Rapid Target Detection in High Resolution Remote Sensing Images Using YOLO Model

Wu Zhihuan [1,2]*, Chen Xiangning [1], Gao Yongming [1], Li Yuntao [1]

[1] Space Engineering University, Beijing, China -wuzhihuan@hotmail.com
[2] 63883 Troops, Luoyang, China -wuzhihuan@hotmail.com

Commission VI, WG VI/4

KEY WORDS: Object Detection, High Resolution, Remote Sensing, Deep learning, YOLO

ABSTRACT:

Object detection in high resolution remote sensing images is a fundamental and challenging problem in the field of remote sensing imagery analysis for civil and military application due to the complex neighboring environments, which can cause the recognition algorithms to mistake irrelevant ground objects for target objects. Deep Convolution Neural Network(DCNN) is the hotspot in object detection for its powerful ability of feature extraction and has achieved state-of-the-art results in Computer Vision. Common pipeline of object detection based on DCNN consists of region proposal, CNN feature extraction, region classification and post processing. YOLO model frames object detection as a regression problem, using a single CNN predicts bounding boxes and class probabilities in an end-to-end way and make the predict faster. In this paper, a YOLO based model is used for object detection in high resolution sensing images. The experiments on NWPU VHR-10 dataset and our airport/airplane dataset gain from GoogleEarth show that, compare with the common pipeline, the proposed model speeds up the detection process and have good accuracy.

1. INTRODUCTION

1.1 General Instructions

Object detection is an important task for understanding high-resolution images and has very important military value. The purpose of target detection is to determine whether a given remote sensing image contains a target of interest type and determine the position information of the predicted target. "Target" usually refers to man-made objects, such as buildings, vehicles, airplanes, ships, etc., which have boundaries independent of the background environment, and feature information that is part of the background environment. With the rapid development of Remote Sensing (RS) technology, the RS imagery produced by high-resolution remote sensing satellites (such as IKONOS, SPOT-5, WorldView and Quickbird) have more abundant information to extract features and detect ground object than the low-resolution remote sensing imagery. Many artificial objects that are difficult to be detected in the past are now available to be detected. Since the 1980s, Object detection in remote sensing image is widely studied, mainly using shallow features that were hand-engineered by skilled people who have experience in the field and also often required domain-expertise. This also means that if the conditions change even slightly, a framework which works well in a given task may fail in another task. So that the whole feature extractor might have to be rewritten from scratch, which is very time consuming and expensive. These disadvantages led researchers in the field looking for a more robust and effective approach.

In 1998, Yan LeCun et al. proposed a handwritten digital recognition method using neural networks which allowed them to achieve more than 99% accuracy in the digit recognition task. The result re initiated interest in researches of using the neural network for image recognition application. But limited by the lack of computational power and effective techniques to train neural network model for more complex tasks, correlation researches were largely abandoned for many years. In 2012, Alex Krizhevsky et al. won the ILSVRC competition by a model with 7-layer convolutional neural network named AlexNet, and that result opened the floodgates to new research in the field with the name "deep learning". Concept of convolution layer introduced from Convolutional neural network (CNN) uses two methods to greatly reduce the number of parameters: local receptive field and parameter sharing. Local receptive field is a principle learned from human visual that the pixels in an image have more relevant with the adjacent pixels. Instead of having each neuron receive connections from all neurons in the previous layer, CNNs use a receptive field-like layout in which each neuron receives connections only from a subset of neurons in the previous (lower) layer. The use of receptive fields in this fashion is thought to give CNNs an advantage in recognizing visual patterns when compared to other types of neural networks. Parameter sharing scheme is used in Convolutional Layers to control the number of parameters. Generally, the statistical features of part of the image is considered same with the other parts. So that the convolution kernel is used as a feature extraction method regardless of the position, pool layer is used to represent the implied invariance from transformation of the image. A preliminary set of experiments fusing CNN obtains state-of-the-art results for the well-known UCMerced dataset. The researches show that the CNN model can be generalized to the field of remote sensing imagery and obtain better results than the traditional methods.

* Corresponding author

2. OBJECT DETECTION FOR REMOTE SENSING IMAGERY

2.1 Machine Learning based object detection

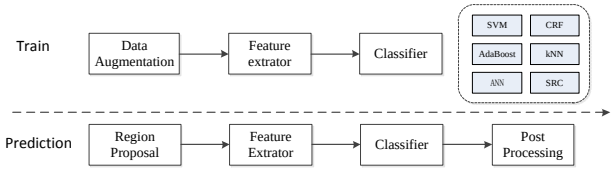

Figure 1. Framework of machine Learning based object detection

With the development of machine learning methods, especially feature extraction and classification methods, a major breakthrough has been made in the field of computer vision to translate the target detection problem into a classification problem for machine learning. The diagram shows a typical machine learning-based target detection method implementation flow. The training process learns a classifier from the training data set in a supervised, semi-supervised or weakly supervised manner. In the detection process, a series of regional image blocks extracted by a sliding window or candidate region extraction algorithm are extracted by a feature extraction algorithm as a classifier input, and a corresponding prediction class label is output (for example, whether the target is for a binary classification output). Usually, the detection is performed. The results also need to be corrected and optimized by post-processing operations. Candidate area extraction algorithms include sliding windows, superpixel segmentation, EdgeBox, SelectSearch, Bing, etc. Feature extraction algorithms include HOG, BoW, texture features, sparse representation features, Haar features, and CNNs-based methods; classifiers include SVM, AdaBoost , k-nearest neighbors (kNN), conditional random fields (CRF), sparse representations (SRC) and artificial neural networks (ANN); post-processing methods include NMS, Box-fusion, and Bounding-box Regression.

2.2 CNN based object detection

Traditional CNNs are usually used for classification task, and the output of classification is discrete class label. In object detection tasks for the remote sensing image, more information such as location of object needs to be obtained. Each pixel in the image should be predicted and labeled by classification algorithm. In the field of computer vision, object detection methods for image based on CNN can be divided into two types: object localization and semantic segmentation. As shown in Fig 1, the target of the object localization is to obtain the bounding box of object in input image, and the target of scene segmentation is to obtain a predicted mask for input image. RCNN (proposed by Girshick, Ross, et al. in 2014) is a representative method of object location methods which built a framework to solve problem of object location. RCNN consist of a Region Proposal algorithm (such as Selective Search, BING and so on) to extract a large number of ROIs (Region Of Interest) from the input image, a CNN model to extract the features of the ROIs, a classifier to classify the ROIs, and a regression algorithm to correct the border position of bounding box.

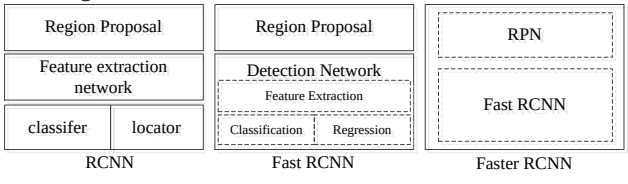

Figure 2. Framework of machine Learning based object detection

The main disadvantages of RCNN are the computational complexity and the accuracy of border position. Spatial pyramid pooling networks (SPPnet) implement computational sharing acceleration R-CNN. SPPnet calculates the convolution map of the entire image. The classifier extracts the feature vectors of the candidate regions from the feature map for classification. The input image of the R-CNN algorithm is limited to a fixed size, which limits the flexibility of the algorithm. The adjustment of the image size is mostly through clipping or deformation. This operation will result in a certain degree of information loss. SPPnet can pool any size image into a fixed-length feature representation. SPPnet accelerates RCNN10-100 times. The training process of SPPnet is divided into multiple steps like R-CNN, including extracting features, tuning the network, training the SVM, and finally the bounding box regression. The tuning algorithm cannot update each convolutional layer weight before SPP, limiting the depth of the convolutional layer.

The speed of area extraction becomes the bottleneck of the detection system. To overcome this problem, Faster RCNN replaces the selective search module in the RCNN/fast RCNN with a region proposal network (RPN), which further improves the accuracy and speed of detection, and can be achieved without considering the region extraction. Near real-time detection. You Only Look Once(YOLO) solves the object detection as a regression problem. After an input image passes an Inference, it can obtain the position, category and corresponding confidence probability of all objects in the image, further improving the detection speed and realizing the real-time target detection.The RCNN model (as shown in the figure) uses the SelectSearch algorithm to extract candidate regions, CNN as a feature extraction, SVM as a classifier, and the Bounding-box Regression algorithm to perform candidate box regression. This framework has also become a target detection model based on deep learning. Foundation. In the RCNN algorithm, each candidate image needs to be separately classified, and the time overhead is large. For the shortcomings of classic RCNN models with low efficiency, Ross B. Girshick and others successively proposed Fast-RCNN to reduce the repetitive calculation after transferring the region extraction to Feature Map; Faster-rcnn uses the Region Proposal Network (RPN) to complete the region extraction function; YOLO directly In the output layer, the location of the regression bounding box and its category, real-time target video detection. Based on Faster R-CNN, Han X [] proposed that R-P-Faster R-CNN achieves high-efficiency target detection for high-resolution remote sensing images.

3. OBJECT DETECTION BASED ON YOLO FRAMEWORK

Type text single-spaced, **with** one blank line between paragraphs and following headings. Start paragraphs flush with left margin.

3.1 Architecture

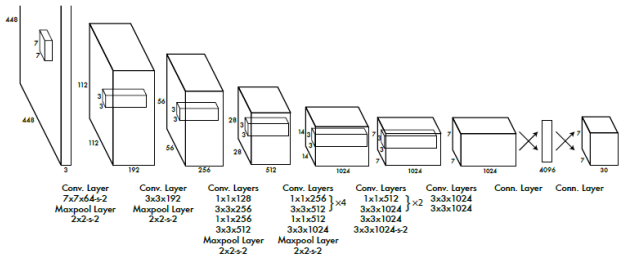

Figure 3. Framework of YOLO

The YOLO model is a CNN based framework. At the beginning, the convolutional image extraction feature predicts the output probability through the all-connected layer. Through similar to GoogleNet, also trained YOLO quick version, producing Fast-YOLO at the final output of 7 * 7 * 30 tensor. Figure 1 shows the

yolo network structure, which has 24 convolutional layers and 2 fully connected layers. The alternating 1 × 1 convolution layer reduces the feature space of the previous layer, enhancing the resolution of detection by pretraining the convolutional layers on the ImageNet classification. The final output of our network is the 7 × 7 × 30 tensor of predictions.

3.2 Model Training

(1) Data augmentation
The purpose of data augmentation is to generate new sample instances, and when the training samples are few, data augmentation is very useful for improving the robustness of the network. For remote sensing imagery, Methods such as rotating 45 degrees, scaling 15-25%, cutting, switching frequency band, vertical/horizontal flipping and other image operations is used in this paper to increase the generalization capacity of the network. The above operations are not used to the validation set and the test set.
(2) Optimization
In CNNs, SGD (Stochastic Gradient Descen) is the most common optimizer, but it has some problems in finding the appropriate learning rate and easily converging to local optimum. In this paper, Adam (Adaptive Moment Estimation) optimizer is used as learning strategy to calculate the adaptive learning rate of each parameter. Adam optimizer adjusts the learning rate of each parameter according to the first moment estimation and the second moment estimation of its gradient function of loss function. Adam optimizer runs fast and can correct the problems of learning rate disappears, slow convergence speed and great fluctuation of the loss function. Adam optimizer is shown as below formulas.

$$\hat{m}_t = m_t / 1-\beta_1^t \tag{1}$$

$$\hat{v}_t = v_t / 1-\beta_2^t \tag{2}$$

$$\theta_{t+1} = \theta_t - \frac{\eta}{\sqrt{\hat{v}_t} + \varepsilon}\hat{m}_t \tag{3}$$

where c = focal length
x, y = image coordinates
X_0, Y_0, Z_0 = coordinates of projection center
X, Y, Z = object coordinates

Formula 1 represents first moment deviation, formula 2 represents second moment deviation. η is step, ε is a little constant (default value is 10^{-8}).
(3) Loss function
In mathematical optimization, statistics, econometrics, decision theory, machine learning and computational neuroscience, a loss function or cost function is a function that maps an event or values of one or more variables onto a real number intuitively representing some "cost" associated with the event. An optimization problem seeks to minimize a loss function. An objective function is either a loss function or its negative (in specific domains, variously called a reward function, a profit function, a utility function, a fitness function, etc.), in which case it is to be maximized. In YOLO, the mean square error is used as the loss function to optimize the model parameters, and the mean square error of the vector output from the network and the vector corresponding to the real image. As shown below, where coordError is the coordinate error, iouError is the IOU error, and classError is the classification error.

$$loss = \sum_{i=0}^{S^2} coordError + iouError + classError \tag{4}$$

$$\begin{aligned}
&\lambda_{coord}\sum_{i=0}^{S^2}\sum_{j=0}^{B}1_{ij}^{obj}\left[(x_i-\hat{x}_i)^2+(y_i-\hat{y}_i)^2\right]\\
&+\lambda_{coord}\sum_{i=0}^{S^2}\sum_{j=0}^{B}1_{ij}^{obj}\left[(\sqrt{w_i}-\sqrt{\hat{w}_i})^2+(h_i-\sqrt{\hat{h}_i})^2\right]\\
&+\sum_{i=0}^{S^2}\sum_{j=0}^{B}1_{ij}^{obj}(C_i-\hat{C}_i)^2\\
&+\lambda_{noobj}\sum_{i=0}^{S^2}\sum_{j=0}^{B}1_{ij}^{obj}(C_i-\hat{C}_i)^2\\
&+\sum_{i=0}^{S^2}1_{ij}^{obj}\sum_{c\in classes}(p_i(c)-\hat{p}_i(c))^2
\end{aligned} \tag{5}$$

where x, y, w, C, p = network prediction
$\hat{x}, \hat{y}, \hat{w}, \hat{C}, \hat{p}$ = groud truth
1_i^{obj} means object fall into the lattice i
1_{ij}^{obj} means object fall into the the bounding box j in lattice i
1_{ij}^{noobj} means object not fall into the the bounding box j in lattice i

3.3 Post processing

The last step of YOLO is to perform non-max suppression (NMS) on S*S*(B*5+C) vectors. The purpose of the NMS is to eliminate redundant boxes and find the best object detection location. NMS algorithms are used in well-known target detection frameworks such as RCNN and SPPnet.

3.4 Fine-tuning

Transfer learning strategies depend on various factors, but the two most important ones are the size of the new dataset, and its similarity to the original dataset. Fine-tuning allows us to bring the power of state-of-the-art DCNN models to new domains where insufficient data and time/cost constraints might otherwise prevent their use.

3.5 Evaluation metric

In order to evaluate performance of the algorithm, results of the model should be compared to the ground truth. In this paper, Precision and recall are used to evaluate the similarity and diversity between detection results and ground truth in test dataset.

$$\mathrm{P} = TP/(TP+FP) \tag{6}$$

$$\mathrm{R} = TP/(TP+FN) \tag{7}$$

where TP = true positive
FP = false positive
FN = false negative.

3.6 Advantages and disadvantages

Advantages:
1) Faster. YOLO solves object detection as a regression problem, that simplify the entire pipeline of detection network.
2) Lower background false positive rate. During the training and prediction process, the information of the entire image can be perceived. The RCNN-based method can only perceive local information from the candidate frame.

3) Strong versatility. YOLO can be applied in object detection for unnatural images.
Disadvantages:
1) The model uses multiple downsampling layers. The characteristics of the objects learned on the Internet are not fine, which affects the detection effect.
2) Poor recognition of object position accuracy.
3) The detection of small targets and dense targets is poor.
4) Lower recall rate.

4. EXPERIMENTS

4.1 Hardware and software environment

The model is implemented by Keras with Tensorflow backend. Keras is a high-level neural networks API, written in Python and capable of running on top of TensorFlow, CNTK, or Theano. Tensorflow developed by the Google Brain team is an open-source software library for dataflow programming across a range of tasks. Many 3rd party libraries are required such as Tifffle for reading remote sensing imagery, OpenCV for basic image processing, Shapely for handling polygon data, Matplotlib as visualization tool, Imglabel for dataset construction. The experiments run on a Sugon W560-G20 Server with E5-2650 v3 CPU, 32GB memory, and Quora k2000 GPU.

4.2 Experimental process

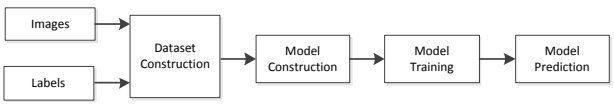

Figure 4. Experimental Process
The experimental process can be divided into four steps:
Dataset construction. The remote sensing images (gain from open source data source such as GoogleEarth, USGS, DigitalGlobe and so on) are animated using imgLabel to obtain standard PASCAL_VOC format dataset. Divide labeled dataset into training, validation, and test sets.
Model construction. Constructing a CNN structure and setting its hyperparameters.
Model training. Training with training sets and validation sets.
Model prediction. Testing with the test set, and the result is used for evaluates model.

4.3 Dataset

(1) Airport dataset and airplane dataset gain from GoogleEarth and manually annotated using Imglabel software. Airport dataset contains 1893 remote sensing images. Airplane dataset contains 250 remote sensing images. Figure 5 shows two demos of image label.

Figure 5. Demo of image label
(2) NWPU VHR-10 dataset. It is a publicly available 10-class geospatial object detection dataset.These ten classes of objects are airplane, ship, storage tank, baseballdiamond, tennis court, basketball court, ground track field, harbor, bridge,and vehicle. This dataset contains totally 800 very-high-resolution (VHR)remote sensing images that were cropped from Google Earth and Vaihingen dataset and then manually annotated by experts.

4.4 Results

Figure 6 shows airports detected using our method.

Figure 6. Results of airport detection

TF	miss detection rate	Recall	FP	false detection rate	TP	all
23	11.5%	88.5	8	4%	169	200

Table 1. Performance of airport detection
Figure 7 shows airplane detected using our method.

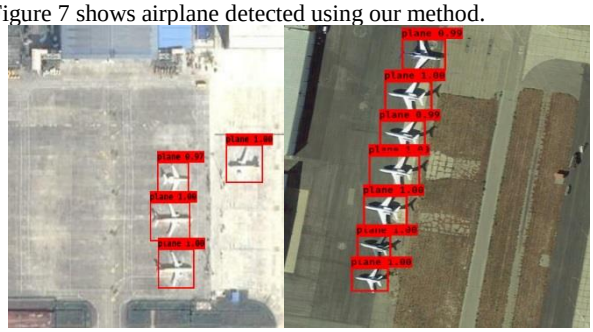

Figure 7. Results of airplane detection

TF	miss detection rate	Recall	FP	false detection rate	TP	all
16	7.34%	92.66%	29	13.30%	206	218

Table 2. Performance of airplane detection.

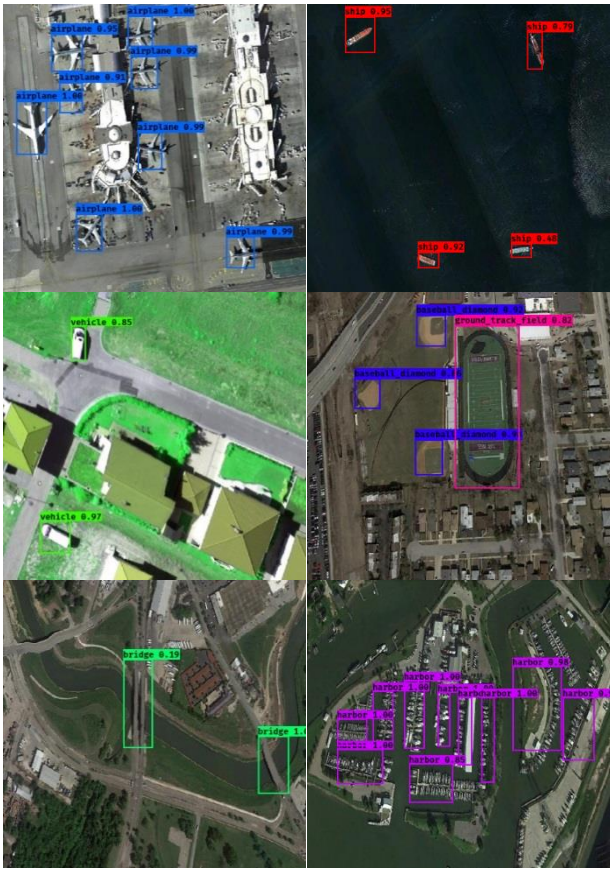

Figure 8. Detection results of NWPU VHR-10 dataset

Results and performances of airport detection and airplane detection are shown in Figure 6-7 and Table 1-2. Results of object detection of NWPU VHR-10 dataset are shown in Fighre 7.

R-CNN	Fast R-CNN	Faster R-CNN	YOLO
64.8	3.3	0.9	0.1

Table 3. Testing time of four object detection methods.

As is shown in Table 3, YOLO model greatly improved the speed of detection and can reach the requirement of real time

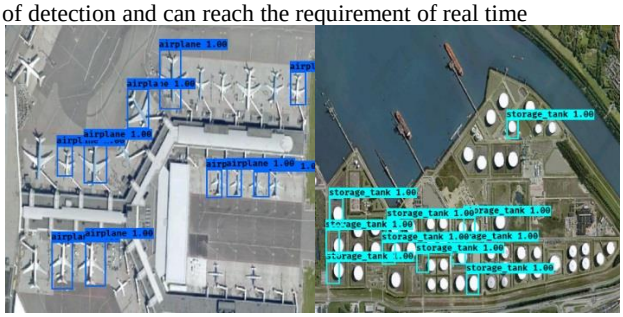

Figure 9. Results of object detection with small and dense objects.

As is shown in Figure 9. YOLO do not perform well for objects that are very close to each other (middle points of more than one objects fall into the same grid), and for small object group. Because there are only two boxes belong to one category are predicted in a grid. There are problems of bad training approximation and generalization for test image of unusual aspect ratio. Due to the problem of loss function, positioning error is one of the most important reason that affects the detection effect, specially for large or small objects.

5. CONCLUSION

This work addresses the problem of rapid object detection for high-resolution remote sensing image with CNNs. A YOLO model is used in this paper for object detection in high resolution remote sensing images. Experiments on NWPU VHR-10 dataset, our airport dataset and airplane dataset gain from GoogleEarth demonstrate that YOLO model has a strong applicability for remote sensing image, especially in speed of prediction. The main disadvantages of YOLO are its poor positioning accuracy, bad training approximation and generalization for images of unusual aspect ratio and objects that are very close to each other. It needs a large number of high quality Ground Truth labels for the model training, which relies on professional interpretation experiences and lots of manual work. Therefore, to solve these problems is orientation of the future research.

REFERENCES

Penatti O A B, Nogueira K, Santos J A D, Do deep features generalize from everyday objects to remote sensing and aerial scenes domains?, Computer Vision and Pattern Recognition Workshops. IEEE, 44-51, 2015.

Azayev T, Object detection in high resolution satellite images, Czech Technical University, 2016.

Zhong Y, Fei F, Liu Y, et al, SatCNN: satellite image dataset classification using agile convolutional neural networks, Remote Sensing Letters, Vol.8, No.2, 136-145, 2017.

Ball, John E., and C. S. Chan., Comprehensive survey of deep learning in remote sensing: theories, tools, and challenges for the community." Journal of Applied Remote Sensing 11.4(2017).

Redmon, J., Divvala, S., Girshick, R., & Farhadi, A., 2016. You Only Look Once: Unified, Real-Time Object Detection. Computer Vision and Pattern Recognition (pp.779-788). IEEE.

Ren, Shaoqing, et al., Faster R-CNN: towards real-time object detection with region proposal networks, International Conference on Neural Information Processing Systems MIT Press, 2015:91-99.

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

Long, Yang, et al., Accurate Object Localization in Remote Sensing Images Based on Convolutional Neural Networks, IEEE Transactions on Geoscience & Remote Sensing 55.5(2017):2486-2498.

Cao, Yu She, X. Niu, and Y. Dou, Region-based convolutional neural networks for object detection in very high resolution remote sensing images, International Conference on Natural Computation, Fuzzy Systems and Knowledge Discovery IEEE, 2016:548-554.

Kingma D P, Ba J, Adam: A Method for Stochastic Optimization, Computer Science, 2014.

Wang, Jia Qi, et al, Aircraft Detection in Remote Sensing Images via CNN Multi-scale Feature Representation, smce(2017).

Han, Xiaobing, Y. Zhong, and L. Zhang, An Efficient and Robust Integrated Geospatial Object Detection Framework for High

Spatial Resolution Remote Sensing Imagery, Remote Sensing 9.7(2017):666.

Shi, Shaohuai, et al, Benchmarking State-of-the-Art Deep Learning Software Tools, (2016).

Cheng, Gong, et al, Object detection in VHR optical remote sensing images via learning rotation-invariant HOG feature, International Workshop on Earth Observation and Remote Sensing Applications IEEE, 2016:433-436.

Cheng, Gong, et al, Multi-class geospatial object detection and geographic image classification based on collection of part detectors, Isprs Journal of Photogrammetry & Remote Sensing 98.1(2014):119-132.

Cheng, Gong, P. Zhou, and J. Han, Learning Rotation-Invariant Convolutional Neural Networks for Object Detection in VHR Optical Remote Sensing Images, IEEE Transactions on Geoscience & Remote Sensing 54.12(2016):7405-7415.

Revised October 2017