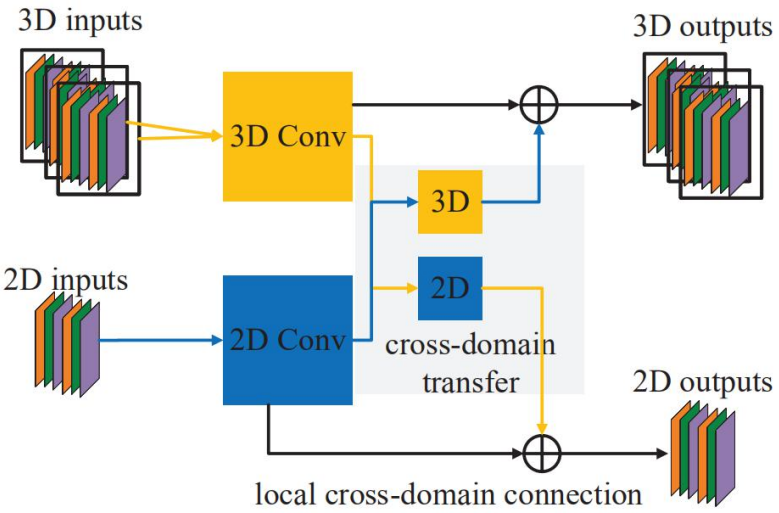

Figure 1 The framework of the proposed H-CNN which contains 2D CNN, 3D CNN and the data interaction module for hyper-spectral image classification

We propose the H-CNN network to classify hyperspectral images. In the pro- posed H-CNN, the 2D CNN component and the 3D CNN component are mixed together. Different from the conventional 3D CNNs that stack up 3D convolution layer by layer, the proposed H-CNN, as shown in Fig. 1, integrates 3D CNNs with 2D CNNs to learn the salient features. Then, a data interaction module is proposed which fuses the 2D features and 3D features together. Experimental results on one benchmark dataset have demonstrated that the proposed H-CNN outperform several state-of-the-art hyperspectral image classifiers. Our major research contributions are summarized as follows.

(1). We design an end-to-end hyperspectral image classification framework - H- CNN by integrating data interaction module into the mixture model to generate richer feature map.

(2). Rigorous experiments have been conducted on one dataset and the promising results demonstrate that the proposed H-CNN is superior to other state-of- the-art hyperspectral image classifiers.

2. RELATED WORK

Referring to existing literature, hyperspectral image classification is a very com- mon research problem. However, previous work mainly explored conventional computational methods. In this section, we briefly review the latest deep learning- based approaches which are roughly categorized into 2D CNN and 3D CNN based approaches in this paper.

2D CNN-based approaches. To extract the spectral-spatial information contained in hyperspectral images, a 2D CNN-based approach was proposed in [13] where 2D CNN was utilized to explore *the band* selection results generated by the AdaBootst SVM [13]. Based on the band selection results, several methods were proposed to fuse some 2D CNN networks for hyperspectral images classification [11]. Furthermore, [10] proposed a semi-supervised 2D CNN model consisted of the original encoder, the corrupted encoder, and the decoder. A deep 2D CNN model was designed [19] to label each pixel in the hyperspectral image. Some attempts were made to adapt the 2D CNN model for action recognition in video data, and this could be considered as 3D image data to some extent. Previous works also examined a two-stream like framework and each of which was a pre-trained 2D CNN model [2,16]. The 2D CNN-based approaches demonstrated their superior performance but a large training dataset was a prerequisite. Apparently, this is a serious limitation for most hyperspectral image classification applications because large training datasets were usually unavailable.

3D CNN-based approaches. In [7], the 3D CNN approach was first pro- posed to learn discriminative features for action recognition on spatial-temporal datasets. [1] proposed a deep 3D CNN network which stacked up 3D convolutional layers to extract spectral-spatial feature maps for classification. Similarly, a deep fully convolutional network (FCN) with a focus on 3D data was pro- posed in [8]. Different from [8], [9] proposed a 3D CNN network which stacked up 3D convolutional layers without the pooling layer. This model could capture the changes of local signals contained in the spectral-spatial data. The pooling layer could also be replaced by the spectral-spatial 3D convolutional layer [4]. Furthermore, there were some hybrid models that combined 2D CNNs with 3D CNNs. Obviously, the 3D CNN-based methods involved a much larger number of parameters than that of the 2D CNN models. Therefore, both the model complexity and the memory consumption of 3D CNN model are huge. Consequently, [14] tried to replace the 3D convolutional layer by a mixture of a 2D spatial convolutional layer and a 1D temporal convolutional layer which could largely alleviate the aforementioned problem. However, it only extracted the spatial features and spectral features separately and failed to explore the spectral-spatial interactions. To cope with this issue, we propose the H-CNN model, which synergistically trains a 2D CNN and a 3D CNN. Our approach not only enables the 3D CNN to invoke fewer 3D convolution operations, but also achieving better performance by taking into account the outputs from the 2D CNN.

3. THE PROPOSED H-CNN APPROACH

In this section, we first introduce the 3D convolution

operation. Then, we intro- duce the proposed H-CNN with more details. Finally, we propose a deep H-CNN network for hyperspectral image classification task.

3.1. The Proposed Synergistic Convolutional Neural Network

The proposed H-CNN framework is illustrated in Fig. 1. Instead of training a 2D CNN (or 3D CNN) separately, H-CNN is a deep neural network which is composed of 2D CNNs and 3D CNNs. At each round of the model fusion process, the 2D CNNs and 3D CNNs are equally weighted and fused together to generate deeper and more sophisticated feature maps, and then data interaction is invoked to produce the cross-domain transfer. In addition, local cross-domain concatenate operations are conducted to generate the new 2D features and 3D features as the next inputs. We use I to denote the input data. The first step in H-CNN is to generate 2D and 3D input data, let I_2 denote the 2D input data and I_3 denote the 3D input data. Let $v_2(x)$ and $v_3(x)$ denote the two output features after invoking the 2D CNN and the 3D CNN. In this paper, $o_2(x)$ and $o_3(x)$ are convolutional representations of I before the final fully-connected layer $f(\cdot)$ that classifies x to one of the pre-defined categories. Let ψ_2 and ψ_3 denote the 2D convolutional and 3D convolutional operations, and D is the data interaction operation. Accordingly, we use the standard cross entropy loss function:

$$
\begin{aligned}
L(x, y) &= H(y, f(o_2(x) + o_3(x))), \\
(o_2(x), o_3(x)) &= D(v_2(x), v_3(x)), \\
v_2(x) &= \psi_2 \otimes I_2, \\
v_3(x) &= \psi_3 \otimes I_3,
\end{aligned}
$$

for any data (x, y) in I where y is the real label for x and $H(\bullet)$ is the cross entropy function. The proposed H-CNN model integrates 2D CNNs and 3D CNNs to generate deeper feature maps at each round of the spectral-spatial fusion process, and invokes the data interaction module to provide sufficient training samples for the 3D convolution operation.

3.2. Deep H-CNN Network

For hyperspectral image classification, we design a hybrid deep model by stacking 2D convolution and 3D convolution which is synergistically trained. As shown in Fig. 2, We design a simple yet efficient deep H-CNN network (H-CNN is short only have one 3D convolution) by stacking the H-CNN together. In fact, the proposed H-CNN model is an end-to-end network and takes the hyperspectral images as input. The proposed deep H-CNN network consists of three H-CNN which only involves 3D convolutions. Furthermore, process the input data is not a trivial task. Note that there are BN-inception [15] and the ReLU function [3] layers after each convolutional block of the proposed models. For simplicity reason, this BN and ReLu layers are omitted after each convolutional block. In order to allow the input images of any length, we use a global pooling layer as the last layer of the network.

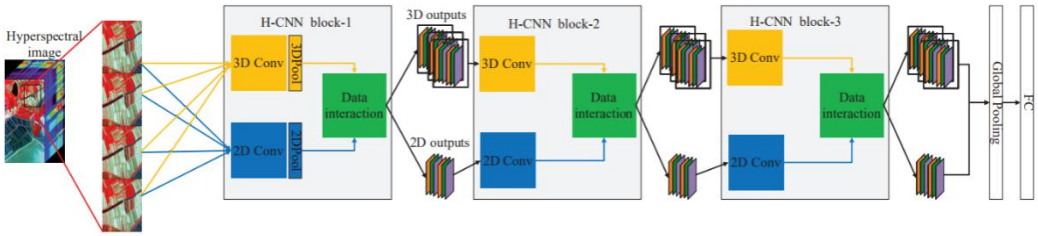

Figure 2 Illustration of the proposed models the deep H-CNN model. Yellow blocks and blue blocks refer to 3D convolution and 2D convolution. Green blocks refer to data interaction module

The difference between the proposed approach and the state-of-the-art 3D CNN models [5, 19] is that the deep H-CNN requires fewer 3D convolution operations for the spectral-spatial fusion stage, and yet it can generate deeper and richer feature maps. Moreover, different from the conventional 3D CNN based models, the deep H-CNN approach can take full advantage of 2D CNN approaches, and yet it can be trained using a much smaller image dataset.

4. EXPERIMENTAL RESULTS

To evaluate the performance of the proposed deep H-CNN, we conduct a comparative evaluation by comparing its performance with that of several state- of-the-art baseline models on benchmark hyperspectral image dataset. The experimental settings as well as evaluation criteria are illustrated in the following subsections. Note that the extended version of the proposed model with data interaction module is called DH-CNN.

4.1. Experimental Settings

In the experiments, one widely adopted benchmark hyperspectral image dataset, i.e., Indian Pines Scene, is chosen for comparative evaluation. The Indian Pines Scene dataset has 16 classes and contains 145 × 145 pixels in spatial dimension in the image, and 200 pixels in spectral dimension.

We compare the proposed models with nine conventional and the state-of-the-art methods. Three widely used evaluation criteria including average accuracy (AA), the F1 score and Kappa coefficient (K) are adopted for measuring the

performance of each model. We have evaluated all aforementioned models on the benchmark dataset, and the following experimental results are recorded.

The visualization of our experimental results is plotted in Fig. 3. It can be observed from Fig. 3 that the proposed DH-CNN model achieves the best performance among when compared to other methods. The possible reasons might be as follows: (i) the model is trained in a synergistic manner, which can use 2D convolutions and 3D convolutions to simultaneously generate deeper and richer features; (ii) it makes full use of the 2D and 3D features by the data interaction module.

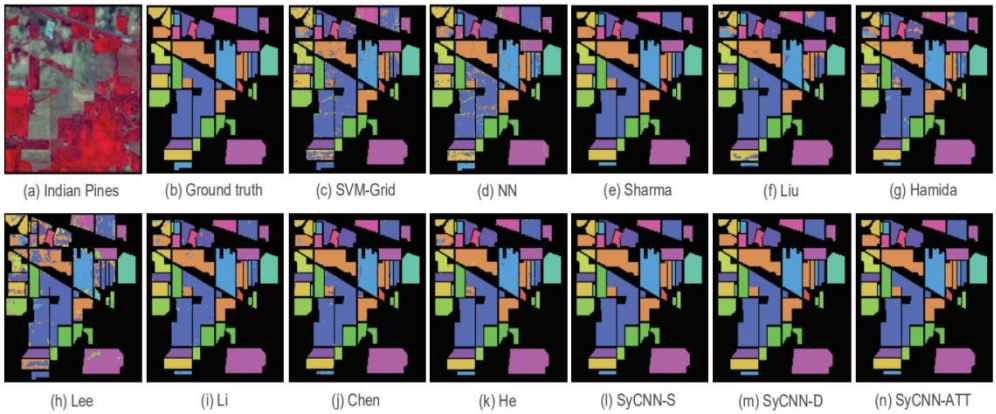

Figure 3 Visualization of the experimental results on Indian Pines dataset: (a) Indian Pines, (b) Ground truth, (c) SVM-Grid, (d) NN, (e) Sharma, (f) Liu, (g) Hamida, (h) Lee, (i) Li, (j) Chen, (k) He, (l) DH-CNN. It is observed that the results of our proposed models are very close to the ground truth.

It is noticed that the Sharma approach is the second-best model and is superior to the 3D CNN-type approaches like Hamida, Lee, Li, Chen, and He's approaches. The reason is that the 3D CNN models, having more parameters, usually need more training samples. Therefore, their performance might be worse when the training examples are few. It is also observed that the model performance of NN is the worst. The reason is that the NN model only has one 2D convolution and 4 fully connected layers for classifying the images. This simple architecture fails to generate deep and rich features for classifying hyperspectral images.

5.CONCLUSIONS

With the advance of deep learning, more and more related techniques have been adapted to the hyperspectral image classification task with superior model performance. However, the state-of-the-art 2D CNNs or 3D CNNs seldom capture the spectral-spatial information of hyperspectral image simultaneously. In this paper, we propose a hybrid Convolutional Neural Network which contains a module of 2D/3D CNNs, and a data interaction module to fuse the spectral- spatial data. We evaluate the proposed H-CNN as well as a number of baseline and state-of-the-art approaches on the widely adopted benchmark dataset. The promising experimental results have demonstrated that the proposed approach is superior to the compared methods with respect to evaluation criteria, i.e., average accuracy, F1 score and Kappa coefficient.

Table 1 Comparative evaluation based on the Indian Pines Scene dataset

Method	AA(%)	F1(%)	K(×100)
SVM-Grid	87.93	87.40	86.2
NN	87.57	89.07	85.8
Sharma	95.64	97.48	95.1
Liu	89.56	81.08	88.1
Hamida	86.99	90.16	85.2
Lee	87.87	83.42	86.1
Li	94.22	96.71	93.4
Chen	93.20	95.51	92.3
He	91.87	92.21	90.8
DH-CNN	**96.13**	**98.08**	**95.6**

ACKNOWLEDGMENTS

This paper is partially supported by Shenzhen Science and Technology Pro- gram under Grant No. JCYJ20170811153507788, and the Guangdong Province Science and Technology Department Project under Grant NO.20178090901022. This work is also partially supported by the National Science Foundation of China under grant No.61872108.

REFERENCES

[1] Chen, Y., Jiang, H., Li, C., Jia, X., Ghamisi, P. Deep feature extraction and classification of hyperspectral images based on convolutional neural networks. IEEE Transactions on Geoscience and Remote Sensing 54(10), 6232–6251 (2016).
[2] Diba, A., Sharma, V., Van Gool, L. Deep temporal linear encoding networks. In: 2017 CVPR. pp. 1541–1550. IEEE (2017).
[3] Glorot, X., Bordes, A., Bengio, Y. Deep sparse rectifier neural networks. In: Proceedings of the fourteenth international conference on artificial intelligence and statistics. pp. 315–323 (2011).
[4] Hamida, A.B., Benoit, A., Lambert, P., Amar, C.B. 3-d deep learning approach for remote sensing image classification. IEEE Transactions on Geoscience and Remote Sensing (2018).
[5] He, M., Li, B., Chen, H. Multi-scale 3d deep convolutional neural network for hyperspectral image classification. In: ICIP), 2017. pp. 3904–3908. IEEE (2017).
[6] Hughes, G. On the mean accuracy of statistical pattern recognizers. In: IEEE Trans. Inf. Theory 1968. pp. 55–63 (1968).
[7] Ji, S., Xu, W., Yang, M., Yu, K. 3d convolutional neural networks for human action recognition. IEEE transactions on pattern analysis and machine intelligence 35(1), 221–231 (2013).
[8] Lee, H., Kwon, H. Contextual deep CNN based hyperspectral classification. In: IGARSS, 2016 IEEE International. pp. 3322–3325. IEEE (2016).
[9] Li, Y., Zhang, H., Shen, Q. Spectral–spatial classification of hyperspectral imagery with 3d convolutional neural network. Remote Sensing 9(1), 67 (2017).
[10] Liu, B., Yu, X., Zhang, P., Tan, X., Yu, A., Xue, Z. A semi-supervised convolutional neural network for hyperspectral image classification. Remote Sensing Letters 8(9), 839–848 (2017).
[11] Lorenzo, P.R., Tulczyjew, L., Marcinkiewicz, M., Nalepa, J. Band selection from hyperspectral images using attention-based convolutional neural networks. arXiv preprint arXiv:1811.02667 (2018).
[12] Scholkopf, B., Smola, A.J. Learning with kernels: support vector machines, regularization, optimization, and beyond. MIT press (2001).
[13] Sharma, V., Diba, A., Tuytelaars, T., Van Gool, L. Hyperspectral CNN for image classification & band selection, with application to face recognition (2016).
[14] Sun, L., Jia, K., Yeung, D.Y., Shi, B.E. Human action recognition using factorized spatial-temporal convolutional networks. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 4597–4605 (2015).
[15] Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., Wojna, Z. Rethinking the inception architecture for computer vision. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 2818–2826 (2016).
[16] Wang, L., Xiong, Y., Wang, Z., Qiao, Y., Lin, D., Tang, X., Van Gool, L.Temporal segment networks: Towards good practices for deep action recognition. In: European Conference on Computer Vision. pp. 20–36. Springer (2016).
[17] Wang, Q., Lin, J., Yuan, Y. Salient band selection for hyperspectral image classification via manifold ranking. IEEE transactions on neural networks and learning systems 27(6), 1279–1289 (2016).
[18] Wang, Q., Meng, Z., Li, X. Locality adaptive discriminant analysis for spectral– spatial classification of hyperspectral images. IEEE Geoscience and Remote Sensing Letters 14(11), 2077–2081 (2017).
[19] Yang, X., Ye, Y., Li, X., Lau, R.Y., Zhang, X., Huang, X. Hyperspectral image classification with deep learning models. IEEE Transactions on Geoscience and Remote Sensing (2018).
[20] Yuan, Y., Lin, J., Wang, Q. Dual-clustering-based hyperspectral band selection by contextual analysis. IEEE Transactions on Geoscience and Remote Sensing 54(3), 1431–1445 (2016).