# FULLY CONVOLUTIONAL NETWORK BASED SHADOW EXTRACTION FROM GF-2 IMAGERY

LI Zhi-qiang[1,2], CAI Guo-yin[1,2,*], Ren Hui-qun[1,2]

[1]School of Geomatics and Urban Spatial Informatics, Beijing University of Civil Engineering and Architecture, Beijing, China
[2]The Key Laboratory for Urban Geomatics of National Administration of Surveying, Mapping and Geoinformation, Beijing University of Civil Engineeringand Architecture, Beijing, 100044, China

**KEY WORDS:** Fully Convolutional Network, Shadow Extraction, Deep Learning, GF-2, Building Shadows

## ABSTRACT

There are many shadows on the high spatial resolution satellite images, especially in the urban areas. Although shadows on imagery severely affect the information extraction of land cover or land use, they provide auxiliary information for building extraction which is hard to achieve a satisfactory accuracy through image classification itself. This paper focused on the method of building shadow extraction by designing a fully convolutional network and training samples collected from GF-2 satellite imagery in the urban region of Changchun city. By means of spatial filtering and calculation of adjacent relationship along the sunlight direction, the small patches from vegetation or bridges have been eliminated from the preliminary extracted shadows. Finally, the building shadows were separated. The extracted building shadow information from the proposed method in this paper was compared with the results from the traditional object-oriented supervised classification algorihtms. It showed that the deep learning network approach can improve the accuracy to a large extent.

## 1. INTRODUCTION

### 1.1 Summary

There are high density of buildings in the urban region which presents clearly on the high resolution satellite imagery. Building shadows are consequently a common phenomenon on the images. The building shadows, on the one hand, has become a major problem in thematic map production of land cover or land use, on the other hand, they have been served as one of the auxiliary information for the extraction of buildings（GAO Xianjun et al., 2017）. Therefore, the extraction of shadows from urban areas with high spatial resolution imagery is of practical significance. Traditional high spatial resolution image classification methods usually need to segment the satellite images, and then perform the object oriented classification (ZHANG Meng et al., 2017). The quality of segmentation results will also affect the classification results. Moreover, traditional classification methods need to manually define features, which requires people's expertise and experience at expert level to choose better features for classification. In recent years, the development of deep learning has provided a new way to support technological help for the field of remote sensing imagery information extraction. By training parameters of convolutional neural network hidden layer, the final features applicable to different categories are obtained (Zeiler, M. D. et al., 2014), such that it saves the process of artificial selection of characteristics, and extracts more diverse features. The fully convolutional network designed in this paper is an end-to-end network, that is, the input is the original image, and the output is the result of the same size with the original image. It predicts the category of each pixel. The use of the fully convolutional network to extract thematic information saves the image segmentation process.

### 1.2 Study area and data

The Chinese Gaofen -2 (GF-2) satellite imagery is employed in this study. The GF-2 satellite, launched in August 19, 2014, is the first Chinese civil land satellite with a spatial resolution of less than 1 meters, with a spatial resolution of 0.81 meters (panchromatic) and 3.24 meters (multispectral). The GF-2 satellite imagery has 1 panchromatic band and 4 multispectral bands, including blue, green, red and near infrared. In this paper, we collected the GF-2 satellite 1A multispectral and panchromatic images on June 15, 2015. 30 m ASTGTM data are used as the height datum for the GF-2 image ortho-rectification. The corrected images are fused with multispectral data and panchromatic data by Gram-Schmidt algorithm, and finally the multispectral remote sensing data with 0.8 m spatial resolution are obtained.

## 2. NETWORK STRUCTURE

Convolutional neural network used for image recognition is to input a fixed size picture to the network, and then through the alternation of the convolution layer and the pooling layer, and finally generate a fixed number of feature maps. Then the feature maps are passed through several fully connected layers, and the eigenvectors of the fully connected layer output are inputted into the classifier to get the corresponding scalar of the category. The fully connected layer in the convolutional neural network is transformed into a convolution layer that makes the output 1×1, so that the fully convolutional network is obtained. The input of the network is not limited by the fully connected layer and can be inputted into any size of the picture after the fully connected layer is converted to the convolution layer, finally a heat map is outputted (Long, J. et al.,2015). The heat map is restored to the same input image as the same size of the output data through the upper sampling layer, that is, the fully convolutional network for semantic segmentation. In this study, the end to end, pixel to pixel full convolution network is

[*] Corresponding author: CAI Guo-yin - cgyin@bucea.edu.cn

This contribution has been peer-reviewed.
https://doi.org/10.5194/isprs-archives-XLII-3-985-2018 | © Authors 2018. CC BY 4.0 License.

employed to extract the shadows of the urban areas from GF-2 satellite imagery.

This research uses the classic VGG model to sample images downwards using a staggered 3*3 convolution layer and a 2*2 max-poolling layer (Simonyan, K. et al.,2014). We replace the 3*3 convolution structure of single scale into Inception structure of multi-scale, and use batch normalization to standardize input data and output data of each layer (Szegedy, C. et al., 2014). The small size output of the network down sampling represents only the semantic information of the whole picture, which leads to the lack of local location information in the ascending sampling process. Therefore, through cross layer structure, coarse and semantic information in ascending sampling process can be combined with local positional information in the process of descending sampling, so as to achieve an improvement of spatial prediction (Long, J. et al., 2015). The network framework is shown in figure 1. The network structure shows that the data shape of input image is (100,100,3), and the 3 is the number of input image channels. Three bands of R, G and B of GF-2 image are used. In the process of descending sampling, the number of channels of the characteristic graph, that is, the number of convolution kernel is gradually increased. In the process of upwards sampling, the number of convolution kernel decreases gradually. Finally, the classification results of (100,100,1) are obtained by the sigmoid classifier.

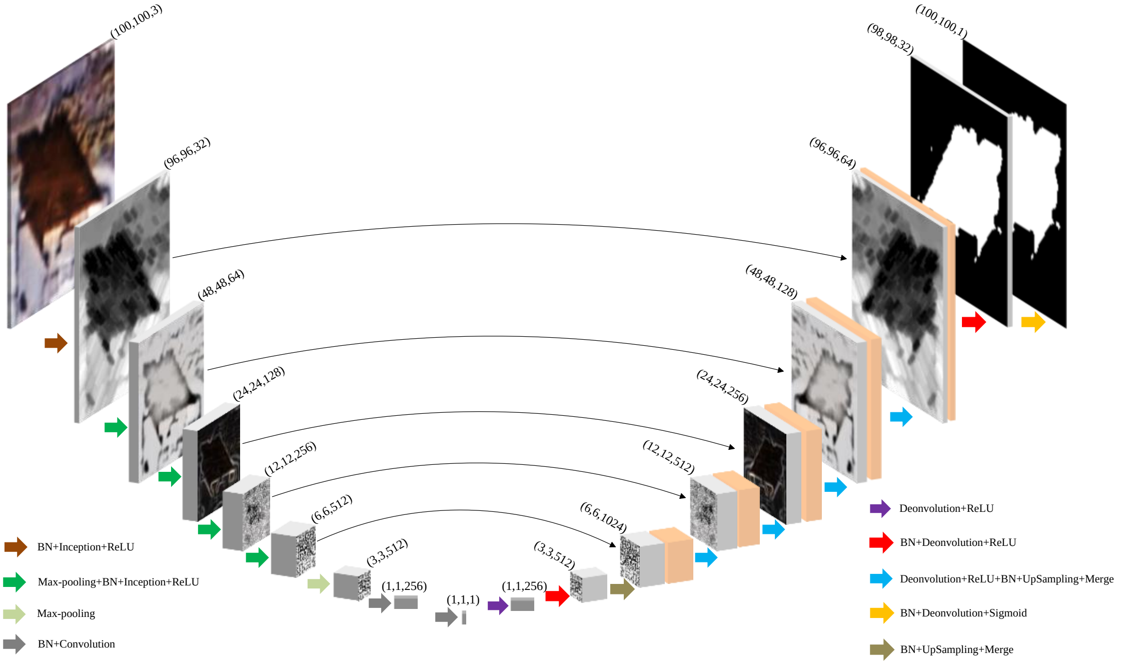


Figure 1. The network framework employed in this study

**Batch Normalization Layer(BN)** Batch normalization can standardize each small batch training data into the distribution of zero mean and unit variance. It can stabilize the distribution of output activation values of each layer of neural network and avoid the occurrence of gradient vanishing problem, thereby greatly improving the convergence speed of training and reducing the use of dropout layer (Ioffe, S. et al., 2015). In this paper, we use batch normalization layer to carry out standardized preprocessing for input data. After that, we perform batch normalization standardization processing for output data of every convolution layer and deconvolution layer.

**Convolution Layer** The convolution layer can perceive local region characteristics by the sparse connection between the convolution kernel and the image. Each convolution layer can extract features of different abstraction levels. Shallow convolution layer extracts local features of images, and deep convolution layer extracts global features of images. In this paper, convolution layer is designed as 3×3 convolution, and 3×3 convolution is decomposed into two layers convolution of 1×3 and 3×1. The purpose of decomposition is to reduce the number of parameters and increase the nonlinear operation of network (Szegedy, C. et al., 2016).

**ReLU Layer** Rectified linear unit（ReLU）is one of the most popular activation functions in neural network in recent years. It does not activate all neurons input values of non-positive number, and linear activate to neurons input values of positive number. It can avoid the problem of gradient disappearance, and it accelerates the convergence speed of network training.

**Max-pooling Layer** Max-pooling only keeps the largest pixel value in the window and removes other pixels, so as to achieve the purpose of image down sampling and effectively reduce data redundancy. The max-pooling layer window used in this paper is 2*2 size, and the step length is assigned as 2.

**Inception** The Inception structure is proposed in GoogleNet. It has four branches, which are convolution branches of three different scale convolution kernel and a pooling branch. It increases the depth and width of the network, and reduces the amount of parameters, and increases the adaptability of the network to the same scale. Five Inception structures are used in the down sampling process in this study. First, the input image of 100×100 is sampled to 96×96, the structure is shown in figure 2 (a), and the remaining four Inception structures are the same, and the output size is equal to the input size, and the structure is presented in figure 2 (b).

**Deconvolution Layer** In this paper, alternating deconvolution and UpSampling layer are used in the process of ascending sampling of the network. In essence, the forward propagation process of deconvolution layer is the reverse propagation process of convolution layer (Zeiler, M. D. et al., 2010).

**UpSampling Layer** UpSampling layer is used to magnify the output of deconvolution in size dimension, which is to repeat pixels on the corresponding dimension according to the set magnification.

**Merge Layer** Merge layer is a layer that is used by cross layer connection. It connects the output data of ascending sampling process and the output data of the same size of the process of descending sampling on the channel dimension. In addition, at the end of the Inception structure, the merge layer is used to fuse the output of the four branches into a data body in the channel dimension.

**Sigmoid Layer** The Sigmoid layeris used for the final classification, which can map the activation values of the previous layer to a probability between 0 and 1 to make a category prediction for each pixel (ZHANG Xiao et al., 2011). In this paper, the probability is greater than 0.5, then the pixel is classified as shadow, otherwise it is classified as the background.

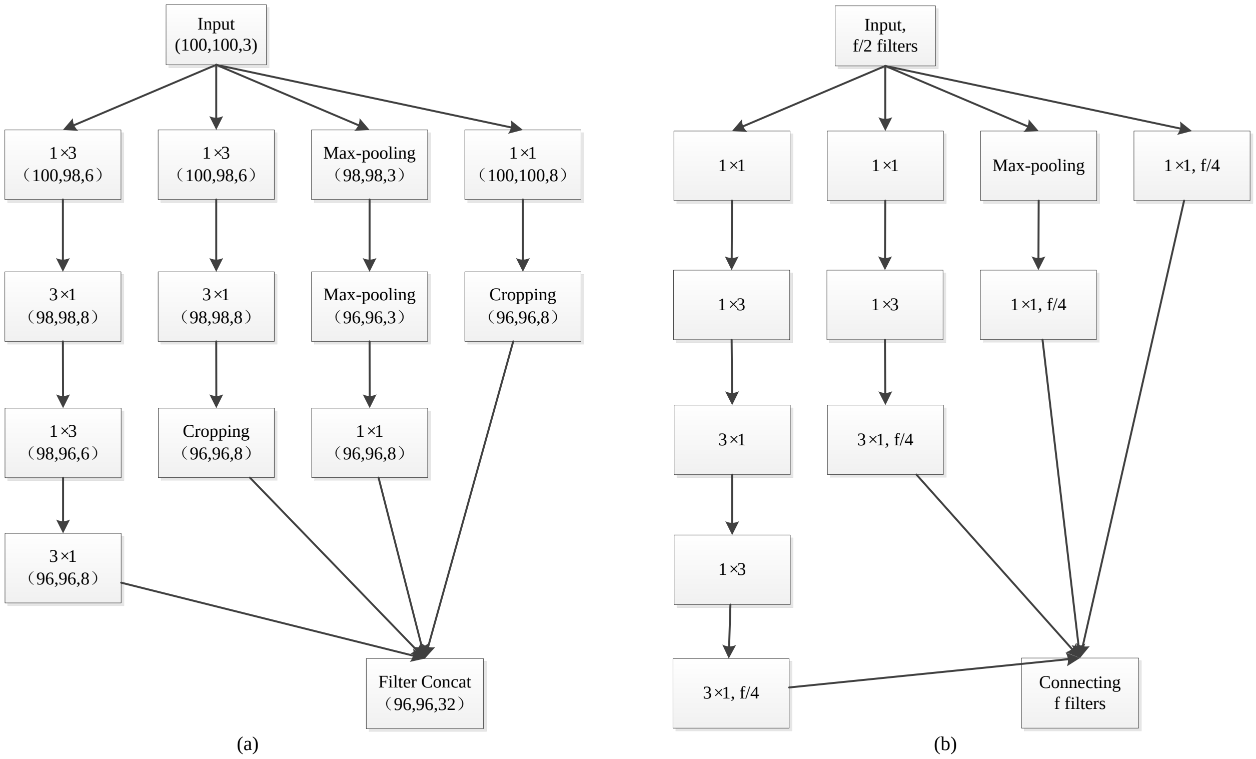


Figure 2. The designed inception structure in this study

## 3. TRAINING AND TESTING

The extraction of shadow is also the binary classification problem. In this paper, two categories of shadow and background are classified. According to the spatial resolution of GF-2 satellite shadow, the size of 100*100 is cutted out as training data and test data. We expanded the training data set by 7.5 times, and finally obtained 1500 training images. Two methods of data expansion are used in this paper. One is to force all the original images to make size jitter, to randomly narrow and enlarge the original image to five square sizes between 100*100 and 200*200, and then to tailor the middle 100*100 size picture as training set, so that the shadow and background training set is expanded by 5 times.

Another way is to split and reorganize the shadow training data set, indicated in figure 3(a), and divide the picture into four parts: 0, 1, 2, 3, and then reorganize the four part in a way shown in figure 3(b), so that the shadow training set is enlarged by 5 times. Accordingly, we execute the same disassembling and reassembling procedures for each ground truth corresponding to the graph. The purpose of this article is to artificially increase a number of shadow edges without reducing the number of shadow pixels.

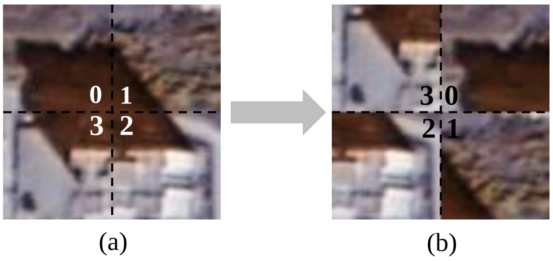


Figure 3. The illustration of disassembly and reorganization procedure

The training data set input network trained 50 epoch, tested the accuracy rate of each epoch through the test data set, saved the highest accuracy results, and carried out subsequent experiments. The accuracy rate is calculated by dividing the number of all the pixels that are correctly classified by the number of all pixels. Finally, the accuracy rate of test sets of the forty-first epoch is the highest, which is 0.9637. Considering the small proportion of the shadow in the image, resulting in the imbalance of the number of shadow and background pixels in the training set, we made qualitative analysis of the test set test results, which is shown in figure 4. It can be observed from the figure 4 that the pixels of the wrong points are mostly concentrated on the edge of the category, whether in the shadow or in the background. When the artificial production of pixel labels, a lot of details is difficult to control, such as the shadow boundary, there is some small patches in the form of shadow (shadows of car, individual trees), so when a pixel level label is made, the shadow can't be depicted by one hundred percent. The principle that we make pixel level labels is that large area shadows must be selected, and small area shadows are selected as appropriate. Due to the restriction of label making conditions, the results show that the pixels that are misclassified are concentrated on the edge of category, and the other parts are misclassified pixels gathered in small patches. But it can be seen from figure 4 that the gross outline of shadow of large area and shape rule is well extracted.

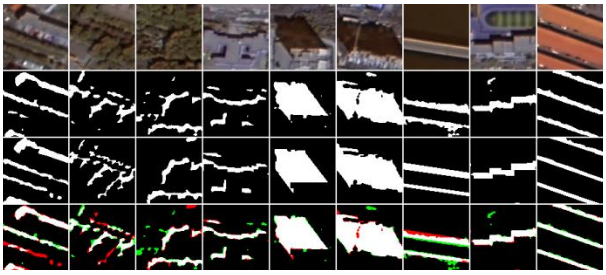


Figure 4. The analysis of misclassified pixel regions

The pictures in first line of Figure 4 is the original pictures of the test data set. The second line is the result of the network's prediction of the test data set. The third line is the ground truth. The fourth line shows the part that is misclassified on the basis of the ground truth, in which red represents the area that originally belongs to the shadow in the ground truth, but is predicted as the background area, and the green is the area which originally belongs to the background in the ground truth, but is predicted to be the shadow area.

## 4. EXTRACTION OF BUILDING SHADOWS IN URBAN AREAS

### 4.1 Extraction process

We use the model and the trained weight and bias parameters saved in the section of 4 to use the image in the larger region to verify its performance. We cutted out two images with size of 644*964, extracted the shadow through the trained fully convolutional network, and processed the shadow extraction results according to the analysis method and the the specific situation of the image, and finally carried out the accuracy evaluation. Figure 5 shows the specific process of shadow extraction and accuracy evalution. In some areas, there is no large area of vegetation shadow. Therefore, as shown in the dotted line in figure 5, we can directly remove small patches from the shadow results extracted from fully convolutional network.

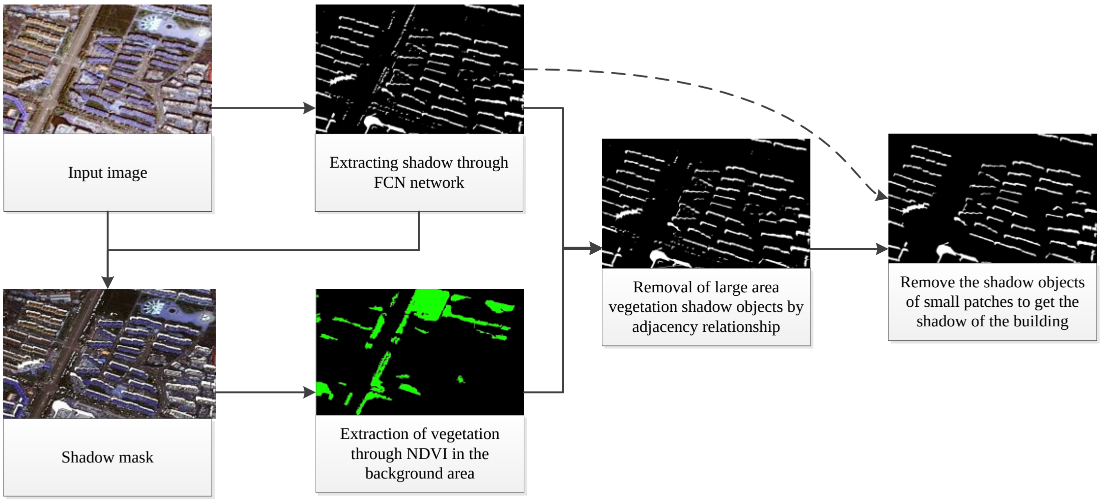

Input image
Extracting shadow through FCN network
Removal of large area vegetation shadow objects by adjacency relationship
Remove the shadow objects of small patches to get the shadow of the building
Shadow mask
Extraction of vegetation through NDVI in the background area

Figure 5. The procedure of building shadow extraction

### 4.2 Post-processing

According to the analysis in section 3, we can see that the shadow of small scale surface exists as the form of small patches in the results extracted from fully convolutional network, while the building shadow exists generally with larger patches. So we can remove these small patches by segmenting and then setting the threshold with the number of pixels of the small patch shaded objects. In experimental image number 1, there is also a problem of interference with building shadows. It is the shadow of patches of a large area of trees. Their sizes are similar to those of some small buildings. It is difficult to remove these shadows with the method of removing small patches. We solved this problem well through the relationship between the sun illumination direction and the adjacent objects.

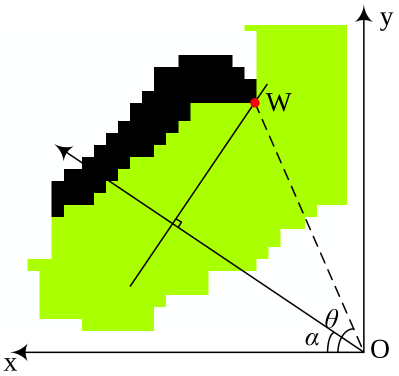


Figure 6. Adjacency diagram of sun illumination direction and adjacent objects

This contribution has been peer-reviewed.
https://doi.org/10.5194/isprs-archives-XLII-3-985-2018 | © Authors 2018. CC BY 4.0 License.

As shown in figure 6, the green object in the coordinate system represents the vegetation, the black object represents the shadow of the vegetation, and the $\alpha$ angle represents the sun illumination direction. In this paper, the sun illumination is estimated by the connection of the corner point of the building and the corresponding corner of the building's shadow. We judge whether the shadow belongs to the vegetation shadow by the relationship between the shadow and the adjacent objects in the direction of the sun light. By looking for the point W of the most outward direction of the shadow object in the opposite direction of the sunlight direction, and then judging whether the shadow object is a vegetation shadow by whether another object, including the public point W, is vegetation. It is known from figure 6 that the projection of the distance between the W point and the origin of the coordinates of O in the direction of the sunlight is the shortest of all the points that make up the shadow object, so we should first find W. The angle of $\theta$ is the azimuth of the point in the coordinate system, which can be obtained by the cosine formula (1):

$$\theta = \arccos\frac{x_W\text{-}x_O}{\sqrt{(x_W\text{-}x_O)^2+\left(y_W\text{-}y_O\right)^2}} \qquad (1)$$

where $\theta$ = the angle of the vector $\overrightarrow{OW}$ and the x axis
$x_W, y_W$ = coordinates of the W point
$x_O, y_O$ = coordinates of the O point

Then the absolute value of the difference between the θ and the α is obtained, and the length L that the point is mapped in the direction of the sunlight is obtained by the formula (2):

$$L=\sqrt{(x_W\text{-}x_O)^2+\left(y_W\text{-}y_O\right)^2}\cos(|\theta\text{-}\alpha|) \qquad (2)$$

where $L$ = the mapping length of the vector $\overrightarrow{OW}$ in the direction of the sunlight
$\alpha$ = angle of the direction of the sunlight and the x axis

### 4.3 Accuracy evaluation

After post-processing, we perform the accuracy evaluation for the shadow results extracted by fully convolutional network and the shadow of the building. The former is used to superposition and statistics the confusion matrix by using the pre-selected vector point label and the fully convolutional network shadow extraction result, and calculates the error detection rate and missing detection rate of the shadow. In the latter part, we overlay the real label and post-processing result of the building shadow, and calculate the error detection rate and missing detection rate of the building shadow through the confusion matrix, and use the IoU (Intersection-over-Union) index to make the accuracy evaluation of the building shadow. IoU is an evaluation index commonly used in semantic segmentation. It reflects their overlap by calculating the ratio of the intersection and union of the segmentation results to the real label. The formula of error detection rate, missing detection rate and of IoU are as follows.

$$ED= \frac{C_{01}}{T} \qquad (3)$$

$$MD= \frac{C_{10}}{T} \qquad (4)$$

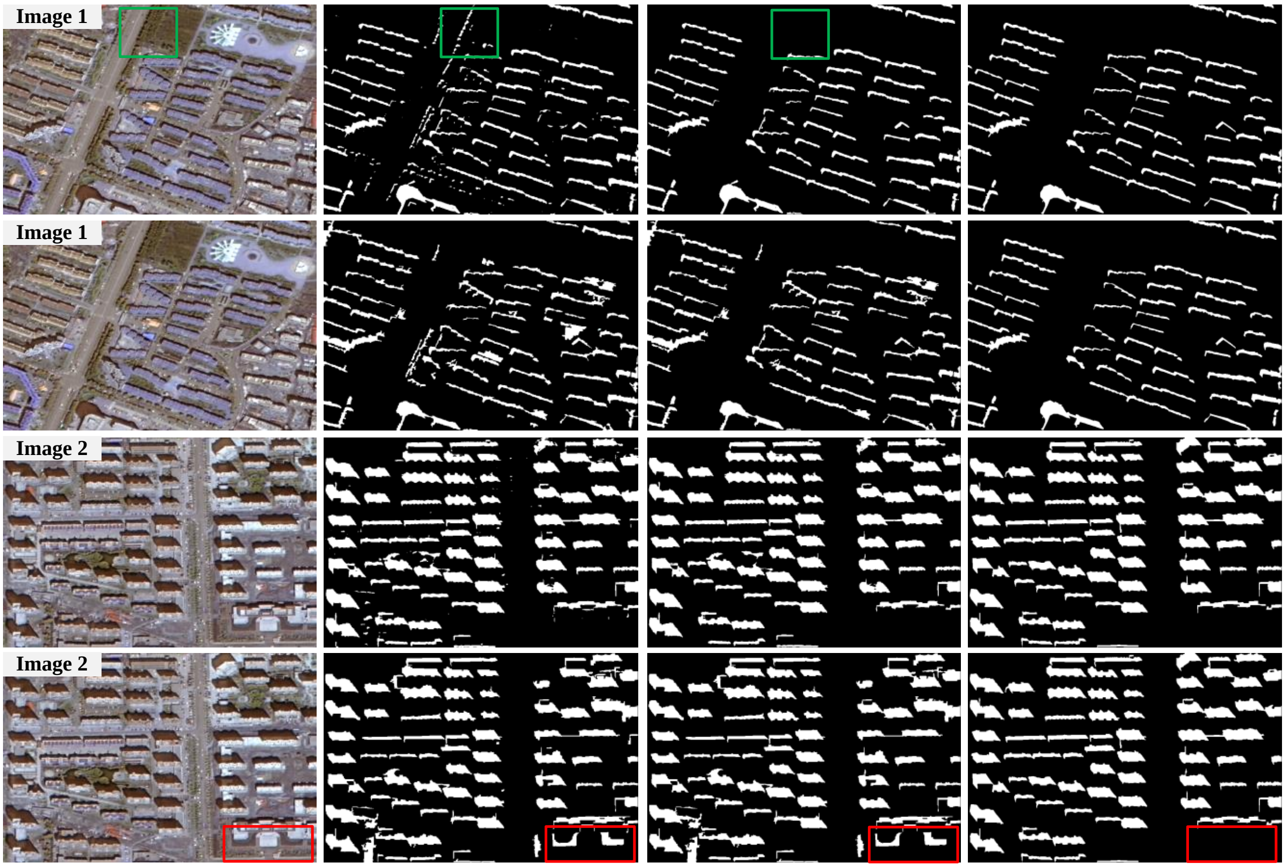
$$IoU= \frac{C_{11}}{C_{11}+C_{01}+C_{10}} \qquad (5)$$

where $ED$ = error detection rate
$MD$ = missing detection rate
$IoU$ = Intersection-over-Union
$C_{01}$ = the number of background pixels which is divided into shadow
$C_{10}$ = the number of shadow pixels which is divided into background
$C_{11}$ = the number of shadow pixels that are correctly classified
$T$ = number of shadow pixels in ground truth

In section 3, we have drawn a conclusion that most of the misclassified pixels are concentrated in the edge of the category. In order to reduce the influence of the boundary part on the accuracy evaluation, the building shadow and the background object edge of the extracted results were corroded by a pixel unit with radius, and then the accuracy index calculation of the building shadow extraction result was made in the non-corrosive area.

## 5. RESULTS AND CONCLUSIONS

In order to compare with traditional methods, the two images are classified as object oriented supervised classification. Figure 7 is the extraction result and post-processing result of image 1 and image 2. Compared with the ground truth, fully convolutional network and object-oriented supervised classification can extract the building shadow at the location degree. However, due to object-oriented supervised classification, image segmentation is needed, because the limitation of segmentation results makes the shadow extracted from object-oriented supervised classification more wrong than fully convolutional network. For example, the dark bare ground near the building in the image 2 red frame is misclassified as the shadow. In this paper, we propose a method to combine vegetation information and sunlight direction to detect shadow adjacent objects in the opposite direction of light direction, and effectively remove large area and small patches of vegetation shadow. As shown in the area of the figure 7 image 1 green frame, the shadow of the building is better extracted.



(a) Image (b) Shadow (c) Building shadow (d) Ground truth

Figure 7. The shadow extraction results of two methods

As shown in Figure 4, the second column and the third column in Figure 7 are the results of this paper. The first and third lines are the fully convolutional network extraction results and post-processing results of image 1 and image 2. The second and fourth lines are object oriented supervised classification results and post-processing results of image 1 and image 2.

We evaluate the results of shadow extraction and building shadow extraction such as table 1:

| Data | Method | Shadow | | Building shadow | | |
|---|---|---|---|---|---|---|
| | | ED/ % | MD/ % | ED/ % | MD/ % | IoU |
| Image 1 | FCN | 0.26 | 1.28 | 1.87 | 3.01 | 0.9521 |
| | OOSC | 1.28 | 5.87 | 20.77 | 6.99 | 0.7701 |
| Image 2 | FCN | 0.28 | 1.41 | 1.69 | 1.61 | 0.9676 |
| | OOSC | 5.07 | 5.35 | 10.64 | 9.23 | 0.8204 |

Table 1. Results of the accuracy evaluation

*FCN: Fully convolutional network
OOSE: Object-oriented supervised classification
ED: Error detection rate MD: missing detection rate

The powerful analysis ability of deep learning for a large number of data makes it have great potential in the field of image processing. Through this experiment, we can see that the fully convolutional neural network for shadow information extraction is much better than the traditional object-oriented supervised classification method. Table 1 shows that the error detection rate and missing detection rate of the two methods are very low through the vector point sample, which indicates that the two methods are effective in extracting the shadows from the location degree. However, the error detection rate and missing detection rate of the fully convolutional network shadow extraction results are greatly reduced compared with the object oriented supervised classification. The accuracy of the evaluation results from the building shadows, the error detection rate and missing detection rate of building shadow of object oriented supervised classification with post-processing than fully convolutional network with post-processing results is larger, and the IoU index of the former is much lower than the latter, indicating that fully convolutional network in a single shadow object comparison for object oriented supervised classification can be extracted more like complete, and error detection and omission detection.

## REFERENCES

GAO Xianjun, ZHENG Xuedong, SHEN Dajiang, YANG Yuanwei, ZHANG Jiahua, 2017. Automatic building extraction based on shadow analysis from high resolution images in suburb areas. *Geomatics and information science of Wuhan university*, 42(10), pp. 1350-1357.

Ioffe, S., Szegedy, C., 2015. Batch normalization: accelerating deep network training by reducing internal covariate shift. pp. 448-456.

Long, J., Shelhamer, E., Darrell, T., 2015. Fully convolutional networks for semantic segmentation. *Computer Vision and Pattern Recognition* , Vol.79, pp.3431-3440.

Simonyan, K., Zisserman, A. 2014. Very deep convolutional networks for large-scale image recognition. *Computer Science*.

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., et al., 2014. Going deeper with convolutions. pp. 1-9.

Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., Wojna, Z., 2016. Rethinking the inception architecture for computer vision. *Computer vision and pattern recognition*. pp. 2818-2826.

Zeiler, M. D., Krishnan, D., Taylor, G. W., Fergus, R., 2010. Deconvolutional networks. *Computer vsion and pattern recognition*, Vol.238, pp. 2528-2535.

Zeiler, M. D., Fergus, R., 2014. Visualizing and understanding convolutional networks. *European conference on computer vision*, Vol.8689, pp. 818-833.

ZHANG Meng, ZENG Yongnian, ZHU Yongsen, 2017. Wetland mapping of Donting lake basin based on time-series MODIS data and objet-oriented method. *Journal of remote sensing*, 21(3), pp, 479-492.

ZHANG Xiao, HUANG Xi, ZHONG Weihan, ZHANG Liang, 2011. Implementation of sigmoid function and its derivative on FPGA. *Journal of Fujian normal university*, 27(2), pp, 62-65.