



Built-Up Area Extraction from Landsat 8 Images Using Convolutional Neural Networks with Massive Automatically Selected Samples

Tao Zhang and Hong Tang^(✉)

Beijing Key Laboratory of Environmental Remote Sensing and Digital Cities,
Beijing Normal University, Beijing 100875, China
201721480054@mail.bnu.edu.cn,
{hongtang, tanghong}@bnu.edu.cn

Abstract. Extraction of built-up area (e.g., roads, buildings, and other Man-made object) from remotely sensed imagery plays an important role in many urban applications. This task is normally difficult due to complex data in the form of heterogeneous appearance with large intra-class variations and lower inter-class variations. In order to extract the built-up area of 15-m resolution based on Landsat 8-OLI images, we propose the convolutional neural networks (CNN) which built in Google Drive using Colaboratory-Tensorflow. In this Framework, Google Earth Engine (GEE) provides massive remote sensing images and preprocesses the data. In this proposed CNN, for each pixel, the spectral information of the 8 bands and the spatial relationship in the 5 neighborhood are taken into account. Training this network requires lots of sample points, so we propose a method based on the ESA's 38-m global built-up area data of 2014, Open Street Map and MOD13Q1-NDVI to achieve rapid and automatic generation of large number of sample points. We choose Beijing, Lanzhou, Chongqing, Suzhou and Guangzhou of China as the experimentation sites, which represent different landforms and different urban environments. We use the proposed CNN to extract built-up area, and compare the results with other existing building data products. Our research shows: (1) The test accuracy of the five experimental sites is higher than 89%. (2) The classification results of CNN can be very good for the details of the built-up area, and greatly reduce the classification error and leakage error. Therefore, this paper provides a reference for the classification and mapping of built-up areas in large space range.

Keywords: Extract built-up area · CNN · Landsat 8 image · Generate samples
Google Earth Engine · Google Colaboratory

1 Introduction

Built-up area refers to the land of urban and rural residential and public facilities, including industrial and mining land, energy land, transportation, water conservancy, communication and other infrastructure land, tourist land, military land and so on.

This work was supported by the National Key R&D Program of China (No. 2017YFB0504100).

© Springer Nature Switzerland AG 2018

J.-H. Lai et al. (Eds.): PRCV 2018, LNCS 11257, pp. 492–504, 2018.

https://doi.org/10.1007/978-3-030-03335-4_43

Built-up is one of the most important elements of land use, and plays an extremely important role in urban planning. In the process of urban development, on the one hand, the expansion of the built-up area provides suitable sites for industrial production, economic activities, and people living. On the other hand, the built-up area has profoundly changed the natural surface of the region, and then affects the natural processes of heat exchange, hydrological process and ecosystem [1].

Built-up area extraction based on remote sensing image has always been a research hotspot. In [2–4], normalized difference building index (NDBI), index-based build-up index (IBI) and texture-derived built-up presence index (PanTex) are separately proposed in order to extract buildings. But these methods have strong dependence on threshold selection, and how to find the suitable threshold is very difficult. In recent years, very high resolution and hyperspectral images have been gradually used in building extraction. Many methods which are based on morphological filtering [5], spatial structure features [6], grayscale texture features [7], image segmentation [8] and geometric features [9] have been applied to building extraction. But these methods are difficult to consider spectral information and spatial structure information at the same time. With the rapid development of convolution neural network and deep learning, especially the excellent performance of deep convolution neural network on ImageNet contest [10–13], CNN has shown great advantages in image pattern recognition, scene classification [14], object detection and other issues. More and more researchers have applied CNN to remote sensing image classification. In [15–17], CNN of different structures have been used for building extraction. In [24], Deep Collaborative Embedding (DCE) model integrates the weakly-supervised image-tag correlation, image correlation and tag correlation simultaneously and seamlessly for understanding social image. These studies were based on open data sets within small range and ground truth, which made great progress in the methodology. However, there are few researches on image classification in large areas. For 15-m built-up area extraction based on CNN in large region, there are three key issues: (1) Acquisition and preprocessing of huge Landsat 8-OLI images. (2) Generate a large number of accurate sample points applied to train CNN. (3) CNN framework using multiband-spectral and spatial information simultaneously. In this paper, we study these three points and try to find a solution.

A typical pipeline for thematic mapping from satellite images consists of image collection, feature extraction, classifier learning and classification. This pipeline works more and more difficult along with increasing of mapping geographical scope or spatial resolution of satellite images. The difficulty originates from the inefficiency of data collection and processing and the ineffectiveness of both classifier learning and prediction. One could be partly relieved from the inefficiency of the traditional pipeline by using the data on the Google Earth Engine (GEE). A multi-petabyte catalog of satellite imagery and geospatial datasets have been collected in GEE and freely available to everyone [18]. Some functions of geo-computation are also available to users by online interface or offline application programming interfaces. However, many advanced functions of machine learning are still out of the GEE, for example deep learning technologies.

So we introduce an effective way to rapid extract built-up areas by making full use of huge data on the GEE, high speed computation of the Google Cloud Storage (GCS) and the contemporary machine learning technologies within Google Colaboratory (Colab). As shown in Fig. 1, GCS can be used as a bridge between the GEE and

Colab. First of all, Landsat 8 images on the GEE are selected and preprocessed, e.g., cloud mask and mosaicking. Massive training samples are automatically selected from low-resolution land-cover production using a preset rule. Then, a Convolutional Neural Network (CNN) is designed based on the TensorFlow using Colab. Finally, the CNN is trained and utilized to rapid mapping of built-up areas within the GCS using the Landsat 8 images and training samples transferred from the GEE.

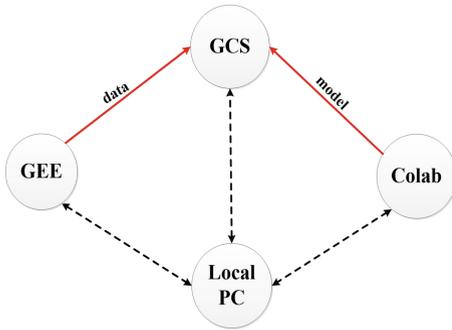


Fig. 1. Flowchart of extracting built-up area

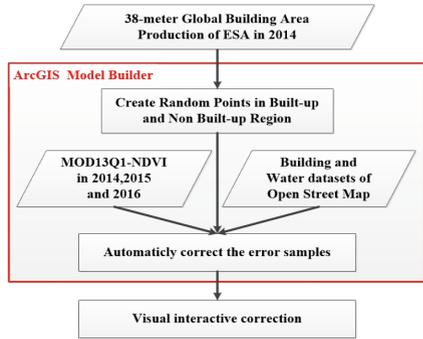


Fig. 2. Sample generation and correction

2 Study Area and Data

China has vast territory, complex and changeable climate types, and different landforms in different regions. In order to train a CNN with strong universality and robustness, the choice of Landsat 8 images needs to consider multiple regions and multi time period. In this paper, we choose Beijing, Lanzhou, Chongqing, Suzhou and Guangzhou as the experimental region, as shown in Table 1: the five regions represent the situation of the cities under different climates and landforms in China. We chose Landsat 8 images on GEE, taking into account the large amount of cloud cover in spring and autumn, so the date of image is mainly in summer. And in order to ensure data quality, the cloud coverage of each image is less than 5%. The image of each region was preprocessed by mosaicing and cutting on GEE. Finally, in each region, the size of image at 15-m resolution is 12000 * 12000. As shown in Fig. 3, A, B, C, D and E are the five experimental region numbers selected respectively. The false color (7, 6, 4 band combination) shows that the quality of the data is good and meets the requirements.

The training and testing samples are automatically selected from the 38-m global built-up production of ESA in 2014 [21]. Based on the ArcGIS-Model Builder tool (Fig. 4 is the model created), a large number of sample points are automatically generated, filtered and corrected. As shown in Fig. 2, the detailed process includes three steps: (1) Randomly select 20 thousand sample points in each experimental area. (2) Use the buildings and water data sets of Open Street Map (OSM) in China and

Table 1. The information of Landsat 8 images in the five experimental region

Code	Test sites	Characteristics	Climates	Path/row	Date	Cloud cover (%)
A	Beijing	The political center, the mega city, the buildings are densely distributed	The semi humid continental monsoon climate in the north temperate zone	122/32	2015-08-15	0.1
				122/33	2015-06-12	0.43
				123/32	2015-04-16	0.25
				123/33	2015-05-18	0.44
				124/31	2015-09-14	0.05
				124/32	2015-05-25	0.06
				124/33	2015-08-13	0.72
B	Lanzhou	City in the gobi. Buildings are distributed along the river	The temperate continental climate	130/34	2015-10-10	0.27
				130/35	2015-10-10	0.07
				130/36	2015-07-06	1.48
				131/34	2015-06-11	0.37
				131/35	2015-06-11	0.76
C	Chongqing	The terrain is undulating. Buildings are distributed in valley and low-lying land	Subtropical monsoon humid climate	127/39	2016-05-16	0.92
				127/40	2016-06-17	2.11
				128/39	2014-08-06	2.92
				128/40	2014-08-06	3.76
D	Suzhou	Too many rivers and lakes. The distribution of buildings is broken	Subtropical monsoon ocean climate	118/38	2015-08-03	0.33
				118/39	2015-08-03	0.5
				119/38	2015-10-13	4.37
				119/39	2015-10-13	1.01
E	Guangzhou	Pearl River Delta, Developed economy, Complex building types	Subtropical oceanic monsoon climate	122/43	2015-10-18	0.03
				122/44	2015-10-18	1.15
				123/43	2015-04-16	0.14
				123/44	2015-04-16	0.09

MOD13Q1-NDVI data to filter and correct the selected sample points. The aim is to modify the built-up sample points in the vegetation area and the water body into non-built-up sample points, and to modify the non-built-up sample points in the built-up area into the built-up sample points. (3) Combine with ArcGIS Online Image for manual correction. Finally, accurate sample points of built-up area and non-built-up area are obtained. For each experimental region, the corrected sample points were hierarchically divided into training samples and test samples at 50% ratio. Table 2 shows the number of samples that will eventually be used for training and testing.

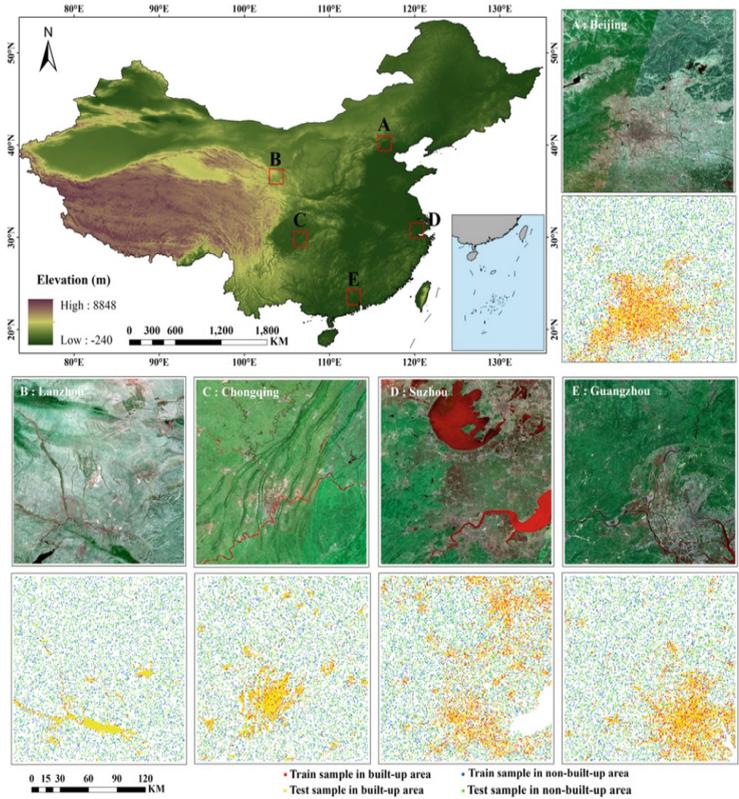


Fig. 3. Location of the five experimental sites and the all sample points (Color figure online)

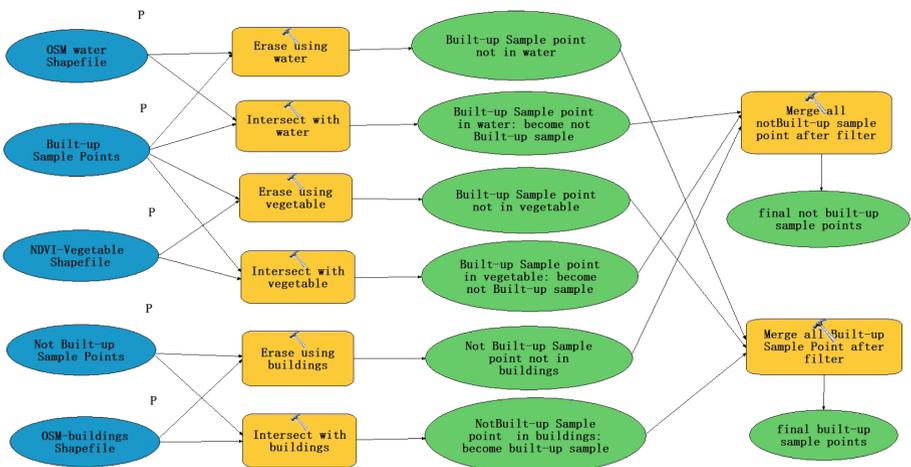


Fig. 4. Correct error sample points by ArcGIS Model Builder

Table 2. The final number of training and testing samples

Test sites	Training samples		Testing samples		Sum
	Built-up area	Non built-up area	Built-up area	Non built-up area	
A	4973	5026	4973	5026	19998
B	4760	5000	4762	4999	19521
C	4906	5017	4905	5017	19845
D	4950	5042	4950	5043	19984
E	4955	5038	4954	5037	19984

3 The Proposed CNN

There are multiple spectral bands in Landsat 8 OLI sensors. Many empirical indexes have been shown to be effective to characterize land-cover categories by a combination of a subset Landsat bands, for example the Normalized Urban Areas Composite Index [19] and normalized building index [3]. In addition, Yang et al. have shown that the combination of a subset of spectral bands can promote the classification accuracy of convolutional neural networks [20]. This motivate us to use the networks as shown in Fig. 5, which consists of input layer, convolution layer, full-connected layer and output layer.

As for the input layer, all of the first seven bands of the Landsat 8 OLI images are up-sampled to 15 m using the nearest neighborhood sampling. Then the up-sampled seven bands are stacked with the panchromatic band. In the 15-m resolution image, the size of the building is generally less than 5 pixels. For each pixel, the 5-neighborhood is considered, which means the size of image patch is $5 * 5 * 8$. So an image patch with 8 bands and $5 * 5$ neighborhood centered on each sample is inputted into the neural

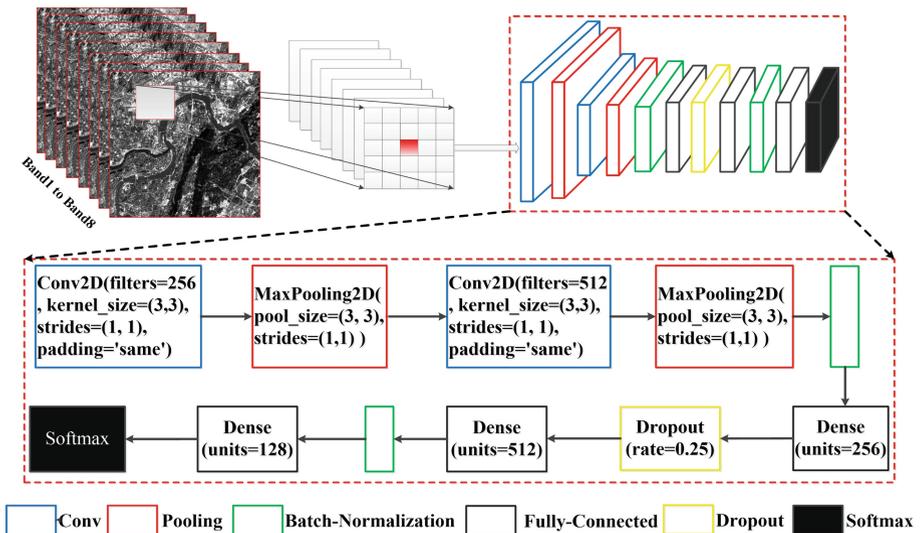


Fig. 5. The proposed CNN framework

network. Within the convolutional layer, there are two convolution and two max-pooling layers, which aim to extract spectral features and spatial features, even more high-grade features. In the fully-connected layer, we use three fully-connected layers. In order to prevent over fitting, one batch-normalization layer and a dropout layer are added. The output layer consists of a soft max operator, which outputs two categories. In the whole network, we use the popular function called Rectified Linear Unit (ReLU) solving the vanishing gradient problem for training epochs in the network.

4 Experiments and Data Analysis

There are a total of 44717 training sample points, including 24544 built-up area samples and 20173 non built-up area samples. We divide the training samples into training set and validation set according to the ratio of 7:3. For the CNN, we input training data and set up some hyper-parameters. The batch size was set to 128 and the epoch was set to 200. We use cross entropy as the loss function and use stochastic gradient descent (SGD) as the optimizer to learn the model parameters. The initial learning rate of SGD is set to 0.001. When the learning rate is no less than 0.00001, the learning rate decreases exponentially with the increase of training epoch-index. The calculation formula of the learning rate is as follows:

$$\begin{aligned} & \text{initial_lrate} = 0.001 \\ & \text{if}(\text{lrate} < 0.00001) : \text{lrate} = 0.00001 \\ & \text{else} : \text{lrate} = \text{initial_lrate} * 0.5^{\frac{1+n}{10}} \end{aligned}$$

In this formula, ‘Initial_lrate’ represents the initial learning rate, ‘lrate’ represents the learning rate after each iteration, and ‘n’ indicates the epoch of iteration.

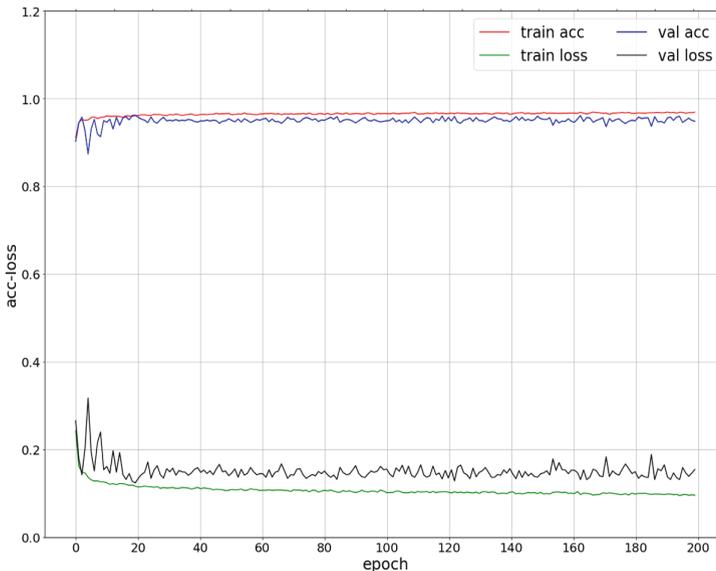


Fig. 6. Accuracy and loss in the training process

As shown in Fig. 6, after the thirtieth epoch, the model tends to converge, and the values of training accuracy and verification accuracy are stable at 0.96 and 0.95, but the value of training loss and verification loss still slows down slowly. At the thirtieth epoch, the time is 50 min. When the 200 epochs is completed, it takes 368 min. Using the trained network to classify the five experimental sites, and calculate the classification accuracy and loss value of the test samples. As shown in Table 3, the test accuracy of the five experimental sites is higher than 0.89, and it can be seen from Fig. 7 that the result of extracting built-up area is very good. The test accuracy of Chongqing and Lanzhou is the highest, 0.95 and 0.94 respectively, while the test accuracy of Suzhou and Guangzhou is lower, and their values are 0.92 and 0.91 respectively. Beijing has the lowest test accuracy, only 0.89. According to our priori knowledge, Beijing is a mega city, and the buildings are densely distributed and ground surface coverage is mixed and complex. So there are great difficulties to classify built-up and non-built-up area accurately.

Table 3. The test accuracy loss in the five experimental sites

Experimental site	A	B	C	D	E
Test accuracy	0.89	0.95	0.94	0.92	0.91
Test loss	0.37	0.20	0.24	0.23	0.34

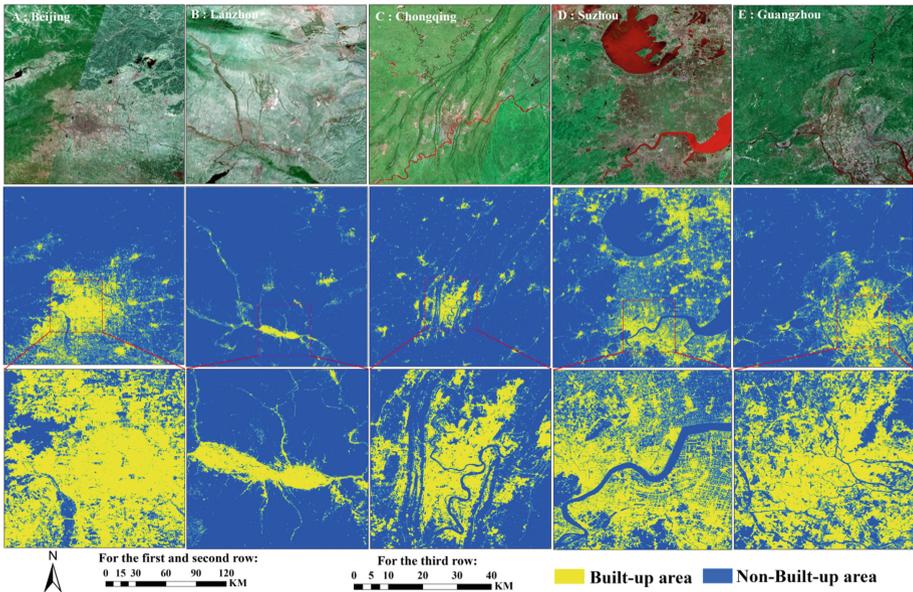


Fig. 7. Results of built-up area by CNN in the five experimental sites

5 Discussion

5.1 Qualitative Comparison

For the sake of qualitative comparison, we compare the results of the proposed CNN with that of Global-Urban-2015 [22], GHS-BUILT38 produced by ESA [21] and GlobalLand30 [23] in the C experimental site: Chongqing. Three small regions with the size of 1000 * 1000 pixels are separated from the Chongqing region, numbered as 1, 2 and 3 in Fig. 8.

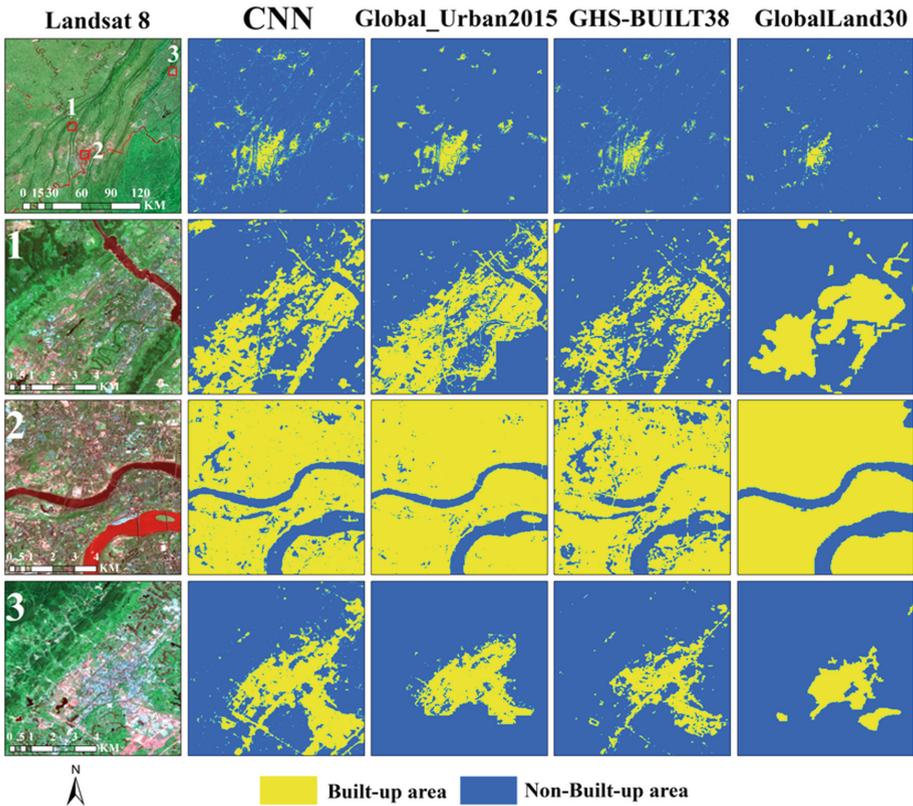


Fig. 8. The classification results. The first column shows the experimental images and their three sub-regions. The other columns exhibit the classification results of CNN, Global_Urban 2015, GHS-BULT 38, and GlobalLand 30, respectively.

As shown in Fig. 8, we can find many details in the results of the CNN, which are missing in the other productions. One of reasons is that the result of the CNN is produced from satellite images with higher spatial resolutions. Consequently, within the urban area, non-built-up area, e.g., water and vegetation in the dense buildings, can be discriminated from built-up area within the Landsat 8 images. Meanwhile, in the

suburbs, small size of built-up areas and narrow roads become distinguishable from the background. Another reason is due to the higher classification accuracy of the CNN.

5.2 The Proposed CNN VS VGG16

We choose Beijing as experimental site, and compare the results of the proposed CNN with VGG16 [11]. As shown in Fig. 9, We reserve the weight of the convolution layers and the pooling layers of VGG16, and reset the top layers including the full connection layers, BatchNormalization layer and softmax layer. We use Colab-keras to achieve transfer learning and fine-tuning of VGG16. As for VGG16, the channels of input data must be 3, and the size must be greater than $48 * 48$, so the neighborhood of $5 * 5$ is up-sampled to $50 * 50$ by nearest neighbor sampling. Since the original image has 8 bands and cannot be directly input to VGG16, we fuse panchromatic and multispectral bands by Gram-Schmidt Pan Sharpening to obtain the fusion image with 15 m resolution having 7 bands. Then we take three bands in two ways: (1) The first three principal components are taken after principal component analysis; (2) The 432 bands representing RGB are taken directly.

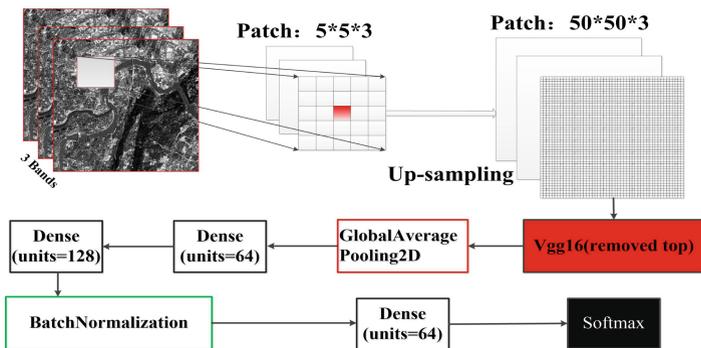


Fig. 9. Transfer learning and fine-tuning of VGG16

We set the ratio of training samples and validation samples to 7:3 for training the proposed CNN and VGG16. The accuracy and loss of the training process are shown in Fig. 10.

We recorded training accuracy, training loss, test accuracy, test loss, and the training time of 200 epochs. We can see from Table 4 that the accuracy of the proposed CNN is significantly better than that of VGG16, and the training time is greatly shortened. In Fig. 11, the classification effect of the proposed CNN is obviously greater than that of VGG16, and the extraction of built-up area is more detailed and accurate.

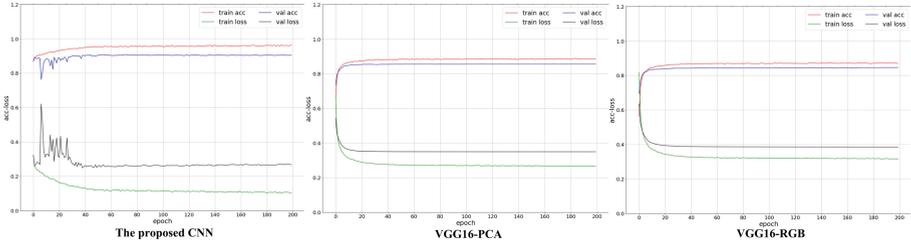


Fig. 10. Accuracy and loss in the training process of the proposed CNN, VGG16-PCA and VGG16-RGB

Table 4. The accuracy and loss of the proposed CNN, VGG16-PCA and VGG16-RGB

CNN-strategy	Train accuracy	Train loss	Test accuracy	Test loss	Training time (s)
The proposed CNN	0.968	0.102	0.901	0.262	4000
VGG16-PCA	0.886	0.266	0.842	0.321	36000
VGG16-RGB	0.873	0.316	0.816	0.357	34000

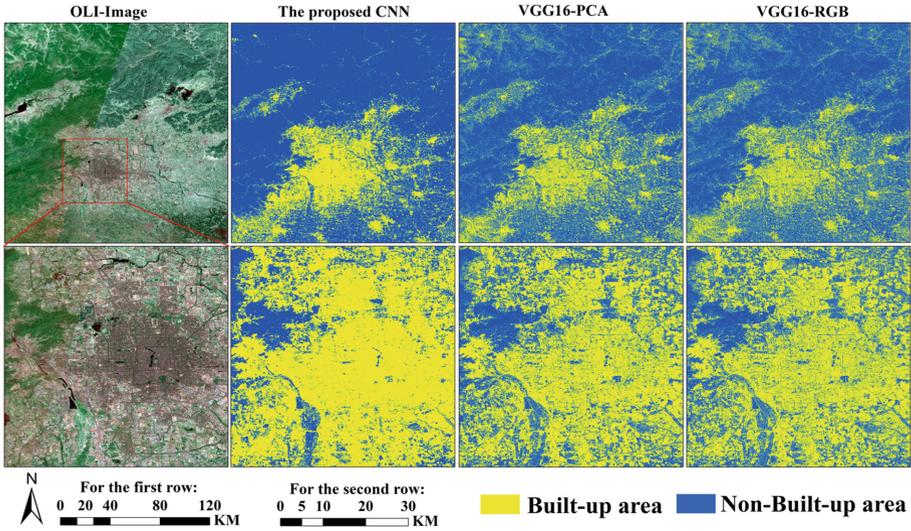


Fig. 11. Results of built-up area by the proposed CNN, VGG16-PCA and VGG16-RGB

6 Conclusion

In this paper, we build a simple and practical CNN framework, and use the automatic generation and correction of a large number of samples to realize the 15-m resolution built-up area classification and mapping based on the Landsat 8 image. We selected five typical experimentation sites, trained a better universal network, and classified the data of each experimentation site. The results show that all the test accuracy is higher than 89%, and the classification effect is good. We compared the CNN classification results with the existing built-up data products, which showed that the classification results of CNN can be very good for the details of the built-up area, and greatly reduced the classification error and leakage error. We also compared the results of the proposed CNN with VGG16, which indicated that the classification effect of the proposed CNN is obviously greater than that of VGG16, and the extraction of built-up area is more distinct and accurate. Therefore, this paper provides a reference for the classification and mapping of built-up areas in a wide range. At the same time, this paper also has the shortage that the choice of experimentation site is only in China and we fails to carry out experiments and verification on a global scale.

References

1. Chen, X.H., Cao, X., Liao, A.P., et al.: Global mapping of artificial surface at 30-m resolution. *Sci. China Earth Sci.* **59**, 2295–2306 (2016)
2. Zha, Y., Gao, J., Ni, S.: Use of normalized difference built-up index in automatically mapping urban areas from TM imagery. *Int. J. Remote Sens.* **24**(3), 583–594 (2003)
3. Xu, H.: A new index for delineating built-up land features in satellite imagery. *Int. J. Remote Sens.* **29**(14), 4269–4276 (2008)
4. Pesaresi, M., Gerhardinger, A., Kayitakire, F.: A robust built-up area presence index by anisotropic rotation-invariant texture measure. *IEEE J. Sel. Top. Appl. Earth Obs. Remote. Sens.* **1**(3), 180–192 (2009)
5. Chaudhuri, D., Kushwaha, N.K., Samal, A., et al.: Automatic building detection from high-resolution satellite images based on morphology and internal gray variance. *IEEE J. Sel. Top. Appl. Earth Obs. Remote. Sens.* **9**(5), 1767–1779 (2016)
6. Jin, X., Davis, C.H.: Automated building extraction from high-resolution satellite imagery in urban areas using structural, contextual, and spectral information. *EURASIP J. Adv. Signal Process.* **2005**(14), 745309 (2005)
7. Pesaresi, M., Guo, H., Blaes, X., et al.: A global human settlement layer from optical HR/VHR RS data: concept and first results. *IEEE J. Sel. Top. Appl. Earth Obs. Remote. Sens.* **6**(5), 2102–2131 (2013)
8. Goldblatt, R., Stuhlmacher, M.F., Tellman, B., et al.: Using Landsat and nighttime lights for supervised pixel-based image classification of urban land cover. *Remote Sens. Environ.* **205** (C), 253–275 (2018)
9. Yang, J., Meng, Q., Huang, Q., et al.: A new method of building extraction from high resolution remote sensing images based on NSCT and PCNN. In: *International Conference on Agro-Geoinformatics*, pp. 1–5 (2016)
10. Krizhevsky, A., Sutskever, I., Hinton, G.E.: ImageNet classification with deep convolutional neural networks. In: *International Conference on Neural Information Processing Systems*, vol. 60, no. 2, pp. 1097–1105 (2012)

11. Russakovsky, O., Deng, J., Su, H., et al.: ImageNet large scale visual recognition challenge. *Int. J. Comput. Vis.* **115**(3), 211–252 (2015)
12. Szegedy, C., et al.: Going deeper with convolutions. In: 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, pp. 1–9 (2015)
13. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, United States, pp. 770–778 (2016)
14. Castelluccio, M., Poggi, G., Sansone, C., et al.: Land use classification in remote sensing images by convolutional neural networks. *Acta Ecol. Sin.* **28**(2), 627–635 (2015)
15. Vakalopoulou, M., Karantzas, K., Komodakis, N., et al.: Building detection in very high resolution multispectral data with deep learning features. In: *Geoscience and Remote Sensing Symposium*, vol. 50, pp. 1873–1876 (2015)
16. Huang, Z., Cheng, G., Wang, H., et al.: Building extraction from multi-source remote sensing images via deep deconvolution neural networks. In: *Geoscience and Remote Sensing Symposium*, pp. 1835–1838 (2016)
17. Makantasis, K., Karantzas, K., Doulamis, A., Loupos, K.: Deep learning-based man-made object detection from hyperspectral data. In: Bebis, G., et al. (eds.) *ISVC 2015. LNCS*, vol. 9474, pp. 717–727. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-27857-5_64
18. Gorelick, N., Hancher, M., Dixon, M., et al.: Google earth engine: planetary-scale geospatial analysis for everyone. *Remote Sens. Environ.* **202**, 18–27 (2017)
19. Liu, X., Hu, G., Ai, B., et al.: A normalized urban areas composite index (NUACI) based on combination of DMSP-OLS and MODIS for mapping impervious surface area. *Remote Sens.* **7**(12), 17168–17189 (2015)
20. Yang, N., Tang, H., Sun, H., et al.: DropBand: a simple and effective method for promoting the scene classification accuracy of convolutional neural networks for VHR remote sensing imagery. *IEEE Geosci. Remote Sens. Lett.* **5**(2), 257–261 (2018)
21. Martino, P., Daniele, E., Stefano, F., et al.: Operating procedure for the production of the global human settlement layer from landsat data of the epochs 1975, 1990, 2000, and 2014. *JRC Technical report EUR 27741 EN*. <https://doi.org/10.2788/253582>
22. Liu, X., Hu, G., Chen, Y., et al.: High-resolution multi-temporal mapping of global urban land using landsat images based on the Google earth engine platform. *Remote Sens. Environ.* **209**, 227–239 (2018)
23. Chen, J., Chen, J., Liao, A., et al.: Global land cover mapping at 30 m resolution: a POK-based operational approach. *ISPRS J. Photogramm. Remote. Sens.* **103**, 7–27 (2015)
24. Li, Z., Tang, J., Mei, T.: Deep collaborative embedding for social image understanding. *IEEE Trans. Pattern Anal. Mach. Intell.* **PP**(99), 1 (2018)