EasyChair Preprint

№ 877

Fully Convolution Networks for Buildings Segmentation in High Spatial Resolution Remote Sensing Images

Linxiang Li

EasyChair preprints are intended for rapid dissemination of research results and are integrated with the rest of EasyChair.

April 5, 2019

Fully Convolution Networks for Buildings Segmentation in High Spatial Resolution Remote Sensing Images

Linxiang Li
School of Electrical Engineering and Automation
Henan Polytechnic University
Jiaozuo, China
311608070330@home.hpu.edu.cn

***Abstract*—Segmentation with high precision of buildings in high resolution remote sensing images is an important issue for the construction of digital city and classification the military targets. The main technology of this paper is based on FCN-VGG (Fully Convolution Networks-Visual Geometry Group) to implement the building segmentation in remote sensing image, in which the features are extracted by multiple convolution and the transposed convolution is employed to recover the original image size. The valid experiment results show that the method proposed could overcome disturbance produced by roads, trees, pedestrians, the cross-bedding of buildings effectively. Meanwhile, the multi-experiment demonstrates the generalization of the buildings segmentation in multi-resolutions.**

Keywords—fully convolution network, buildings segmentation, remote sensing image

I. INTRODUCTION

Buildings segmentation is one of the key techniques in the forefront of image processing, and it plays an important issue in resource exploration and planning, geological disaster monitoring and the national defense security.

Since the earth surface is too complex to extract the special targets from the remote sensing image, the segmentation accuracy of traditional methods for remote sensing image is unsatisfactory, and the generalization is also limited, so the building segmentation could not meet the requirement of what we want.

In recently, the deep convolution neural networks [11, 2] have obtained a series of breakthroughs for image classification [2, 7, 8], since the ability of extraction features automatically, so as to be widely researched and applied. In segmentation of remote sensing images, the features could be segmented exactly due to the semantic segmentation of deep convolutional neural network.

In 2014, Long et al. [5] proposed fully convolutional networks for semantic segmentation, and the experiment confirmed that a fully convolutional network could realize the prediction of end-to-end and pixels-to-pixels on semantic segmentation. In 2017, the reference [2] adopted VGG [6] and Encoder-Decoder framework to extract the features of images and restore images, and excellent results were gotten in the end.

In the same year, the reference [7] applied FCN to high resolution Image-3 and the polarization decomposition of H-A-α, completed the classification and identification of waterbody, vegetation and buildings. Meanwhile, the reference [1] proposed a neural network structure named single patch-based convoluted neural network, which can extract the features of roads and buildings in remote sensing image, and the segmentation of roads and buildings are realized accurately. The reference [4] applied FCN to detect ships near the coast and the reference [3] put forward a deep convolution neural network model with spectrum feature extraction of multiscale space for hyperspectral remote sensing image segmentation, achieved the features extraction and targets segmentation automatically.

In this paper, the FCN-VGG is proposed to segment buildings in high spatial resolution remote sensing image. The experiment we conducted could overcome disturbance produced by roads, trees, pedestrians, the cross-bedding of buildings effectively and segment the buildings precisely.

II. FRAMWORK OF FCN-VGG

The methodology of this paper is based on FCN-VGG. The process of buildings segmentation in high spatial resolution remote sensing image shown in Figure 1.

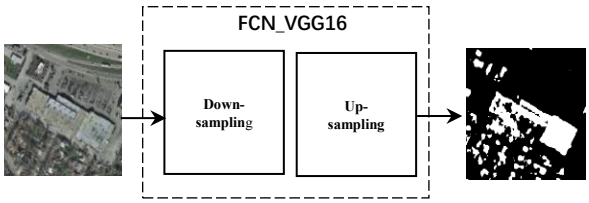


Fig.1. Framework of building segmentation

The fully convolution networks are adopted in this paper, because it could be applied to the high spatial resolution remote sensing image with any size, and the features of buildings is extracted exactly. Two steps are included in the whole process, down-sampling and up-sampling. Down-sampling includes convolution and pooling, and there are eight convolution layers and five pooling layers in it, but only two transposed convolution layers are included in up-sampling. The fully connected layers of VGG are translated to convolution layers and two transposed convolution layers are designed in the end

of the last convolution layer. The framework of remote sensing image building segmentation is shown in Table I. The function of down-sampling is to extract the features of buildings in high spatial resolution remote sensing image, and the up-sampling is designed to restore the size of original image.

TABLE I.1. THE FRAMEWORK OF REMOTE SENSING IMAGE BUILDING SEGMENTATION

Input Image	
Conv1(k=7,c=64,s=2)+BN	
MaxPool(k=3,c=64,s=2)	
Conv2	Conv2_1(k=1,c=64,s=1)+BN
	Conv2_2(k=3,c=64,s=1)+BN
	Conv2_3(k=1,c=256,s=1)+BN
Conv3	Conv3_1(k=1,c=128,s=2)+BN
	Conv3_2(k=3,c=128,s=1)+BN
	Conv3_3(k=1,c=512,s=1)+BN
Conv4	Conv4_1(k=1,c=256,s=2)+BN
	Conv4_2(k=3,c=256,s=1)+BN
	Conv4_3(k=1,c=1024,s=1)+BN
Conv5	Conv3_1(k=1,c=512,s=2)+BN
	Conv3_2(k=3,c=512,s=1)+BN
	Conv3_3(k=1,c=2048,s=1)+BN
AveragePool(k=2,c=2048,s=1)	
T_conv1	T_conv1(k=3,s=2,c=2)
Conv	Conv4 => Conv(k=1,s=2,c=2)
T_conv1 + Conv	
T_conv2	T_conv2(k=32,s=16,c=2)
Result	SoftMax

It should be noted that convolution and transposed convolution are opposite process. Convolution is input image I multiplied left by weight matrix T, $h = T I$. And the transposed convolution is the output feature h multiplied left by weight matrix T^T, $I_s = T^T h$. They would be introduced in next section.

In the case of limited training samples to make the network stronger and prevent the occurrence of over-fitting, Batch-normalization is adopted in this paper to further improve the neural network, weaken the degree of dependence between neurons and improve the robustness of the neural network.

III. DOWN-SAMPLING

A. *Convolution*

Generally, the pixels in a remote sensing image are numerous and bring huge information, which result in unmanageable extremely. Multiple convolution layers are adopted to extract features of buildings. Let the input image be a vector as $X = x_1, x_2, x_3, ..., x_9^T$, and the output vector as $Y = y_1, y_2, y_3, y_4^T$, then the convolution process can be expressed as (1). To make the convolution process clear, the diagram of convolution is shown in Figure 2.

$$Y = CX, C = \begin{bmatrix} \omega_1 & \omega_2 & 0 & \omega_3 & \omega_4 & 0 & 0 & 0 & 0 \\ 0 & \omega_1 & \omega_2 & 0 & \omega_3 & \omega_4 & 0 & 0 & 0 \\ 0 & 0 & \omega_1 & \omega_2 & 0 & \omega_3 & \omega_4 & 0 & 0 \\ 0 & 0 & 0 & \omega_1 & \omega_2 & 0 & \omega_3 & \omega_4 & 0 \end{bmatrix} \quad (1)$$

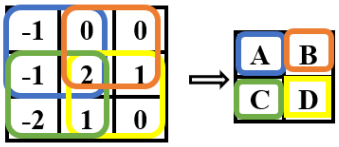


Fig.2. The diagram of convolution

In the process of convolution, kernel likes a filter to extract the local features. Kernel is described with square in Figure 2, kernel= [*H*, *W*, *C1*, *C2*], H and W are the size of kernel, and *C1* is the channel of input layer, *C2* is the channel of output layer. The one other important parameter is stride, kernel moved in feature map by stride, stride= [1, *S1*, *S2*, 1], *S1* and *S2* are the stride size.

If large kernel size and stride are chosen, the feature map we got would be very blurry. If small kernel size and stride are adopted, although the feature map would be clear, the computational speed would be very slow. In this paper, a part of kernel size we adopted is the trained results [12] of VGG16, and stride = [1, 1, 1, 1] is employed in convolution layer.

B. *Pooling*

Pooling is similar to convolution, but convolution spans the entire input layer, while the pooling operation moves only in one layer. The kernel of pooling is a filter with single channel, kernel= [1, *W*, *H*, 1], W and H are size of the kernel.

The most important function of pooling layer is to reduce the input size effectively. The function is realized by controlling stride, stride = [1, 2, 2, 1] is employed in this paper. It should be remarkable that the kernel of pooling is to get maximum or average, and the max-pooling is adopted in this paper.

In the case of ensuring undistorted, the size of input image could be reduced effectively by pooling. Pooling layer can not only accelerate the training speed, but also can prevent overfitting.

C. *Batch-normalization*

In the process of image processing, different input image data have different distribution states. And the network parameters are constantly adjusted to adapt to their changes, which wastes a lot of resources. As a result, the parameter distribution of each layer of the neural network is complex, which leads to difficulties in training and convergence of the neural network. Batch - normalization is a kind of optimization techniques, it could control the output parameters of each layer of the neural network output parameters to (0, 1). Ensuring distribution in each layer of the neural network parameters are consistent, effectively prevented excessive fitting, reduced the happening of the gradient disappeared, achieved the goal of accelerate network convergence, and could also use big learning rate to speed up the training. The specific algorithm process is shown in Figure 3.

$$x_i \rightarrow y_1 = W \cdot x_i + b \rightarrow y_2 = \frac{h_1 - \mu}{\sqrt{\sigma^2 + \varepsilon}}$$
$$\rightarrow y_3 = \gamma \cdot h_2 + \beta \rightarrow y_i = Relu(h_3)$$
$$\mu = \frac{1}{m}\sum_{i=1}^{m} x_i$$
$$\sigma = \frac{1}{m}\sum_{i=1}^{m} (x_i - \mu)$$
$$x_i = \frac{x_i - \mu}{\sqrt{\sigma^2 + \varepsilon}}$$
$$y_i = \gamma \cdot x_i + \beta, h_i = Relu(y_i)$$

Fig.3. The algorithm of Batch-normalization

D. *Down-sampling*

In the first, the whole process in down-sampling is shown in Figure 4. The main framework of down-sampling is based on VGG16. The primary function of it is to extract the feature we want to get. Meanwhile, the channel of feature map is increased and the size becomes more and more small.

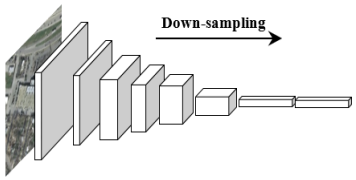


Fig.4. Down-sampling Framework

In the process of down-sampling, eight convolution layers and five pooling layers are designed, and there is one convolution layer included in Conv6, Conv7 and Conv8, two subsidiary convolution layers included in Conv1, Conv2 and Conv5, three subsidiary convolution layers included in Conv3 and Conv4. Pooling layers are only imported in the behind of Conv1 to Conv5. After multiple convolution and pooling, a three-dimensional image becomes to a small multi-channel feature map, and the feature extraction is completed.

According to the methodology, a multi-channel feature map is gotten which includes main information we want to extract, and some needless information is abandoned. Because five pooling layers are included in down-sampling, and the feature map could reduce by half after one pooling layer, the final feature map is 1/32 of original input image and the feature map is very blurry. To get satisfactory results, up-sampling is imported in this paper to restore the features map to original image size. The process of up-sampling would be included in next section.

IV. UP-SAMPLING

Transposed convolution has been mentioned that the process is opposite to convolution. It is the output feature *h* multiplied left by weight matrix T^T, $I_s = T^T h$. The diagram of transposed convolution is shown in Figure 5.

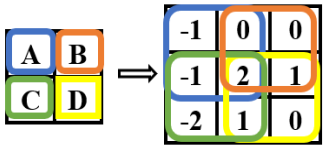


Fig.5. The diagram of transposed convolution

The parameters are same to convolution, them would not be introduced here. But the function is different to convolution, transposed convolution is to restore image rather than to extract features. Some important features are lost unavoidably because of the pooling operations in down-sampling, so the loss has to be recalled by up-sampling. The process of up-sampling is shown in Figure 6.

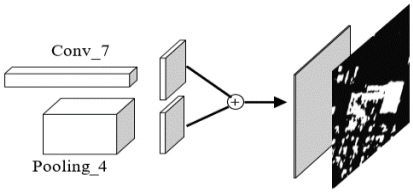


Fig.6. Up-sampling Framework

If the feature map of pooling 1 or pooling 2 is recalled, there is a lot of useless information in them. If the result of pooling 5 is recalled, a lot of important information we want to extract would be missed. The results of pooling 3 and pooling 4 are appropriate to be recalled. The feature map of pooling 4 is adopted in this paper.

The kernel size and stride depend on which feature map of pooling is recalled. Because the feature map of pooling 4 is recalled, kernel= [4, 4, *C1*,*C2*] and stride= [1,2,2,1] are designed in T-Conv1, and kernel= [16,16,*C1*,*C2*] and stride=[1,32,32,1] are designed in T-Conv2, *C1* is the channel of input layer, *C2* is the channel of output layer.

After two transposed convolution layers, the feature map size is same to original input image and the last output feature is the segmentation result. The buildings segmentation in high spatial resolution remote sensing image is realized.

V. RESULT

A. *Image Preprocessing*

In order to improve the efficiency of training, all pixels of samples are conducted to subtract the mean before training. First, to get the pixel mean of the whole remote sensing image, and then let all pixels of RGB channels of the image subtract the corresponding mean.

In addition, the building labels need to be converted into data of two channels. The pixels of buildings are marked as 1 in one channel, and 0 is marked for the rest pixels. The other one channel is marked with opposite data.

B. *Segmentation Performance Index*

To verify the effectiveness of the building segmentation, the experiment is conducted in IAILD (Inria Aerial Image Labeling Dataset) which is a database of remote sensing images.

Stochastic gradient descent is adopted in this paper. The cross loss of the experiment is shown in Figure 7. It could be seen that the cross loss is convergent.

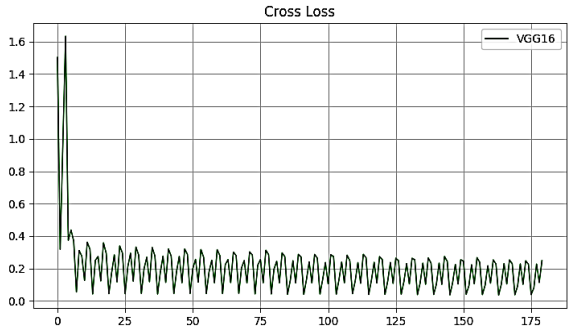


Fig.7. Cross loss of training

In order to quantify the results of buildings segmentation, *Recall*, *Precision*, and *F-measure* are imported to evaluate the results in this paper.

$$Recall = \frac{B_{seg}}{B_{seg} + I_{unseg}} \tag{2}$$

$$Precision = \frac{B_{seg}}{B_{seg} + I_{wseg}} \tag{3}$$

$$F\text{-}measure = \frac{2Recall \times Precision}{Recall + Precision} \tag{4}$$

- B_{seg}: the pixels of buildings are divided.
- I_{unseg}: the pixels of buildings are not divided.
- I_{wseg}: the pixels of background are divided as buildings in mistake.

Recall is the rate of the pixels are divided as buildings to the real buildings, *Precision* is the rate of pixels are divided accurately as buildings to the pixels are divided as buildings, and *F-measure* is a conformance of *Recall* and *Precision*. A part of experimental results are shown in Table II.

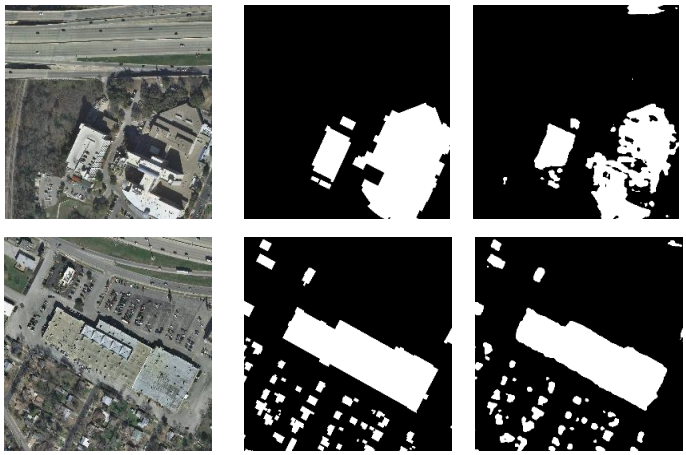
TABLE II. SEGMENTATION PERFORMANCE INDEX

Precision	Recall	F-measure
0.5662	0.7348	0.6396
0.7935	0.6881	0.7370
0.7031	0.8170	0.7558
0.6955	0.8358	0.7592
0.6858	0.8230	0.7482
0.6839	0.8503	0.7581
0.7973	0.8361	0.8162
0.7835	0.7456	0.7614
0.7567	0.7852	0.7707
0.7220	0.7371	0.7259
0.7159	0.7567	0.7106
0.7421	0.8161	0.7258
0.6794	0.8474	0.7035
0.6759	0.7490	0.7106
0.7177	0.8220	0.7663
0.7218	0.7299	0.7258
0.6331	0.7922	0.7037
0.7888	0.7760	0.7824
0.7136	0.8318	0.7682
0.3501	0.8249	0.4914

Because of the constraint of our experimental platform, the results are not enough precise. The data of our experiment shown that if the train steps are increased and smaller learning rate is chosen, the *Precision*, *Recall* and *F-measure* would be higher, and the results would be more precise.

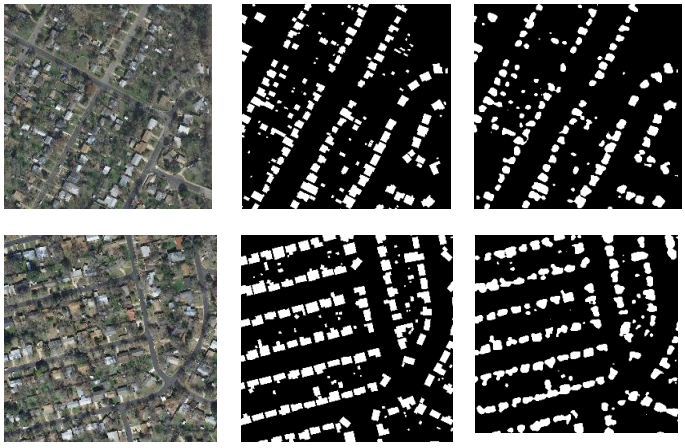
C. *The Results of Image Segmentation*

For the larger buildings with complex structure and distraction caused by shadows, the segmentation results are shown in Figure 8. The results are very consistent with the real structure of the buildings and it overcomes the influence caused by shadows effectively. For the buildings with uniform distribution, simple structure and clear boundary, the segmentation results are shown in Figure 7. The buildings are segmented effectively and precisely.



(a) Imanges (b) Labels (c) Results

Fig. 7. The simple single-storey staggered buildings



(a) Imanges (b) Labels (c) Results

Fig. 8. The group of small buildings affected by trees

VI. CONCLUSIONS

This paper proposed a FCN-VGG structure for buildings segmentation in high spatial resolution remote sensing images. Down-sampling is designed to extract the features of high spatial resolution remote sensing images. Up-sampling is adopted to restore the size to original input image. *Recall*, *Precision*, and *F-measure* are imported to evaluate the results of the buildings segmentation.

Experimental results of IAILD shown that the methodology this paper proposed could overcome the disturbance caused by roads, trees, buildings cross-bedding and shadows effectively in remote sensing images, and segment buildings accurately.

But the outline of buildings is not precise, and some small buildings are missed in the process. It could not do well in the case that the colors of buildings are similar to background. In future, we will devote ourselves to overcome the imperfections and improve the segmentation accuracy.

ACKNOWLEDGMENT

The author would like to acknowledge the help with my teachers and friends. And the author would like to acknowledge the Henan Polytechnic University.

REFERENCES

[1] Alshehhi R, Marpu P R, et al. Simultaneous extraction of roads and buildings in remote sensing imagery with convolutional neural networks. ISPRS Journal of Photogrammetry and Remote Sensing. 2017,130: 139-149. J. Clerk Maxwell, A Treatise on Electricity and Magnetism, 3rd ed., vol. 2. Oxford: Clarendon, 1892, pp.68–73.

[2] A. Krizhevsky, I. Sutskever, and G. Hinton. Imagenet classification with deep convolutional neural networks. In *NIPS*, 2012.

[3] Badrinarayanan V, Kendall A, et al. SegNet: A Deep Convolutional Encoder-Decoder Architecture for Scene Segmentation. IEEE Transactions on Pattern Analysis & Machine Intelligence. 2017, PP (99): 1-1

[4] Jiao L, Liang M, et al.Deep Fully Convolutional Network-Based Spatial Distribution Prediction for Hyperspectral Image Classification. IEEE Transactions on Geoscience and Remote Sensing. 2017,PP(99): 1-15.

[5] Lin H, Shi Z, et al.Fully Convolutional Network With Task Partitioning for Inshore Ship Detection in Optical Remote Sensing Images. IEEE Geoscience and Remote Sensing Letters. 2017,PP(99): 1-5.

[6] Long J, Shelhamer E, et al.Fully convolutional networks for semantic segmentation. IEEE Transactions on Pattern Analysis & Machine Intelligence. 2017,39(4): 640-651.

[7] M. D. Zeiler and R. Fergus. Visualizing and understanding convolutional neural networks. In *ECCV*, 2014.

[8] P. Sermanet, D. Eigen, X. Zhang, M. Mathieu, R. Fergus, and Y. LeCun. Overfeat: Integrated recognition, loclization and detection using convolutional networks. In *ICLR*, 2014.

[9] Simonyan K, Zisserman A. Visual Geometry Group. 2014 from http://www.robots.ox.ac.uk/~vgg/research/very_deep/.

[10] Wang Y, Wang C, et al.,2017. IntegratingH-A-α with fully convolutional networks for fully PolSAR classification. 2017 International Workshop on Remote Sensing with Intelligent Processing (RSIP).

[11] Y. LeCun, B. Boser, J. S. Denker, D.Henderson, R. E. Howard, W. Hubbard, and L. D. Jackel. Backpropagation applied to handwritten zip code recogntion. *Neural computation*, 1989.

[12] https://mega.nz/#!YU1FWJrA!O1ywiCS2IiOlUCtCpI6HTJOMrneN-Qdv3ywQP5poecM