Harbor Ship Detection Based on Channel Weighting and Spatial Information Fusion

Quancheng Zhou\textsuperscript{1,2}, Fei Song\textsuperscript{1,2,3}, Zhenghao Chen\textsuperscript{1}, Rui Zhang\textsuperscript{1,2}, Ping Jiang\textsuperscript{*} and Tao Lei\textsuperscript{1}

\textsuperscript{1} Institute of Optics and Electronics, Chinese Academy of Sciences, Chengdu, Sichuan, 610209, China
\textsuperscript{2} University of Chinese Academy of Sciences, Beijing 100049, China
\textsuperscript{3} University of Electronic Science and Technology of China, Chengdu, Sichuan, 610051, China

Corresponding author’s e-mail: jiangping@ioe.ac.cn

Abstract. Aiming at the problems of false alarms and missed detection caused by large differences among ship types and complex background in harbor remote sensing images, a robust single-stage ship detection method is proposed. First, a Channel Weighting Mechanism is devised, which self-learns the weights of different channel to enhance valid features in channel dimension. Second, a Spatial Information Fusion Module is designed to enhance features in spatial dimension, which extracts more information of ship appearance from shallow feature maps, then excavates potential contextual information in deep semantic features. Finally, to promote the detection ability of ships of various scales, a Multi-stage Weighted Fusion Pyramid is applied to optimize the fusion of high-stage features and low-stage features. Extensive experiments conducted on self-established dataset for harbor ships show that the proposed method provides the performance by 3.08\%mAP compared to RetinaNet.

1. Introduction

In recent years, remote sensing ship detection benefits human society in multiple area such as harbor management, marine law enforcement, military deployment. However, due to high similarity between complex background objects and certain ships and large disparities among various ship types, it’s still challenging to develop a both fast and accurate ship detector.

Convolutional neural networks such as AlexNet\textsuperscript{[1]}, VGG\textsuperscript{[2]} and ResNet\textsuperscript{[3]} have made breakthroughs in image recognition performance, which also promoted the development of object detection. The two-stage detector Faster RCNN\textsuperscript{[4]}, one-stage detectors SSD\textsuperscript{[5]} and RetinaNet\textsuperscript{[6]} have been the milestones and shown great practical value. To improve the performance of SSD, Liu et al. \textsuperscript{[7]} designed a receptive field enhancement module which mimics human vision. RetinaNet surpasses SSD in performance and speed by using Focal Loss and feature pyramid. In pursuit of the speed-accuracy balance, Tan et al. \textsuperscript{[8]} proposed EfficientDet, which follows the same structure as RetinaNet but modified its backbone and neck to new designs. Yang et al. \textsuperscript{[9]} proposed Feature Refinement Module to promote the performance in remote sensing object detection based on RetinaNet.

In optical remote sensing ship detection, many researchers have carried out exploration based on deep learning methods. Huang et al. \textsuperscript{[10]} used convolutional neural networks first to extract features and then used Support Vector Machine for classification. However, its performance is outdated for large
computational cost. Besides, ordinary rectangular label cannot locate densely arranged ships accurately. In order to balance the speed and accuracy of ship detection, Ma et al. [11] designed a feature fusion module to improve SSD. However, it aims at the broad ocean without harbor, where ships are distributed sparsely and easy to locate using ordinary rectangles. Aiming at complex port scenes, Gu et al. [12] proposed a method based on Fisher discrimination and Mask R-CNN[13]. Xiao et al. [14] proposed a paired semantic segmentation network to generate high quality pre-set anchors to enhance robustness. However, using image segmentation to assist detection leads to vast computation and memory cost.

Although many studies have been carried out in remote sensing ship detection, the following limitations still exist: (i) Single stage detectors are more applicable, but there are relatively few applied studies based on RetinaNet in this field. (ii) Although RetinaNet shows fine speed-accuracy balance, it still causes numerous false alarms and missed detections due to high similarity between certain background objects and ships, and large appearance discrepancies among various types of ships. Valid features extracted from the backbone network are insignificant to discriminate background and targets. (iii) Ships are slender, randomly oriented and densely arranged. Standard convolution with square receptive field is unable to sample inclined ship targets completely, which brings in interference when generating deep semantic features. Besides, ships are distributed alongside the coast or grouped, which contain potential contextual information but not fully exploited.

To address the aforementioned problems, a robust single-stage ship detector is proposed based on RetinaNet. First, a Channel Weighting Mechanism is devised, which self-learns the weight of semantic description of each channel and enhances the significance of valid features. Second, a Spatial Information Fusion Module is designed. Deformable convolution[15] is applied to focus on the outline of the ship by extracting more appearance information on shallow features to enhance the significance of deep semantic features, then dilated convolution[16] is applied to exploit the contextual information related to surrounding objects from deep semantic features. Finally, in order to detect ships of various scales, inspired by BiFPN[8], Multi-stage Weighted Fusion Pyramid is introduced to optimize the fusion of multi-stage features from the backbone. On account of lacking high-quality data, a dataset is also established. Experiments show the proposed method increases by 3.08%mAP compared to RetinaNet.

2. Method

2.1. Overall framework

Figure 1 shows the structure of RetinaNet and our proposed method. RetinaNet contains ResNet as the backbone, Feature Pyramid Network as the neck and several classification/regression heads. In our method, features from the backbone are processed by $1\times 1$ convolution and Max Pooling to generate features tensors (P3–P7). First, these features are input into Channel Weighting Mechanism (CWM) to adjust the significance of each channel. Then, only features of P3–P5 stages are processed by Spatial Information Fusion Module (SIFM). Finally, all stages of features are input into Multi-stage Weighted Fusion Pyramid (MWFP).

![Figure 1. (a) RetinaNet (b) Our proposed method](image-url)
2.2. Channel Weighting Mechanism

Each channel of a feature tensor describes specific semantic information. Channel Weighting Mechanism strengthens the significance of valid features by self-learning the weight of each channel during training. As shown in Figure 2, W, H, C represents width, height, channel numbers respectively. After Max Pooling and 1D convolution (kernel size=3), we get a vector with length of C. Then we use sigmoid function to normalize the value in the vector. Last, each element in the vector multiplies the original feature tensor so that we get the optimized feature. It can be expressed by the following formula:

$$B = \sigma(1Dconv_{k=3}[GMP(A)]) \cdot A$$

(1)

Where $A$ is the input feature tensor, $B$ is the result, $GMP$ is Global Max Pooling, $1Dconv_{k=3}$ is 1D convolution (kernel size=3) and $\sigma$ is the Sigmoid function.

Global Max Pooling is adopted for the following analysis. Global Average Pooling can be expressed as:

$$GAP(A') = \frac{1}{W \cdot H} \sum_{(i,j)=(1,1)}^{(W,H)} A'(i,j)$$

(2)

Where $A'$ is a channel of input feature with width W and height H. We can conclude that important semantic information could be inhibited for Average Pooling operation on whole feature map. The proportion of ship area is very small relative to the whole image, so the valid feature area is small while the background feature is relatively significant. Hence, Using Max Pooling can keep the max value so that key information is reserved. Moreover, to decrease computational cost, our method adopts one-dimension convolution which is different from the fully connection layer in SENet[17].

2.3. Spatial Information Fusion Module

To enhance the features in spatial dimension, we consider two steps: First, extract features from ship target itself on shallow layers. Second, excavate potential context information from deep layers. In this way, we design a new structure shown in Figure 3. The first path in this module includes a deformable convolution and a standard convolution. The second path includes a deformable convolution. These two paths cover different scope of feature map. Dilated convolution is adopted on the tail to extract context information of various scope.

Ships are slender, inclined and densely arranged. Standard convolution will bring in information of surrounding background or ships when sampling shallow features, which causes confusion in deeper semantic features, as shown in Figure 4(a) first row. Therefore, we consider inclined ships as geometric deformation and introduce deformable convolution. It can learn spatial geometric deformation, and the position of the sampling point will be adaptively adjusted according to the image content so that it can adapt to the distortion of ship appearance. Its receptive field focuses on the outline of ships so that it can generate more significant deep semantic features, as shown in Figure 4(a) second row.

A brief introduction of deformable convolution are as follows.

The position set P of 3×3 square is:

$$P = \{p_i, \ i \in [1, 9]\}$$

(3)

Standard convolution is:

$$y = \sum_{p_i \in P} K(p_i) \cdot A(p_i)$$

(4)
Where \( K(p_i) \) is the value of the convolution kernel at the position of \( p_i \), and \( A(p_i) \) is the value of the feature map at the position of \( p_i \). In deformable convolution, position offset \( \Delta p_i \) is introduced:

\[
y = \sum_{p_i \in P} K(p_i) \cdot A(p_i + \Delta p_i)
\]

(5)

As shown in Figure 4(b), the offset \( \Delta p_i \) is obtained by training a convolution layer. The number of channels is \( 2K=18 \), which respectively correspond to the values in the x and y directions. We believe it performs better when applied to low-stage features. This is verified by experiments.

![Figure 4](image)

**Figure 4.** (a) Sampling point of standard and deformable convolution (b) Structure of deformable convolution

### 2.4. Multi-stage Weighted Fusion Pyramid

Feature Pyramid Networks in RetinaNet sums high-level stages after upsampling and low-level stages, which can easily cause ambiguity between high-level and low-level features in the same spatial location of the feature map. Inspired by BiFPN[8], we introduce Multi-stage Weighted Fusion Pyramid to enhance the detection ability of different scales, which self-learns the weight of each stage. Then it adds all stages by weight. In this paper, our Multi-stage Weighted Fusion Pyramid includes 3 steps: first, the number of channels in the pyramid is adjusted to 256 by a 1x1 convolution, then all features are processed by Channel Weighting Mechanism and Spatial Information Fusion Module. Finally, all stages are weighted and sum up shown in Figure 5(b).

![Figure 5](image)

**Figure 5.** (a) Feature Pyramid Network (b) Multi-stage Weighted Fusion Pyramid

### 3. Dataset and Details

#### 3.1. Dataset

To validate the effectiveness of our proposed method, a dataset named Optical Remote Sensing Ship Detection is established. All images with a resolution of 1m are collected from Google Earth, which contains 401 patches and about 3800 samples. The size ranges from 800 to 2500 pixels. The dataset is collected from multiple harbors around the globe. It contains various types of ship with complex
appearance to promote difficulties, such as aircraft carrier and submarines. 201 images are for training and 200 for testing. Considering the uniqueness of ship appearance, we adopt rotated quadrilateral label to locate ships for ground truth.

3.2. Evaluation Criteria and Training details
We use mean Average Precision(mAP)[18] for evaluation. Intersection over Union = 0.5 is adopted to determine to be positive or negative. In our training phase, images are cropped to $512 \times 512$ with an overlap of 200 pixels. We also adopt backbone network pre-trained on ImageNet[19] with initial learning rate of 0.01. Loss function is Focal Loss and Smooth L1 Loss. The network is trained for 25 epochs. Software environment is: Ubuntu 18.04, Pytorch 1.4, Python 3.7. Hardware environment is: Intel I7-7700HQ, Nvidia GTX 1060(6GB), 16GB memory.

4. Experiment Analysis

4.1. Multistage Weighted Fusion Pyramid.
To validate the effectiveness of Multistage Weighted Fusion Pyramid, we compare our method with RetinaNet, which employs Feature Pyramid Network. Table 1 shows Multistage Weighted Fusion Pyramid can improve performance. Note that r50 and r101 represent ResNet-50 and ResNet-101.

| Method Name                | mAP (%) |
|----------------------------|---------|
| RetinaNet(r50)             | 63.57   |
| ResNet50 + MWFP            | 64.08   |
| RetinaNet(r101)            | 65.92   |
| ResNet101+MWFP             | 66.00   |

4.2. Channel Weighting Mechanism
To validate the effectiveness of Channel Weighting Mechanism, we conduct the following experiment based on Multistage Weighted Fusion Pyramid. Table 2 shows Channel Weighting Mechanism elevates performance by 2.32%, but number of parameters only increases from 30.9500M to 30.9510M, which is negligible. Compared with SENet[17], our method shows better performance as well.

| Method Name                              | mAP (%) | #Params(M) |
|------------------------------------------|---------|-----------|
| ResNet50 + MWFP                          | 64.08   | 30.95     |
| ResNet50 + SENet + MWFP                  | 66.29   | 30.97     |
| ResNet50 + CWM + MWFP                    | 66.40   | 30.95     |

To validate the effectiveness of different Pooling operation, Max Pooling is compared with Average Pooling. Table 3 shows Max Pooling will reserve valid feature information of small area while Average Pooling will inhibit valid information of small target area.

| Method Name    | mAP (%) |
|----------------|---------|
| Average Pooling| 65.52   |
| Max Pooling    | 66.40   |
4.3. Spatial Information Fusion Module
In this section, the effectiveness of Spatial Information Fusion Module is validated. The second row of Table 4 is the result of replacing the deformable convolution with standard convolution in Figure 3.

| Method Name                                      | mAP (%) |
|-------------------------------------------------|----------|
| ResNet50 + MWFP                                 | 64.08    |
| ResNet50 + SIFM (standard conv) + MWFP          | 64.40    |
| ResNet50 + SIFM (deformable conv) + MWFP        | 65.33    |

To verify that Spatial Information Fusion Module is more suitable for low-level features, it is placed on different stages of the pyramid. Table 5 shows it performs best on lower stages (P3–P5), while it decreases the performance when applied to all stages (P3–P7).

| P3 | P4 | P5 | P6 | P7 | mAP (%) |
|----|----|----|----|----|---------|
| √  | √  | √  | −  | −  | 65.33   |
| −  | −  | √  | √  | √  | 64.44   |
| √  | √  | √  | √  | √  | 63.15   |

4.4. Overall Performance Comparison with other methods
Table 6 shows our proposed method outperforms baseline method RetinaNet(r50) by 3.08% and the classic two-stage detector Faster RCNN with only losing negligible speed. Besides, our method is superior to R3Det[9], which is also a remote sensing detector based on RetinaNet.

| Method Name     | mAP (%) | Speed (FPS) |
|-----------------|---------|-------------|
| RetinaNet(r50)  | 63.57   | 16.8        |
| RetinaNet(r101) | 65.92   | 13.0        |
| Faster RCNN(r50)| 65.54   | 10.6        |
| R3Det           | 66.30   | 9.6         |
| Ours(r50)       | 66.65   | 15.5        |

4.5. Visualized Results
Figure 6 shows part of the detection results of baseline method and ours. It shows our method eliminates the false alarms and missed detection in certain cases and therefore promotes the performance.

Figure 6. The first row is the result of RetinaNet(r50). The bottom row is the result of our method.

5. Conclusion
To reduce the missed detection and false alarms in complex optical remote sensing harbor scenes, a fast and accurate ship detection convolution neural network based on RetinaNet is proposed. This method includes Channel Weighting Mechanism and Spatial Information Fusion Module, which enhance discrimination of features in channel and spatial dimension respectively. Experiments conducted on the
self-established dataset demonstrate the effectiveness of our proposed method, which increases by 3.08\% mAP compared to RetinaNet. In the future, fine-grained recognition of ship types will be studied.

Acknowledgments
This work was supported by Youth Innovation Promotion Association, Chinese Academy of Sciences (Grant No. 2016336).

References
[1] Krizhevsky, A., Sutskever, I., & Hinton, G.. (2012). Imagenet classification with deep convolutional neural networks. In: Advances in neural information processing systems, 25(2).
[2] Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. https://arxiv.org/pdf/1409.1556.pdf.
[3] He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770-778.
[4] Ren, S., He, K., Girshick, R., & Sun, J. (2015). Faster r-cnn: Towards real-time object detection with region proposal networks. In Advances in neural information processing systems. pp. 91-99.
[5] Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C. Y., & Berg, A. C. (2016). Ssd: Single shot multibox detector. In: European conference on computer vision. pp. 21-37. Springer, Cham.
[6] Lin, T. Y., Goyal, P., Girshick, R., He, K., & Dollár, P. (2017). Focal loss for dense object detection. In: Proceedings of the IEEE international conference on computer vision. pp. 2980-2988.
[7] Liu, S., & Huang, D. (2018). Receptive field block net for accurate and fast object detection. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 385-400.
[8] Tan, M., Pang, R., & Le, Q. V. (2020). Efficientdet: Scalable and efficient object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10781-10790.
[9] Yang, X., Liu, Q., Yan, J., Li, A., Zhang, Z., & Yu, G. (2019). R3det: Refined single-stage detector with feature refinement for rotating object. https://arxiv.org/abs/1908.05612.
[10] Huang, J., Jiang, Z., Zhang, H., & Yao, Y. (2017). Ship object detection in remote sensing images using convolutional neural networks. J. of Beijing University of Aeronautics and Astronautics, 2017, 43 (9): 1841-1848. (in Chinese)
[11] Ma, J., Shi, W., Bao S. (2019) Ship target detection in remote sensing images based on feature fusion SSD. J. of Computer Applications, 2019, (S2 vo 39): 253-256. (in Chinese)
[12] Gu, Z., Jiang, W. Improved remote sensing image ship detection based on Mask R-CNN. Computer Engineering and Applications, 2020, 56(8): 171-176. (in Chinese)
[13] He, K., Gkioxari, G., Dollár, P., & Girshick, R. (2017). Mask r-cnn. In: Proceedings of the IEEE international conference on computer vision. pp. 2961-2969.
[14] Xiao, X., Zhou, Z., Wang, B., Li, L., & Miao, L. (2019). Ship detection under complex backgrounds based on accurate rotated anchor boxes from paired semantic segmentation. Remote Sensing, 11(21), 2506.
[15] Dai, J., Qi, H., Xiong, Y., Li, Y., Zhang, G., Hu, H., & Wei, Y. (2017). Deformable convolutional networks. In: Proceedings of the IEEE international conference on computer vision. pp. 764-773.
[16] Yu, F., & Koltun, V. (2015). Multi-scale context aggregation by dilated convolutions. https://arxiv.org/abs/1511.07122.
[17] Hu, J., Shen, L., & Sun, G. (2018). Squeeze-and-excitation networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 7132-7141.
[18] Everingham, M., Van Gool, L., Williams, C. K., Winn, J., & Zisserman, A. (2010). The pascal visual object classes (voc) challenge. International journal of computer vision, 88(2), 303-338.
[19] Deng, J., Dong, W., Socher, R., Li, L. J., Li, K., & Fei-Fei, L. (2009). Imagenet: A large-scale hierarchical image database. In: 2009 IEEE conference on computer vision and pattern recognition. pp. 248-255.