Hyperspectral Image Classification Using Residual 2D and 3D Convolutional Neural Network Joint Attention Model

Qinglie Yuan[a,b], Yuhao Ang[a], Helmi Zulhaidi Mohd Shafri[a]

[a] Department of Civil Engineering and Geospatial Information Science Research Centre (GISRC), Faculty of Engineering, Universiti Putra Malaysia, 43400 Serdang, Selangor, Malaysia-yuanqinglie8236@gmail.com
[b] Faculty of Civil and Architecture Engineering, Panzhihua University ,617000 Panzhihua, China

KEY WORDS: Hyperspectral image classification (HISC), convolutional neural network (CNN), deep learning, Channel attention, machine learning.

ABSTRACT:

Hyperspectral image classification (HSIC) is a challenging task in remote sensing data analysis, which has been applied in many domains for better identification and inspection of the earth surface by extracting spectral and spatial information. The combination of abundant spectral features and accurate spatial information can improve classification accuracy. However, many traditional methods are based on handcrafted features, which brings difficulties for multi-classification tasks due to spectral intra-class heterogeneity and similarity of inter-class. The deep learning algorithm, especially the convolutional neural network (CNN), has been perceived promising feature extractor and classification for processing hyperspectral remote sensing images. Although 2D CNN can extract spatial features, the specific spectral properties are not used effectively. While 3D CNN has the capability for them, but the computational burden increases as stacking layers. To address these issues, we propose a novel HSIC framework based on the residual CNN network by integrating the advantage of 2D and 3D CNN. First, 3D convolutions focus on extracting spectral features with feature recalibration and refinement by channel attention mechanism. The 2D depth-wise separable convolution approach with different size kernels concentrates on obtaining multi-scale spatial features and reducing model parameters. Furthermore, the residual structure optimizes the back-propagation for network training. The results and analysis of extensive HSIC experiments show that the proposed residual 2D-3D CNN network can effectively extract spectral and spatial features and improve classification accuracy.

1. INTRODUCTION

Hyperspectral imaging has a wide variety of real-world applications, including land cover analysis, urban analysis, environmental and agricultural analysis, and anomaly identification. Hyperspectral remote sensing classification is an effective way to distinguish different features and provide critical decision-making and reference information for different fields. Some advanced remote sensing platforms, such as airborne, space satellite, and UAV platforms, can achieve hyperspectral and high-resolution remote sensing data. Therefore, abundant spectral features and accurate spatial information can improve the classification accuracy (Paoletti et al., 2019; Chen et al., 2019; Seydi et al., 2020).

Abundant research has been carried out on precise feature classification with hyperspectral imagery, adopting two primary classification strategies (Zhong et al., 2017; Zhang et al., 2018). The first strategy solely uses spectral features. Another method combines spectral and spatial information to distinguish features. However, the exceptionally high spatial resolution could cause severe spectral variability and heterogeneity. For instance, rice filed and forests have similar spectral curves, as illustrated in Figure 1, which brings some obstacles to the classification task. Hence, there are still some challenges when applying the previous classification strategies to hyperspectral and high-resolution imagery.

In the few decades, numerous traditional classification methods such as minimum distance, maximum likelihood, and spectral angle mapper have proven the ability to classify the features based on the advantage of the rich spectral information. Moreover, advanced machine learning algorithms such as support vector machine, decision tree, and random forest (RF) have the stability for the hyperspectral classification task. For instance, the support vector machine (SVM) seeks to separate two-class data by learning an optimal decision that separates the training samples in a kernel-included high dimensional feature space. Some studies using SVM for hyperspectral image classification can improve result performance (Mountrakis et al., 2011; Li, Bioucas-Dias & Plaza 2013; Song et al., 2020).

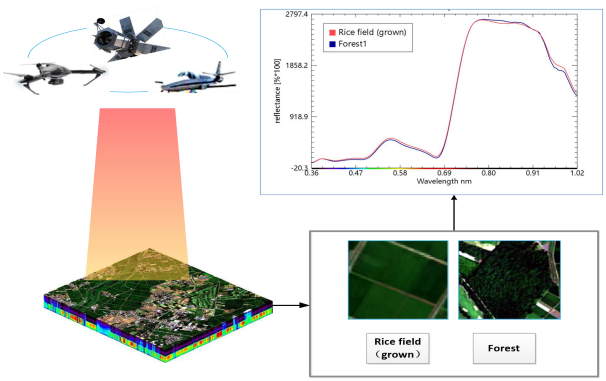


Figure 1. Remote sensing image with high spectral resolution and high spatial resolution.

However, most traditional approaches presenting hand-designed feature descriptions based on expertise knowledge probably limit the application potential for the precise classification. Some classifiers have limited representation capacity to utilize the abundant spectral and spatial features

fully. It is challenging to find appropriate parameters to generate features for different classification tasks. Handcrafted feature as shallow properties based on expertise knowledge probably limits the application potential for the precise classification.

Recently, deep learning-based methods reached powerful performance in many applications where visual information is required, such as image classification and object detection. CNN plays a promising role in processing feature extraction. The convolutional neural network is one of the most commonly used in the hyperspectral classification task because of its superior performance to hand-designed features. In recent years, considerable development is also worked in deep learning for the HIS analysis.

Zhong et al. (2017) have researched the proposed spectral-spatial residual network (SSRN). The residual block in SSRN was used to determine a feature map that aids in propagating previous information to the following units to improve the backward step by promoting gradient propagation. Roy et al. (2019) proposed hybridSN, a joint spectral-spatial 3D CNN network followed by spatial 2-D-CNN. 3D-CNN facilitates the focus on the joint spatial-spectral feature representations from a stack of spectral bands, then followed by 2D-CNN to further learn the more abstract level spatial representation. Compared with other handcrafted methods and deep learning-based methods on Indian pines, the university of Pavia and Salinas scene. The proposed model achieved promising results with 98.39%, 99.72%, and 99.98, respectively.

It is noticeable from the previous studies that using 2D and 3D alone had a few disadvantages, such as very complex models or missing channel relationships, respectively. The main intention is because HIS comprises both volumetric data and has spectral dimension. Comparatively, a deep 3-D-CNN is more computationally complicated, and this mechanism alone sometimes tends to perform worse for classes having similar textures over many spectral bands. This paper proposed a novel method to combine the advantages of 2D CNN and 3D CNN. The proposed deep convolution network model uses 3D CNN to extract spectral information and ensure network propagation effectiveness. The residual structure is applied to the 3dcnn module to optimize the network structure, and an improved channel attention mechanism refines features to improve the efficiency of the model. The network model adopts 2D depth-wise convolution extract spatial features following the 3D residual module to obtain multi-scale spatial information and reduce model parameters.

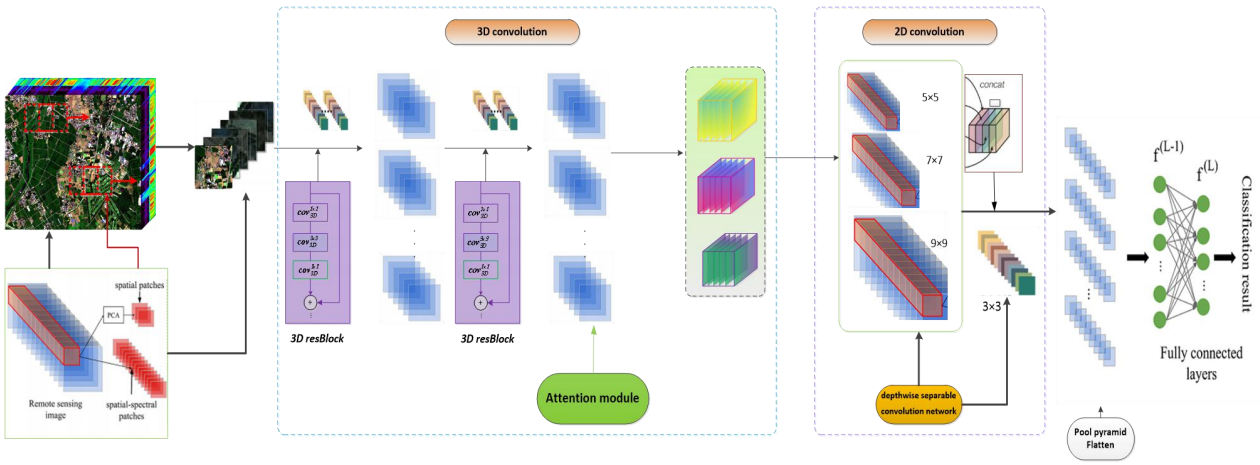


Figure 2. The proposed network framework.

2. THE PROPOSED NETWORK MODEL

As illustrated in Figure 2, The proposed model framework consists of four parts: Firstly, the hyperspectral remote sensing images were fed into the network and divided into hyperspectral image cubes with fixed pixel size (In this study, the size was set at 15 ×15×15). Then the image cube is transmitted into the residual 3D convolution module to extract the spectral feature information along channel direction, and the channel attention model using squeeze and excitation network is applied for spectral feature optimization. The obtained feature map is transferred to the depth-wise separable convolution module to extract multi-scale spatial context information. Finally, the pooling pyramid feature interaction is carried out to aggregate features, and the fully connection layers complete the final prediction.

2.1 3D residual convolutional module

3D CNN can extract features of three dimensions along spatial and channel dimensions through 3D convolution layers, which is suitable for hyperspectral image data. The network, stacking 3D convolution layers, can simultaneously learn the spatial correlation and spectral characteristic of ground objects. 3D CNN is defined in formula (1):

$$f_{ij}^{xyz} = \varphi\left(\sum_{h=0}^{H_{i-1}}\sum_{v=0}^{V_{i-1}}\sum_{c=0}^{C_{i-1}} w_{ijk}^{hvc} f_{(i-1)k}^{(x+h)(y+v)(z+c)} + b_{ij}\right), \quad (1)$$

where f_{ij}^{xyz} represents the value of the neuron at (x, y, z), i is the neural network layer index, j is the feature sample index, m is the feature map index of i-1th layer network; h and v are the width and length of 2D spatial convolution kernel, respectively, and C is the size of channel dimension. w_{ijk}^{hvc} denotes the weights of the convolutional kernel at position (h, v, c) connected to the m-th feature map, b_{ij} is the bias. φ is the activation function of neurons.

Residual learning is applied to the 3D CNN blocks to avoid network degradation as the increase of network depth. As shown in Figure 3, the residual structure can alleviate the issues of network gradient degradation when the training samples are not enough (He et al., 2016). In the residual convolution module, as shown in the figure, a short-cut connection path can be established, skipping some 3D convolutional modules, and is used to fuse the features from the previous layers. Through the skip-connection path, the error generated in the training of the network can be propagated, which can solve the issues of gradient dispersion caused by stacking many convolutional layers and accelerate the update and iteration of weight parameters. Hence, the residual structure can improve efficiency for the network architecture.

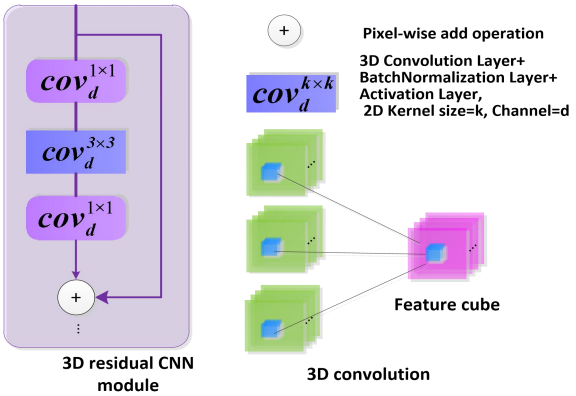


Figure 3. 3D residual convolution block.

If 3D CNN is applied to hyperspectral cubes with large convolution kernels, network parameters will significantly increase due to hundreds of spectral bands for the hyperspectral image. After several residual modules are contacted, it can cause a computational burden and affect model training efficiency. Some traditional methods use data dimension reduction methods, such as principal component analysis (PCA), which may lose the correlation and some features between spectral channels. Therefore, inspired by residual 2D CNN, a bottleneck structure is introduced into 3D CNN modules, making the network learn spectral features and reduce the model parameters. The bottleneck structure contains two convolution layers using kernels with the size of $1 \times 1 \times d$, where d is the number of spectral bands. In Figure 3, the first convolution layer is used to reduce hyperspectral cube dimensions and capture spectral information, where r is the reduction rate. Then, features are transmitted into a convolution of $3 \times 3 \times d$ to extract spectral and spatial features. Finally, the last convolutional layer recovers feature dimensions. Therefore, a 3D residual module unit consists of a bottleneck structure and a layer of 3D convolution, where two 3D residual modules are contacted to generate multiple feature cubes.

In the proposed 3D residual module, the upper layer of convolution features via skip-connection are fused by the three layers of convolution in the residual module. The ReLu activation function is applied to these features before they are transmitted into the subsequent convolutional module. This enables the residual module to learn new features based on the input features to improve the feature representation and reduce computational cost.

2.2 Spectral feature optimization using the channel attention mechanism

Although the deep 3D convolutional neural network can learn to extract different spectral-spatial information levels, these features may not be the optimal results for the classifier to recognize different objects. On the one hand, the convolution filter extracts fusion information of space and channel in the local receptive field. With the addition of the nonlinear activation layer and downsampling layer, CNN can obtain the hierarchical pattern with the large receptive field to capture image features. However, this process requires stacking enough convolution layers, which undoubtedly increases the difficulty and computational complexity of network model training. On the other hand, the features extracted by the 3D residual convolution module have many different representations. These features are not filtered with redundant information, especially for a large number of spectral bands, which may cause classification ambiguity and affect the network efficiency. Attention mechanism can assist the model to assign different weights to each channel or spatial features, making the model filter redundant information without bringing more calculation complexity and memory consumption. Channel attention can model the correlation and dependence of different spectral features. Therefore, we introduce a channel attention model to learn global representation and optimize 3D convolutional features.

Squeeze and exception network (SEnet) is an effective channel attention model that aims to improve the representation ability of the network by modeling the dependency of each channel and recalibrate the features channel-wise (Hu et al., 2018). The network can learn to selectively enhance the efficient information and suppress useless features through establishing global representations weights. The basic structure of the SEnet block is shown in Figure 3. First, the squeeze operation obtains the global spatial features of each channel as the representation parameters to form the global descriptors. Second, the excitation operation constructs the dependency for each channel by two layers of the fully connected neural network, which adjusts the feature map based on the learning parameters.

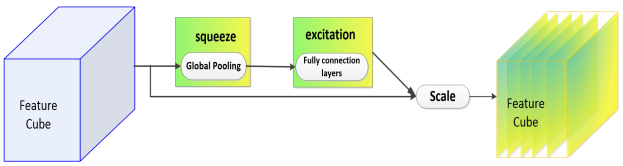


Figure 4. SEnet network structure.

As shown in Figure 5, the proposed channel attention model is similar to SEnet, including squeeze, activation, and scaling process. Different from SEnet, in the squeeze operation, two pooling layers, including global average pooling (GAP) and maximum pooling (MAP), are adopted to enhance the learning ability of global representation that can be defined in formula (2) and (3):

$$f_{avg}^{k} = \frac{1}{H \times V} \sum_{i=1}^{H} \sum_{i=1}^{V} f_{ijk}, \quad (2)$$

$$f_{max}^{k} = max(f_{ijk}), \quad i, j = 1, 2, 3, \cdots, HV, \quad (3)$$

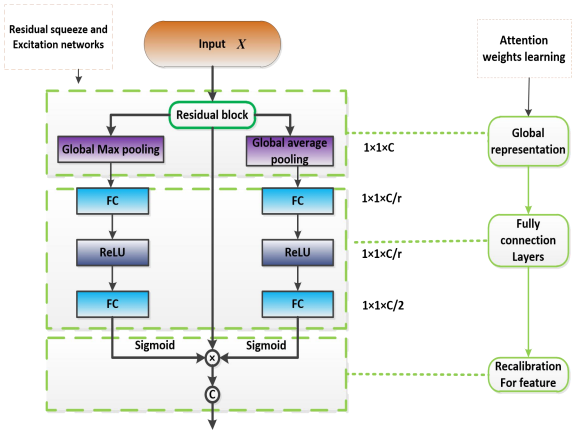


Figure 5. The framework of proposed channel attention model.

where f_{ijk} is the value at position (i, j, k) of the feature map in the k-th channel. H and V are the length and width of the feature map, respectively. f_{avg}^{k} and f_{max}^{k} are the feature value of GAP and MAP, respectively. Attention weight parameters can be obtained by two layers of the fully connected neural network as formula (4) and (5):

$$\lambda_{avg} = sigmoid(F(\phi(F(f_{avg}^{k})))\,, \tag{4}$$

$$\lambda_{max} = sigmoid(F(\phi(F(f_{max}^{k})))\,, \tag{5}$$

where, λ_{avg} and λ_{max} are the global scaling parameters, ϕ is the ReLu activation function, F is dense layers weight operation. Finally, the 3D residual block features can be calibrated by global scaling parameters, as shown in formula (6). $contcat(\bullet)$ denotes concatenation operation, $\otimes$ denotes pixel-wise multiplication operation, f_{re} is the fused feature.

$$f_{re} = contcat(\lambda_{max} \otimes f_{ij}^{xyz}, \lambda_{ave} \otimes f_{ij}^{xyz})\,, \tag{6}$$

In the activation process, two layers of dense connections are used to scale the global characterization parameters by setting different numbers of neurons. The first dense connection layer reduces the global representation parameters to $1 / r$ of the original channel number, r is the reduction ratio, while the second dense connection layer recovers the parameters to 1 / 2. The two new features via GAP and MAP can be generated and are concatenated to construct the optimized spectral-spatial features. This operation aims to enable the channel representations to capture different aspects of globally spatial position in each channel-wise and enhance spectral feature attention ability. In figure 1, the proposed channel attention model is applied to the network, following by each 3D residual CNN module to refine features.

2.3 Multi-scale feature fusion network by 2D CNN

Due to the computational complexity and parameter limitation, 3D convolution uses the convolution kernels with a small size to extract spatial-spectral information. However, objects have different scale characteristics that are presented in a variety of sizes in remote sensing images. If the convolution operation uses a fixed receptive field, the performance is inconsistent with the features of different scales, which probably causes the loss of some spatial information, such as the boundaries and corners of different categories. Therefore, in the model, the 2D CNN network following with the 3D residual convolution module is applied to extract multi-scale spatial information to improve the spatial accuracy of classification.

In 2D CNN, deep separable convolution is an effective operation (Chollet, F., 2017). Firstly, the multi-channel features from the upper layer are divided into each feature map channel-wise, and then they are convoluted using different kernels, respectively, and are fused through point convolution. This channel level splitting operation only adjusts the size of the feature maps from previous layers, but the number of channels does not change. Therefore, the depth-separable convolution can ensure the channel feature dependency and extract the spatial features on different channels. Besides, it significantly reduces the model training parameters by channel separation. In the proposed model, as shown in figure 1, a set of convolution kernels with different sizes, including 5 × 5, 7 × 7, and 9 × 9, are applied to obtain multi-scale spatial features in the proposed model 2D depth separable convolution blocks. Finally, these multi-scale features are concatenated and fused by 2D convolution blocks.

3. EXPERIMENTAL RESULTS AND DISCUSSION

A series of experiments are conducted to test the superiority of the proposed model. The results are compared with state-of-the-art models such as SVMs, RF, 2D-CNN, 3DCNN. The model is trained using Adam optimizer with a learning rate of 0.001 for 2000 epochs over each HSI data set. The categorical cross-entropy loss is minimized using back-propagation. Batch normalization (BN) and 50% of dropout are used to deal with the over-fitting problem. Accuracy metrics were used to evaluate the experimental results, including Overall accuracy (OA), Kappa coefficient, Recall, and F1 score.

3.1 Data Description

In this paper, experiments are completed on two representative hyperspectral data sets with different settings, including the Chikusei dataset and the Pavia dataset.

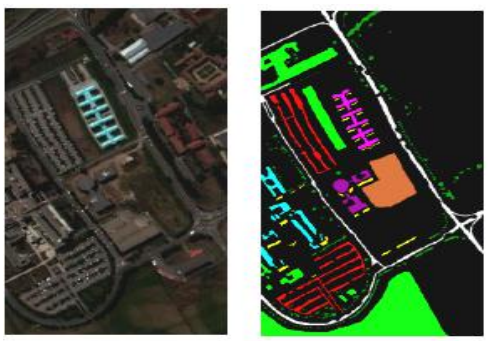
The first dataset is airborne hyperspectral remote sensing by the Hyperspec imaging sensor over agricultural and urban areas in Chikusei, Japan (Yokoya & Iwasaki, 2016). The dataset comprises 128 bands in the spectral range of 0.363 μm to 1.018 μm and have 2.5 m spatial resolution. The whole image has been geometrically and radiometrically corrected. In the experiment, the study area was a subset with the size of 991 pixels ×1121 pixels in the yellow zone, as shown in Figure 5. Ground truth of 19 classes was collected via a field survey and visual inspection using high-resolution color images.

Stratified systematic samplings are applied for sampling techniques. A total of 34820 points from the whole samples based on stratified systematic samplings were created and consisted of 19 classes (including water, bare soil farmland, forest, grass, the rice field-grown, plastic house, manmade non-dark, manmade dark, manmade grass, paved ground, and asphalt, etc.).

Pavia university data contains a part of urban site scenes acquired in 2013 by the ROSIS spectral sensor over the University of Pavia, Italy. The hyperspectral imagery has 115 wavebands in the wavelength range of 0.43-0.86 μm, and the spatial resolution is 1.3 m. Due to the influence of noise, 12 bands are eliminated, and 103 bands are left as classification data. The image size is 610 × 340, and the pixels are be classified into nine categories. The pseudo color image and the

ground truth map of the hyperspectral data are shown in Figure 6.

Figure 5. Chikusei dataset. The experimental subset is located in the yellow rectangle.



a b

Figure 6. Pavia dataset. a is the pseudo color image, b is the ground truth map.

3.2 Classification results for the proposed network

Table 1 displays data from the Chikusei dataset's experimental results. Our proposed residual 2D-3D CNN method achieved an overall classification accuracy of 99.24% with the F1 score of 99.78% and kappa score of 99.48% and outperformed 3D-CNN with an overall accuracy of 91.23%, kappa coefficient of 92.17%, F1 score of 82.47%, the precision score of 87.24% and Recall of 95.34%. 2D-CNN with 87.21% overall classification accuracy, F1 score of 85.24%, and kappa score of 85.34%.

Generally, there is a tendency that the 2D-CNN model provides the lowest values due to the kappa (85.24%), F1(82.34%), precision (81.36%), recall (83.21%), and therefore overall accuracy, common classifiers such as random forest and support vector machine performed better than 2D-CNN. Comparatively, the proposed residual 2D-3D CNN has almost a similar performance with random forest, support vector machine, and random forest.

The overall accuracy of the proposed method is similar to that of random forest, with 99.21% overall classification accuracy and similar to that of support vector machine with 99.53% overall classification accuracy. Besides, the precision score of the proposed 2D-3D CNN (99..24%) is achieved equivalent to random forest (99.21%), support vector machine (99.53%). F1 and Recall of proposed residual 2D-3D CNN are slightly higher than random forest and support vector machine, indicating that the proposed method is slightly better at the true positive and contributes to more balanced predictions. The proposed model's training time is the highest than other classifiers, which cost only 53.98 seconds per 50 epochs, whereas the traditional classifiers such as random forest and support vector machine recorded 307.12 seconds and 974.25 seconds, respectively. 2D-CNN achieved the lowest training time, which is 312.59 seconds.

Methods	Overall accuracy %	Kappa %	F1%	Precision%	Recall%
RF	99.21	99.41	99.56	99.47	99.41
SVM	**99.53**	99.36	99.57	**99.54**	99.53
2D-CNN	87.21	85.24	82.34	81.36	83.21
3D-CNN	91.23	92.17	82.47	87.24	95.34
Proposed model	99.24	**99.48**	**99.78**	**99.54**	**99.58**

Table 1. Comparison Accuracy on Chikusei data with another state-of-the-art method. The bold values denote the best result.

Table 2 shows the accuracy of the proposed methods and other classifiers for the experimental results on the Pavia dataset. Overall accuracy for the proposed model was 97.57 percent, with a Kappa coefficient of 97.42 percent. However, for the tree class, the proposed model performs slightly worse than 3D-CNN. Furthermore, SVM outperformed the proposed model in classification.

Category	2D CNN	3D CNN	RF	SVM	Proposed model
Asphalt	88.56	91.68	86.91	93.45	**97.46**
Meadows	84.15	84.73	84.59	93.78	**96.01**
Gravel	56.82	62.17	38.21	82.53	**99.24**
Trees	94.09	**99.80**	94.94	99.38	99.12
Metalsheet	99.70	99.93	99.35	99.6	**100.00**
Baresoil	45.93	98.27	98.57	97.38	**99.92**
Bitumen	64.11	93.38	90.38	94.19	**100.00**
Bricks	99.08	98.45	97.39	98.31	**98.57**
Shadows	99.4	97.79	94.09	**99.86**	98.52
Overall accuracy/%	80.27	89.48	87.04	94.68	**97.57**
Kappa/%	73.90	86.46	87.23	92.92	**97.42**

Table 2. Comparison accuracy on Pavia dataset with another state-of-the-art method. The bold values denote the best result.

Obviously, the proposed model achieved higher accuracy than the others on both datasets. 3D convolution layer can facilitate gradient back-propagation, whereas the 2D convolution block aims to extract spatial features abundantly. Therefore, The combined 3D and 2D have effectively extract refined features and enhance the classification accuracy. Moreover, the multi-scale information improved classification performance by fusing different level representations from 3D residual blocks. Besides, depth-wise separable convolution was used in the convolution filter for the input channel to ensure fewer parameters.

The attention gate layers support the network to obtain global context by squeezing the operation and obtaining channel-wise representations to calibrate features. For instance, the global

average pooling maintains the information in the global context, whereas the maximum pooing extracts remarkable features. The implementation of multiple dependencies for the global feature correlations between channels affirms that the final excitation scores should not be biased towards local spatial information. Therefore, the proposed model can learn more distinct and powerful spectral-spatial feature representations. Our result also has proven in agreement with the previous works of Yu et al. (2020) and Yang et al. (2020).

Figures 7 and 8 show the prediction maps of the proposed model and other classifiers. Visually, The classification map generated through SVM and RF is better than 2D-CNN and 3D-CNN, but still exist some artifacts within the class boundaries and false classification of the pixel. 2D CNN has worse performance than other methods, which indicates 2D-CNN cannot sufficiently extract spectral representation to predict the target pixels. The proposed model can learn more discriminative and robust spectral-spatial feature representations by refining the feature while suppressing the ineffective feature.

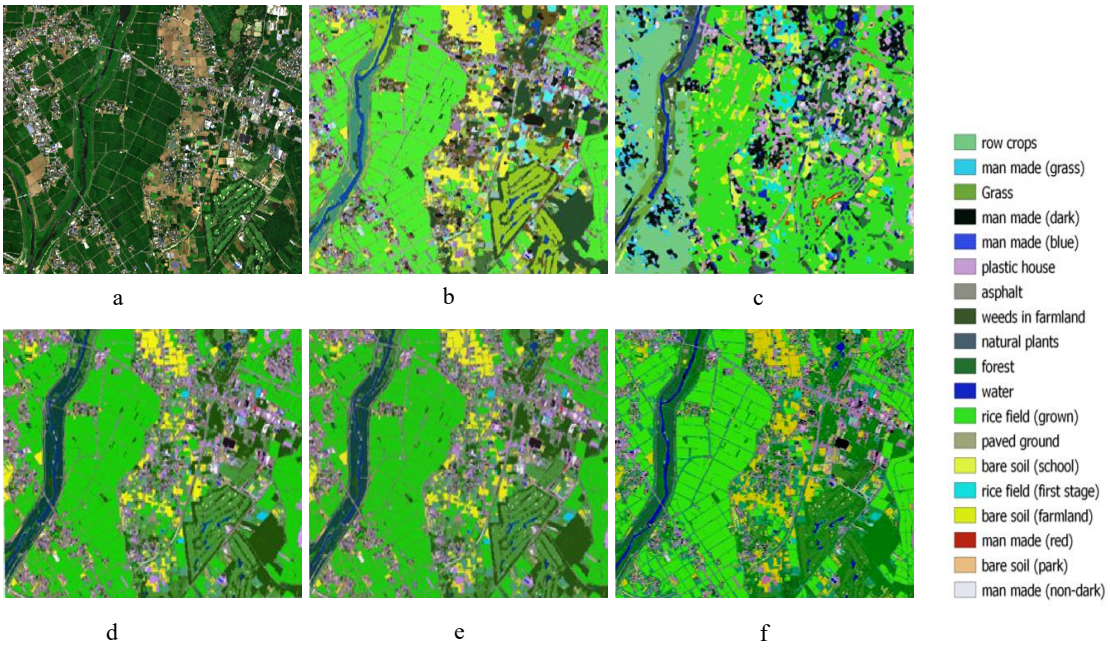


Figure 7. The classification map for Chikusei data subset. a. Study area, b. The proposed model, c. 2D-CNN, d.3D-CNN, e. Support vector machine, f. Random forest.

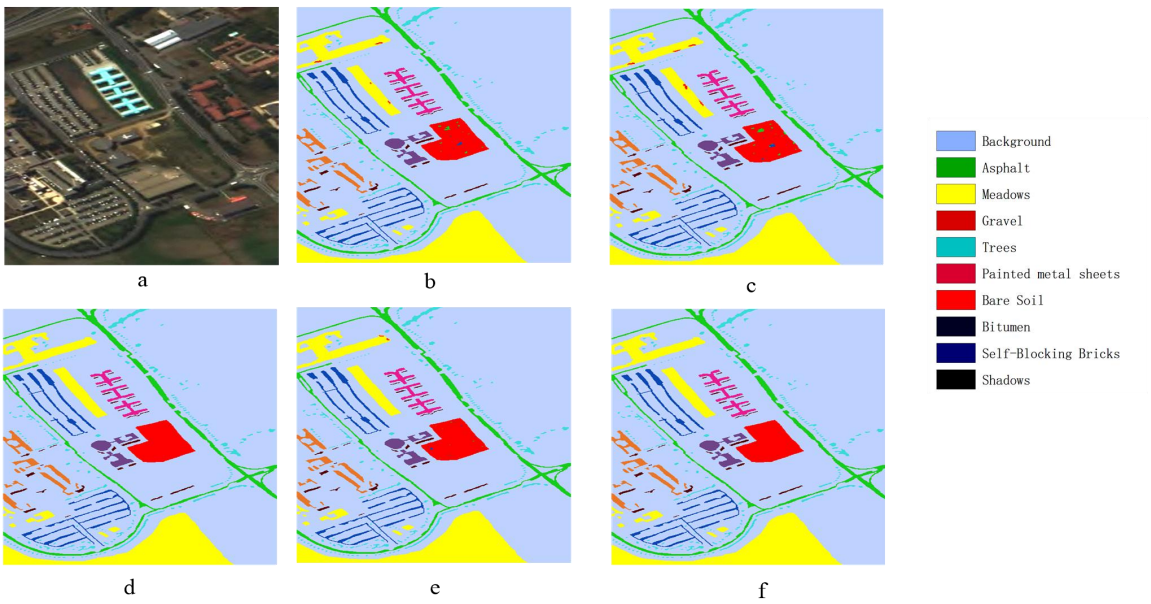


Figure 8. The classification map for Pavia dataset. a. Study area, b. 2D-CNN, c. 3D-CNN, d. Random forest, e. Support vector machine, f. Proposed model.

4. CONCLUSION

This paper studies the hyperspectral remote sensing image classification method based on deep learning and proposes a 2D-3D hybrid convolutional neural network algorithm to achieve the effective end-to-end classification of hyperspectral images. Instead of handcrafted feature extraction using hyperspectral images, the network uses 3D hyperspectral cube

data as input and combines 2D and 3D convolution advantages for the classification. The proposed algorithm adopts the residual learning network structure, which can extract the spectral and spatial features of hyperspectral images while deepening the network and reduce the gradient degradation problem. 2D depth separable convolution uses multiple convolution kernels with different sizes to extract multi-scale spatial features. In addition, the proposed channel attention model can effectively learn the global spatial representation and optimize the spectral features to improve the operation efficiency and classification accuracy of the model. The experimental results show that the performance of the algorithm is better than the traditional classification algorithm, 2D-CNN algorithm, and 3D-CNN algorithm.

In conclusion, compared with the current convolution neural network model, the proposed network model can optimize the extracted features and effectively fused the spectral and spatial information of hyperspectral remote sensing images. The experimental results demonstrated that the proposed classification framework has stronger feature extractability and effectively improves the classification accuracy for hyperspectral imagery.

ACKNOWLEDGEMENTS

The authors gratefully acknowledge Space Application Laboratory, Department of Advanced Interdisciplinary Studies, the University of Tokyo to provide the hyperspectral Chikusei data. Besides, the author also thankfully acknowledges telecommunications and remote sensing laboratory Pavia University (Italy) for providing the hyperspectral Pavia data. Comments from anonymous reviewers towards improving this paper are also acknowledged.

REFERENCES

Audebert, N., Le Saux, B., & Lefevre, S., 2019. Deep learning for classification of hyperspectral data: A comparative review. *IEEE Geoscience and Remote Sensing Magazine*, 7(2), 159–173.

Chen, Y., Zhu, K., Zhu, L., He, X., Ghamisi, P., & Benediktsson, J. A., 2019. Automatic design of convolutional neural network for hyperspectral image classification. *IEEE Transactions on Geoscience and Remote Sensing*, 57(9), 7048-7066.

Chollet, F., 2017. Xception: Deep learning with depthwise separable convolutions. *In Proceedings of the IEEE conference on computer vision and pattern recognition*, 1251-1258.

Chollet, F. (2017). Xception: Deep learning with depthwise separable convolutions. *In Proceedings of the IEEE conference on computer vision and pattern recognition,*1251-1258.

Dang, L., Pang, P., & Lee, J., 2020. Depth-Wise Separable Convolution Neural Network with Residual Connection for Hyperspectral Image Classification. *Remote Sensing*, 12(20), 3408.

He, K., Zhang, X., Ren, S., & Sun, J., 2016. Identity mappings in deep residual networks. *In European conference on computer vision*, 630-645. Springer, Cham.

Hu, J., Shen, L., & Sun, G., 2018. Squeeze-and-excitation networks. *In Proceedings of the IEEE conference on computer vision and pattern recognition* 7132-7141.

Li, J., Bioucas-Dias, J. M., & Plaza, A., 2013. Spectral-spatial classification of hyperspectral data using loopy belief propagation and active learning. *IEEE Transactions on Geoscience and Remote Sensing*, 51(2), 844–856.

Man, Q., Dong, P., Yang, X., Wu, Q., & Han, R. (2020). Automatic Extraction of Grasses and Individual Trees in Urban Areas Based on Airborne Hyperspectral and LiDAR Data. *Remote Sensing*, 12(17), 2725.

Masarczyk, W., Głomb, P., Grabowski, B., & Ostaszewski, M., 2020. Effective Training of Deep Convolutional Neural Networks for Hyperspectral Image Classification through Artificial Labeling. *Remote Sensing*, 12(16), 2653.

Mountrakis, G., Im, J., & Ogole, C., 2011. Support vector machines in remote sensing: A review. *ISPRS Journal of Photogrammetry and Remote Sensing*, 66(3), 247–259.

Paoletti, M. E., Haut, J. M., Plaza, J., & Plaza, A. (2019). Deep learning classifiers for hyperspectral imaging: A review. *ISPRS Journal of Photogrammetry and Remote Sensing*, 158, 279-317.

Roy, S. K., Krishna, G., Dubey, S. R., & Chaudhuri, B. B., 2019. HybridSN: Exploring 3D-2D CNN Feature Hierarchy for Hyperspectral Image Classification. *ArXiv*, 17(2), 277–281.

Song, S., Qin, H., Yang, Y., Zhang, Z., & Zhou, H., 2020. Hyperspectral anomaly detection via graphical connected point estimation and multiple support vector machines. *IEEE Access*, 8, 94152–94164.

Seydi, S. T., Hasanlou, M., & Amani, M. (2020). A New End-to-End Multi-Dimensional CNN Framework for Land Cover/Land Use Change Detection in Multi-Source Remote Sensing Datasets. *Remote Sensing*, 12(12), 2010.

Yang, X., Zhang, X., Ye, Y., Lau, R. Y. K., Lu, S., Li, X., & Huang, X., 2020. Synergistic 2D/3D convolutional neural network for hyperspectral image classification. *Remote Sensing*, 12(12), 1–19.

Yu, C., Han, R., Song, M., Liu, C., & Chang, C. I., 2020. A simplified 2D-3D CNN architecture for hyperspectral image classification based on spatial-spectral fusion. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*,13,2485–2501.

Yokoya, N., & Iwasaki, A., 2016. Airborne hyperspectral data over Chikusei. Space Appl. Lab., Univ. Tokyo, Tokyo, Japan, Tech. Rep. SAL-2016-05-27.

Zhong, Zilong, et al., 2017. Spectral–spatial residual network for hyperspectral image classification: A 3-D deep learning framework. *IEEE Transactions on Geoscience and Remote Sensing*, 56.2, 847-858.

Zhong, Z., Li, J., Luo, Z., & Chapman, M., 2017. Spectral–spatial residual network for hyperspectral image classification: A 3-D deep learning framework. I*EEE Transactions on Geoscience and Remote Sensing*, 56(2), 847-858.