Hierarchical Shrinkage Multi-Scale Network for Hyperspectral Image Classification with Hierarchical Feature Fusion

Hongmin Gao, *Member, IEEE,* Zhonghao Chen, *Student Member, IEEE,* and Chenming Li

***Abstract*—Recently, deep learning (DL) based hyperspectral image classification (HSIC) has attracted substantial attention. Many works based on the convolutional neural network (CNN) model have been certificated to be significantly successful for boosting the performance of HSIC. However, most of these methods extract features by using a fixed convolutional kernel and ignore multi-scale features of the ground objects of hyperspectral images (HSIs). Although some recent methods have proposed multi-scale feature extraction schemes, more computing and storage resources were consumed. Moreover, when using CNN to implement hyperspectral image classification, many methods only use the high-level semantic information extracted from the end of the network, ignoring the edge information extracted from shallow layers of the network. To settle the preceding two issues, a novel HSIC method based on hierarchical shrinkage multi-scale network (HSMSN) and the hierarchical feature fusion (HFF) is proposed, with which the newly proposed classification framework can fuse features generated by both of multi-scale receptive field and multiple levels. Specifically, multi-depth and multi-scale residual block (MDMSRB) is constructed by superposition dilated convolution to realize multi-scale feature extraction. Furthermore, according to the change of feature size in different stages of the neural networks, we design a hierarchical shrinkage multi-scale feature extraction network by pruning MDMSRB to reduce the redundancy of network structure. In addition, to make full use of the features extracted in each stage of the network, the proposed network hierarchically integrates low-level edge features and high-level semantic features effectively. Experimental results demonstrate that the proposed method achieves more competitive performance with a limited computational cost than other state-of-the-art methods.**

***Index Terms*—Hyperspectral image classification (HSIC), convolutional neural network (CNN), hierarchical shrinkage multi-scale network (HSMSN), multi-depth and multi-scale residual block (MDMSRB), hierarchical feature fusion (HFF)**

Ⅰ. INTRODUCTION

With the rapid development of imaging spectrometers and platforms, imaging spectroscopy (also called hyperspectral imaging) has gradually occupied a significant and central position in lots of fields of visual data analysis. This is due to the hyperspectral image (HSI) has hundreds of continuous narrow-band spectral information (high spectral resolution), which reflects the absorption and reflection of solar reflection light in different wavelength bands (i.e., visible light, infrared or near-infrared) by the imaged ground object [1], [2]. These rich features, like fingerprint information of imaged objects, can realize accurate detection (pixel level) of objects. Therefore, HSI has been widely used in many fields and achieved promising results, such as agricultural application [3], water quality monitoring [4], mineral distribution and composition analysis [5], [6], natural disaster prediction [7], and military field [8]. Among them, hyperspectral image classification (HSIC) is an extremely important research direction in HSI analysis [9]-[12].

In the early research of HSIC, due to different ground objects that can reflect corresponding spectral differences in the HSI, a large number of researchers put forward plenty of classification methods based on the spectral characteristics of a single-pixel and achieved good classification results [12]-[14]. However, the high-dimensional data characteristics of HSI not only bring rich features but also lead to the Hughes phenomenon [15], which significantly limited the performance of HSIC [16]. Therefore, the extreme learning machine (ELM) [17] and support vector machine (SVM) [18] were put forward, which solved this problem to a certain extent. However, the spectral features contained in single-pixel are easily blurred by environmental factors (i.e. salt-pepper noise) and the consequent problem of intraclass variability (and interclass similarity), which seriously hinders the classification performance of the above methods [19]. In addition, a single-pixel independently will inevitably lose the spatial correlation information in the image, so the spatial features of hyperspectral images have been paid more and more attention by researchers. Considering that adjacent pixels in hyperspectral images generally have similar spectral characteristics and prone to be the same category, a series of classifiers with neighborhood blocks centered by labeled pixels as input have been proposed [20], [21]. Experiments show that the complementary information of spatial features makes the classifiers can achieve better classification performance. Although these traditional machine learning methods have achieved good classification performance, the shallow

This paragraph of the first footnote will contain the date on which you submitted your paper for review. This work was supported in part by the National Natural Science Foundation of China under Grant 62071168. in part by the National Key Research and Development Program of China under Grant 2018YFC1508106, and in part by the Fundamental Research Funds for the Central Universities of China under Grant B200202183. *(Corresponding author: Chenming Li.)*

Hongmin Gao, Zhonghao Chen, and Chenming Li are with the College of Computer and Information, Hohai University, Nanjing 211100, China (e-mail: gaohongmin@hhu.edu.cn; chenzhonghao@hhu.edu.cn; lcm@hhu.edu.cn).

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JSTARS.2021.3083283, IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing

handcrafted features utilized by machine learning are seriously lacking in representation capacity.

In recent years, the excellent performance of deep-learning (DL) has made it widely studied in many fields. For instance, natural language processing (NLP), computer vision (CV), and emotion analysis. In the early exploration of applying DL-based methods to HSIC, the deep belief network (DBN) [22], the stacked autoencoder (SAE) [23], and their derivative models [24], [25] were introduced for HSIC. However, these methods only use the spectral information of the HSI, which is based on the premise that each pixel of the HSI is pure. In comparison, the convolutional neural network (CNN) is the most widely studied and applied structure. Its translational invariance when extracting image features (linear and non-linear features) has the parameters sharing characteristics that make it powerful in image feature excavation. In the literature, there is a massive amount of works that applied CNN to hyperspectral image feature extraction to enhance the performance of HSIC [26]-[28]. On the one hand, due to the powerful extraction capabilities of the CNN for the features of complex hyperspectral data, the HSI classifier based on the CNN achieves excellent classification performance. On the other hand, the CNN-based classification models have an immensely flexible structure, which makes them well adapted to the data input of different structures and the flexible extraction of different dimensional features (spatial and spectral domains) [29]-[31].

Therefore, considering the spatial texture information of hyperspectral images, the 2D-CNN model is widely introduced to HSIC [32], [33]. For example, Li et al. [32] proposed a deep 2D-CNN model, whose a large number of parameters are trained by pixel pairs, and finally, the final classification results were obtained by voting strategy. Paoletti et al. [33] designed a deep pyramidal residual network, which not only gets more feature maps by increased the dimension of the feature maps gradually but also reduces the computational burden of the model. Then, pan et al. [28] used 1D and 2D CNN to extract spectral and spatial features respectively and fused the extracted features for classification. In addition, Chen et al. [26] and Ben et al. [34] introduced 3D-CNN based model for HSIC, which uses the 3D convolution kernel to extract spectral-spatial joint features robustly. However, although the 3D-CNN model achieved good classification performance, 3D convolution operations will greatly increase the computational complexity and consume a lot of computing resources. For this reason, 3D and 2D hierarchical extraction strategies of spectral-spatial features were proposed [35]-[37]. These 3D/2D hybrid models can extract spectral and spatial features in turn, which can alleviate the computational burden caused by only using full 3D convolution to some extent. With the deepening of the network model, simply deepening the network cannot further improve the classification performance, but will lead to the gradient vanishing problem [38]. Moreover, due to the scarcity of HSI annotated samples, a large number of learnable parameters of the depth model cannot be fully trained, which results in an overfitting phenomenon [39]. To solve these problems, residual structure [40] is widely used in deep CNN to alleviate the gradient vanishing problem. For instance, in [41], a spectral-spatial residual network (SSRN) was proposed with two consecutive residual blocks in the spectral and spatial extraction module respectively, which can perform well under the condition of small samples especially. In the meantime, song et al. [42] constructed a deep feature fusion network (DFFN) with three different stages of residual modules to achieve complementary feature extraction, and finally fused the three levels of features before the classification. Inspired by the residual structure, a dense network [43] based on more sufficient bridging is proposed, which can not only effectively alleviate the problems such as the disappearance of the gradient, but also make use of the front layer information of the network repeatedly. Then, in [44], a dual-channel dense network was proposed to extract spectral and spatial features consecutively with several dense modules.

Although the models described above have achieved promising classification performance, single input size and receptive field may ignore the diversity of spatial feature sizes, which limits the further improvement of classification accuracy, especially under the condition of small samples. As a result, to fully extract more discriminative spatial features, a series of multi-scale feature extraction methods have been proposed [45]-[49]. In [45], multi-scale pyramid images were used as the input of the model called MCNN, which can explore the spatial features of different scales. However, MCNN results in serious redundancy of spatial features, which leads to unnecessary computational and memory costs. Lee and Kwon [46] proposed a spatial contextual deep convolution neural network (CD-CNN), in which three different sizes of the convolution kernel extracted multi-scale spatial features at the start of the model and fused subsequently. Similarly, Gong et al. [47] designed an HSIC method based on a three-channel multi-scale convolution neural network (MS-CNNs), which can consider multi-scale features of HSI in the spectral and spatial dimensions by 1D, 2D, and 3D multi-scale convolution kernels. However, [46] and [47] only extract and fuse spatial multi-scale features in the shallow layer of the network, which is not enough to fully obtain more abstract and comprehensive features. Furthermore, in order to make full use of the spatial structure information of ground objects, Zhang et al. [48] constructed a diverse region input convolution neural network (DR-CNN), which can provide considerable performance due to more neighborhood information be considered. Recently, Li et al. [49] proposed a two-stream deep feature fusion model based on global and local spatial features，which according to the different amount of information contained in the global and local, used two extraction networks with different depth, and finally realized the extraction of global and local features effectively. Particularly, the attention mechanism based on the squeeze and excitation network (SEN) was used to enhance the confidence of spatial useful features. Although [48] and [49] can get good classification results, the multi-branches structure inevitably makes the model larger and has a vast number of learnable parameters, which significantly increases the computing and storage burden.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JSTARS.2021.3083283, IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing

Admittedly, although these CNN-based methods have proved their powerful capacity for feature extraction, there are still some drawbacks that need to be overcome. Firstly, with the deepening of the network, the features extracted by the neural network gradually change from specific edge features to abstract semantic features. However, the sensitivity to the scale of these features is diverse at different depths of the network. As a result, only extracting multi-scale features at the input of the model or designing multi-scale feature extraction in the whole model will cause the loss or redundancy of features. Secondly, although the abstract semantic information extracted by deep neural networks plays an important role in the classification task, a series of convolution and pooling operations will seriously injure the boundary information [50]. Thirdly, although the model with a large number of training parameters can improve the classification accuracy to some extent, which will significantly consume time and storage resources [51].

In order to solve the above problems, a novel hierarchical shrinkage multi-scale network with a hierarchical feature fusion (HSMSN-HFF) is proposed in this paper. Different from the aforementioned multi-scale feature extraction method, considering that the features change from concrete to abstract with the depth of the network increases, the scale of the HSMSN fusion feature gradually decreases. In addition, in order to make full use of different levels of features, especially the fusion of shallow edge information and deep semantic information, a hierarchical feature fusion strategy is applied to the HSMSN. The main contributions of this paper are summarized as follows:

1) In order to generate fewer training parameters to improve the classification speed when expanding the receptive field and extracting multi-scale features. In the feature extraction stage, we use continuous dilated convolution kernels to light the weight of the network.
2) In order to extract multi-scale features, a multi-depth and multi-scale residual block (MDMSRB) is introduced. Specifically, it is realized by stacking different numbers of dilated convolution kernels with different dilation rates, which can obtain different scale receptive fields and suppress the gridding problem [52] caused by dilated convolution. Especially, different scale MDMSRB is used for multi-scale feature extraction under different network depths, which makes the network extract features more efficient. To effectively alleviate the over-fitting problem, the residual structure is applied to each MDMSRB.
3) In order to improve the classifier's utilization of the features extracted from the HSMSN, a hierarchical feature fusion (HFF) strategy is introduced to extract the features of different stages of the network. Specifically, we fuse the shallow edge features and deep abstract semantic features extracted by HSMSN to get clear edge information and accurate abstract information. Finally, two different levels of features are fused to generate more discriminative features.

The rest of this paper is summarized as follows. Section II introduces the related work of dilated convolution and multi-scale feature fusion. Section Ⅲ describe proposed method. Section Ⅳ gives experimental results. Finally, the conclusion of this article is drawn in Ⅴ.

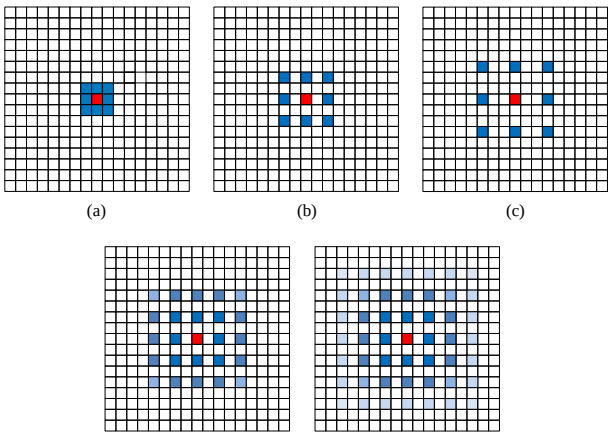


Fig. 1. Illustration of dilated convolution and gridding phenomenon. (a) Traditional 3×3 convolution, Dilation rate=1. (b) Dilation rate=2. (c) Dilation rate=3. (d-e) Gridding problem caused by dilated convolution.

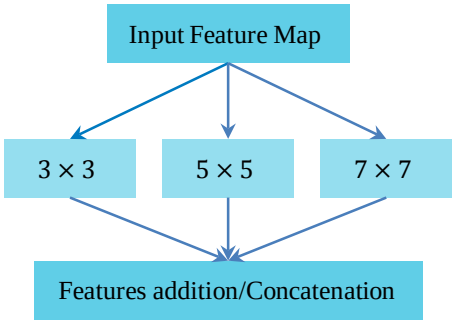


Fig. 2. Illustration of traditional multi-scale feature extraction.

Ⅱ. Related Work

A. Receptive Fields and Dilated Convolution

It is universally known that the size of the receptive field is significant for feature extraction of CNN, mainly because that the size of the receptive field determines the amount of neighbor information. Especially, in HSIC, many methods have proved that a large receptive field can improve the ability of global spatial feature extraction. However, the general method to increase the receptive field is usually realized by using a larger convolution filter, which will greatly increase the computational complexity of the model. Assuming that the number of input and output channels is the same, it is C ($C_{in} = C_{out} = C$). The convolution kernel size is $k \times k$ and the output characteristic graph size is $W \times H$. According to equations (1) and (2), the training parameter P and the number of floating-point operations (FLOPs) consumed by 5×5 convolution are 2.8 times that of 3×3convolution. To alleviate the problem of parameter explosion with the improvement of the receptive field, the dilated convolution was first proposed by Chen et al. [53]. As shown in Fig. 1 (b-c), a 3×3

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JSTARS.2021.3083283, IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing

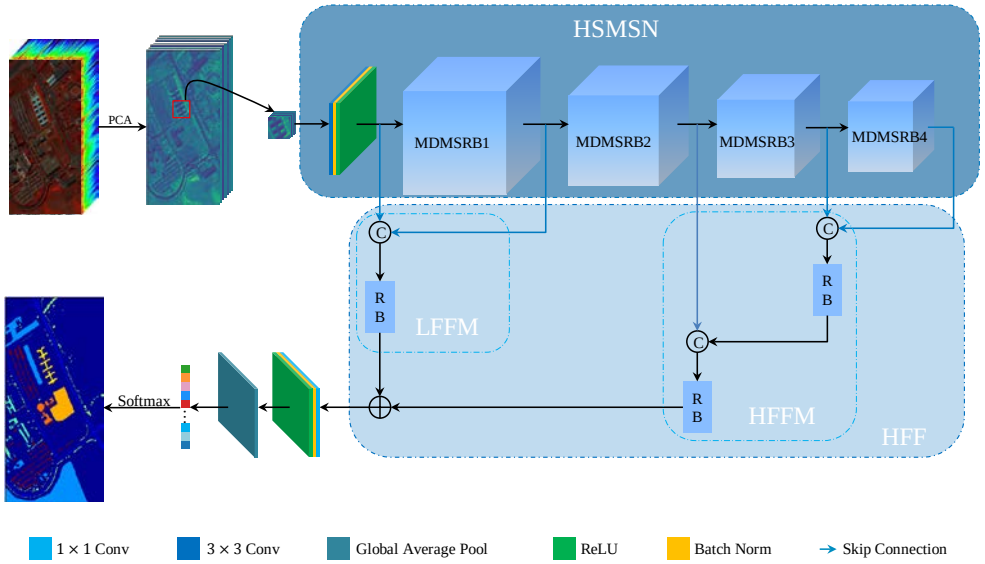


Fig. 3. Schematic of the proposed HSMSN-HFF. © denotes feature concatenation, and ⊕ denotes feature addition. 'RB' means 'residual block'. 'LFFM' and 'HFFM' represent the low-level feature fusion module and high-level feature fusion module. 'HSMSN' denotes multi-depth multi-scale residual network.

convolution operation has a larger receptive field when the dilation rate is 2/3, but the generated training parameters do not change than the original 3×3 convolution. In addition, the dilated convolution will not lead to the decline of the output resolution. Therefore, a series of hyperspectral image classification methods based on dilated convolution are proposed [54], [55]. However, it can be seen from Fig. 1 (d) and (e) that the dilated convolution extracts feature by sparsely sampling the feature maps, which will lead to the damage of information continuity and the loss of local feature information (gridding problem). Therefore, when implementing pixel-level classification of hyperspectral images with poor spatial resolution, the reduction of information continuity caused by the dilated convolution seriously limits the performance of the classification.

$$P = k \times k \times C_{in} \times C_{out} \#(1)$$
$$F = W \times H \times P \#(2)$$

B. *Multi-scale Analysis*

In the field of CV, multi-scale feature extraction can often get different scale information of the target, which is highly considered in many tasks (i.e. target detection [56], image segmentation [57], etc.). Generally speaking, as shown in Fig. 2, multi-scale feature extraction is achieved by using convolution filters of different scales [58]. Although it can get promising classification results, it will consume a lot of computing resources. In addition, as the feature map gradually becomes abstract from shallow to deep in the CNN, the diversity of feature scale decreases, so the multi-scale feature extraction should be different in different stages of the CNN.

Ⅲ. Proposed Method

In this section, the proposed multi-depth and multi-scale residual network with hierarchical feature fusion (HSMSN-HFF) will be introduced in detail. In the following subsections, the framework of the proposed model in Fig.3 will be introduced firstly. Then, we explain the reason why MDMSRB be introduced and how does it work. Finally, we introduce the architecture of the HFF module.

A. *Framework for Proposed Model*

Fig.3 shows the framework of the proposed HSIC model, which takes the Pavia of University (PU) data set as an example. Firstly, principal component analysis (PCA) is applied to the original HSI to reduce its spectral dimension to avoid the Hughes phenomenon. In addition, PCA can effectively retain the main features of the HSI spectral dimension, which can not only reduce the redundant spectral bands but also decrease the burden of model training. Then, in order to fully utilize the spatial features of the HSI, the PCA-processed image is segmented into 3-D image cubes centered on labeled pixels. Subsequently, some of these cubes are used to train the parameters of the proposed HSMSN-HFF. Then, in the HSMSN-HFF, the feature maps with a high representative are obtained by novel feature extraction and fusion mechanism. In the first step of HSMSN-HFF, multi-scale features of the HSI are extracted by a hierarchical shrinkage multi-scale feature extraction network at different stages of the network. With the deepening of the network, the feature scale and diversity are decreasing, so four gradual shrinkage scale MDMSRBs are used in different stages of HSMSN for multi-scale feature extraction. In the second step, an HFF mechanism is used to take full advantage of the features extracted in different stages of the neural network, especially the complementary fusion of high-level semantic features and low-level margin features. In the HFF, after the low-level features and high-level features are fused step by step respectively, the two different levels of

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JSTARS.2021.3083283, IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing

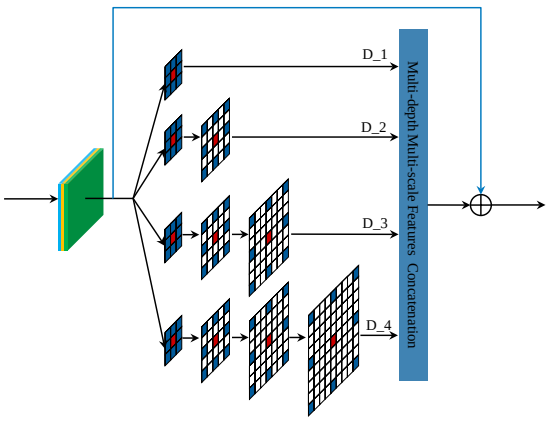


Fig. 4. Architecture of proposed MDMSRB.

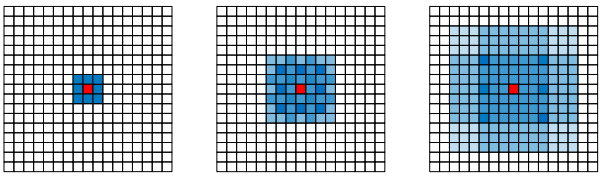


Fig. 5. Illustration of the receptive field of superposition dilated convolution with gradually rising dilation rates. (a) Receptive filed of D_1 branch. (b) Receptive filed of D_2 branch. (c) Receptive filed of D_3 branch.

features are fused to get a feature map with more comprehensive information. Then, a global average pooling is used to transform the feature map into the feature vector. Finally, the classification prediction result is obtained by the feature vector via the softmax function. In addition, as shown in Fig. 3, MDMSRB, HSMSN, and HFF are three key components of our proposed model, which will be described in detail as follow.

B. Structure of MDMSRB

Studies have demonstrated that the size of the receptive field is of great help to improve the performance of HSIC. Significantly, the fusion of multi-scale spatial information and the mining of spatial relationships of different distance pixels in the receptive field can effectively enhance the utilization of spatial information. In this article, we propose MDMSRB as the backbone of the network to achieve HSI multi-scale feature extraction. The structure of MDMSRB is shown in Fig. 4.

As can be seen from Fig.4, it is different from the aforementioned traditional multi-scale feature extraction fusion network, which is composed of four dilated convolution chains with diverse depths (D_1, D_2, D_3, D_4) in parallel, so as to realize multi-scale feature extraction. Specifically, the dilated convolutions with gradually rising dilation rates are stacked in every chain so that the receptive field is expanded. Therefore, multi-scale feature fusion is implemented by fusing the features excavate from four different depth dilated convolution chains. In addition, compared with using a large convolution kernel to expand the receptive field, under the premise of the same receptive field, the dilated convolution with increasing dilation rate consumes fewer training parameters. The formula for calculating the receptive field F of superposition dilated convolution is as follows:

$$S = K + (K-1)(r-1) \#(3)$$

$$F = \sum_{1}^{n} (S_n - 1) + 1 \#(4)$$

where K and r denote convolution kernel size and dilation rate respectively, S is the receptive field size of dilated convolution, and S_n is the receptive field size of the nth convolution kernel in one dilated convolution chain. With D_ 3 branch in Fig. 4 as an example, which can obtain a 13×13 receptive field. According to Formula (1-2), however, using the traditional 13×13 convolution kernel needs 5.26 times more training parameters and FLOPs. Therefore, MDMSRB can achieve a larger receptive field and multi-scale feature fusion with fewer parameters and FLOPs. It is worth mentioning that according to the receptive field state obtained by different depth branches given in Fig. 5, compared with Fig. 1, this dilated convolution superposition method with different dilation rates can effectively solve the gridding problem caused by dilated convolution. Furthermore, considering the problem of gradient vanishing caused by network deepening, we use skip connection to fuse the feature maps before and after multi-scale feature extraction, and an ordinary 1×1 convolution is used to keep the scale of the feature maps consistent.

C. Structure of HSMSN

In the process of image feature extraction by the CNN, the state of feature extracted at different depths of the network is gradually changing from concrete to abstract, and the scale of the feature is gradually decreasing. In this process, therefore, the adaptation of the receptive field to feature scale at different depths of the network is should be considered. In this paper, the HSMSN is proposed to adapt to the changing feature scale in the process of feature extraction.

According to the multi-scale feature extraction structure proposed in section B, which is realized by four branches with different depths. Therefore, we can adjust the scale of feature extraction by using the combination of branches with different depths in different stages of the network. Specifically, in the shallow stage of feature extraction, MDMSRB1 with four scale feature extraction branches (D_1-D_4) is used for feature extraction. Then, the combinations of feature extraction branches owned by MDMSRB2-MDMSRB4 are: D_ 1-D_3, D_ 1-D_2, and D_ 1. As shown in Fig. 3, the number and scale of feature extraction gradually decrease from MDMSRB1 to MDMSRB4. In this way, on the one hand, the adaptability of the model to changing features is improved; on the other hand, the model pruning strategy effectively improves the efficiency of model feature extraction and reduces redundant parameters.

D. Structure of HFF

As discussed in section C, in the process of feature extraction by CNN, the change of feature state will inevitably lead to the loss of low-level boundary texture features extracted in the shallow layers of the network. These low-level features have higher resolution than high-level semantic features, which is of great significance to the extraction of spatial structure features of hyperspectral images. Therefore, in order to improve the classification performance through the fusion of different level features, a hierarchical feature fusion (HFF) module is applied to HSMSN to fuse the features of different stages, especially

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JSTARS.2021.3083283, IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing

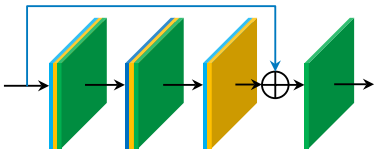


Fig. 6. Illustration of the residual block (RB).

the fusion of high and low-level features. The structure of HFF is shown in Fig. 3, which is composed of a low-level feature fusion module (LFFM) and a high-level feature fusion module (HFFM).

As can be seen from Fig. 3, hierarchical feature fusion is divided into two steps: first, the low-level and high-level features are fused respectively, and then the two features are fused. Specifically, the output feature δ_0 from the first convolution layer of HSMSN is directly connected with the output feature characteristic δ_1 from MDMSRB1 through concatenating. Subsequently, a residual block (RB) is used to fuse the low-level features to get low-level fusion feature: γ_{01}. The structure diagram of the RB is shown in Fig. 6 γ_{01} can be represented as:

$$\gamma_{01} = g(Concat(\delta_0 + \delta_1) + f(Concat(\delta_0 + \delta_1) + \omega))\#(5)$$

where $Concat()$ represents the channel connection, $f()$ represents the residual function, $g()$ represents the ReLU function, and ω represents the weight and bias coefficient of the residual correlation block. As a result, low-level features are effectively preserved. For high-level features, similar to low-level features fusion, we fused the output features δ_3 and δ_4 of MDMSRB3 and MDMSRB4 to obtain γ_{34}, which can be expressed as:

$$\gamma_{34} = g(Concat(\delta_3 + \delta_4) + f(Concat(\delta_3 + \delta_4) + \omega))\#(6)$$

Then, the output feature δ_5 of MDMSRB2 is fused with γ_{34} to obtain the final high-level semantic featureγ_{234}, which can be expressed as:

$$\gamma_{234} = g(Concat(\delta_2 + \gamma_{34}) + f(Concat(\delta_2 + \gamma_{34}) + \omega))\#(7)$$

Finally, the low-level features and high-level features are directly added to obtain high-level and low-level complementary features with a high representative. Therefore, the high-resolution feature of low-level features and the strong semantic feature of high-level features are hierarchically fused for HSIC. The effectiveness of the HFF mechanism proposed in this paper will be demonstrated by ablation experiments in part Ⅳ.

Ⅳ. Experiments and Discussion

In this section, firstly, the characteristics of data sets used will be described. Then, we discuss the details of the experimental design. Finally, the classification performances of the proposed method and some state-of-the-art methods are given and analyzed.

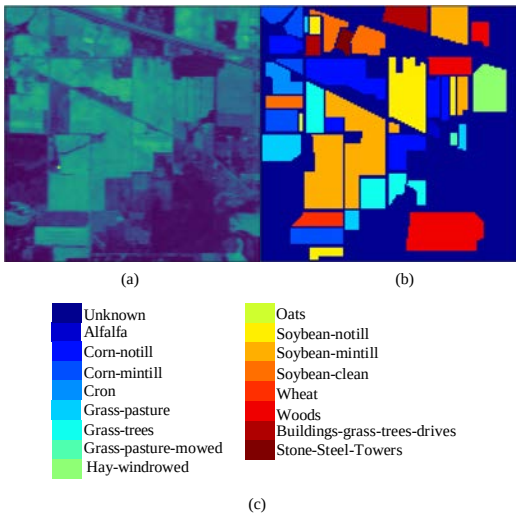


Fig. 7. Indian Pines dataset. (a) False color composite image. (b) Ground truth map. (c) Color code board.

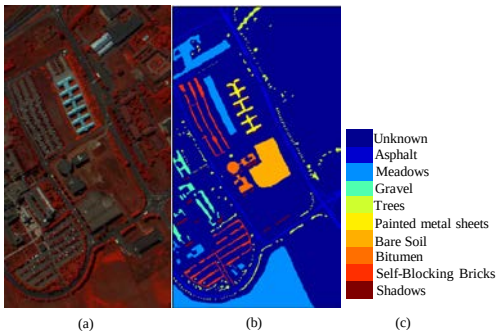


Fig. 8. Pavia University dataset. (a) False color composite image. (b) Ground truth map. (c) Color code board.

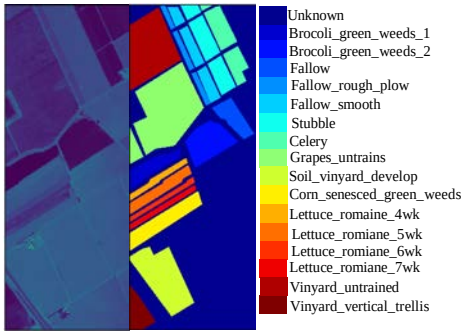


Fig. 9. Salinas dataset. (a) False color composite image. (b) Ground truth map. (c) Color code board.

A. *Data sets Descriptions*

In order to demonstrate the effectiveness of the proposed HSMSN-HFF in HSIC, three real benchmark HSI data sets are used in experiments: Indian Pines (IP), University of Pavia (PU), and Salinas (SA).

The IP image was acquired in 1992 by the AVIRIS sensor over the Indian Pines agriculture experimental area in Indiana. It covers 145×145 pixels with spatial resolution of 20 m and 220 bands across the wavelength scope of $0.4 - 2.5$ μm. There are 16 categories of land objects in the image for classification.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JSTARS.2021.3083283, IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing

TABLE I
LAND-COVER CLASSES AND NUMBERS OF SAMPLES IN THE IP DATA SET

Class	Name	Training_Num	Testing_Num
1	Alfalfa	5	41
2	Corn-notill	143	1285
3	Corn-mintill	83	747
4	Corn	24	213
5	Grass-pasture	48	435
6	Grass-t	73	657
7	Grass-p-m	3	25
8	Hay-w	48	430
9	Oats	2	18
10	Soybean-notill	97	875
11	Soybean-mintill	246	2209
12	Soybean-clean	59	534
13	Wheat	20	185
14	Woods	126	1137
15	Buildings-g-t-d	39	347
16	Stone-s-t	9	84
	Total	1025	9224

TABLE II
LAND-COVER CLASSES AND NUMBERS OF SAMPLES IN THE PU DATA SET

Class	Name	Training_Num	Testing_Num
1	Asphalt	66	6565
2	Meadows	186	18463
3	Gravel	21	2078
4	Trees	31	3033
5	Painted-M-S	14	1331
6	Bare Soil	50	4979
7	Bitumen	13	1317
8	Self-B-B	37	3645
9	Shadows	9	938
	Total	427	42349

TABLE III
LAND-COVER CLASSES AND NUMBERS OF SAMPLES IN THE SA DATA SET

Class	Name	Training_Num	Testing_Num
1	Brocoli_g_w_1	20	1989
2	Brocoli_g_w_2	37	3689
3	Fallow	20	1956
4	Fallow_r_p	14	1380
5	Fallow_s	27	2651
6	Stubble	39	3920
7	Celery	36	3543
8	Grapes_u	113	11158
9	Soil_v_d	62	6141
10	Corn_s_g_w	33	3245
11	Lettuce_r_4wk	11	1057
12	Lettuce_r_5wk	19	1908
13	Lettuce_r_6wk	9	907
14	Lettuce_r_7wk	11	1059
15	Vinyard_u	72	7196
16	Vinyard_v_t	18	1789
	Total	541	53047

Due to the water vapor contamination, 20 noise-affected bands were discarded and remained 200 bands for the experiment.

The PU image was acquired by ROSIS sensor over the University of Pavia in northwestern Italy. It has 610×340 pixels with spatial resolution of 1.3 m and 115 bands cover the spectral wavelength range from 0.43 to 0.86 μm. After removing the 12 noisy bands, the other 103 high signal-to-noise ratio bands were retained for experiments. Furthermore, it contains 42776 labeled pixels and can be divided into nine ground-truth classes.

The third data set is SA, which was captured by the AVIRIS spectral sensor at SA Valley in California. It contains 224 bands and 512×217 pixels with spatial resolution of 3.7 m. Similar to the IP data set, 20 water absorption attenuation bands were removed and remained 204 spectral bands with wavelength coverage range from 0.36 to 2.5 μm. In addition, 16 land cover types with 54129 annotated pixels are available in the SA data set.

Figs. 7-9 show the false-color composite and corresponding ground-truth maps of these thee HIS data sets. The above three data sets are divided into training sets and test sets. Moreover, considering the different sample equilibria of different data sets, we use different proportions to divide the three data sets. Specifically, randomly selected 10% labeled samples of each class in the IP data set for model training and the remaining 90% samples for testing. due to the sample number of each class in PU and SA data sets is relatively balanced, we only select 1% of the labeled samples as the training set and the rest 99% samples as the test set. Tables -Ⅲ inform the detail of the sample division of all data sets.

B. Experimental Setup

The overall accuracy (OA), average accuracy (AA), and kappa coefficient (κ) are used as the standard evaluation metrics of classification performance. Especially, for each experiment, the model will be executed ten times with randomly selected samples to get the mean as the final result. In order to fit the model more efficiently, the weights of the model are initialized, and the Adam optimizer is adopted to update the learnable parameters of the model. For the three datasets, the initial learning rate is 0.01, and to improve the learning efficiency, the learning rate decreases by 1% every 20 training epochs. The batch size for every data set is set to 64. The total training epochs are set as 200 for the IP, PU, and SA data sets, respectively. All experiments are carried with TensorFlow 2.0.0rc1 on a desktop PC with NVIDIA GeForce GTX1660 GPU and 32 GB RAM.

C. Analysis of parameters

As mentioned before, in advance of the hyperspectral image is segmented into the training set and test set, PCA is used to preprocess it to get P spectral principal components. In addition, in order to take advantage of the spatial features of HSI, the samples of the input network are neighborhood blocks with the size of $S \times S \times P$ centered on the label pixel, where $S \times S$ is the spatial size. Consequently, the number of principal components and spatial size of sample blocks are significant hyperparameters that affect the classification results. In this section, the impact of these two hyperparameters on classification performance will be elaborated.

1) Effect of P on classification performance

For the proposed HSMSN-HFF method, the PCA operation was first adapted on the original HSI to decrease the dimension of the spectral and obtain principal components. In this discussion, the size of input samples is set to $15 \times 15 \times P$,

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JSTARS.2021.3083283, IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing

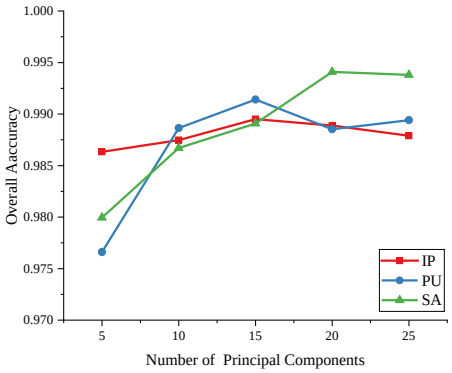


Fig. 10 Effect of *P* on overall accuracies on the three HSI data sets.

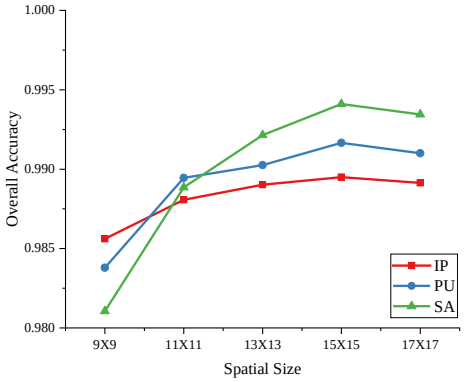


Fig. 11 Effect of Spatial Size on overall accuracies on the three HSI data sets.

TABLE Ⅳ
OAs (%) OF HSMSN-HFF WITH DIFFERENT TRAINING RATIOS

Training Ratio	5%	7%	10%	13%	15%
Indian Pines	96.43	97.69	98.95	99.11	99.39
Training Ratio	0.5%	0.7%	1%	1.3%	1.5%
Pavia University	96.11	97.83	99.16	99.33	99.34
Training Ratio	0.5%	0.7%	1%	1.3%	1.5%
Salinas	97.48	99.14	99.41	99.65	99.66

which the spatial size is fixed. Fig. 10 indicates the OAs obtained by HSMSN-HFF on three data sets according to different numbers (*P*) principal components. It can be observed that with the increase of the number of principal components, OAs will increase in all three datasets, which is due to insufficient spectral information when the principal components are small. However, when the number of principal components is too much, the OAs have a certain downward trend, which is because some unnecessary spectral components affect the classification performance. In addition, too many principal components will inevitably generate more computational and storage pressure. Therefore, the *P* is set to be 15, 15, and 20 for the IP, PU, and SA data sets, respectively.

2) Effect of spatial size on classification performance

For the CNN adopted on HSIC, the input spatial size determines how much information the neural network can obtain from hyperspectral label samples neighborhood, which has a great impact on the classification results. Therefore, in the HSMSN-HFF method, we discuss the classification performance under different spatial sizes. For the sake of fairness, the principal components P of the input samples of the three data sets all choose the optimal value of the above experiment.

Fig. 11 reports the effect of different patch sizes on the OAs of HSMSN-HFF. For the three datasets, the input spatial size is set to 9×9, 11×11, 13×13, 15×15, and 17×17. As can be observed, when the space size is small, the classification performance is relatively poor because it cannot provide enough receptive fields. Within a certain range, as the increase of spatial size, OAs of the three datasets can be greatly improved, especially for PU and SA datasets. However, a larger spatial size will introduce too much noise interference, which will impede the classification performance to a certain extent. Furthermore, considering that a larger input spatial size would also beget higher computational cost, the spatial size is set to be 15×15 for all three data sets.

D. Impact of Training Ratio

In this section, we explore the performance of the proposed HSMSN-HFF with different training ratios. The *P* and Spatial size of the input samples of three data sets are set to the optimal values discussed above. On the IP database, 5%, 7%, 10%, 13%, and 15% of the annotated pixels in each type of land-cover are randomly selected as training sets. For the PU and SA data sets, the training sets portion is set to 0.5%, 0.7%, 1%, 1.3%, and 1.5% of each land-cover category. Table Ⅳ reports the average OAs of different ratios of training samples which are conducted ten times separately. It can be observed that the proposed method can generate robust performance even under the small sample scenario. More specifically, the SA dataset shows better classification performance because of its higher spatial resolution and richer spectral information. All in all, the classification performance of the proposed method in three data sets increases with the increase of training samples.

E. Comparison Results of different Methods

In order to evaluate the classification performance of the proposed HSMSN-HFF, the performance of six state-of-the-art DL-based methods is given in this section to compare with the HSMSN-HFF. The six methods are: 3D-CNN [26], DFFN [43], MSDN [59], HybirdSN [35], MDR-CNN [55], and 2D3D-MBFF [60]. 3D-CNN is a shallow CNN model, which is constructed with two 3D Conv–pooling blocks and classified by logistic regression. HybirdSN performs 3D convolution and 2D convolution in the shallow and deep locations of the model, respectively. DFFN is a very deep CNN model which adopts the residual network to alleviate the overfitting. MSDN exploits features with a dual-direction network (vertical and horizontal) which develops with dense connection architecture. MDR-CNN uses the dilated convolutional kernel to extract multi-scale features. And 2D3D-MBFF is inspired by HybirdSN, which uses the 2D3D mixed convolution structure to form multi-scale branches for feature extraction. Among them, MDR-CNN, 2D3D-MBFF, and our proposed method are multi-scale frameworks.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JSTARS.2021.3083283, IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing

TABLE V
CLASSIFICATION ACCURACIES OF DIFFERENT METHODS FOR THE IP DATA SET (10% SAMPLES FOR TRAINING)

Class	Methods						
	3D-CNN	DFFN	MSDN	HybirdSN	MDR-CNN	2D3D-MBFF	HSMSN-HFF
1	**100±0.00**	98.12±1.41	85.71±3.27	78.66±2.34	88.10±2.49	95.24±0.67	**100±0.00**
2	91.49±1.91	97.73±1.56	95.17±2.75	95.16±1.62	96.18±1.61	97.51±0.16	**99.23±0.15**
3	95.66±0.94	97.24±1.75	82.28±1.79	97.93±0.86	99.21±0.09	99.21±0.04	**99.73±0.07**
4	64.35±0.37	**100±0.00**	90.28±1.46	93.31±2.57	**100±0.00**	97.69±1.06	98.60±1.04
5	84.86±1.29	92.31±3.45	96.18±0.37	99.54±0.07	97.08±1.22	**100±0.00**	**100±0.00**
6	96.19±0.35	96.4±2.34	99.39±0.09	99.85±0.19	99.39±0.07	99.39±0.03	**100±0.00**
7	**100±0.00**	100±0.00	50.00±7.57	96.00±2.23	76.92±3.79	92.31±0.39	88.46±2.33
8	**100±0.00**	99.61±0.26	**100±0.00**	**100±0.00**	**100±0.00**	99.31±0.12	**100±0.00**
9	**100±0.00**	72.3±5.37	0±5.31	84.72±1.47	**100±0.00**	94.44±1.15	**100±0.00**
10	93.64±0.70	97.51±0.32	97.70±0.87	98.20±0.31	**99.20±0.32**	96.78±0.66	98.86±0.7
11	94.84±2.28	98.70±1.14	98.09±0.96	98.70±0.65	98.82±0.57	98.50±0.27	**99.50±0.03**
12	71.18±1.85	94.34±2.39	82.82±3.56	95.46±1.22	92.28±1.47	**95.95±0.87**	93.98±0.79
13	94.97±0.20	94.16±1.67	**100±0.00**	99.73±0.04	**100±0.00**	**100±0.00**	**100±0.00**
14	99.27±1.33	**100±0.00**	99.30±0.07	99.89±0.03	98.87±0.33	**100±0.00**	99.91±0.04
15	97.33±0.99	97.66±0.22	94.83±0.76	98.20±0.17	**100±0.00**	**100±0.00**	**100±0.00**
16	81.32±1.90	97.62±0.43	98.85±0.52	**99.11±0.09**	98.85±0.02	97.7±0.91	97.65±1.37
OA(%)	92.67±0.36	95.85±0.19	95.08±0.39	97.99±0.19	98.15±0.2	98.41±0.13	**98.95±0.16**
AA(%)	91.56±0.49	97.63±0.24	85.66±1.87	95.90±1.04	96.55±0.34	97.75±0.54	**98.49±0.24**
Kappa×100	91.63±0.53	97.31±0.27	94.40±0.53	97.71±0.22	97.90±0.29	98.28±0.39	**98.81±0.19**

TABLE VI
CLASSIFICATION ACCURACIES OF DIFFERENT METHODS FOR THE PU DATA SET (1% SAMPLES FOR TRAINING)

Class	Methods						
	3D-CNN	DFFN	MSDN	HybirdSN	MDR-CNN	2D3D-MBFF	HSMSN-HFF
1	99.97±0.03	96.29±2.74	97.74±039	98.45±0.66	99.02±0.06	**99.44±0.67**	99.21±0.06
2	98.23±1.21	99.81±0.07	99.86±0.06	99.86±0.07	99.98±0.04	99.77±0.29	**99.94±0.07**
3	96.25±0.79	95.39±0.29	80.49±1.35	89.74±1.09	84.79±0.57	91.00±1.33	**96.64±0.33**
4	**96.73±0.49**	94.61±0.69	92.02±0.79	92.20±0.73	96.64±0.37	81.02±0.92	95.14±0.23
5	99.62±0.27	99.25±0.09	**100±0.00**	**100±0.00**	99.62±0.05	90.30±1.43	99.92±0.04
6	95.36±0.38	**100±0.00**	88.88±0.43	96.83±0.63	98.67±0.67	**100±0.00**	**100±0.00**
7	95.96±0.86	88.65±1.39	75.33±1.31	97.11±0.49	99.92±0.07	97.95±1.09	**100±0.00**
8	66.11±1.22	95.20±0.79	79.40±0.64	79.08±0.37	92.44±1.08	93.07±2.03	**98.14±0.11**
9	97.54±0.37	99.36±0.05	97.12±0.37	98.08±0.51	96.26±0.59	81.73±1.69	**99.57±0.24**
OA(%)	95.14±0.39	97.93±0.36	94.15±0.59	96.33±0.50	97.94±0.26	96.64±0.95	**99.16±0.07**
AA(%)	93.97±0.47	96.50±0.27	90.09±0.37	94.59±0.70	96.37±0.23	92.69±1.34	**98.72±0.15**
Kappa×100	93.58±0.61	97.25±0.24	92.17±0.80	95.11±0.67	97.27±0.18	95.54±0.36	**98.89±0.11**

TABLE VII
CLASSIFICATION ACCURACIES OF DIFFERENT METHODS FOR THE SA DATA SET (1% SAMPLES FOR TRAINING)

Class	Methods						
	3D-CNN	DFFN	MSDN	HybirdSN	MDR-CNN	2D3D-MBFF	HSMSN-HFF
1	**100±0.00**	95.45±0.22	**100±0.00**	**100±0.00**	99.95±0.05	98.50±0.17	**100±0.00**
2	**100±0.00**	**100±0.00**	**100±0.00**	**100±0.00**	**100±0.00**	**100±0.00**	**100±0.00**
3	75.36±0.66	97.30±0.47	**100±0.00**	**100±0.00**	98.78±0.12	**100±0.00**	**100±0.00**
4	92.91±1.07	**100±0.00**	99.59±0.04	98.58±0.07	99.78±0.07	95.16±0.76	99.13±0.26
5	**100±0.00**	90.95±0.45	96.97±0.19	99.80±0.05	99.96±0.01	97.93±0.33	99.55±0.33
6	99.97±0.06	**100±0.00**	**100±0.00**	99.90±0.03	99.92±0.04	99.41±0.21	99.95±0.02
7	99.89±0.04	99.75±0.07	**100±0.00**	99.99±0.07	99.94±0.03	98.03±0.31	**100±0.00**
8	98.84±0.39	97.02±1.04	95.40±0.47	97.14±0.36	96.19±0.37	**99.64±0.17**	98.66±0.37
9	99.87±0.09	**100±0.00**	**100±0.00**	**100±0.00**	**100±0.00**	**100±0.00**	**100±0.00**
10	97.93±0.78	**99.44**±0.23	99.31±0.67	98.96±0.22	99.17±0.17	98.89±0.31	98.15±0.54
11	96.60±0.00	97.36±0.31	99.72±0.01	99.31±0.34	96.04±0.77	99.43±0.34	**99.91±0.07**
12	**100±0.00**	**100±0.00**	**100±0.00**	99.86±0.02	**99.89±0.0**5	99.84±0.05	**100±0.00**
13	99.89±0.22	91.61±0.62	96.88±0.08	**99.93±0.01**	99.23±0.17	98.57±0.09	**100±0.00**
14	87.71±1.34	95.10±0.53	**99.09**±0.06	97.79±0.21	86.71±0.35	94.25±0.19	96.98±1.13
15	57.23±1.22	91.47±0.67	91.20±0.12	93.03±0.13	98.44±0.15	**100±0.00**	99.43±0.23
16	**99.72±0.02**	98.26±0.14	99.37±0.05	99.46±0.03	95.57±0.67	97.87±0.21	99.55±0.23
OA(%)	92.46±0.33	97.12±0.29	97.60±0.16	98.28±0.12	98.37±0.14	98.92±0.16	**99.41±0.14**
AA(%)	94.12±0.27	97.10±0.39	98.59±0.07	99.00±0.05	98.09±0.15	98.59±0.22	**99.45±0.21**
Kappa×100	91.57±0.18	96.79±0.25	97.28±0.18	98.08±0.13	98.19±0.32	98.86±0.17	**99.34±0.19**

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JSTARS.2021.3083283, IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing

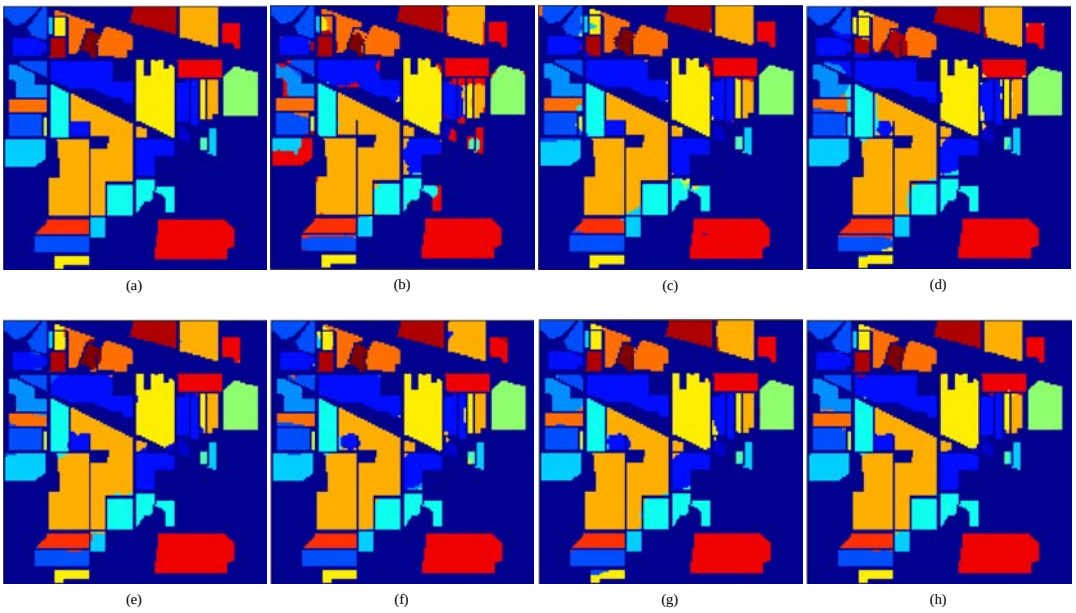


Fig. 12. Classification maps for IP. (a). Ground truth (b)-(h) Predicted classification maps for 3D-CNN (OA=92.67%), DFFN (OA=95.85%), MSDN(OA=95.08%), HybirdSN (OA=97.99%), MDR-CNN (98.15%), 2D3D-MBFF (OA=98.41%), and proposed HSMSN-HFF (98.95%).

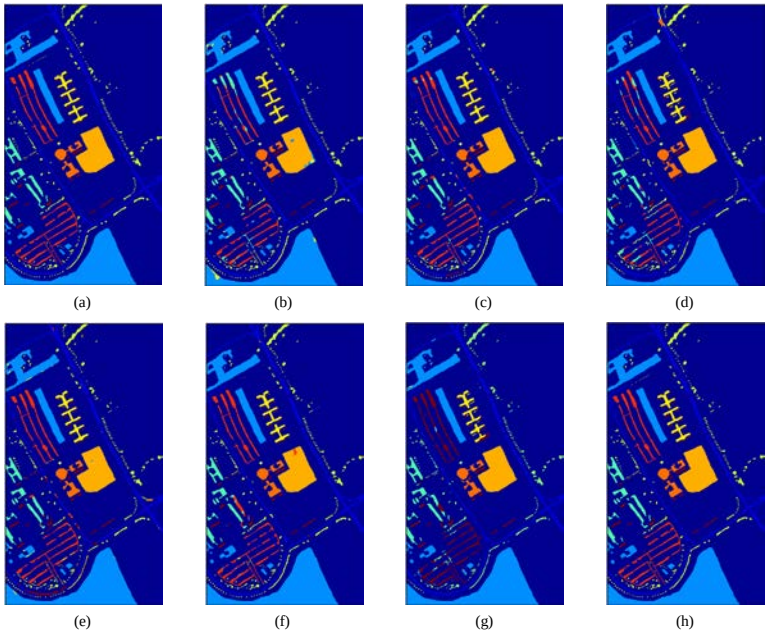
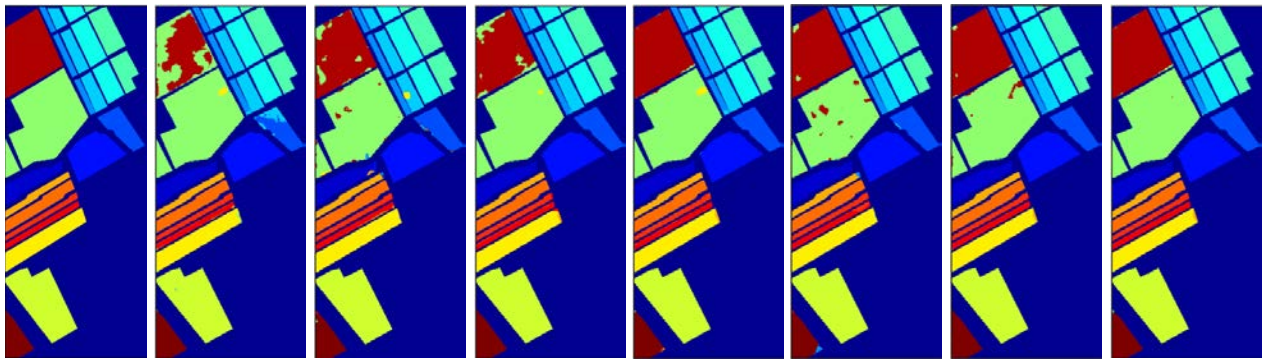


Fig. 13. Classification maps for PU. (a). Ground truth (b)-(h) Predicted classification maps for 3D-CNN (OA=97.93%), DFFN (OA=97.93%), MSDN(OA=94.15%), HybirdSN (OA=96.33%), MDR-CNN 97.94%), 2D3D-MBFF (OA=96.64%), and proposed HSMSN-HFF (99.16%).

To ensure fairness, some parameters affecting feature extraction (such as the number of principal components) in these methods for comparison are consistent with the setting of corresponding references. Besides, other hyperparameters may be affected by different experimental platforms (such as the batch size and learning rate) are adjusted to obtain the optimal result of each method. For the training sample size of the model, the limited training sample can more effectively reflect the performance of each model. Therefore, 10% (IP), 1% (PU), 1% (SA) samples are randomly selected for the training of all models. The classification results for each class and overall evaluation indicators obtained by different methods are reported in Tables -WII, respectively. Statistically, the classification results on three benchmark data sets substantiate that the proposed HSMSN-HFF outperforms other methods. From these tables, it is can be easily discovered that the performances of 3D-CNN are much lower than other methods, which is due to the shallow architecture cannot fully extract features, especially high-level features. Different from 3D-CNN, DFFN and MSDN have very deep structures, which can obtain stronger discriminative semantic features for better classification than 3D-CNN. In addition, the direct connection is widely used for feature reuse in these two methods. As for HybirdSN, 3D and 2D convolution hierarchically exploit

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JSTARS.2021.3083283, IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing



(a) (b) (c) (d) (e) (f) (g) (h)

Fig. 14. Classification maps for SA. (a). Ground truth (b)-(h) Predicted classification maps for 3D-CNN (OA=92.46%), DFFN (OA=97.12%), MSDN(OA=97.60%), HybirdSN (OA=98.28%), MDR-CNN(OA=98.37%), 2D3D-MBFF (OA=98.92%), and proposed HSMSN-HFF (99.41%).

TABLE Ⅷ
TRAINABLE PARAMETERS AND FLOPS OF DIFFERENT MODELS FOR THE SA DATA SET

Methods	3D-CNN	DFFN	MSDN	HybirdSN	MDR-CNN	2D3D-MBFF	HSMSN-HFF
TTPs	9,867,216	370,096	1,535,044	5,122,176	440,280	940,440	165,664
FLOPs (Million)	12,035,326M	2,695M	3,874M	15,942M	2,767M	28,566M	2,282M

TABLE Ⅸ
TRAINING AND TESTING TIMES (IN SECONDS) OF DIFFERENT MODELS FOR THE SA DATA SET

Methods	Train time	Test time	OA
3D-CNN	3987.6	20.37	92.46%
DFFN	3128.44	15.34	97.12%
MSDN	189.32	17.22	97.60%
HybirdSN	152.9	3.72	98.28%
MDR-CNN	46.71	6.82	98.37%
2D3D-MBFF	156.91	33	98.92%
HSMSN-HFF	69.5	13.2	99.41%

spectral and spatial features, which is simple and effective. Considering that multi-scale features are informative for classification, the MDR-CNN and 2D3D-MBFF use dilated convolution kernels and multi-scale convolution kernels to extract more plenteous features. Subsequently, the multi-scale extraction methods achieve OAs 98.15% and 98.41%, with the gains of 6.59% and 6.9%, 2.3% and 2.61, 3.07% and 3.38%, 0.16% and 0.47% over the 3D-CNN, DFFN, MSDN, and HybirdSN method in IP data set, respectively. Furthermore, by comparing the classification performances of the proposed HSMSN-HFF with two other multi-scale methods MDR-CNN and 2D3D-MBFF, it can be noticed that our proposed method also shows superior performance: the mean OAs of the HSMSN-HFF is 0.8%, 1.22%, and 1.04% higher than that of the MDR-CNN, and 0.54%, 2.52%, and 0.51% higher than that of the 2D3D-MBFF. Especially, both the proposed method and MDR-CNN method use dilated convolution as a multi-scale feature extraction tool, but compared with MDR-CNN, the HSMSN-HFF effectively solves the gridding problem caused by dilated convolution. Therefore, the proposed method can achieve better classification results.

In addition to the quantitative classification results report, we visualize the classification maps corresponding to the results reported in Tables Ⅴ–Ⅶ. The classification maps of different methods discussed above are presented in Fig. 12-14. Obviously, as can be observed that the 3D-CNN results in the most misclassified pixels in all classification maps. Furthermore, the deep model and the multi-scale model can improve the classification performance effectively and generate smoother classification maps. In addition, an obvious observation that the classification map of the proposed method is the closest to the reference ground truth, which produces less internal noise and a cleaner boundary.

Moreover, to further evaluate the computational efficiency of the proposed method, Table Ⅷ reports the total trainable parameters (TTPs) and FLOPs for the SA data set for different models. As can be seen, due to the extensive use of dilated convolution to achieve multi-scale feature extraction, the

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JSTARS.2021.3083283, IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing

MDR-CNN and the proposed HSMSN-HFF produce fewer training parameters and floating-point operations than other methods. Then, in order to directly reflect the computational efficiency of the proposed algorithm, Table Ⅸ shows the elapsed time of training and testing of each method. As listed in the table, compared with 3D-CNN, DFFN, MSDN, and 2D3D-MBFF, the MDR-CNN and the proposed HSMSN-HFF need shorter training and testing time. It confirms that the effectiveness of dilated convolution in improving the classification efficiency of hyperspectral images. Compared with MDR-CNN, the proposed method needs to consume longer training and testing time, but the proposed method results in better performance of classification. It can be concluded that the proposed algorithm method is not only competitive on the part of accuracy but also computational cost relative to the state-of-the-art methods.

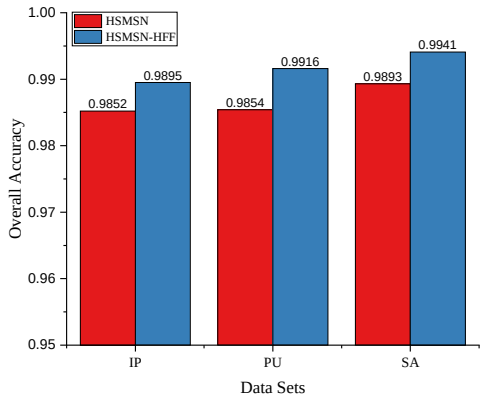


Fig. 15. Effect of HFF module on overall accuracies on the three HSI data

F. *Ablation Study*

In order to verify that the proposed hierarchical feature fusion method is productive in improving the classification performance of hyperspectral images, ablation experiments are applied to three datasets. The model used as a comparison is coherent with the model structure of the proposed HSNSN-HFF method except for removing the HFF module to be validated from the original network. The input size of all experimental samples was set to the optimal value of the above analysis, and the experiment was repeated ten times to take the average classification results.

Fig. 15 shows the classification performances achieved by the proposed HSNSN-HFF and HSNSN. It can be observed that the classification results using the HFF module model in all data sets are more excellent than those without. The main reason for this result is that the use of an HFF module can make full use of the complementary information of high and low-level information generated by the neural network in different stages of the model. It can be concluded that high-level semantic features and low-level texture features can produce more comprehensive features to achieve more precise classification through the hierarchical fusion strategy proposed in this paper.

Ⅴ. Conclusion

In this article, a novel hierarchical shrinkage multi-scale network for hyperspectral image classification with hierarchical feature fusion has been proposed. Specifically, we design a multi-scale feature extraction block MDMSRB by superimposing dilated convolution, in which the dilation rate of dilated convolution of each branch increases gradually and the depth is different. In this way, the multi-scale features can be extracted effectively with a lower computational cost. Moreover, we construct the HSMSN based on the MDMSRB with a hierarchical shrinkage architecture, which can not only achieve multi-scale feature extraction in different stages of the network but also mitigate the model structure. In addition, we introduce an HFF strategy into the HSMSN to fuse the low-level edge information and high-level semantic information to boost the description and representation of the feature map. Experimental results on three benchmark HSI data sets demonstrate that the proposed HSMSN-HFF outperforms several state-the-of-art methods for both classification accuracies and computational efficiency.

References

[1] M. Paoletti, J. Haut, J. Plaza, and A. Plaza, “Deep learning classifiers for hyperspectral imaging: A review,” ISPRS J. Photogram. Remote Sens., vol. 158, pp. 279–317, Dec. 2019.

[2] S. Li, W. Song, L. Fang, Y. Chen, P. Ghamisi and J. A. Benediktsson, "Deep Learning for Hyperspectral Image Classification: An Overview," in IEEE Transactions on Geoscience and Remote Sensing, vol. 57, no. 9, pp. 6690-6709, Sept. 2019, doi: 10.1109/TGRS.2019.2907932.

[3] M. Teke, H. S. Deveci, O. Halilolu, S. Z. Gürbüz and U. Sakarya, "A short survey of hyperspectral remote sensing applications in agriculture," 2013

6th International Conference on Recent Advances in Space Technologies (RAST), Istanbul, 2013, pp. 171-176, doi: 10.1109/RAST.2013.6581194.

[4] L. Olmanson, P. Brezonik, and M. Bauer,"Airborne hyperspectral remote sensing to assess spatial distribution of water quality characteristics in large rivers: The Mississippi River and its tributaries in Minnesota,"in Remote Sensing of Environment, Vol. 130, pp. 254-265, Mar. 2013, dio: 10.1016/j.rse.2012.11.023.

[5] I. C. C. Acosta, M. Khodadadzadeh, L. Tusa, P. Ghamisi and R. Gloaguen, "A Machine Learning Framework for Drill-Core Mineral Mapping Using Hyperspectral and High-Resolution Mineralogical Data Fusion," in IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, vol. 12, no. 12, pp. 4829-4842, Dec. 2019, doi: 10.1109/JSTARS.2019.2924292.

[6] X. Z. Shi, M. Aspandiar, and D. Oldmeadow, "Using hyperspectral data and PLSR modelling to assess acid sulphate soil in subsurface," *J. Soils Sediments*, vol. 14, no. 5, pp. 904–916, 2014.

[7] S. Veraverbeke et al., "Hyperspectral remote sensing of fire: State-of-the-art and future perspectives," Remote Sens. Environ., vol. 216, pp. 105–121, Oct. 2018.

[8] J. Ardouin, J. Levesque and T. A. Rea, "A demonstration of hyperspectral image exploitation for military applications," 2007 10th International Conference on Information Fusion, Quebec, Que., 2007, pp. 1-8, doi: 10.1109/ICIF.2007.4408184.

[9] B. Rasti et al., "Feature Extraction for Hyperspectral Imagery: The Evolution from Shallow to Deep (Overview and Toolbox)," in IEEE Geoscience and Remote Sensing Magazine, doi: 10.1109/MGRS.2020.2979764.

[10] D. Hong, L. Gao, J. Yao, B. Zhang, A. Plaza, and J. Chanussot. Graph Convolutional Networks for Hyperspectral Image Classification. IEEE Transactions on Geoscience and Remote Sensing, 2020, DOI: 10.1109/TGRS.2020.3015157.

[11] D. Hong, X. Wu, P. Ghamisi, J. Chanussot, N. Yokoya, and X. X. Zhu. Invariant Attribute Profiles: A Spatial-Frequency Joint Feature Extractor for Hyperspectral Image Classification. IEEE Transactions on Geoscience and Remote Sensing, 2020, 58(6): 3791-3808.

[12] L. Gao et al., "Subspace-based support vector machines for hyperspectral image classification," IEEE Geosci. Remote Sens. Lett., vol. 12, no. 2, pp. 349–353, Feb. 2015.

[13] Y. Bazi and F. Melgani, "Gaussian Process Approach to Remote Sensing Image Classification," in IEEE Transactions on Geoscience and Remote Sensing, vol. 48, no. 1, pp. 186-197, Jan. 2010, doi: 10.1109/TGRS.2009.2023983.

[14] Y. Chen, Z. Lin and X. Zhao, "Riemannian manifold learning based k-nearest-neighbor for hyperspectral image classification," 2013 IEEE International Geoscience and Remote Sensing Symposium - IGARSS, Melbourne, VIC, Australia, 2013, pp. 1975-1978, doi: 10.1109/IGARSS.2013.6723195.

[15] G. Hughes, "On the mean accuracy of statistical pattern recognizers," IEEE Trans. Inf. Theory, vol. IT-14, no. 1, pp. 55–63, Jan. 1968.

[16] D. Hong, N. Yokoya, J. Chanussot, J. Xu, and X. X. Zhu, "Learning to propagate labels on graphs: An iterative multitask regression framework for semi-supervised hyperspectral dimensionality reduction," ISPRS J. Photogramm. Remote Sens., vol. 158, pp. 35–49, Dec. 2019.

[17] F. Cao, Z. Yang, J. Ren, W. Chen, G. Han, and Y. Shen, "Local block multilayer sparse extreme learning machine for effective feature extraction and classification of hyperspectral images," IEEE Trans. Geosci. Remote Sens., vol. 57, no. 8, pp. 5580–5594, Aug. 2019.

[18] 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.

[19] D.Hong, N. Yokoya, J. Chanussot and X. X. Zhu, "CoSpace: Common Subspace Learning From Hyperspectral-Multispectral Correspondences," in IEEE Transactions on Geoscience and Remote Sensing, vol. 57, no. 7, pp. 4349-4359, July 2019, doi: 10.1109/TGRS.2018.2890705.

[20] M. Fauvel, Y. Tarabalka, J. A. Benediktsson, J. Chanussot, and J. C. Tilton, "Advances in spectral spatial classification of hyperspectral images," Proc. IEEE, vol. 101, no. 3, pp. 652–675, Mar. 2013.

[21] L. Fang, S. Li, W. Duan, J. Ren, and J. A. Benediktsson, "Classification of hyperspectral images by exploiting spectral–spatial information of superpixel via multiple kernels," IEEE Trans. Geosci. Remote Sens., vol. 53, no. 12, pp. 6663–6674, Dec. 2015.

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

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

[24] P. Zhong, Z. Gong, S. Li, and C.-B. Schönlieb, "Learning to diversify deep belief networks for hyperspectral image classification," IEEE Trans. Geosci. Remote Sens., vol. 55, no. 6, pp. 3516–3530, Jun. 2017.

[25] C. Zhao, X. Wan, G. Zhao, B. Cui, W. Liu, and B. Qi, "Spectral–spatial classification of hyperspectral imagery based on stacked sparse autoencoder and random forest," Eur. J. Remote Sens., vol. 50, no. 1, pp. 47–63, Jan. 2017.

[26] Y. Chen, H. Jiang, C. Li, X. Jia and P. Ghamisi, "Deep Feature Extraction and Classification of Hyperspectral Images Based on Convolutional Neural Networks," in IEEE Transactions on Geoscience and Remote

Sensing, vol. 54, no. 10, pp. 6232-6251, Oct. 2016, doi: 10.1109/TGRS.2016.2584107.

[27] H. Yu et al., "Global Spatial and Local Spectral Similarity-Based Manifold Learning Group Sparse Representation for Hyperspectral Imagery Classification," in IEEE Transactions on Geoscience and Remote Sensing, vol. 58, no. 5, pp. 3043-3056, May 2020, doi: 10.1109/TGRS.2019.2947032.

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

[29] J. Wang, X. Song, L. Sun, W. Huang and J. Wang, "A Novel Cubic Convolutional Neural Network for Hyperspectral Image Classification," in IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, vol. 13, pp. 4133-4148, 2020, doi: 10.1109/JSTARS.2020.3008949.

[30] D. Hong et al., "More diverse means better: Multimodal deep learning meets remote-sensing imagery classification," IEEE Trans. Geosci. Remote Sens., to be published, doi: 10.1109/TGRS.2020.3016820.

[31] D. Hong, N. Yokoya, G. Xia, J. Chanussot, and X. Zhu, "X-ModalNet: A semi-supervised deep cross-modal network for classification of remote sensing data," ISPRS J. Photogramm. Remote Sens., vol. 167, pp. 12–23, 2020.

[32] W. Li, G. Wu, F. Zhang and Q. Du, "Hyperspectral Image Classification Using Deep Pixel-Pair Features," in IEEE Transactions on Geoscience and Remote Sensing, vol. 55, no. 2, pp. 844-853, Feb. 2017, doi: 10.1109/TGRS.2016.2616355.

[33] 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," in IEEE Transactions on Geoscience and Remote Sensing, vol. 57, no. 2, pp. 740-754, Feb. 2019, doi: 10.1109/TGRS.2018.2860125.

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

[35] S. K. Roy, G. Krishna, S. R. Dubey and B. B. Chaudhuri, "HybridSN: Exploring 3-D–2-D CNN Feature Hierarchy for Hyperspectral Image Classification," in IEEE Geoscience and Remote Sensing Letters, vol. 17, no. 2, pp. 277-281, Feb. 2020, doi: 10.1109/LGRS.2019.2918719.

[36] C. Yu, R. Han, M. Song, C. Liu and C. Chang, "A Simplified 2D-3D CNN Architecture for Hyperspectral Image Classification Based on Spatial–Spectral Fusion," in IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, vol. 13, pp. 2485-2501, 2020, doi: 10.1109/JSTARS.2020.2983224.

[37] X. Yang, et al, "Synergistic 2D/3D Convolutional Neural Network for Hyperspectral Image Classification," in Remote Sens., vol. 12, no. 12, pp. 2033, Jun. 2020.

[38] R.K. Srivastava, K. Greff and J. Schmidhuber, 2015. "Training very deep networks, " in Advances in Neural Information Processing Systems 28. Curran Associates, pp. 2377–2385.

[39] G. Licciardi, F. Del Frate, and R. Duca, "Feature reduction of hyper-spectral data using autoassociative neural networks algorithms," in Proc. IEEE Int. Geosci. Remote Sens. Symp., vol. 1, Jul. 2009, pp. 176–179.

[40] K. He, X. Zhang, S. Ren and J. Sun, "Deep Residual Learning for Image Recognition," 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, 2016, pp. 770-778, doi: 10.1109/CVPR.2016.90.

[41] Z. Zhong, J. Li, Z. Luo and M. Chapman, "Spectral–Spatial Residual Network for Hyperspectral Image Classification: A 3-D Deep Learning Framework," in IEEE Transactions on Geoscience and Remote Sensing, vol. 56, no. 2, pp. 847-858, Feb. 2018, doi: 10.1109/TGRS.2017.2755542.

[42] W. Song, S. Li, L. Fang and T. Lu, "Hyperspectral Image Classification with Deep Feature Fusion Network," in IEEE Transactions on Geoscience and Remote Sensing, vol. 56, no. 6, pp. 3173-3184, June 2018, doi: 10.1109/TGRS.2018.2794326.

[43] G. Huang, Z. Liu, L. Van Der Maaten and K. Q. Weinberger, "Densely Connected Convolutional Networks," 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 2017, pp. 2261-2269, doi: 10.1109/CVPR.2017.243.

[44] Z. Li et al., "Deep Multilayer Fusion Dense Network for Hyperspectral Image Classification," in IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, vol. 13, pp. 1258-1270, 2020, doi: 10.1109/JSTARS.2020.2982614.

[45] W. Zhao and S. Du, "Learning multiscale and deep representations for classifying remotely sensed imagery," ISPRS J. Photogramm. Remote Sens., vol. 113, pp. 155–165, Mar. 2016.

[46] H. Lee and H. Kwon, "Going Deeper with Contextual CNN for Hyperspectral Image Classification," in IEEE Transactions on Image Processing, vol. 26, no. 10, pp. 4843-4855, Oct. 2017, doi: 10.1109/TIP.2017.2725580.

[47] Z. Gong, P. Zhong, Y. Yu, W. Hu and S. Li, "A CNN With Multiscale Convolution and Diversified Metric for Hyperspectral Image Classification," in IEEE Transactions on Geoscience and Remote Sensing, vol. 57, no. 6, pp. 3599-3618, June 2019, doi: 10.1109/TGRS.2018.2886022.

[48] M. Zhang, W. Li and Q. Du, "Diverse Region-Based CNN for Hyperspectral Image Classification," in IEEE Transactions on Image

Processing, vol. 27, no. 6, pp. 2623-2634, June 2018, doi: 10.1109/TIP.2018.2809606.

[49] X. Li, M. Ding and A. Pižurica, "Deep Feature Fusion via Two-Stream Convolutional Neural Network for Hyperspectral Image Classification," in IEEE Transactions on Geoscience and Remote Sensing, vol. 58, no. 4, pp. 2615-2629, April 2020, doi: 10.1109/TGRS.2019.2952758.

[50] X. Li, D. Song and Y. Dong, "Hierarchical Feature Fusion Network for Salient Object Detection," in IEEE Transactions on Image Processing, vol. 29, pp. 9165-9175, 2020, doi: 10.1109/TIP.2020.3023774.

[51] D. Hong, N. Yokoya, N. Ge, J. Chanussot, and X. X. Zhu, “Learnable manifold alignment (LeMA): A semi-supervised cross-modality learning framework for land cover and land use classification, ISPRS J. Photogramm. Remote Sens., vol. 147, pp. 193–205, Jan. 2019.

[52] P. Wang et al., "Understanding Convolution for Semantic Segmentation," 2018 IEEE Winter Conference on Applications of Computer Vision (WACV), Lake Tahoe, NV, 2018, pp. 1451-1460, doi: 10.1109/WACV.2018.00163.

[53] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, “Semantic image segmentation with deep convolutional nets and fully connected CRFs,” 2014, arXiv:1412.7062. [Online]. Available: http://arxiv.org/abs/1412.7062

[54] K. Pooja, R. R. Nidamanuri and D. Mishra, "Multi-Scale Dilated Residual Convolutional Neural Network for Hyperspectral Image Classification," 2019 10th Workshop on Hyperspectral Imaging and Signal Processing: Evolution in Remote Sensing (WHISPERS), Amsterdam, Netherlands, 2019, pp. 1-5, doi: 10.1109/WHISPERS.2019.8921284.

[55] B. Pan, X. Xu, Z. Shi, N. Zhang, H. Luo and X. Lan, "DSSNet: A Simple Dilated Semantic Segmentation Network for Hyperspectral Imagery Classification," in IEEE Geoscience and Remote Sensing Letters, vol. 17, no. 11, pp. 1968-1972, Nov. 2020, doi: 10.1109/LGRS.2019.2960528.

[56] Y. Li, Y. Chen, N. Wang, and Z.-X. Zhang, “Scale-aware trident networks for object detection,” in Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV), Oct. 2019, pp. 6054–6063.

[57] X. Zhu, H. Hu, S. Lin, and J. Dai, “Deformable ConvNets v2: More deformable, better results,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2019, pp. 9308–9316.

[58] Z. Liao and G. Carneiro, “Competitive multi-scale convolution,” 2015, arXiv:1511.05635. [Online]. Available: http://arxiv.org/abs/1511.05635

[59] C. Zhang, G. Li and, S. Du, "Multi-Scale Dense Networks for Hyperspectral Remote Sensing Image Classification," IEEE Trans. Geosci. Remote Sens., vol. 57, no. 11, pp. 9201-9222, Nov. 2019.

[60] Z. Ge, G. Cao, X. Li and P. Fu, "Hyperspectral Image Classification Method Based on 2D–3D CNN and Multibranch Feature Fusion," in IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, vol. 13, pp. 5776-5788, 2020, doi: 10.1109/JSTARS.2020.3024841.

Hongmin Gao (M’21) received the Ph.D. degree in computer application technology from Hohai University, Nanjing, China, in 2014.

He is currently a Professor with the College of Computer and Information, Hohai University. His research interests include deep learning, information fusion, and image processing in remote sensing.

Zhonghao Chen (S’21) received the B.S. degree in electronics and information engineering from West Anhui University, Luan, China, in 2019.

He is a Graduate Student with the College of Computer and Information, Hohai University. His research interests include deep learning and image processing.

Chenming Li received the B.S., M.S., and Ph.D. degrees in computer application technology from Hohai University, Nanjing, China, in 1993, 2003, and 2010, respectively.

He is a Professor and the Deputy Dean of the College of Computer and Information, Hohai University. His research interests include information processing systems and applications, system modeling and simulation, multisensor systems, and information processing.

Dr. Li is a Senior Member of the China Computer Federation and the Chinese Institute of Electronics.