DOI: 10.1051/itmconf/20171201006

# Ship Detection Using Transfer Learned Single Shot Multi Box Detector

Gu-Hong Nie[1], Peng Zhang[2], Xin Niu[3], Yong Dou[4], Fei Xia [5]

[1]*National Lab for Parallel and Distributed Processing, National University of Defense Technology, Changsha, China*
[2]*National Lab for Parallel and Distributed Processing, National University of Defense Technology, Changsha, China*
[3]*National Lab for Parallel and Distributed Processing, National University of Defense Technology, Changsha, China*
[4]*National Lab for Parallel and Distributed Processing, National University of Defense Technology, Changsha, China*
[5] *Institute of Electronic Information Warfare, Naval University of Engineering, Wuhan, China*
[1] nieguhong15@nudt.edu.cn,[2] zhangpeng14f@nudt.edu.cn, [3]niuxin@nudt.edu.cn, [4]yongdou@nudt.edu.cn, [5] xcyphoenix@nudt.edu.cn

**Abstract:** Ship detection in satellite images is a challenging task. In this paper, we introduce a transfer learned Single Shot MultiBox Detector (SSD) for ship detection. To this end, a state-of-the-art object detection model pre-trained from a large number of natural images was transfer learned for ship detection with limited labeled satellite images. To the best of our knowledge, this could be one of the first studies which introduce SSD into ship detection on satellite images. Experiments demonstrated that our method could achieve 87.9% AP at 47 FPS using NVIDIA TITAN X. In comparison with Faster R-CNN, 6.7% AP improvement could be achieved. Effects of the observation resolution has also been studied with the changing input sizes among 300 x 300, 600 x 600 and 900 x 900. It has been noted that the detection accuracy declined sharply with the decreasing resolution that is mainly caused by the missing small ships.

## 1. Introduction

Ship detection from satellite imagery plays an important role in the maritime surveillance, e.g. traffic monitoring, fishing management, oil pollution control etc. In the past few years, with the development of remote sensing technologies, the optical satellites can provide optical imagery with high resolution, making ship detection in optical satellite images a hotspot. There is some prior knowledge about the ships in satellite images: (1) the shape of ship is rectangle, with small width and large height, (2) the length of ships is in large various, (3) complex surroundings, including open water and harbor, (4) ship directions are almost in a circle unit. All the above make the ship detection become a challenging task.

The object detection task is usually composed of feature extraction and classification. Like other object detection tasks, the ship detection performance is proportional to the description ability of ship features. According to the feature extraction method, the previous work on ship detection can be classified into two groups, handcraft feature and machine learning feature.

Over the handcraft feature, researchers proposed some adaptive feature extraction methods according to the expert knowledge about the ship in satellite imagery. Some ships are just a few pixels in size in satellite images, and the ships may be highly blurred, leaving only object outlines as differentiable. So the handcraft feature approaches for ship detection are often based on shape, edge, and texture features [1]. However, the description ability of the handcraft feature is limited to the human cognition. There are still many deviations in the human cognition over the underlying feature. Besides, the handcraft feature approaches have poor generalization ability. When the application scene is not uniform or the images are polluted by noise, the detection performance may be affected. The HOG-based classifiers [2] [3] [4] may break down in crowded or complex scenes. Due to the variety of ship size, the handcraft feature approaches may also lead to poor performance for ship detection.

On the other side, machine learning feature makes a great profit for object detection. These approaches take advantage of some prior knowledge, automatically learn image features, and find potential object characters and the distribution rules over the object which cannot be described by human cognition. Recent years, the machine learning has opened up prospects for superior image classification and detection, especially the success of deep CNN since 2012 [5]. Zhu et al. [6] explored orientation robust features from combined layers of DCNN for object detection in aerial image. Tang and Deng et al. [7] proposed compressed domain ship detection system using deep neural network and extreme learning machine (ELM) [8]. However, the existing ship detection methods using DNN utilize the features from top layer. While the ships are only a few pixels in size or blurred, the model may generate false candidates confused by small islands, harbors etc.

The deep learning (DL) algorithm automatically extracts the representative and discriminative feature from bottom level (e.g. edge and texture) to top level (underlying feature) in a hierarchical manner and turns to be a hotspot in machine learning [9]. Current state-of-the-art object detection systems using deep learning are variants of the following approach: hypothesize bounding boxes, resample pixels or features for each box, and apply a high quality classifier. The methods of generating candidate boxes in Region-Based Convolutional Networks (R-CNN) [10] can be Selective Search [11], EdgeBox [12] or BING [13], but they are similar to sliding window essentially. Though SPP-NET [14], Fast R-CNN [15] and Faster R-CNN [16] extract features from convolutional layer and avoid the large amount of repeated convolution computation for every candidate region. There are still many unnecessary search, resulting in a large amount of computation. The Fast and Faster R-CNN adapt regression learning method to amend the object position. YOLO [17] and SSD [18] generate object position and category from the network directly.

There are some differences between satellite images and natural images. Firstly, the range of satellite images usually covers a large area, which is much larger than natural images, and the object size in the satellite images is much smaller. Secondly, the satellite images are simpler. The satellite images are taken in nearly constant zenith view angle along with simple surroundings. Besides, the direction of objects in satellite images is in circle. Nevertheless, the optical satellite images still share common visual features in low and middle abstract levels with natural images.

SSD, a fully convolutional network, discretizes the output space of bounding boxes into a set of default boxes over different aspect ratios and scales per feature map location [18]. The SSD achieved state-of-the-art detection performance on the PASCAL VOC, MS COCO, and ILSVRC datasets. During the ship detection task, the shape of ships is simple, but the size scale is large. Combined with the feature map from different layers, we can project the ship feature to different layers according to the ship size automatically.

To the best of our knowledge, this could be one of the first studies which introduce SSD into ship detection on satellite images. The rest of this paper is organized as follows. In section II, we detail the ship detection methodology. In Section III, Experiments and analysis are shown. Finally, we conclude this paper in section IV.

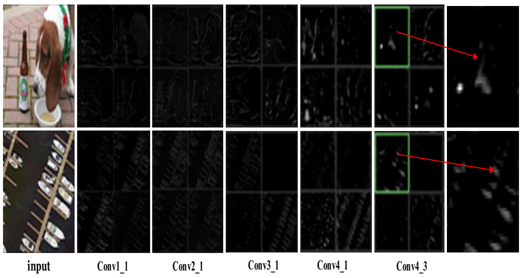

Figure 1. Feature maps from low to middle layers for natural and satellite image using pre-trained VGG-16

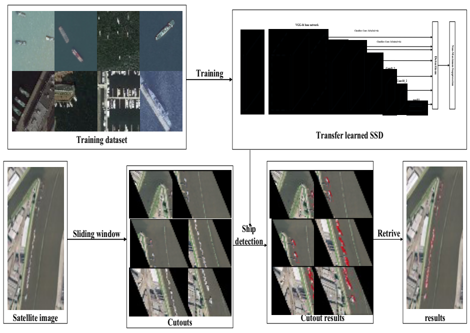

Figure 2. Framework of the ship detection

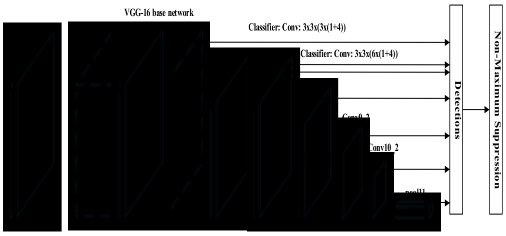

Figure 3. Transfer learned SSD architecture

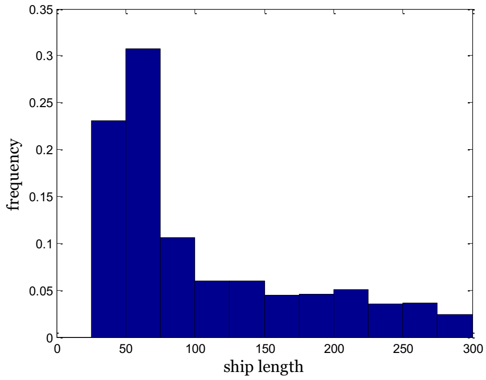

Figure 4. statistic of the ship length in satellite images.

## 2. Methodlogy

### 2.1 Framework

Though the satellite images are abundant, the original images are usually megapixel and the ships are small compared to the image range, which makes labeling work complex and complicated. Training an effective DNN model means to estimate millions of parameters, which requires a large number of labeled data. There is a fact that the convolutional neural networks provide means to learn rich low and middle level features transferrable to a variety of visual recognition tasks [19]. As is shown in Fig. 1, there are feature maps among conv1_1, con2_1, con3_1, con4_1 and conv4_3 for natural image and satellite image, using the pre-trained VGG-16 [21] model. We hypothesize that there are many common low and middle level visual features between satellite images and natural images. Transfer learning aims to transfer knowledge between related source and target domains [20]. Therefore, we transfer the SSD from natural images to satellite images domain. We use the transfer learned SSD for ship detection. The framework is shown in Fig. 2. The details of the transfer learned SSD are described later.

## 2.2 Details

As is shown in Fig. 3, our transfer learned SSD architecture is based on a feed-forward CNN. The number of labeled satellite images is limit and is not sufficient to train an effective DNN. We extract the low and middle level visual feature with the pre-trained base network of VGG-16. Fig. 4 shows the statistic on ship length in the satellite images. The size of ships is multi-scale. We concatenate the end of truncated base network with convolutional feature layers, which decrease in size. The reception filed of those layers is different. The multi-scale problem can be solved by utilizing the feature map from different layers. For each Extra feature layer (or the top layer in the base network), we can produce a fixed set of detection prediction, including the scores for each class, and the offsets of the bounding boxes, using a set of small convolutional filters. In the architecture, the filter size is 3 x 3. For a layer with m x n x p feature map, a 3 x 3 x p small kernel can produce a score for a class or the bounding box offset. The ships in the satellite images usually moored in a circle unit. We choose some default aspect ratios for ship detection. As is shown in Fig. 3, we use Conv4_3, Conv6 (FC6), Conv7 (FC7), Conv8_2, Conv9_2, Conv10_2 and pool11 to generate the ship confidence and offsets relative to the default box. In the end, we use the non-maximum suppression (NMS) step to produce the final detections.

# 3. Experiment

Our database contains 206 images covering the harbor or open water. All the satellite images were collected from Google Earth with 0.54m ground sample distance (GSD). And the images are taken near the Pacific and the Atlantic States. The image size is various, ranging from 900 x 900 to 3600 x 5400. For each of the original satellite images, we cut it in a fixed size using sliding window. A total of 138 original images and 3898 ships were used as a training set. A total of 34 original images and 1265 ships were used as the validation set. The others containing 34 original images and 1383 ships were used as the test set.

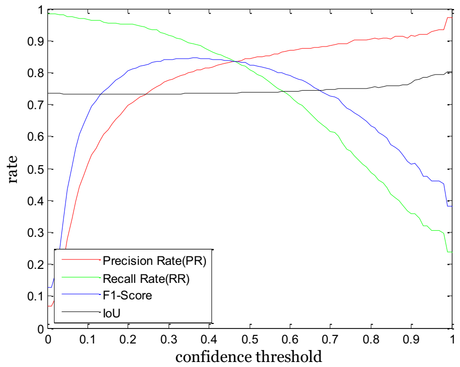

Figure 5. Ship detection result in our ship test set with original pixel.

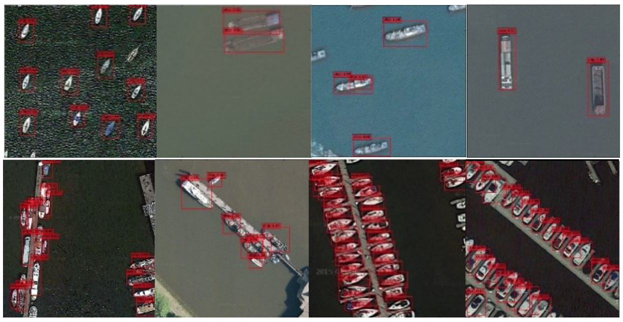

Figure 6. Some detection result by transfer learned SSD in open water

and harbor. The four images in the first row are the detection results in the open water, and the other four images in the second row is the detection results in the harbor.

Table1. Detection Result Using Transfer Learned Ssd

| Detectio n result | Given Recall Rate | | | | | |
|---|---|---|---|---|---|---|
| | *95.0 %* | *90.0 %* | *85.0 %* | *80.0 %* | *75.0 %* | *70.0 %* |
| Precisio n Rate | 71.0 % | 79.2 % | 82.6 % | 84.5 % | 85.9 % | 87.5 % |
| F1-Score | 0.813 | 0.843 | 0.838 | 0.822 | 0.801 | 0.778 |
| IOU | 0.732 | 0.734 | 0.736 | 0.737 | 0.739 | 0.743 |

## 3.1 Ship Detection Using Transfer Learned SSD

The input size of the transfer learned SSD in our framework is 300 x 300. All the input images in our framework will be resized to the fixed size. We use the original image pixels and cut the satellite images with 300 x 300 scale using sliding window method. The precision rate (PR), recall rate (RR) and F1-Score are defined as (1).

$$\begin{cases} PR = \frac{number\ of\ detected\ ships}{number\ of\ detected\ objects} \\ RR = \frac{number\ of\ detected\ ships}{number\ of\ ships} \\ F1 - Score\ = \frac{2 \times PR \times RR}{PR + RR} \end{cases} \quad (1)$$

A ship is detected when the intersection-over-union (IoU) between candidate bounding box and the actual object rectangle is bigger than 0.45. Fig. 5 shows the ship detection results in the test set, including precision rate, recall rate, F1-Score and IoU. The IoU is high, which is above 0.73. Given the confidence threshold = 0.45, we have PR = 82.8%, RR = 84.9%, F1-Score = 0.838 and IOU = 0.736. Fig. 6 shows some detection results by our transfer learned SSD in open water and harbor. Table I lists the details of the detection results given the recall rate.

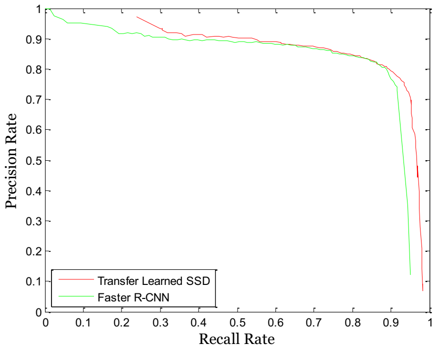

Figure 7. The Recall-Precision-Curves (RPC) of Transfer learned SSD and Faster R-CNN in our ship test set.

Table2. Detection Results of Transfer Learned Ssd and Faster R-Cnn

| Method | AP | FPS |
|---|---|---|
| Transfer learned SSD | 87.9% | 47 |
| Faster R-CNN | 81.2% | 6 |

Fig. 6 shows the Recall-Precision-Curve of Faster R-CNN and Transfer learned SSD in our ship test set. Table II shows the comparison between transfer learned SSD and Faster R-CNN. Both methods use the VGG-16 as base network. Our computing platform consists of a CPU of Intel i7-4970K and a GPU of NVIDIA TITAN X. In our experiment, our transfer learned SSD method outperformed the Faster R-CNN in both accuracy and speed. Our transfer learned SSD achieved 87.9% average performance (AP) at 47 FPS. In comparison with Faster R-CNN, 6.7% AP improvement could be achieved.

For the ship detection task in the satellite images, the ship size is varied, and a large number of ships are only a few pixels in size. Faster R-CNN only uses the feature map from the top layer. Because of the pooling and up-sampling, the information of small ship may vanish and be lost, leading to some missing detections. On the contrary, our framework utilizes the feature map from multi-layers and deals well with the multi-scale problem. The faster R-CNN framework is composed of region proposal generation, feature extraction, classification and location refine. Our framework eliminates the proposal generation, pixel resampling or feature resampling stage, and utilizes small convolutional filter to predict category and offsets of the default bounding boxes, which reduce the computation.

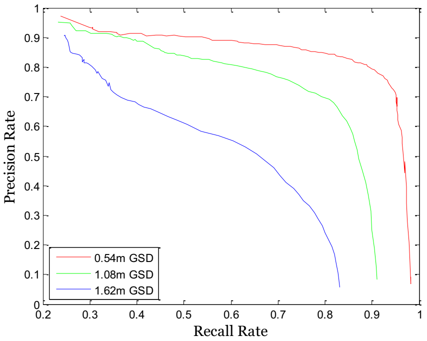

Figure 8. The Recall-Precision-Curves (RPC) of Transfer learned SSD with images in different spatial resolution.

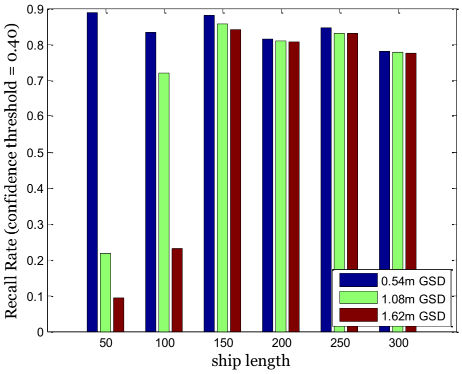

Figure 9. Given the confidence threshold = 0.40, the recall rate for each ship group in different resolution.

Table3. The ap in different spatial resolution

| Detection result | 0.54m GSD | 1.08m GSD | 1.62m GSD |
|---|---|---|---|
| AP | 87.9% | 72.6% | 57.9% |

## 3.2 Effect of Resolution on Ship Detection Accuracy

The input size of the transfer learned SSD in our framework is 300 x 300. All the input images will resize to the fixed size. In order to examine the effect of resolution on ship detection, we form the database by cutting the original satellite image with the size with 300 x 300, 600 x 600 and 900 x 900 individually. The original collected images are 0.54m GSD. Thus, after the resize operation, the 600 x 600 and 900 x 900 cutouts are 1.08m and 1.62m GSD individually. Fig. 7 shows the Recall-Precision-Curve (RPC) of our ship detection framework with images in different resolution. Table III shows the AP of ship detection in different spatial resolution. The detection accuracy declined sharply with the decreasing resolution. In order to further explore this phenomenon, we divided the ships into 6 groups according to the ship length, and then count recall rate for each group. Fig. 8 shows the comparison of recall rate for each group in different spatial resolution. From the chart, we can see that the recall rate declines sharply when the ship length is less than 100 and declines gradually when the ship length is over 100. Though our method predicts the category and offsets of bounding box with the feature map from different layers, the lowest layer for prediction is Conv_4_3 and the feature map size is 38 x 38 x 512. The reception field in this layer is 92 x 92. The ship size declines with decreasing resolution. The visual feature of ships with few pixels may vanish and lose when it passes through many layers, resulting in low recall rate.

## 4. Conclusions

This paper introduces a transfer learned SSD that transfers visual knowledge between natural image and satellite image. We use the feature map from different layers to predict category and offsets of default box, which deals well with the multi-scale problem. Compared with some other DNN methods, our framework leverages a fully convolutional network and eliminates the proposal generation, pixel resampling or feature resampling stage, which reduces the computation. Experiments demonstrated that our method achieved state-of-the-art ship detection performance and could satisfy the real-time requirements. We investigated the effects of resolution on ship detection performance. There is a sharp degradation for small ships and almost no change for big ships.

## Acknowledgment

We acknowledge support by the National Natural Science Foundation of China under Grants of U1435219 and 61402507.

## References

[1] C. Zhu, H. Zhou, R. Wang, and J. Guo, “A novel hierarchical method of ship detection from spaceborne optical image based on shape and texture features,” Geoscience and Remote Sensing, IEEE Transactions on, vol. 48, no. 9, pp. 3446–3456, 2010.

[2] N. Dalal and B. Triggs, “Histograms of oriented gradients for human detection,” in Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on, vol. 1. IEEE, 2005, pp. 886–893.

[3] X. X. Bao, S. S. Zinger, D. P. P. With, R. R. Wijnhoven, and J. J. Han, “Water region and multiple ship detection for port surveillance,” in Proc. the 33rd WIC Symposium on Information Theory in the Benelux, 2012.

[4] Gan, Lu, P. Liu, and L. Wang. "Rotation Sliding Window of the Hog Feature in Remote Sensing Images for Ship Detection," International Symposium on Computational Intelligence and Design IEEE, 2015:401-404.

[5] Krizhevsky, Alex, I. Sutskever, and G. E. Hinton. "ImageNet classification with deep convolutional neural networks," International Conference on Neural Information Processing Systems Curran Associates Inc. 2012:1097-1105.

[6] Zhu, H., et al. "Orientation robust object detection in aerial images using deep convolutional neural network," IEEE International Conference on Image Processing IEEE, 2015:3735-3739.

[7] J. Tang, C. Deng, and G. B. Huang. "Compressed-Domain Ship Detection on Spaceborne Optical Image Using Deep Neural Network and Extreme Learning Machine," IEEE Transactions on Geoscience & Remote Sensing 53.3(2015):1174-1185.

[8] G. B. Huang, Q. Y. Zhu, and C. K. Siew, “Extreme learning machine: Theory and applications,” Neurocomputing, vol. 70, no. 1–3, pp. 489–501, Dec. 2006.

[9] L. Zhang, L. Zhang, and B. Du. "Deep Learning for Remote Sensing Data: A Technical Tutorial on the State of the Art," IEEE Geoscience & Remote Sensing Magazine 4.2(2016):22-40.

[10] R. Girshick, J. Donahue, T. Darrell, et al. “Region-Based Convolutional Networks for Accurate Object Detection and Segmentation,” IEEE Transactions on Pattern Analysis & Machine Intelligence 38.1(2016):142-158.

[11] J. Uijlings, K. van de Sande, T. Gevers, and A. Smeulders. “Selective Search for Object Recognition[J],” International Journal of Computer Vision, 2013, 104(2):154-171.

[12] C. L. Zitnick and P. Dollár. "Edge Boxes: Locating Object Proposals from Edges," 8693(2014):391-405.

[13] M. M. Cheng, Z. Zhang, W. Y. Lin, et al. Cheng, Ming Ming, et al. "BING: Binarized Normed Gradients for Objectness Estimation at 300fps," (2014):3286-3293.

[14] K. He, X. Zhang, S. Ren, et al. “Spatial Pyramid Pooling in Deep Convolutional Networks for Visual Recognition,” He, Kaiming, et al. "Spatial Pyramid Pooling in Deep Convolutional Networks for Visual Recognition." IEEE Transactions on Pattern Analysis & Machine Intelligence 37.9(2015):1904-1916.

[15] R. Girshick. “Fast R-CNN,” Proceedings of the IEEE International Conference on Computer Vision. 2015: 1440-1448.

[16] S. Ren, K. He, R . Girshick, et al. “Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks[J],” Computer Science, 2015:1-1.

[17] J. Redmon, S. Divvala, R. Girshick, et al. “You Only Look Once: Unified, Real-Time Object Detection[J],” Computer Science, 2015.

[18] W. Liu, D. Anguelov, D. Erhan, et al. “SSD: Single Shot MultiBox Detector[J]. 2016. "SSD: Single Shot MultiBox Detector[J]," (2016).

[19] M. Oquab, L. Bottou, I. Laptev, and J. Sivic, “Learning and transferring mid-level image representations using convolutional neural networks,” in Proc. IEEE Comput. Vis. Pattern Recog., 2014, pp. 1717–1724.

[20] S. Pan and Q. Yang. “A survey on transfer learning. Knowledge and Data Engineering,” IEEE Transactions on, 22(10):1345–1359, 2010. 2

[21] K. Simonyan, A. Zisserman. “Very deep convolutional networks for large-scale image re

[22] cognition,” Computer Science, 2015.