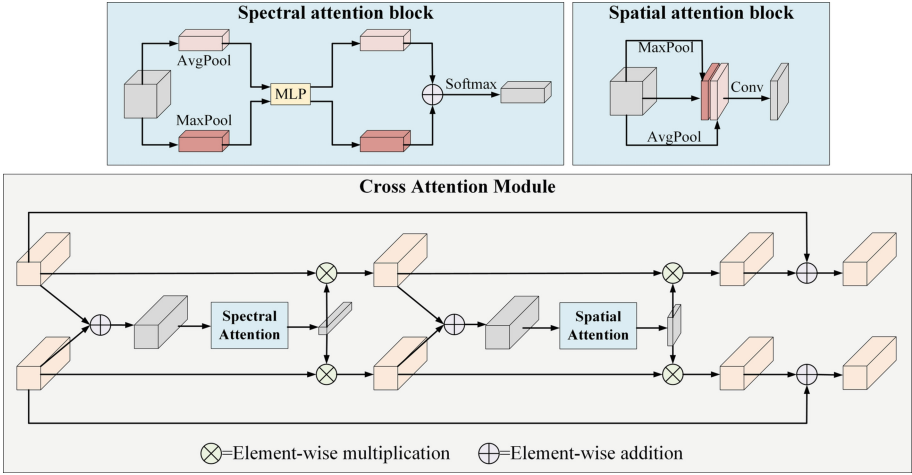


Fig. 4. Specific structure of the cross-attention module

where  $*$  represents the convolution operation,  $add[\cdot]$  represents element-wise addition,  $\sigma$  and  $\delta$  represent the Softmax normalization function and ReLU activation function, respectively.  $W_0$  and  $W_1$  are the weight parameters of the two fully connected layers sharing weights.

Then, the obtained band weights  $W_{spe}$  are used to recalibrate the bands of features  $M$  and  $N$  to highlight the informative bands. Next, we additively fuse the calibrated features to obtain the feature  $F_{spa} \in R^{H \times W \times C}$ , which is fed into the spatial attention.

In spatial attention, firstly, the input features are subjected to global average pooling and max pooling operations in the two branches, respectively, to obtain  $F_{spa,avg}$  and  $F_{spa,max}$ , whose shapes are both  $H \times W \times 1$ , and they are concatenated along the channel dimension to obtain  $F_{avg,max} \in R^{H \times W \times 2}$ . Then, it is fed into a two-dimensional convolution to obtain the spatial weight  $W_{spa}$ . The operations above are described as shown in Eq. (5):

$$W_{spa} = \sigma(W_0 * cat[F_{spa,avg}, F_{spa,max}]) \quad (5)$$

where  $*$  represents the convolution operation,  $cat[\cdot]$  represents concatenation along the channel dimension,  $\sigma$  represents the Sigmoid activation function, and  $W_0$  represents the weight parameter of a  $3 \times 3$  convolution kernel.

Then, the resulting spatial weight  $W_{spa}$  is used to recalibrate the spatial information of the features to highlight the useful information. Next, we add the original features through the skip connections and the features passed through the attention blocks to avoid the loss of information of features with lower weights.

### 2.3 Multi-level Feature Aggregation Module

In deep neural networks, shallow features contain spatial details such as edges and textures, whereas they are lacking of semantic information. Conversely, deep features have

stronger semantic information but lose spatial details. Therefore, fusing complementary information between features at different levels can further improve the performance of HSI classification. In this paper, by stacking MSFE modules, features at different levels can be extracted. To fully utilize these features, we construct the MFAM module, the specific structure is shown in Fig. 5.

Firstly, MFAM aggregates all levels of features through a top-down channel to enhance the feature representation, as shown in Eq. (6).

$$x_i = \delta(W_i * \text{add}[x_i, x_{i+1}]), i = 0, 1, 2 \quad (6)$$

where  $*$  and  $\delta$  represent convolution operation and ReLU activation function respectively,  $W_i$  represents the weight parameter of a  $3 \times 3$  convolution kernel, and  $\text{add}[\cdot]$  represents element-wise addition. Then, the enhanced features of each level are concatenated along the channel dimension, and the  $3 \times 3$  convolution kernel is used to unify the channel number.

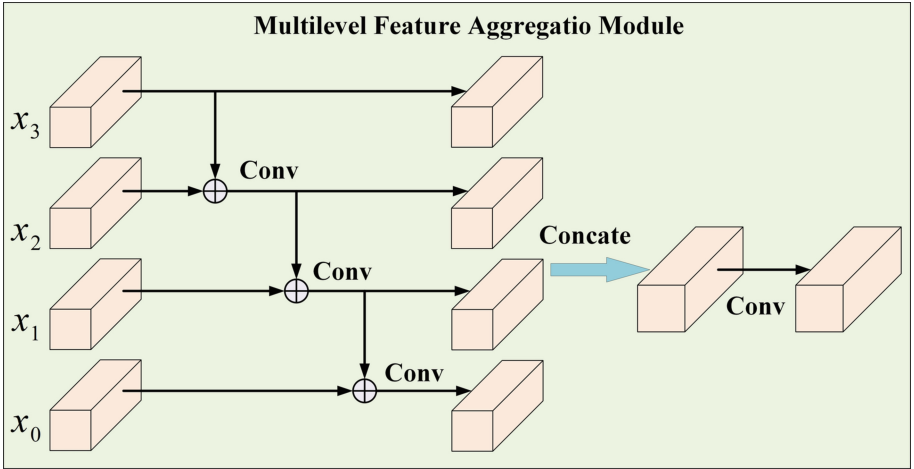


Fig. 5. Specific structure of the MFAM module

### 3 Experiment and Analysis

#### 3.1 Dataset Description and Experiment Setup

**Dataset Description.** To evaluate the performance of the proposed method, two classic datasets are selected for the experiments: Indian Pines (IP) and Pavia University (PU). IP dataset is a hyperspectral remote sensing image with a size of  $145 \times 145$ . It contains 200 available spectral bands and 16 classes of land cover with a total of 10,249 labeled samples. The PU dataset is a hyperspectral remote sensing image with a size of  $610 \times 340$ . It contains 103 available spectral bands and nine classes of land cover with a total of 42,776 labeled samples.

**Evaluation Metrics.** Overall accuracy (OA), average accuracy (AA), and Kappa coefficient are used as evaluation metrics. OA indicates the ratio of the number of correctly classified samples to the total number of samples. AA represents the average of the accuracy values with which the samples of each class are classified. Kappa coefficient can measure the consistency of classification results.

**Experiment Setup.** The Pytorch deep learning framework is used to train the network in the experiments, and epoch and batch\_size are set to 100 and 32, respectively. The learning rate is set to 0.001, and Adam is selected as the optimization method for the experiment. Each group of experiments is performed five times independently, and the average values are taken as the experimental results, and the standard deviations about three metrics are also provided.

The patch sizes (Patch\_Size) in IP and PU are set to 21 and 15, respectively, and the numbers of the principal components of PCA (PCA\_Components) is set to 32. The percent of the training samples are set to 10% and 5%, respectively.

### 3.2 Experiment Results and Analysis

SVM [4], SSRN [11], DFFN [12], SSFTT [18] and GAHT [19] are selected as comparative methods to validate the effectiveness of MSDC-FAN. The experimental results of these methods on the IP and PU datasets are as follows.

**IP Dataset.** Firstly, the IP dataset is used to evaluate the performance of the proposed model, and the experimental results are shown in Table 1. The results of the evaluation metrics indicate that the MSDC-FAN model proposed in this paper performs the best, achieving the highest OA, AA, and Kappa values.

As shown in Fig. 6, SSFTT and GAHT perform poorly in the “Corn-notill” (class 2, in blue) category and at the edges of the region, while MSDC-FAN generates a more accurate classification map. This is because MSDC-FAN not only uses cross-attention to highlight the significant features but also fully utilizes the multi-scale and multi-level features, which leads to better feature representation ability of the model.

**PU Dataset.** We further evaluate the performance of the proposed model on the PU dataset, and the experimental results are shown in Table 2. The PU dataset has a large number of samples and relatively balanced sample sizes for each land cover category, so the classification results of each method are relatively ideal. The evaluation data shows that MSDC-FAN performs the best and has relatively uniform accuracy on each category. As shown in Fig. 7, several comparative methods perform poorly in the “Gravel” (class 3, in orange) category, while MSDC-FAN improves the accuracy of “Gravel” by 6.45% and 3.38% compared to SSFTT and GAHT, respectively, achieving 99.85%. This indicates that MSDC-FAN has a better ability to represent spectral-spatial features and can distinguish spectrally similar classes well.

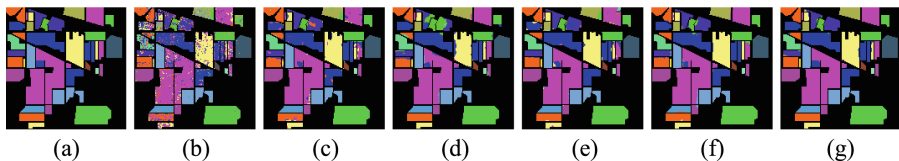
### 3.3 Parametric Analysis

**Impact of Patch\_Size and PCA\_Components on OA.** We analyze the impact of Patch\_size and PCA\_Components on classification performance on IP and PU datasets.



**Table 1** Classification results of IP dataset obtained by different methods

Class	SVM	SSRN	DFFN	SSFTT	GAHT	MSDC-FAN
1	38.10	98.54	93.17	99.02	98.05	96.59
2	77.84	93.21	86.40	95.22	96.17	99.47
3	70.41	86.96	92.34	96.17	93.82	99.12
4	45.79	98.97	99.25	99.53	96.43	99.15
5	90.80	98.80	95.36	98.67	97.43	99.17
6	96.19	99.82	99.21	99.79	99.12	99.06
7	76.92	87.20	84.80	99.20	99.20	100
8	96.29	99.95	99.91	100	99.49	100
9	27.78	78.89	75.56	81.11	93.33	92.22
10	71.78	94.40	92.46	95.59	98.29	98.19
11	83.67	95.82	94.34	96.99	97.90	99.10
12	68.54	86.85	90.60	90.64	94.61	97.79
13	94.05	100	88.76	100	96.54	99.78
14	93.77	99.81	97.70	99.68	99.75	100
15	56.03	88.59	93.78	91.30	94.87	99.31
16	85.71	95.71	95.71	97.62	78.81	98.10
OA	80.81 $\pm 0.01$	95.04 $\pm 1.01$	93.67 $\pm 3.52$	96.82 $\pm 0.24$	97.19 $\pm 0.31$	<b>99.13</b> <b><math>\pm 0.09</math></b>
AA	73.35 $\pm 0.01$	93.97 $\pm 1.64$	92.46 $\pm 6.50$	96.28 $\pm 0.70$	95.86 $\pm 0.32$	<b>98.57</b> <b><math>\pm 0.29</math></b>
Kappa	78.05 $\pm 0.01$	94.33 $\pm 1.15$	92.78 $\pm 4.04$	96.38 $\pm 0.27$	96.80 $\pm 0.35$	<b>99.01</b> <b><math>\pm 0.11</math></b>

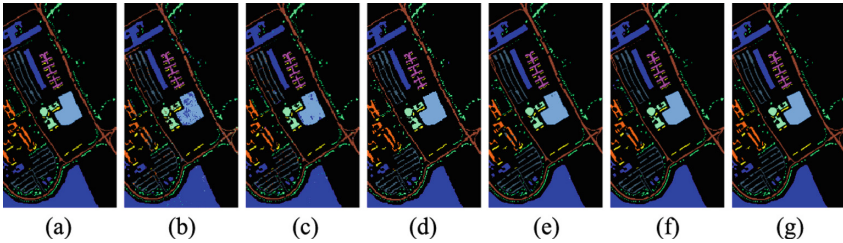


**Fig. 6.** Classification maps of the IP dataset. (a) Ground-true map. (b) SVM (OA = 80.81%). (c) SSRN (OA = 95.04%). (d) DFFN (OA = 93.67%). (e) SSFTT (OA = 96.82%). (f) GAHT (OA = 97.19%). (g) MSDC-FAN (OA = 99.13%)

Among them, the Patch\_size increases in the range of [11,21], and PCA\_Components are among [32,128] and [32,112] on IP and PU datasets, respectively. It can be seen from Fig. 8(a) that the impact of Patch\_size and PCA\_Components on the OA values of IP dataset fluctuates slightly, and a local maximum region can be obtained by selecting appropriate Patch\_size and PCA\_Components. As can be seen from Fig. 8(b),

**Table 2** Classification results of PU dataset obtained by different methods

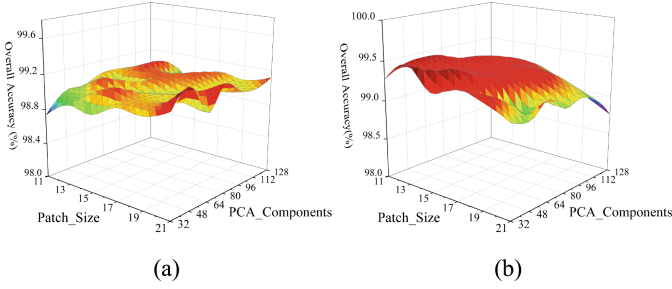
Class	SVM	SSRN	DFFN	SSFTT	GAHT	MSDC-FAN
1	93.37	97.61	98.43	99.40	99.58	99.86
2	98.04	99.58	99.96	99.89	99.95	99.96
3	72.53	86.13	97.28	93.40	96.47	99.85
4	94.40	95.29	92.17	98.83	94.59	97.36
5	98.36	99.98	99.77	100	99.66	100
6	87.02	94.31	99.94	99.65	99.98	100
7	83.23	97.58	99.92	99.70	100	99.95
8	90.82	92.64	99.21	98.92	99.37	99.43
9	99.89	99.31	86.49	99.62	90.71	96.98
OA	93.47 $\pm 0.02$	97.03 $\pm 0.18$	98.66 $\pm 0.63$	99.30 $\pm 0.15$	99.08 $\pm 0.07$	<b>99.65</b> <b><math>\pm 0.04</math></b>
AA	90.85 $\pm 0.01$	95.82 $\pm 0.36$	97.02 $\pm 0.85$	98.82 $\pm 0.24$	97.81 $\pm 0.17$	<b>99.27</b> <b><math>\pm 0.04</math></b>
Kappa	91.31 $\pm 0.01$	96.06 $\pm 0.24$	98.22 $\pm 0.83$	99.07 $\pm 0.20$	98.78 $\pm 0.10$	<b>99.53</b> <b><math>\pm 0.05</math></b>



**Fig. 7.** Classification maps of the PU dataset. (a) Ground-true map. (b) SVM (OA = 93.47%). (c) SSRN (OA = 97.03%). (d) DFFN (OA = 98.66%). (e) SSFTT (OA = 99.30%). (f) GAHT (OA = 99.08%). (g) MSDC-FAN (OA = 99.65%)

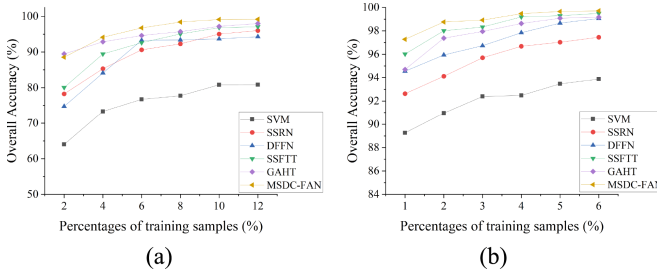
smaller Patch\_size and PCA\_Components are more suitable for PU dataset. The best classification performance is obtained when Patch\_size is 15 and PCA\_Components is 32.

**OA of Different Models Using Different Percentages of Training Samples.** Figure 9 shows the OA values of different models that use different percentages of training samples. From Fig. 9, it can be seen that the OA values of all methods increase as the percentages of training samples increase. Among them, the OA values of SSFTT and



**Fig. 8.** Impact of Patch\_Size and PCA\_Components on OA. (a) IP dataset. (b) PU dataset

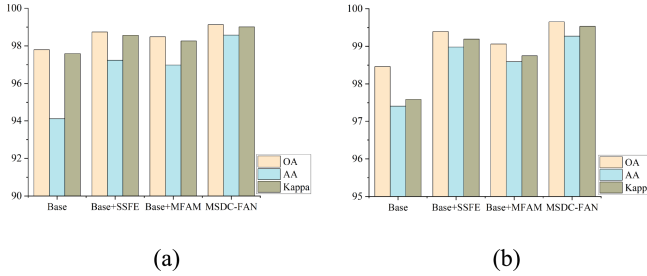
GAHT are close to our MSDC-FAN, demonstrating their good classification performance. On the whole, MSDC-FAN achieves the best results in almost all cases and it can obtain good performance even with extremely few training samples.



**Fig. 9.** OA of different models using different percentages of training samples. (a) IP dataset. (b) PU dataset

### 3.4 Ablation Experiments

We conduct ablation experiments about SSFE and MFAM modules on the IP and PU datasets. Three variants of MSDC-FAN are compared with MSDC-FAN, among which Base represents the network constructed only by MSFE. The experimental results are shown in Fig. 10. The Base network has the worst classification performance on IP and PU datasets. When SSFE or MFAM is added, there is a significant improvement in classification performance compared to the Base network, which verifies the effectiveness of SSFE and MFAM. The MSDC-FAN network has the best classification performance, which reflects that using two modules at the same time can not only fully extract spectral-spatial features, but also make full use of multi-level features, which further enhances the ability of the network to represent the features, thus contributing to the improvement of classification performance.



**Fig. 10.** Ablation experiments on IP and PU datasets. (a) IP dataset. (b) PU dataset

## 4 Conclusion

In this paper, we propose a Multi-scale Densely Connected and Feature Aggregation Network (MSDC-FAN) to improve the performance of hyperspectral image (HSI) classification. The experimental results show that the proposed MSDC-FAN performs better than several state-of-the-art methods in almost all cases, because MSDC-FAN has better ability of spectral-spatial feature representation by fully utilizing multi-scale and multi-level features. MSDC-FAN is able to perform well even with extremely few training samples. In the future, we will investigate how to improve HSI classification performance further with limited samples.

**Acknowledgements.** This work was supported by the National Natural Science Foundation of China (Nos. 62077038, 61672405, 62176196 and 62271374).

## References

1. Zhang, X., Liu, L., Chen, X., Gao, Y., Jiang, M.: Automatically monitoring impervious surfaces using spectral generalization and time series Landsat imagery from 1985 to 2020 in the Yangtze River Delta. *Remote Sens.* **2021** (2021).
2. Avtar, R., Sahu, N., Aggarwal, A. K., et al.: Exploring renewable energy resources using remote sensing and GIS—A review. *Res.* **8**(3), 149(2019)
3. Weiss, M., Jacob, F., Duveiller, G.: Remote sensing for agricultural applications: a meta-review. *Remote Sens. Environ.* **236**, 111402 (2020)
4. Ye, Q., Huang, P., Zhang, Z., Zheng, Y., Fu, L., Yang, W.: Multiview learning with robust double-sided twin SVM. *IEEE Trans. Cybern.* **52**(12), 12745–12758 (2021)
5. Cariou, C., Chehdi, K.: A new k-nearest neighbor density-based clustering method and its application to hyperspectral images. In: 2016 IEEE International Geoscience and Remote Sensing Symposium (IGARSS), pp. 6161–6164(2016).
6. Zhang, Y., Cao, G., Li, X., Wang, B., Fu, P.: Active semi-supervised random forest for hyperspectral image classification. *Remnote Sens.* **11**(24), 2974 (2019)
7. Cui, B., Cui, J., Lu, Y., Guo, N., Gong, M.: A sparse representation-based sample pseudo-labeling method for hyperspectral image classification. *Remote Sensing*, **12**(4), 664 (2020)
8. Cao, X., Xu, Z., et al.: Spectral-spatial hyperspectral image classification via robust low-rank feature extraction and Markov random field. *Remote Sens.* **11**(13), 1565 (2019).

9. Liu, J., Yang, Z., et al.: Hyperspectral remote sensing images deep feature extraction based on mixed feature and convolutional neural networks. *Remote Sens.* **13**(13), 2599 (2021).
10. Li, Y., Zhang, H., Shen, Q.: Spectral–spatial classification of hyperspectral imagery with 3D convolutional neural network. *Remote Sens.* **9**(1), 67 (2017)
11. Zhong, Z., Li, J., Luo, Z., Chapman, M.: Spectral–spatial residual network for hyperspectral image classification: A 3-D deep learning framework. *IEEE Trans. Geosci. Remote Sens.* **56**(2), 847–858 (2017)
12. Song, W., Li, S., et al.: Hyperspectral image classification with deep feature fusion network. *IEEE Trans. Geosci. Remote Sens.* **56**(6), 3173–3184 (2018)
13. Yu, H., Zhang, H., et al.: Dual-channel convolution network with image-based global learning framework for hyperspectral image classification. *IEEE Geosci. Remote Sens. Lett.* **19**, 1–5 (2021)
14. Shi, H., Cao, G., Zhang, Y., Ge, Z., Liu, Y., Fu, P.: H2A2 Net: a hybrid convolution and hybrid resolution network with double attention for hyperspectral image classification. *Remote Sens.* **14**(17), 4235 (2022)
15. Yang, K., Sun, H., et al.: Cross-attention spectral–spatial network for hyperspectral image classification. *IEEE Trans. Geosci. Remote Sens.* **60**, 1–14 (2021)
16. Zhu, M., Jiao, L., et al.: Residual spectral–spatial attention network for hyperspectral image classification. *IEEE Trans. Geosci. Remote Sens.* **59**(1), 449–462 (2020)
17. Hong, D., et al.: SpectralFormer: rethinking hyperspectral image classification with transformers. *IEEE Trans. Geosci. Remote Sens.* **60**, 1–15 (2021)
18. Sun, L., Zhao, G., et al.: Spectral–spatial feature tokenization transformer for hyperspectral image classification. *IEEE Trans. Geosci. Remote Sens.* **60**, 1–14 (2022)
19. Mei, S., Song, C., et al.: Hyperspectral image classification using group-aware hierarchical transformer. *IEEE Trans. Geosci. Remote Sens.* **60**, 1–14 (2022)