CNN-based Large Scale Landsat Image Classification

Xuemei Zhao[1], Lianru Gao[1,*], Zhengchao Chen[1], Bing Zhang[1,2], Wenzhi Liao[3]

1 Key Laboratory of Digital Earth Science, Institute of Remote Sensing and Digital Earth, Chinese Academy of Sciences, Beijing, 100094, China

2 University of Chinese Academy of Sciences, Beijing, 100049, China

3 Department Telecommunications and Information Processing, Ghent University, Ghent 9000, Belgium

***Abstract*— Large scale Landsat image classification is the key to acquire national even global land cover map. Traditional methods typically use only a small set of samples to train the classifier and result in unsatisfied classification results. To improve the performance of large scale Landsat image classification, we apply a convolutional neural network (CNN)-based method named PSPNet in this paper to learn spectral-spatial features from a large training set. By considering the complexities and the various sizes of objects captured in large scale Landsat images, PSPNet can utilize the global information as well as consider the targets with different sizes. In addition, the research area is oversampled with a small offset which can increase the amount of training samples in order to improve the performance of PSPNet on Landsat images. Moreover, PSPNet is finely tuned on the pretrained Resnet50. Experimental results show the efficiency of the CNN based methods for the large-scale land cover mapping. In particular, PSPNet can produce better results even than the provided reference land cover map, with overall accuracy reaching 83%.**

I. INTRODUCTION

Landsat data is one of the most significant resources in medium resolution remote sensing image classification which has made a great contribution to the FROM-GLC [1], GlobeLand30 [2] and some other national or regional land cover products [3, 4]. Generally, the accuracy of the products highly depends on the ability of image classification algorithms. Therefore, tremendous efforts have been made to improve the accuracy of image classification algorithms [5, 6]. Traditional classifiers, such as maximum likelihood classifier [7], decision trees [8] and random forest [9], describe the characteristics of each class by human-designed features. However, when we want to classify the images with large scales, the scenes are always complicate due to the large coverage and the variation in terrains. Human-designed features cannot accurately model the variation of a class in a large scene when used on Landsat images. To solve this problem, machine learning algorithms, including Multilayer Perceptron (MLP) [10] and Support Vector Machine (SVM) [11], have drawn considerable attention in remote sensing image classification. MLP is expert in learning nonlinear spectral features, thus is widely used in high quality image classification. Nevertheless, the fully connected nature restricts its use in complicated scenes. As for SVM algorithm, it aims at finding the hyperplane which has the maximum margin between two classes. So only a few data points representing the boundaries between two classes contribute to the classifier while most samples are not capable to affect the final classification result. It is hard to find such fine samples in Landsat image for training SVM classifer. Therefore, the algorithms mentioned above do not perform well on large scale Landsat image classification.

The outstanding performance of Alexnet [12] in the 2012 ImageNet Large Scale Visual Recognition Competition (ILSVRC) stirs up a passion for research and application of deep learning. A lot of breakthroughs have been made from then on [13-18]. Among them, convolutional neural network (CNN) is one of the most suitable architectures for image classification. CNN uses stacked convolutional kernel to learn the features of images, so not only the spectral but also the texture information in spatial space is learned. Together with the depth of the neural network and the pooling layers, CNN is capable of establishing the connections between the input samples and the output labels. Then the connections can be used to obtain the classification result.

In recent years, CNN has produced state-of-the-art results in remote sensing image classification. However, it is mainly used in high resolution remote sensing images which have fine texture features and fixed shapes as natural images employed in computer vision [19-20]. While texture features of Landsat images are not as fine as high resolution remote sensing images and objects captured with 30m resolution generally do not have fixed shapes. Some researchers have focused on deep learning-based Landsat image object detection [21-22] and classification [23-24]. Ikasari et al. [25] used deep neural networks and 1-D CNN for Landsat images and compared them with Logistic regression, SVM, Random forest and Boost algorithms. The maximum accuracy obtained by deep neural network with batch normalization and dropout layer is 71.79%. Li et al. [26] employed stacked autoencoder to train Landsat images. Compared with random forest, SVM and artificial neural network (76.03%, 77.74%, 77.86%), the stacked autoencoder produces 78.99% with 1% improvement.

Although some deep learning-based algorithms have been introduced and tested for Landsat images, the classification results are still not satisfied. Considering the rough texture features and various sizes of Earth objects in large scale image captured by Landsat, a new deep learning-based algorithm named PSPNet [18] is employed to train the samples of Landsat images. In the early layers, stacked residual blocks are able to learn the spectral and spatial features of Landsat images by concatenating the extracted feature maps before and after the

convolutional layers. Then the pyramid pooling combines the information obtained with various pooling scales. By treating the reference land cover map as the ground truth, we can achieve overall classification accuracy more than 80% for Landsat images covering the Jingjinji Area of China.

The rest of this paper is organized as follows. Section 2 introduces the research area and reference land cover map. The methodology is presented in Section 3. Section 4 gives the experimental results and analyses. Section 5 draw the conclusions.

II. Research Area and Reference Land Cover Map

Jingjinji Area is an important local region in China, which contains Beijing city, Tianjin city and Hebei province. It locates at the North China Plain with the Bohai Sea to the east, the Taihang Mountain in the west and the Yanshan (Mountain) in the north. It is chosen as the research area in this paper because it contains different land-cover and land-use types such as forest, grass, crop, water, urban area, village, bare lands and various terrains such as highland, mountains, hills, basins, plains, seashores, etc. The Landsat image of the Jingjinji Area is shown in Fig. 1 which is composed of 16 Landsat TM images captured in the growing season of 2010. Due to its complexity in land cover types and terrains, objects belonging to the same class present different characteristics in both texture and spectral aspects.

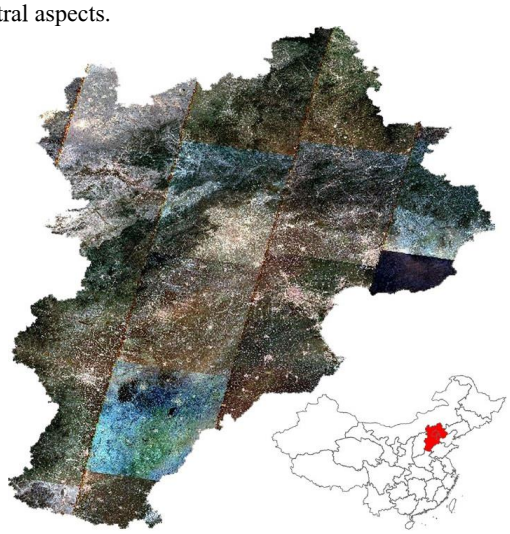

Fig. 1: Study area in the Jingjinji Area of China, Landsat 5 TM images.

"Land Cover Map of the People's Republic of China for 2010" is considered as the reference land cover map to train the CNN since its accuracy reaches 91%, which was validated through 111,356 ground samples all over China. The reference land cover map of the Jingjinji Area was achieved from the National Earth System Science Data Sharing Infrastructure, National Science & Technology Infrastructure of China (http://www.geodata.cn). Analysis on ADE20K shows that even for the entire dataset labeled by the expert annotator, 17.6% pixels in the dataset may still exist as errors [27]. Considering training results of ADE20K and the 30m resolution of Landsat image, the reference land cover map with the accuracy of 91% will meet the training requirements. In this paper, we adopt a part of the first level classification scheme, which contains forest lands, grass lands, water & wetlands, crop lands, built-up lands and bare lands as shown in Fig. 2.

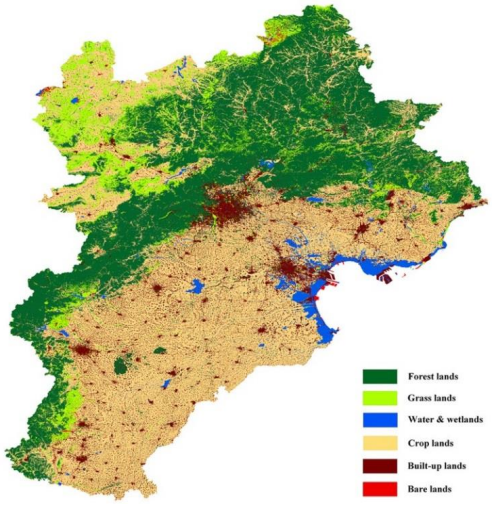

Fig. 2: Reference land cover map with a part of first level classification scheme.

III. Methodology

The flowchart of training PSPNet with Landsat samples is shown in Fig. 3. The training samples fed to PSPNet are oversampled from Landsat images and the corresponding reference land cover map. Then, the residual blocks of PSPNet learn the spatial and spectral features of the input Landsat images and pass it to the pyramid pooling block. Four pooling sizes are applied on the features obtained from residual blocks to utilize the global information and maintain the detail structures at the same time. Finally, the inferenced results are compared with corresponding reference land cover map. By minimizing the loss function between the inference result and reference land cover map, optimal parameters for the employed CNN model are obtained.

A. Overall Architecture

The texture information of Landsat images is not as fine as it appeared in high resolution remote sensing images. Besides, objects captured with 30m resolution are various in sizes. PSPNet [18] is employed in this paper due to the following reasons: 1) The Resnet [15] part in the former layers of PSPNet utilizes the information both before and after convolution, and the stacked residual blocks are able to learn the features of Landsat images efficiently; 2) The pyramid pooling takes the sizes of objects into account, which improves the recognition ability on objects with various sizes, such as built-up lands.

PSPNet is an end to end CNN which takes an image as the input and outputs its classification result. Unlike other deep learning networks (which take a pixel or a small area as input), it is capable of learning the global texture information, which is the key information in recognizing different classes in Landsat images. It uses Resnet to extract features of Landsat images and then down-samples the features with four pyramid

scales. One of the pyramid scales is set to one to capture overall information of the input images. The others are used to learn detailed information of objects with different sizes. To evaluate the differences between inference results and the reference land cover maps, cross entropy is employed as the loss function. Then the weights of convolutional kernels are learned by minimizing the loss function.

B. *Sample Selection Strategy*

Landsat image classification is usually used for a large scale land cover mapping. However, the same kind of objects presents different spectral and texture information in different areas. Therefore, the selection of training samples should take the distribution of objects into account. Grass lands are mainly distributed in the northwest of the Jingjinji Area and cover only a small area. Most part of this area is selected as training samples except the regions which are near to the boundary. As for the regions of the Yanshan, the Taihang Mountain and the North China Plain, which occupy a large proportion in the research area, only small parts of them are selected as training samples.

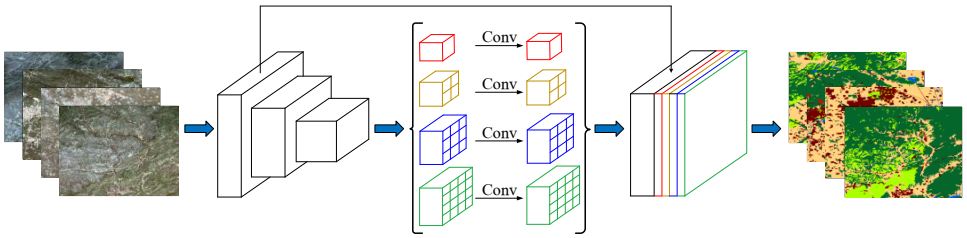

Fig. 3: The flowchart of training PSPNet with Landsat samples.

As the same as other deep learning networks, PSPNet is also a data driven method which performs well only when enough training samples are fed. Besides, the convolution window restricts the usage of pixels near the boundary, and it results in the difference of accuracy between the middle and boundary of the image. To overcome this problem and enlarge the training samples, this paper oversamples the selected area with a small offset between them. By this way, all objects have a chance to be in the center of the input image which means their features can be learned efficiently during training. Meanwhile, the training sample set is also enlarged.

C. *Data Augmentation*

The square window in convolutional layer is sensitive to the orientation of textures, so traditional data augmentation methods rotate the input images with different degrees to ensure that the architecture can effectively learn the features from the input images. Nevertheless, directly rotating an image needs to be padded by zeros. Although randomly cropping a small subimage with different angles from a larger one is equivalent to rotating, the information may not be fully exploited in the corners. Consequently, random flip is employed to increase the quantity of training samples in this paper.

Normalization is another data augmentation method which adjusts the input images to a comparable scale. As shown in Fig. 1, spectral information between scenes of Landsat images are obviously different. After the normalization, differences between images decrease. It helps PSPNet to learn the rightful features of the detected objects and accelerate the process in some extent.

Normalization is performed on every image fed to CNN while flip is random. The mentioned data augmentation method is used in the training process of this paper. Under hundreds of epochs of training, it is equivalent to double the quantity of the training samples.

D. *Parameter Fine Tuning*

In this paper, 1520 images are obtained by the proposed sample selection strategy with the window size of 640×640 pixels. These images are divided into two parts: 1248 images are training samples and 272 images are validation samples. The obtained training sample set of Landsat images is too small compared with 1.26 million pictures in ImageNet. Therefore, parameters of Resnet50 pretrained on ImageNet are employed as the initial parameters and the PSPNet is finely tuned by the selected Landsat samples to make the trained model be suitable for Landsat image classification. In addition, atrous-convolution is used to increase the receptive field in the last two residual blocks. The model is trained on TITAN XP, in which the momentum in the batch normal operation is set to be 0.1, the learning rate is 10^{-9} and the maximum iteration time is set to be 1 million.

IV. RESULTS AND ANALYSES

The PSPNet employed in this paper consumed about a week to train the model. Then the model is applied on all the Landsat images covering the Jingjinji Area. The classification result is shown in Fig. 4. From Fig. 4, it can be clearly seen that forest lands distributed in the Yanshan and the Taihang Mountain. Grass lands locate along with the mountains and distributed in the northwest of the Jingjinji Area. Water & wetland mainly concentrated in the Bohai Bay. Crop lands and built-up lands dominate in the North China Plain.

Accuracy of a classification result is the ratio of the correctly classified number of pixels to the total number of pixels in a class. The Intersection-over-Union (IoU) stands for the ratio of the intersection dividing the union of the inferenced results and reference land cover map. F1 score is a weighted average of the precision and the recall of the model which can be used to evaluate the accuracy of binary classification. To calculate the

above-mentioned evaluation indicator, the refence land cover map is considered as the ground truth. The accuracy, IoU and F1 score of each class and the overall image are shown in Fig. 5.

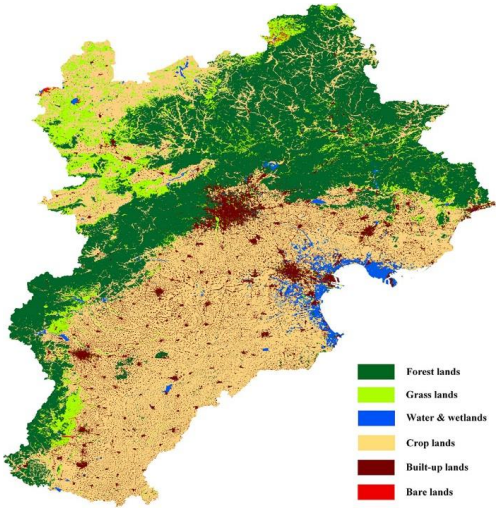

Fig. 4: Classification result of the Jingjingji Area based on PSPNet.

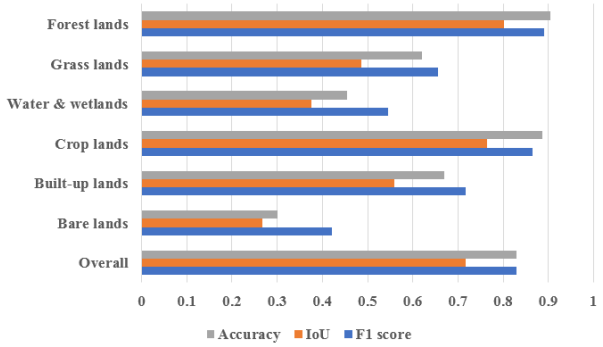

Fig. 5: IoU and F1 score of each class shown in Fig. 4.

Forest lands and crop lands have the first and second highest accuracy, IoU and F1 scores while the accuracy of bare lands is the lowest among all the six classes. The reference land cover map in the research area was composed by forest lands, grass lands, water & wetlands, crop lands, built-up lands and bare lands with their percentages of 33.5%, 9.5%, 2.8%, 43.9%, 9.9% and 0.4%, respectively. Compared with the proportion of each class, it can be found that the classification accuracy is positively correlated with the quantity of training set. The area of water & wetlands and bare lands are less than forest lands and crop lands, even though they are oversampled. Besides, the water & wetlands with large area near the Bohai Sea was not selected because it is near to the boundary in the reference land cover map. If it is selected as the training sample, the features of the images without reference land cover map will obviously affect the features learned by CNN. Similar situation appears to grass lands. Therefore, the classification accuracy of built-up lands is higher than grass lands since it distributes in the middle of reference land cover map while grass lands lay near the boundary. Nevertheless, the overall accuracy, average IoU and average F1 score of the whole image reach 0.83, 0.83 and 0.72, respectively.

The magnification of details with typical regions in Fig. 4 is shown in Fig. 6. Fig. 6 (a1)-(d1) are the prairie in the northwest, the mountain area directed from northeast to southwest, the plain locates in the southeast, and the seashores near the Bohai Sea. Fig. 6 (a2)-(d2) are classification results of the corresponding area. From Fig. 6 it can be found that the classification results can recognize the grass lands, forest lands, built-up lands and crop lands correctly and their boundaries are satisfied. PSPNet learns the features of detected objects through stacked convolution operation, so it is expert in learning the spectral and texture features of areal structured objects. Therefore, the mentioned classes are classified with high accuracy. On the contrary, linear structured objects such as wetlands cannot be accurately recognized. The reasons are the over down sample during the pooling operation and the lack of training examples. In addition, regions near seashore are not selected as the training samples. Features of herbaceous wetland covering large area is not efficiently learned. Therefore, the classification accuracy of water & wetlands is not satisfied as shown in Fig. 6 (a2) and seashore in Fig. 6 (d2).

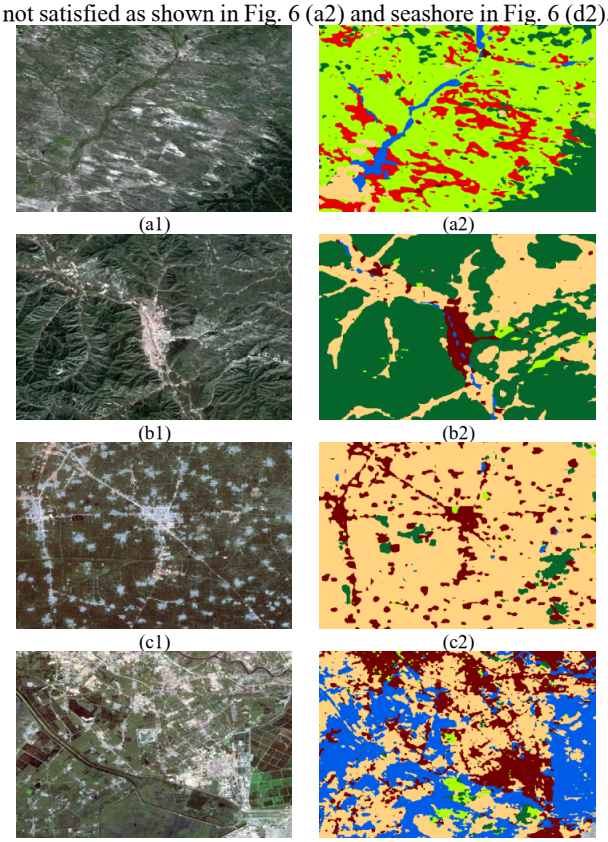

Fig. 6: Magnification of details with typical terrain, in which (a1)-(d1) are original Landsat images and (a2)-(d2) are corresponding classification results achieved by PSPNet.

In the experiments we also found that PSPNet has a strong ability of generality if enough training samples are fed. In the research area, a lot of samples representing the forest lands and

crop lands are selected to support the PSPNet to learn the features of forest lands with various forms. Even though the classification result in the reference land cover map confuses them, it is correctly recognized by PSPNet as shown in Fig. 7. In Fig. 7, (a1)-(c1) are the original Landsat images, (a2)-(c2) are their corresponding references land cover maps, and (a3)-(c3) are classification results achieved by PSPNet. Fig. 7 (a2) contains forest lands, crop lands, built-up lands and water & wetlands. However, some parts of forest lands have no differences with the crop lands such as areas visually represented by dark green. Some of the brighter areas along the road in the right part of the image are labeled as built-up lands while the object in the box is recognized as bare land in the reference land cover map. Although PSPNet did not recognize the small built-up lands in the box, it correctly classifies the crop lands which are labeled as forest land in the reference land cover map. Fig. 7 (b1) is a part of mountain area, the valley of the mountains is built-up lands which are correctly recognized in the reference land cover map and classification results of PSPNet. Nevertheless, grass lands obtained from reference land cover map as shown in Fig. 7 (b2) cannot be distinguished by human eyes from the forest lands. On the other hand, grass lands obtained from PSPNet have distinctly different features from forest lands. Fig. 7 (c1) is a mosaic Landsat image without uniform color. Somehow, the reference land cover map considered some of crop lands as forest lands as shown in the dark green part in Fig. 7 (c2). Since this part is selected as the training sample, the learned features in PSPNet have to balance the conflict between the features of this part and its surroundings. Therefore, the misclassified areas of forest lands are less than the training sample selected from the reference land cover map. In conclusion, the objects in the reference land cover map are more fragment than those in classification results achieved by PSPNet, especially for the built-up lands. The boundaries of built-up lands are hackly while boundaries in the classification results of PSPNet are much smoother. Besides, it is believed that PSPNet is capable of putting right some small misclassification in the training samples once enough training samples are fed. However, it is not good at recognizing objects with linear structure and will miss objects positioned in complex surroundings such as the part in the box in Fig. 7 (a1).

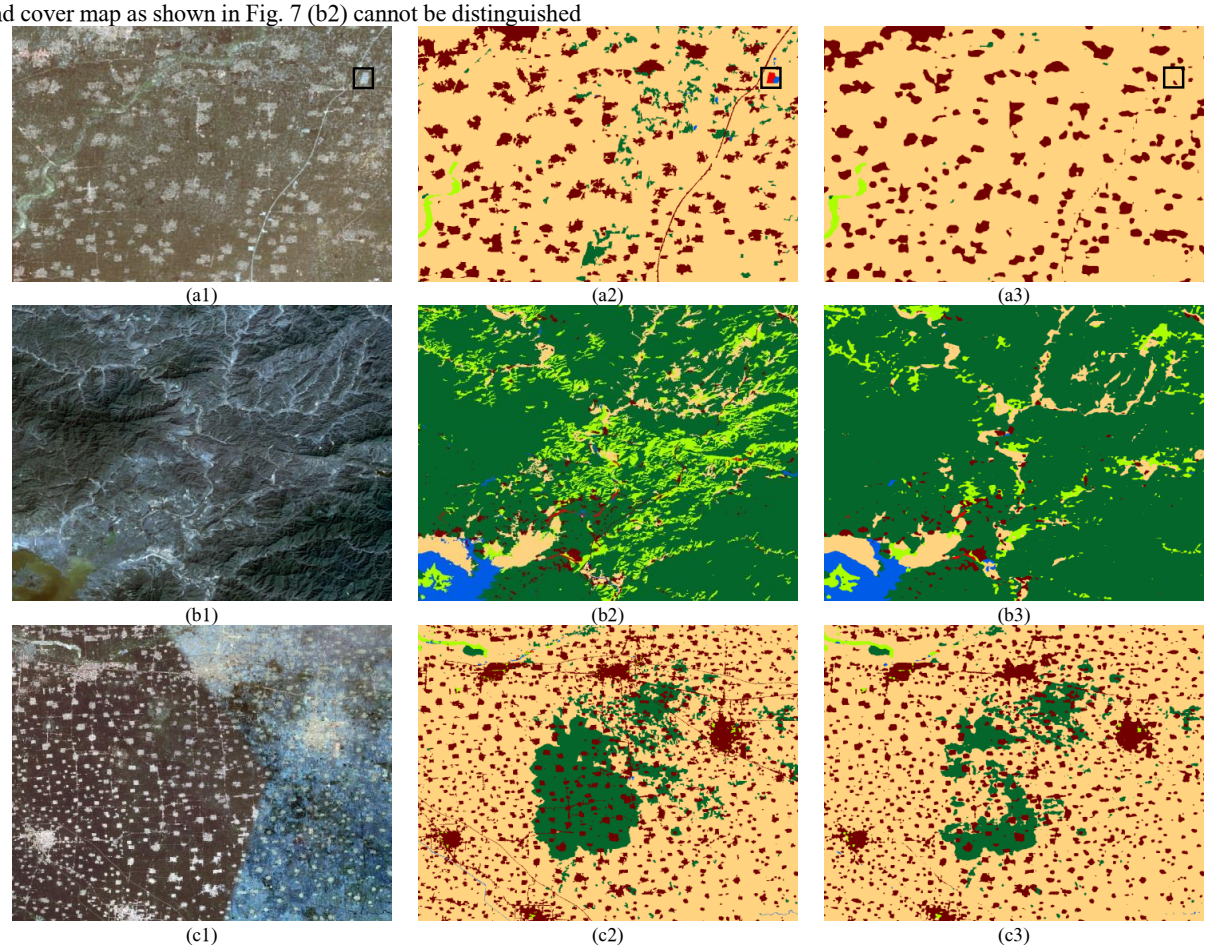

Fig. 7: Examples of classification results which are better than reference land cover map. (a1)-(c1) are original Landsat images; (a2)-(c2) are corresponding reference land cover maps; (a3)-(c3) are classification results of PSPNet.

V. CONCLUSIONS

This paper proposed a new deep learning-based algorithm named PSPNet for large scale Landsat image classification. From the comparison of the classification results between PSPNet and the reference land cover map, several conclusions can be achieved as follows: 1) PSPNet is capable to learn good feature representation for classification if the training samples are enough and correct; 2) The classification accuracy of linear objects is not high since the windows employed in convolution operation is rectangle; 3) It obtains correct classification results even if there exist some errors in the training sample; 4) The domain of error training samples may affect the accuracy of the classification result if it occupies a large area in one of the training samples since classification accuracy of each image is considered in the loss function.

From the experiments of the Jingjinji Area with PSPNet, it also can be found that PSPNet is able to well learn the features of Landsat image and obtains satisfied classification result when the reference land cover map is accurate enough, as well as the training sample is sufficient. However, we should also note that this paper just made an attempt to classify large scale Landsat images by using deep learning techniques. The classification accuracy and results are still not satisfied. In the future, multi-temporal features, new sample selection strategy and multi-algorithm fusion will be considered. The proposed framework will also be tested by using other typical Landsat images over other areas in China.

ACKNOWLEDGMENT

This research was supported by the Strategic Priority Research Program of the Chinese Academy of Sciences under Grant No. XDA19080302, and by the 62-class General Financial Grant from the China Postdoctoral Science Foundation under Grant No. 2017M620947.

Acknowledgement for the data support from "National Earth System Science Data Sharing Infrastructure, National Science & Technology Infrastructure of China. (http://www.geodata.cn)".

REFERENCES

[1] Gong P, Wang J, Yu L, Zhao Y, Zhao Y, Liang L, et al. Finer resolution observation and monitoring of global land cover: first mapping results with Landsat TM and ETM+ data, International Journal of Remote Sensing, 2013, 34(7): 2607-2654.

[2] Chen J, Chen J, Liao A, Cao X, Chen L, Chen X, et al. Global land cover mapping at 30m resolution: a POK-based operational approach, ISPRS Journal of Photogrammetry and Remote Sensing, 2015, 103: 7-27.

[3] Zhao Y, Feng D, Yu L, Wang X, Chen Y, Bai Y, et al. Detailed dynamic land cover mapping of Chile: accuracy improvement by integrating multi-temporal data, Remote Sensing of Environment, 2016,183: 170-185.

[4] Zhang Z, Wang X, Zhao X, Liu B, Yi L, Zuo L, et al. A 2010 update of national land use/cover database of China at 1:100000 scale using medium spatial resolution satellite images, Remote Sensing of Environment, 2014, 149: 142-154.

[5] Crosby M K, Matney T G, Schultz E B, Evans D L, Grebner D L, Londo H A, et al. Consequences of Landsat images strata classification errors on bias and variance of inventory estimates: a forest inventory case study, IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 2017, 10(1): 234-251.

[6] Goldblatt R, Stuhlmacher M F, Tellman B, Clinton N, Hanson G, Georgescu M, et al. Using Landsat and nighttime lights for supervised pixel-based image classification of urban land cover, Remote Sensing of Environment, 2018, 205: 253-275.

[7] Bruzzone L, Prieto D F. Unsupervised retraining of a maximum likelihood classifier for the analysis of multitemporal remote sensing images, IEEE Transactions on Geoscience and Remote Sensing, 2001, 39(2): 456-460.

[8] Polat K, Gunes S. A novel hybrid intelligent method based on C4.5 decision tree classifier and on-against-all approach for multi-class classification problems, Expert System with Applications, 2009, 36(2): 1587-1592.

[9] Belgiu M, Dragut L. Random forest in remote sensing: a review of applications and future directions, ISPRS Journal of Photogrammetry and Remote Sensing, 2016, 114: 24-31.

[10] Skakun S, Kussul N, Shelestov A Y, Lavreniuk M, Kussul O. Efficiency assessment of multitemporal C-band radarsat-2 intensity and Landsat-8 surface reflectance satellite imagery for crop classification in Ukraine, IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 2016, 9(8): 3712-3719.

[11] Zhao F, Huang C, Zhu Z. Use of vegetation change tracker and support vector machine to map disturbance types in Greater Yellowstone ecosystems in a 1984-2010 Landsat time series, IEEE Geoscience and Remote Sensing Letters, 2015, 12(8): 1650-1654.

[12] Krizhevsky A, Sutskever I, Hinton G. ImageNet classification with deep convolutional neural networks. In Proceedings of Advances in Neural Information Processing Systems, 2012, 25: 1090–1098.

[13] Szegedy C, Liu W, Jia Y, Sermanet P, Reed S, Anguelov D, et al. Going deeper with convolutions, IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 1-9.

[14] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition, arXiv preprint arXiv:1409.1556, 2014.

[15] He K, Zhang X, Ren S, Sun J. Deep Residual Learning for Image Recognition, IEEE Conference on Computer vision and Pattern Recognition, 2016, pp. 770-778.

[16] Long J, Evan S, Trevor D. Fully Convolutional Networks for Semantic Segmentation, IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 3431-3440.

[17] Chen L, Papandreou G, Kokkinos I, Murphy K, Yuille A L. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs, IEEE Transactions on Pattern Analysis and Machine Intelligence, 2018, 40(4): 834-848.

[18] Zhao H, Shi J, Qi X, Wang X, Jia J. Pyramid Scene Parsing Network, IEEE Conference on Computer Vison and Pattern Recognition, 2017, pp. 2881-2890.

[19] Langkvist M, Kiselev A, Alirezaie M, Loutfi A. Classification and segmentation of satellite orthoimagery using convolutional neural networks, Remote Sensing, 2016, 8:1-21.

[20] Zhao W, Du S. Spectral-spatial feature extraction for hyperspectral image classification: A dimension reduction and deep learning approach, IEEE Transactions on Geoscience and Remote Sensing, 2016, 54(8): 4544–4554.

[21] Yu L, Wang Z, Tian S, Ye F, Ding J, Kong J. Convolutional neural networks for water body extraction from Landsat imagery, International Journal of Computational Intelligence and Applications, 2017, 16(1): 1750001.
[22] Kussul N, Lavreniuk M, Skakun S, Shelestov A. Deep learning classification of land cover and crop types using remote sensing data, IEEE Geoscience and Remote Sensing Letters, 2017, 14(5): 778-782.
[23] Perez A, Yeh C, Azzari G, Burke M, Lobell D, Ermon S. Poverty Prediction with Public Landsat 7 Satellite Imagery and Machine Learning, 2017, arXiv preprint arXiv:1711.03654.
[24] Kussul N, Shelestov A, Lavreniuk M, Butko I, Skakun S. Deep learning approach for large scale land cover mapping based on remote sensing data fusion, 2016 IEEE International Geoscience and Remote Sensing Symposium, 2016, pp.198-201.
[25] Ikasari I H, Ayumi V, Fanany M I, Mulyono S. Multiple regularizations deep learning for paddy growth stages classification from LANDSAT-8, 2016 International Conference on Advanced Computer Science and Information Systems, 2016, pp. 512-517.
[26] Li W, Fu H, Yu L, Gong P, Feng D, Li C, et al. Stacked autoencoder-based deep learning for remote-sensing image classification: a case study of African land-cover mapping, International Journal of Remote Sensing, 2016, 37(23): 5632-5646.
[27] Zhou B, Zhao H, Puig X, Fidler S, Barriuso A, Torralba A. Semantic understanding of scenes through the ADE20K dataset, arXiv: 1608.05442, 2016.