Targets mask U-Net for Wind Turbines Detection in Remote Sensing Images

Min Han [12,*], Huabin Wang [12], Guanghui Wang[2], Yu Liu [2]

[1] School of Geomatics, Liaoning Technical University, Fuxin, People's Republic of China - abcdefskylight@gmail.com
[2] Satellite Survey and Mapping Centre, Beijing, People's Republic of China – (whb, wanggh, liuy)@sasmac.cn

ICWG II/III

KEY WORDS: object detection, VHRRSI, wind turbines, shadow, wide-field detector, U-Net

ABSTRACT:

To detect wind turbines precisely and quickly in very high resolution remote sensing images (VHRRSI) we propose target mask U-Net. This convolution neural network (CNN), which is carefully designed to be a wide-field detector, models the pixel class assignment to wind turbines and their context information. The shadow, which is the context information of the target in this study, has been regarded as part of a wind turbine instance. We have trained the target mask U-Net on training dataset, which is composed of down sampled image blocks and instance mask blocks. Some post-processes have been integrated to eliminate wrong spots and produce bounding boxes of wind turbine instances. The evaluation metrics prove the reliability and effectiveness of our method for the average F1-score of our detection method is up to 0.97. The comparison of detection accuracy and time consuming with the weakly supervised targets detection method based on CNN illustrates the superiority of our method.

1. INTRODUCTION

Wind turbines, which convert the wind's kinetic energy into electrical energy, are important clean energy facilities in modern society. As the rapidly developing of wind power industry, a proper solution is needed to check the number and location of wind turbines fast and cheaply for evaluation of installed capacity, automatic cartography and environmental surveys. With the high development of sensor technology and active human earth observation campaigns, remote sensing has become an important monitoring means for special targets on the ground. With the help of very high resolution optical satellite remote sensing images, it is easy to find special targets with prominent image features, like planes or wind turbines, thereby determining their location and quantity.

Many approaches have been proposed to detect specific targets automatically through the monocular remote sensing image. Considering the basis of detection model, these methods can be categorized into five groups: template matching based methods, knowledge based methods, OBIA based methods, machine learning based methods and deep learning based methods.

In a template matching based method, a template for each target should be generated first and is used to match the image at each possible position to find the best matches. Weber J. et al (2012) defined a new morphological hit-or-miss transform and illustrated its potential as a template matching operator for coastline extraction and petroleum tank detection. Sirmaçek B. et al (2009) extracted buildings in urban area using a multiple sub graph matching method with scale invariant feature transform features (SIFT) calculated from two template building images. Although simple and powerful for some researches, template matching based methods are sensitive to shape and viewpoint change and need prior information and parameters for template designing. Knowledge based object detection approaches generally translate object detection problems into hypotheses testing problems by establishing various knowledge and rules. Huertas A. et al (1988) assumed the buildings are composed of rectangular components and detected buildings using a genetic model of the shapes. Ok et al. (2013) modeled the spatial relationship between buildings and their shadows to automatically detect buildings with arbitrary shapes from monocular very high resolution (VHR) remote sensing images. Similar to the template-based method, knowledge-based methods also require prior knowledge to define detection rules. Object-based image analysis (OBIA) partitions remote sensing imagery into meaningful image-objects and assessing their characteristics through spatial, spectral and temporal scale which is the basic conception of OBIA based target detection methods. Stumpf A. et al (2011) employed multi-resolution segmentation algorithm to get proposal objects and calculated color and shape metrics to prepare a sample database with all objects assigned either as landslide objects or non-landslide objects. After that, a random forest classifier was trained to detect landslide areas in the test image. Although the OBIA based methods are consistent with the basic knowledge of human beings, they still require prior knowledge to obtain proper segmentation results and group them into meaningful objects. Machine learning based object detection approaches can be performed by learning a classifier which captures the variation of object appearances and views from the prepared training dataset. Sun H. et al (2012.) combined spatial sparse coding bag-of-words representation with linear support vector machine for target detection. Han J. et al (2014) adopted the deep Boltzmann machine to learn high-level features and trained the object detector on Bayesian framework. With advanced machine learning classifiers and high-level features, machine learning based methods have high detection accuracy on the training dataset. But most of the features are handcrafted or shallow learning-based, whose capability become limited for more and more complex detection tasks.

Deep learning based detection methods take the advantage of deep learning to extract deep features directly from data via convolutional neural networks (CNN). Zhou P. et al (2016) developed a transferred deep model and integrated negative bootstrapping scheme into detector training. Li S. et al (2018)

* Corresponding author

divided the image into grids and predicted class label and position of each grid by a CNN model. Cheng G. et al (2016) introduced a new rotation-invariant layer on the basis of existing CNN architectures and imposed a regularization constraint, which explicitly enforces the feature representations of training samples insensitive to rotating, on the object function. Most of current deep learning based methods regard the target detection as a classification on sliding windows or proposal objects, where the CNN is the classifier. The way sliding windows or proposal objects generated affect the detection efficiency.

In this paper, target mask U-Net is proposed to detect wind turbines automatically from full size high resolution Gaofen-2 fused images. First pixel-wise binary masks of training images, which cover the wind turbines and their shadows, are generated. Next, both the images and masks are clipped and down sampled into blocks with specific scale. The training dataset is composed of these image blocks and mask blocks. Thereafter, the target mask U-Net, a predefined 44 layers end-to-end CNN, is trained on the training dataset. Then, test images with wind turbines to be detected are scanned by the target mask U-Net model to produce coarse segmentation results. Finally, some post processes are executed to eliminate errors and produce bounding boxes.

2. METHODOLOGY

The proposed automatic wind turbines detection using target mask U-Net has three main steps: (Fig.1)

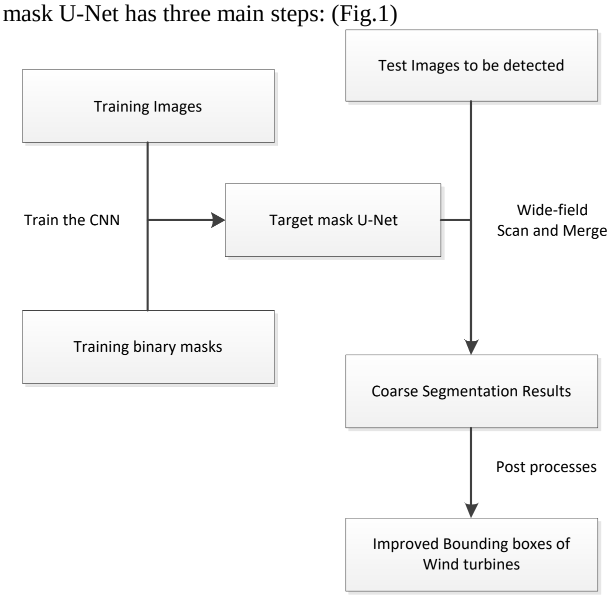


Figure 1. Proposed method for wind turbine detection

2.1 Step1: Target mask U-Net training

Wind turbines in remote sensing images are special for their shadows, which could be seen clearer than themselves. In this study, we considered the wind turbine and its shadow as a whole instance. Therefore, we first prepared some full size fused high resolution remote sensing images and drew polygons to outline the wind turbines and their shadows in these images. After rasterizing the polygons, the binary masks which indicate the location and shape of wind turbines and their shadow were generated for the corresponding remote sensing image. Then, each image was clipped into several blocks. Each block covered a wide-enough area with a plurality of wind turbines. We down sampled those blocks using affine transformation and bilinear interpolation algorithm. The binary masks were processed at the same time. Those down sampled image blocks and mask blocks together formed the training data for target mask U-Net.

The target mask U-Net, which is inspired by semantic segmentation researches in computer vision field, is the core mathematic model of detection. It is a typical encode-decode CNN which classify every pixel in images. The architecture of target mask U-Net is shown in Fig.2.

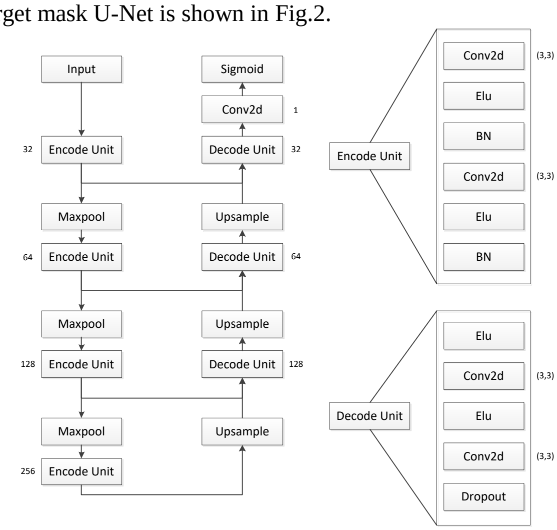


Figure 2. The architecture of target mask U-Net

In every convolution layer, a spatial convolution is performed to produce a set of feature maps. The spatial convolution over images is defined as

$$y_{i,j,d} = \sum_{i'=1}^{H} \sum_{j'=1}^{W} \sum_{d'=1}^{D} f_{i',j',d',d} \times x_{i'+i-1,\, j'+j-1,\, d'} + b_d \quad (1)$$

where $x_{i,j,d}$ and $y_{i,j,d}$ are the input and output values at position (i,j) in the dth channel image. The convolution kernel is f, sized 3x3 in study, and b_d is the bias in the dth channel. The exponential linear activation function

$$f(x) = \begin{cases} \alpha e^{x} - 1 & x < 0 \\ x & x \geq 0 \end{cases} \quad (2)$$

where x is the input to a neuron, model the activation mechanism of neural cells in human brain. To normalize the activation of previous layer and speed up training, the batch normalization (Ioffe S.et al 2015) layer applies a transformation that maintains the mean activation close to 0 and the activation standard deviation close to 1.

We put two groups of convolution, activation and batch normalization layers together to construct an encode unit. Four encode units mixed with max pooling layer were employed to extract different scale deep features of wind turbines in training dataset. Then we used three up-sample layers and decode units, which was composed of one Dropout (Srivastava, N. et al 2014) layer followed with two convolution and activation layers, to restore background information. The numbers of output filters of convolution layers in 7 units were 32,64,128,256,128,64,32. Moreover, skip connections were added to refine edges and speed up training. The final decode unit output was fed to a binary sigmoid classifier to classify each pixel independently.

In order to solve the unknown parameter in the target mask U-Net. The loss function was defined as (Long, J. 2015)

$$loss = \sum y \log(y') + (1-y)\log(1-y') \quad (3)$$

where *y* was the true binary value of pixels in mask blocks and *y'* was the predicted probability of target mask U-Net for the same pixel. We introduced Adam (Kingma P. et al 2014), an algorithm for first-order gradient-based optimization of stochastic objective functions based on adaptive estimates of lower-order moments, to optimize the loss function. During training, inner cross validation was used for model selection.

2.2 Step2: Wide-field Scan and Merge

After training the target mask U-Net, the next step is to input the test image to that model in a proper way. We scanned the image by a wide-field sliding window. While the window was sliding, every image in that window was down sampled using bilinear interpolation algorithm. The down sampled image was then fed into target mask U-Net to get coarse target score map in pixel-level. We up sampled the predicted result to make sure every pixel in the original window had predicted score. The score map window moved as the sliding window steps along axis. The values of pixels in overlapped area are averaged. Fig.3 shows a simple schematic of step 2.

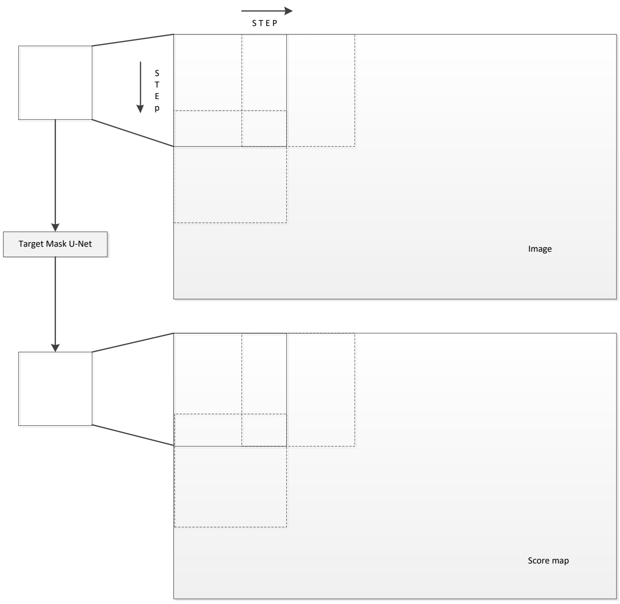


Figure 3. Wide-field scans and merges

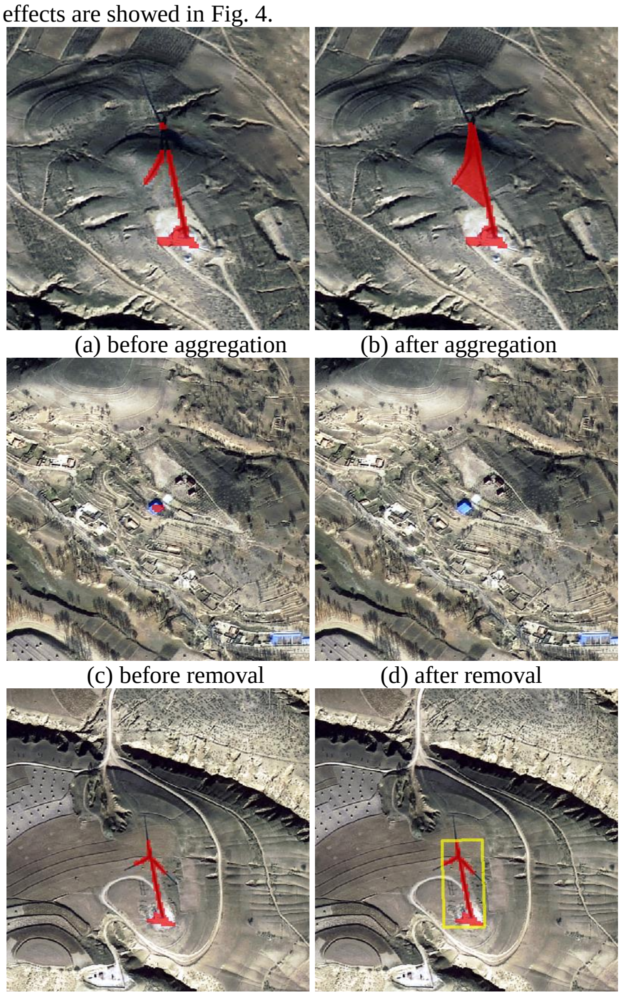
2.3 Step3: Post-processes

The step two produced a target score map for the input remote sensing image. A threshold was used to convert the continuous probability of every pixel to binary value (y_b).

$$y_b = \begin{cases} 0 & y' < \sigma \\ 1 & y' \geq \sigma \end{cases} \quad (4)$$

where y_b was the binary value, σ was the threshold. However, this operation produced many small objects and separate spots along the wind turbines in the binary mask. Some post-processes were employed to eliminate these errors.

In order to aggregate spots along the wind turbines, we combined polygons within a specified distance into a new polygon. Connected components smaller than the specified size were removed to eliminate solitary objects. The bounding box of each remaining connected area was generated to label the wind turbine in remote sensing image. The post-processes effects are showed in Fig. 4.



(a) before aggregation (b) after aggregation
(c) before removal (d) after removal
(e) before envelope (f) after envelope

Figure 4. Post-processes effects

3. EXPERIMENTAL RESULTS

3.1 Image Datasets and Experiment Parameters

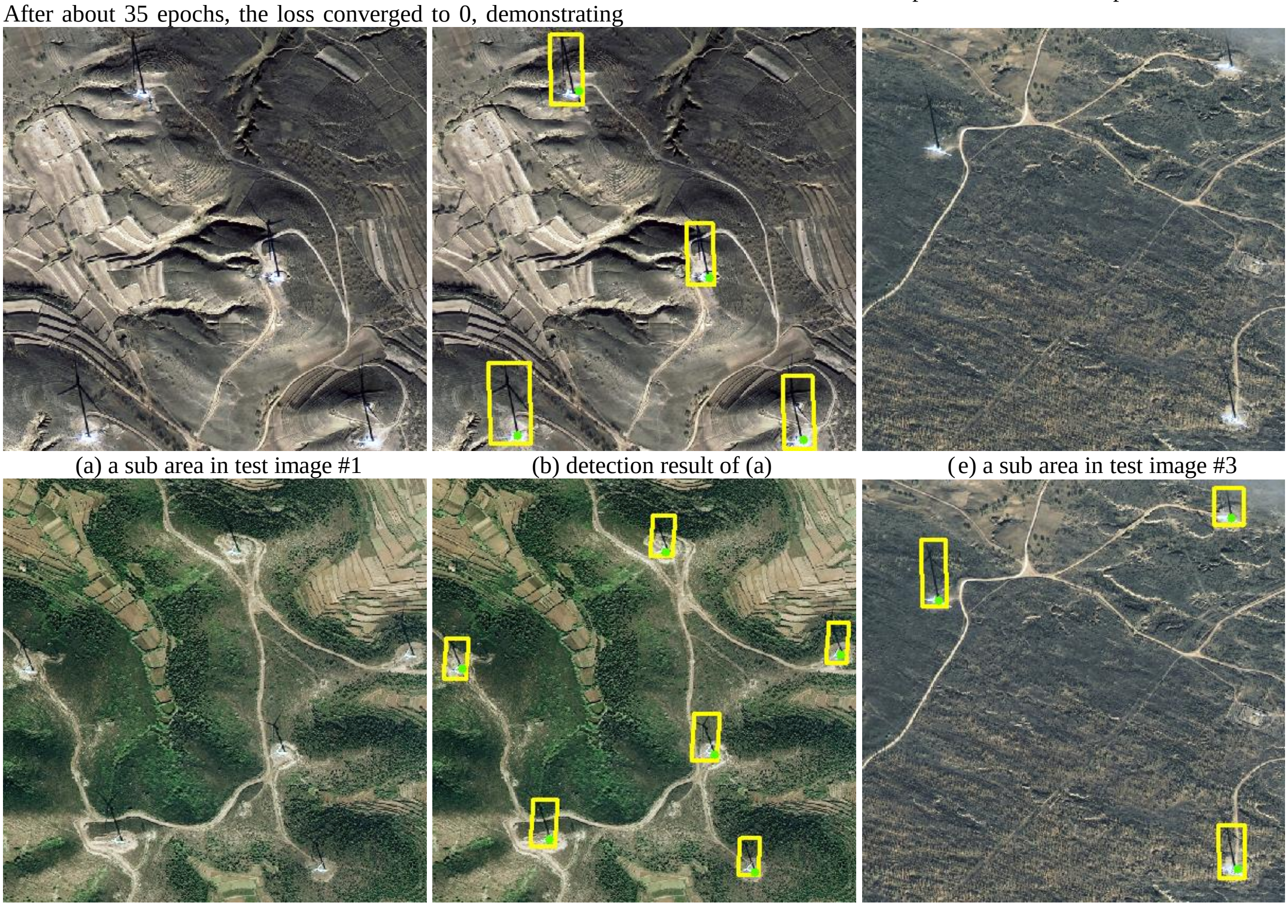
We have prepared five full size Gaofen-2 fused images, which have three bands (RGB) in 1 meter spatial resolution, to test our proposed method. These images are acquired from Shanxi Province and Shandong Province, R.P. China. There are more than fifty wind turbines in every image. The images are specially selected to diversify the characteristics of wind turbines such as shapes and sizes. Some areas sampled from the original image are shown in Fig. 5.

Item	Size (in pixel)	
	width	height
Training image #1	11993	12351
Training image #2	13853	6987
Clipped block	2048	2048
Down sampled block	512	512

Table 1. Parameters in training step

In this experiment, we choose two images for training and others for test. Parameters in training step can be found in Table 1.

After about 35 epochs, the loss converged to 0, demonstrating the unknown parameters have been solved. The size of sliding window was the same with down sampled block and the window slides 1024 pixels every step. σ in formula (4) was set to 0.5. The distance to be satisfied between polygon boundaries for aggregation to happen was 60 metres. The smallest allowable connected component size was 200 pixels.



(a) a sub area in test image #1 (b) detection result of (a) (e) a sub area in test image #3
(c) a sub area in test image #2 (d) detection result of (c) (f) detection result of (e)

Figure 5. Sample areas in test dataset and their detection results

3.2 Accuracy Assessment Strategy

In order to objectively evaluate the performance of our detection method, we have selected three commonly used metrics: precision, recall and F1-score.

We first labelled the location of wind turbines in test images (green points in Fig. 5). The metrics were defined as follows:

$$Precision = \frac{TP}{TP + FP} \tag{4}$$

$$Recall = \frac{TP}{TP + FN} \tag{5}$$

$$F_1 = \frac{2 \times precision \times recall}{precision + recall} \tag{6}$$

Where TP = total number of bounding boxes with label points inside.

FP = total number of bounding boxes with no label point inside.

FN = total number of label points outside of bounding boxes.

3.3 Results and Discussion

Visual interpretations of the detection results illustrated in Fig. 5 show that our method is robust and representative by detecting most of the wind turbines. In addition to visual illustration, the numerical metrics are listed in Table 2.

The precision rates of our detection method on test images are all more than 95%. The average precision is up to 0.98, demonstrating the exactness of detection results. The good rates of recall show that the target mask U-Net has satisfactory robustness. Further, the average F1-score is up to 0.97. Over all, these metrics indicate the reliability and applicability of our method. Furthermore, the metrics of another detection method (Zhou, P. et al, 2016.) are listed for comparison, which demonstrate the superiority of our method.

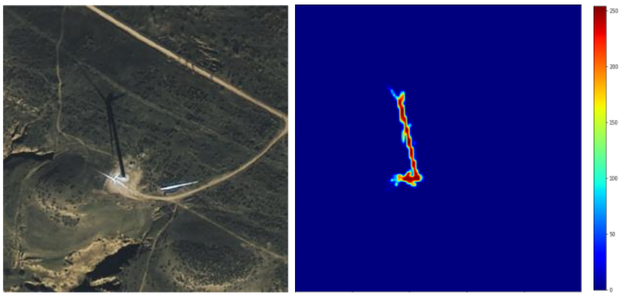


Figure 6. The score map of a wind turbine in test image

Fig. 6 shows a single wind turbine and its score map. Since the wind turbine is small thus diffcult to indentify, the shadow of it is very easy to recongnize in the image. One of the high lights of the proposed method is that we have forced target mask U-Net to learn the context relationship between them. The score map shows that the target mask U-Net has regarded the wind turbine and its shadow as one instance. The high prediction score of pillar section helps us to determine the location of the wind turbine. The thin and well-connected prediction of shadow enhances the reliability of detection. The low prediction score of separate vane on the ground, which has no shadow, indicates the context relationship has been used to distinguish similar objects in wind turbines detection.

The weakly supervised target detection based on CNN (Zhou, P. et al, 2016.) essentially translates the target detection problem into a classification problem in window level. Thus, both the size of sliding window and sliding step affect the accuracy and effectiveness of detection. It can be very time consuming to detect targets in an entire remote sensing image in that way. However, target mask U-Net, as a wide-field detector, not only is able to produce the pixel-level visualized prediction, but also has higher detection efficiency. Fig. 7 has revealed the differences in detection efficiency of these two methods.

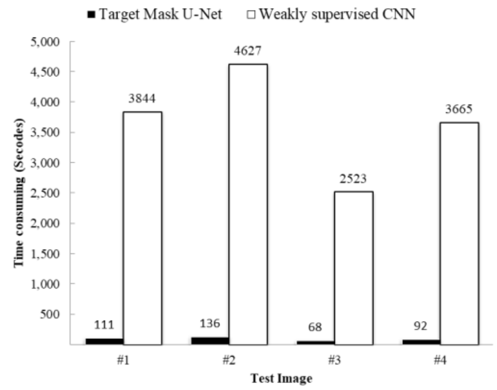


Figure 7. Time consuming of two detection methods

Test Image			Target Mask U-Net			(Zhou, P. et al, 2016.)		
	Width	Height	P.	R.	F_1	P.	R.	F_1
#1	22531	9823	0.98	1.0	0.99	0.76	0.82	0.79
#2	17047	15665	1.00	0.97	0.99	0.37	0.69	0.48
#3	11993	12351	0.97	0.95	0.96	0.81	0.62	0.70
#4	17375	12048	0.95	0.91	0.93	0.46	1.00	0.63
Average			0.98	0.96	0.97	0.60	0.78	0.65
Min			0.95	0.91	0.93	0.37	0.62	0.48
Max			1.00	1.00	0.99	0.81	1.00	0.63

Table 2. Numerical results of proposed method and Zhou’s method (P. means precision, R. means recall)

4. CONCLSION AND FUTURE WORKS

This study has explored a practical wind turbine detection method which introduces target mask U-Net as a wide-field detector to detect objects in pixel level. The context information between wind turbines and their shadows has been learnt by the network to improve the accuracy of detection. Moreover, the wide-field detector speed up the detection, saving much more time than weakly supervised target detection based on CNN. This study has implications for the study of rapid detection of significant targets with complex contextual information in high resolution remote sensing images. However, the detection results of our method are bounding boxes of wind turbines. The fine segmentation and location of wind turbines in the bounding boxes should be the topic of next researches.

REFERENCES

Cheng, G., & Han, J., 2016. A survey on object detection in optical remote sensing images. *ISPRS Journal of Photogrammetry and Remote Sensing*, 117, pp. 11-28.

Cheng, G., Zhou, P., & Han, J., 2016. Learning rotation-invariant convolutional neural networks for object detection in VHR optical remote sensing images. *IEEE Transactions on Geoscience and Remote Sensing*, 54(12), pp. 7405-7415. https://doi.org/10.1109/TGRS.2016.2601622

Han, J., Zhang, D., Cheng, G., et al, 2014. Object Detection in Optical Remote Sensing Images Based on Weakly Supervised Learning and High-Level Feature Learning. *IEEE Geoscience and Remote Sensing Society*, 53(6), pp. 3325-3337. https://doi.org/10.1109/TGRS.2014.2374218

Huertas, A., Nevatia, R., 1988. Detecting buildings in aerial images. *Computer Vision Graphics Image Processing*, 41, pp. 131-152.

Inglada, J., 2007. Automatic recognition of man-made objects in high resolution optical remote sensing images by SVM classification of geometric image features. *ISPRS journal of photogrammetry and remote sensing*, 62(3), pp. 236-248. https://doi.org/10.1016/j.isprsjprs.2007.05.011

Ioffe, S.,Szegedy, C., 2015. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *In International conference on machine learning*, pp.448-456.

Kingma, P., Ba, J., 2014.Adam: A Method for Stochastic Optimization. *arXiv preprint* arXiv:1412.6980

Li, S., Zhang, Z., Li, B., 2018.A Plane Target Detection Algorithm in Remote Sensing Images based on Deep Learning Network Technology. *J. Phys.: Conf. Ser.* 960 012025 https://doi.org/10.1088/1742-6596/960/1/012025

Lizy, A.;Sasikumar, M. (2014):Vehicle detection and classification from high resolution satellite images In: ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences, II-1,pp. 1-8. https://doi.org/10.5194/isprsannals-II-1-1-2014

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

Ok, A.O., 2013. Automated detection of buildings from single VHR multispectral images using shadow information and graph cuts. *ISPRS Journal of photogrammetry and remote sensing*, 86, pp. 21-40.

Sirmacek, B., Ünsalan, C., 2011. A probabilistic framework to detect buildings in aerial and satellite images. *IEEE Transactions on geoscience and remote sensing*, 49, pp. 211-221.

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R., 2014. Dropout: A simple way to prevent neural networks from overfitting. *The Journal of Machine Learning Research*, 15(1), pp. 1929-1958.

Stumpf, A., Kerle, N., 2011. Object-oriented mapping of landslides using Random Forests. *Remote Sensing of Environment.* 115, pp. 2564-2577.

Sun, H., Sun, X., Wang, H., et al, 2012.Automatic target detection in high-resolution remote sensing images using spatial sparse coding bag-of-words model. *IEEE Geoscience and Remote Sensing Letters*, 9(1), pp. 109-113. https://doi.org/10.1109/LGRS.2011.2161569

Weber, J., LefèVre, S.,2012. Spatial and spectral morphological template matching. *Image and Vision Computing*, 30(12), 934-945 https://doi.org/10.1016/j.imavis.2012.07.002

Zhou, P. et al, 2016. Weakly supervised target detection in remote sensing images based on transferred deep features and negative bootstrapping. *Multidimensional Systems and Signal Processing*, 27(4), pp. 925 https://doi.org/10.1007/s11045-015-0370-3