remote sensing

MDPI

Article

Arbitrary-Oriented Inshore Ship Detection based on Multi-Scale Feature Fusion and Contextual Pooling on Rotation Region Proposals

Tian Tian [1,*], Zhihong Pan [2], Xiangyu Tan [1] and Zhengquan Chu [1]

[1] Key Laboratory of Geological Survey and Evaluation of Ministry of Education, School of Computer Science, China University of Geosciences, Wuhan 430074, China; xiangyu_tan@cug.edu.cn (X.T.); chuzhengquan@cug.edu.cn (Z.C.)

[2] School of Artificial Intelligence and Automation, Huazhong University of Science and Technology, Wuhan 430074, China; zpanhust@gmail.com

* Correspondence: tiantian@cug.edu.cn

Received: 30 November 2019; Accepted: 17 January 2020; Published: 20 January 2020

Abstract: Inshore ship detection plays an important role in many civilian and military applications. The complex land environment and the diversity of target sizes and distributions make it still challenging for us to obtain accurate detection results. In order to achieve precise localization and suppress false alarms, in this paper, we propose a framework which integrates a multi-scale feature fusion network, rotation region proposal network and contextual pooling together. Specifically, in order to describe ships of various sizes, different convolutional layers are fused to obtain multi-scale features based on the baseline feature extraction network. Then, for the purpose of accurate target localization and arbitrary-oriented ship detection, a rotation region proposal network and skew non-maximum suppression are employed. Finally, on account of the disadvantages that the employment of a rotation bounding box usually causes more false alarms, we implement inclined context feature pooling on rotation region proposals. A dataset including port images collected from Google Earth and a public ship dataset HRSC2016 are employed in our experiments to test the proposed method. Experimental results of model analysis validate the contribution of each module mentioned above, and contrast results show that our proposed pipeline is able to achieve state-of-the-art performance of arbitrary-oriented inshore ship detection.

Keywords: inshore ship detection; multi-scale feature fusion; rotation region; region proposal network; context feature pooling

1. Introduction

With the development of remote sensing technology, remote sensing data and techniques have been widely applied to marine monitoring and surveys with the purpose of national defense and resource exploitation [1–5]. As major transportation modes and typical targets in seas, ships have been paid more and more attention in applications based on remote sensing images. The detection of ships plays an important role in both civilian and military applications, such as port management, maritime rescue, battlefield surveillance and strategic deployment [6,7].

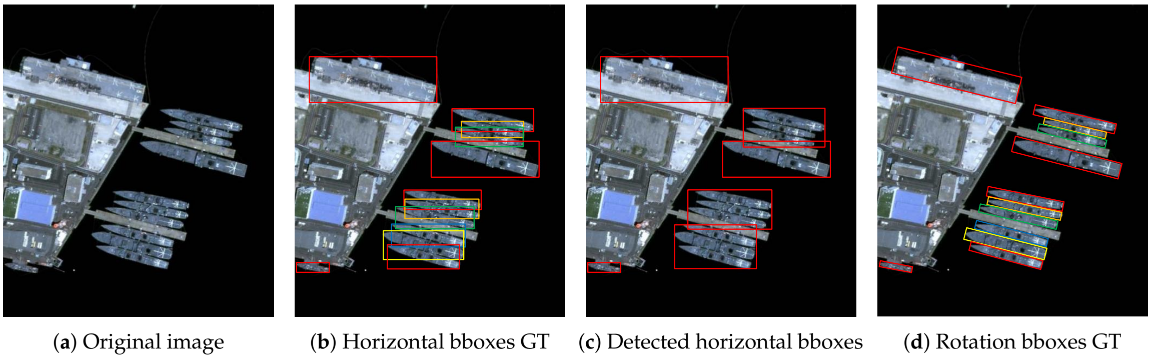
Many ship detection methods have been proposed based on images produced by different sensors. Synthetic aperture radar (SAR) images were used earlier and technologies are relatively mature. Compared to SAR images, high-resolution images contain more detailed structure and texture information, which have attracted more and more research interests in recent years [8,9]. According to the location and background of ships, ship detection methods using high-resolution images can be divided into offshore ship detection and inshore ship detection. Offshore ships are located in the sea, whose background is almost the sea surface only. Since offshore ship targets usually differ significantly from backgrounds on grey levels and textures, this kind of study has already made great progress [10,11]. On the contrary, inshore ships lie in ports which have complex land environments and a great number of disturbance objects similar to ships. Moreover, inshore ships in very high-resolution images exhibit various sizes and shapes, and some of them are densely arranged near the coast. As a result, the detection of inshore ships is still a challenging task [12].

In the traditional methods of inshore ship detection, two major steps are usually employed [13]. First, sea–land segmentation based on texture and shape features is implemented to eliminate land interference and improve target searching efficiency, and then, ship targets are detected with feature extraction and machine learning approaches. However, most of these methods are carried out based on artificially designed features [14], which show limitations on performance in practical applications [15].

With the advancement of deep learning methods, Convolutional Neural Networks (CNNs) have achieved great success in image processing, especially in object detection tasks, where a great number of CNN-based methods have been proposed. Region proposals with CNN (R-CNN) [16] is the first benchmark framework that successfully applies CNN to object detection. Based on the idea of Region Proposal Network (RPN), Spatial Pyramid Pooling Network (SSP-Net) [17], Fast R-CNN [18] and Faster R-CNN [19] are presented in succession to improve the performance and efficiency of R-CNN. All of these methods first generate region proposals by RPN, then use Non-Maximum Suppression (NMS) to select candidates with high confidence, and finally implement classification and location regression on these region proposals to obtain accurate detection results. With the preference to speed, some other one-step structures (without RPN step) are proposed as well, representatives of which include YOLO (You Only Look Once) [20] and SSD (Single Shot Multibox Detector) [21]. They take the object detection task as a problem of regression, and directly employ classification and regression on the feature map without consideration of region proposals.

Since the networks mentioned above have been proven to be effective in object detection of natural images, researchers in remote sensing have made efforts to utilize them in the application of inshore ship detection. Zhang et al., realize ship detection based on a Faster R-CNN framework [22]. Wu et al., employ ship head searching, RPN, mutli-task network and NMS for inshore ship detection [23]. The advantage of a deep neural network is its strong capability of feature description, which is exactly essential for the challenging task of inshore ship detection. Therefore, many model-transplanting methods easily outperform the traditional ones by means of deep learning [24].

The above-mentioned methods all utilize horizontal bounding boxes to mark the location of ship targets. Although a horizontal bounding box is applicable for most object detections in natural images, it cannot locate inshore ships accurately because of their unique shapes. Since ships usually have a large aspect ratio and obvious inclined angle, a horizontal bounding box is not able to estimate the orientation of a ship. Moreover, after region proposals are generated, non-maximum suppression is usually applied as a vital step to reduce the number of candidates and increase detection efficiency. If inclined inshore ships lie densely near the port, the overlap of their horizontal bounding boxes will be very large, which will result in targets being missed when NMS is applied to screen region proposals. Figure 1 shows the differences between horizontal bounding boxes and rotated bounding boxes when detecting densely arranged inshore ships based on region proposals. It is seen from the figure that the horizontal bounding box ground truth (Figure 1b) covers more redundancy regions than the rotated one (Figure 1d), and large overlaps of region proposals may cause side effects of NMS; target regions are not separated appropriately or some are even missed out (Figure 1c).



(**a**) Original image (**b**) Horizontal bboxes GT (**c**) Detected horizontal bboxes (**d**) Rotation bboxes GT

Figure 1. Differences between horizontal bounding boxes (bboxes) and rotation bounding boxes of densely arranged inshore ships.

Rotation bounding boxes are first presented and studied in the research of text detection, which has a similar requirement of detecting text of arbitrary orientations. Rotational Region CNN (R^2CNN) [25] introduces inclined angle information in classification and regression networks and employs skew NMS to select region proposals. Rotation RPN (RRPN) [26] creates rotation anchors to improve the quality of proposals at the stage of RPN, and improves the performance of R^2CNN.

With the achievement these methods have made, rotated region proposals have been applied to ship detection as well. Liu et al., introduce pooling of a Rotation Region of Interest (RRoI) and rotation bounding box regression in the framework [27]. Yang et al., integrate a dense feature pyramid network, RRPN and Fast R-CNN to inclined results [6]. Zhou et al., use a semantic segmentation network to recognize part of ships for rotated region proposals [28,29]. Although methods adopting a rotation bounding box have already been proposed during the past two years, the problem of accurate rotated region proposal and the consequent information loss of employing rotated boxes are not yet well solved. To generate appropriate rotated region proposals for inshore ships of various locations and sizes with a concise end-to-end model is still challenging. Moreover, although a horizontal box is inferior to a rotated one, redundancy regions that a horizontal box contains can provide beneficial contextual information in some sense. The RRoI pooling of a rotated box aims to obtain an accurate description of objects for detection, however, it excludes very useful contextual information, which will result in more false alarms than traditional RPN methods. In consideration of all the above, in this paper, we propose a concise end-to-end framework for arbitrary-oriented inshore ship detection. Our method uses the baseline backbone feature extraction network in accordance with popular deep learning work of object detection. Nevertheless, state-of-the-art feature pyramid processing based on backbone network is employed to implement multi-scale feature fusion. Then, a rotation region proposal network is adopted to generate rotated region proposals for ships of various sizes and locations. Finally, a rotated contextual RRoI pooling is applied to correct the information loss of the rotated box for an accurate description of target detection. The main contributions of this paper are as follows:

1. We propose a novel and concise framework for arbitrary-oriented inshore ship detection which can handle complex port scenes and targets of different sizes.
2. For better detection of various target sizes, we design a multi-scale feature extraction and fusion network to build the multi-scale features, which contribute to the promotion of precision on different ships.
3. In order to obtain accurate target locating and an arbitrary-oriented bounding box, we adopt the rotation region proposal network and skew non-maximum suppression. Consequently, densely moored ships are able to be distinguished, which results in an increase of recall rates.
4. We implement rotated contextual feature pooling of a rotation region of interest to better describe ship targets and their surrounding backgrounds. As a result, the description weakness of rotation bounding boxes is improved with a decrease of false alarms.

Experimental results on remote sensing port images collected from Google Earth validate the effectiveness of the proposed method. It is robust for ship targets of various sizes and types, outperforming other methods on indicators of precision, recall and false alarm criteria. Results on the public ship dataset HRCS2016 provide a parallel comparison of the proposed pipeline, which show its state-of-the-art performance compared to other advanced approaches. The rest of this paper is organized as follows: Section 2 introduces more detailed state-of-the-art studies concerning this application. Section 3 describes the overall framework and details of each part of our method. Section 4 presents experiments of model analysis and contrast methods to validate the effectiveness of the proposed method. Finally, Section 5 concludes the paper.

2. Related Work

From the perspective of different scenes that targets lie in, ship detection methods can be roughly divided into offshore and inshore ship detection; by the types of methodology of detection framework, they can be divided into traditional methods and deep learning ones. An offshore ship scene usually contains small at-sea targets and the detection task mainly aims to resist disturbance of sea surface background. Inshore ship targets are usually relatively larger, where complex similar land confusions of ports are the major challenges. Traditional ship detection methods usually design shape and texture features for classifying the target and background, which may be more applicable to offshore ship detection since offshore scenes are relatively simple compared to coasts and ports. Representative features adopted for offshore ship detection include shape and texture features [10], S-HOG features [30], edge and contour features [11], and topology structure and LBP features [31,32]. Inshore ship detection with traditional methods usually focuses on special characteristics of ships to design features and recognize targets. He et al., adopt weighted voting and rotation-scale-invariant pose to detect ships [12], and Bi et al. [8] employ an omnidirectional intersected two-dimension scanning strategy and decision mixture model to detect ships. Since artificially designed features cannot cope with complex port background, as a result, more and more deep learning methods are proposed in the inshore ship detection task during the most recent three years while non-deep-learning ones are seldom presented.

Object detection frameworks using deep neural networks have been applied to remote sensing applications [33,34] since they were proved to outperform traditional methods in natural images. In the field of inshore ship detection, many researchers have delved into the development of target detection with deep learning methods. A fully convolutional network is applied to detect ships in the literature [35], but facing the problem that localization is not accurate enough. Methods based on R-CNN and its improved variants (especially Fast R-CNN and Faster R-CNN) are highly favored for their better detection effects [22]. Moreover, various improvements on the detection framework are employed, including the addition of ship head detection [23] and contextual information [1].

In consideration of the characteristics of inshore ships that targets are always inclined or even closely arranged, rotation bounding boxes are proposed to further improve the localization accuracy of ships. Liu et al. [27] bring in a rotated region CNN and rotation ROI pooling to detect inclined ships, but the skew proposals are generated by the method like a selective search presented by [11] (not end-to-end model). Li et al. [36] employ a five-box method to produce rotation proposals, which is not a region proposal network. In order to build an end-to-end model, Zhang et al., introduce a rotation region proposal network instead of the above rotation proposal generation methods to generate rotated bounding boxes [37]. However, this work is a direct application from text detection [26] to the remote sensing field. Yang et al. [6] further supplement a dense feature pyramid network based on RRPN and RoI align to improve performance. However, the authors declare that this framework using rotation region proposals has a defect of higher false alarms. Zhou et al. [28,29] propose approaches to use semantic segmentation networks to recognize parts of ships for rotated region proposal, where additional semantic pixel labels of ship parts have to be provided. On the basis of different RPN design to produce rotated bounding boxes, many approaches are being used to improve the performance of inshore ship detection. As far as we know, very few studies, except for [6,28], have simultaneously employed a multi-scale feature, rotation region proposal network and contextual pooling in one uniform end-to-end model.

3. Proposed Method

In this section, we will present our detection method based on multi-scale feature fusion, rotation region proposal network and contextual pooling. The overall framework is shown in Figure 2, which mainly consists of three modules: Multi-scale Feature Extraction module, Rotation Region Proposal Network (RRPN) module, and Contextual Rotation Region-of-Interest (RRoI) Pooling module. In this framework, the first module is used to fuse multi-scale features of different hierarchies, the RRPN module is applied to generate rotation bounding boxes of arbitrary orientations, and the Contextual RRoI pooling module pools the contextual information and implements the classification of rotation bounding boxes and the position regression.

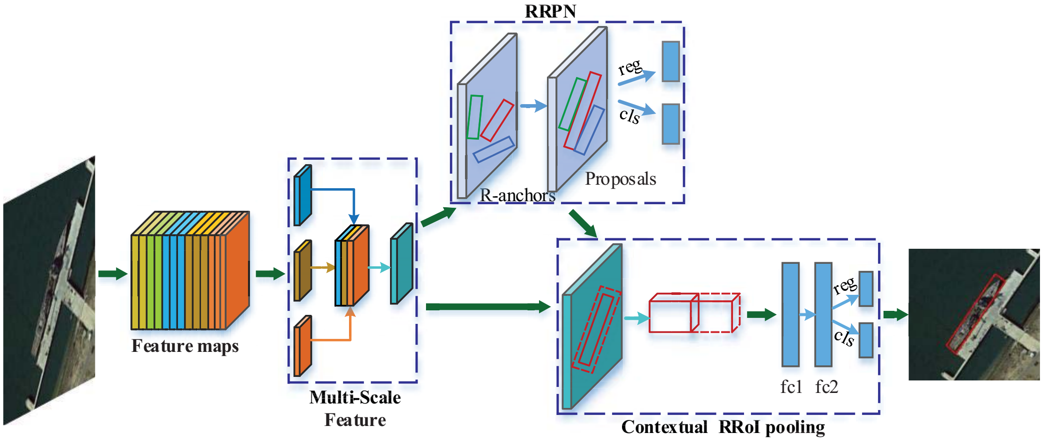


Figure 2. Overall framework of our proposed network.

3.1. Multi-Scale Features

Different layers of a hierarchical convolutional neural network contribute differently to the extraction of features. The lower levels of CNNs have smaller receptive fields and higher spatial resolutions, which can provide more detailed local structure information, whereas the higher levels have larger receptive fields and prefer to extract advanced information concerning semantics. As for the inshore ship targets we focus on, although they usually have similar slender shapes, their sizes vary a lot according to different types of ships. In order to adapt to various scales of targets, we propose to fuse features of different layers and construct a multi-scale feature extraction network structure.

Different convolutional layers naturally constitute a feature pyramid. Inspired by Feature Pyramid Network (FPN) [38], we can fuse different layers of this feature pyramid to achieve multi-scale description with balance of speed and precision. Since many ship detection tasks adopt Fast R-CNN or Faster R-CNN as the baseline detection network, we can also employ VGG-16 as the backbone feature extraction network in our pipeline for parallel comparison. Similarly, other benchmark networks such as ResNet can also be employed as the basic structure. For feature fusion, the backbone network is used as the basic structure with removing its fully connected layers and the last pooling layer. The convolutional layers of the 3rd, 4th and 5th convolutional module are named conv3, conv4 and conv5 for convenient description. As shown in Figure 3, we fuse the features coming from these three layers by pooling features of a low level and deconvoluting features of a high level at the same time. By means of this, feature maps of different layers are converted to the same size, which provides the prerequisite of feature fusion. Since the feature value ranges of low and high levels are often different, for the sake of better model generalization, feature values should be adjusted to a similar range before feature fusion. Therefore, we adopt L2 normalization to process them respectively, and then concatenate them together. Finally, a 1 × 1 convolution is applied to make the fused feature map have the same depth of conv5 layer.

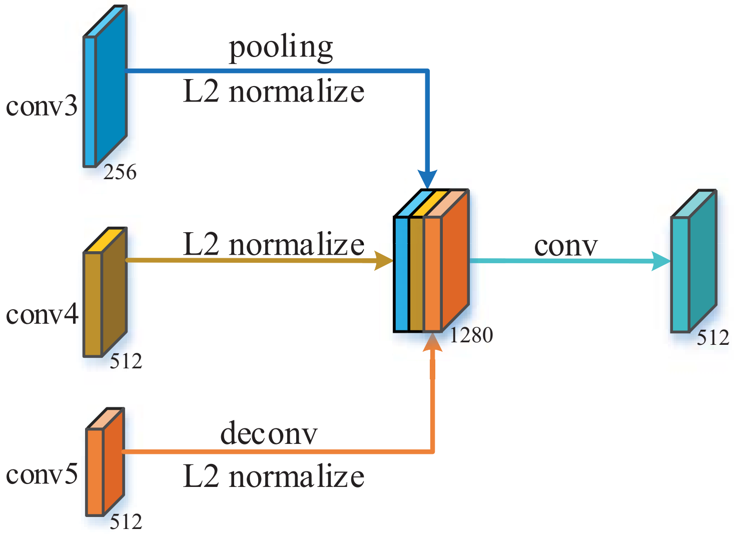


Figure 3. Multi-scale feature fusion scheme.

L2 normalization is applied to each pixel in the feature maps, which can be formulated by:

$$\hat{\mathbf{x}} = \frac{\mathbf{x}}{\|\mathbf{x}\|_2} \tag{1}$$

$$\|\mathbf{x}\|_2 = \left(\sum_{i=1}^{d} |x_i|^2 \right)^{1/2} \tag{2}$$

where **x** and $\hat{\mathbf{x}}$ represent the original pixel vector and normalized vector respectively, and d denotes the depth of the feature map.

When simple normalization as the above-mentioned is implemented on feature maps, the processed results may lie within a different value range from the original feature maps, which will reduce the learning rate. Inspired by the Batch Normalization (BN) layer, we bring in a learnable scaling factor γ to affect the normalized results. Let y_i denote the value after scaling:

$$y_i = \gamma_i \hat{x}_i \tag{3}$$

According to the Chain Rule, the back-propagation gradient of L2 normalization layer in the training process will be:

$$\frac{\partial l}{\partial \hat{\mathbf{x}}} = \frac{\partial l}{\partial \mathbf{y}} \gamma \tag{4}$$

$$\frac{\partial l}{\partial \mathbf{x}} = \frac{\partial l}{\partial \hat{\mathbf{x}}} \left(\frac{I}{\|\mathbf{x}\|_2} - \frac{\mathbf{x}\mathbf{x}^{\mathrm{T}}}{\|\mathbf{x}\|_2^3} \right) \tag{5}$$

$$\frac{\partial l}{\partial \gamma_i} = \sum_{y_i} \frac{\partial l}{\partial y_i} \hat{x}_i \tag{6}$$

where $\mathbf{y} = [y_1, y_2, \ldots, y_d]$.

3.2. Rotation Region Proposal Network

The function of a rotation region proposal network is to generate rotation bounding boxes of arbitrary orientations. It has a similar architecture as the region proposal network (RPN) of Faster R-CNN, whereas the generation of rotation bounding boxes asks for some modifications on the representation and pipeline. More specifically, representation of boxes and anchors, non-maximum suppression, loss function design as well as rotation angles all need to be taken into account.

3.2.1. Rotation Bounding Box

A horizontal bounding box is simple to represent; the top-left and bottom-right corners are sufficient to locate it with four coordinates $(x_{min}, y_{min}, x_{max}, y_{max})$. Since rotation bounding box can not be represent by two corners, a new scheme should be adopted. Let (x, y, w, h, θ) denote a rotation bounding box, which is depicted in Figure 4, where x and y represent the center of it, w and h represent the longer side and shorter side, and θ indicates the angle between the longer side and the horizontal direction with a range set to be $[-\frac{\pi}{4}, \frac{3\pi}{4})$.

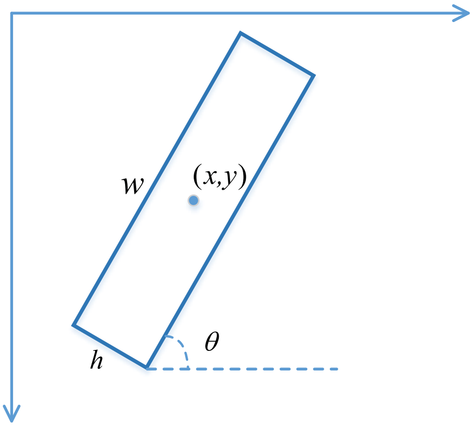
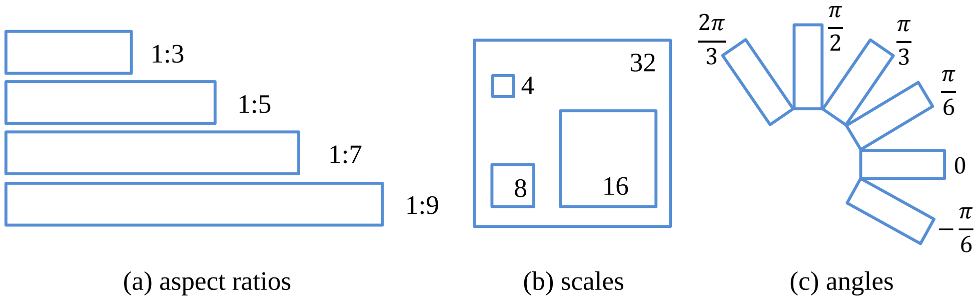


Figure 4. Representation of a rotation bounding box.

3.2.2. Rotation Anchors

The RPN network in Faster R-CNN provides k candidate windows for each pixel in the feature map, which are also called the anchors. Original anchors only concern the aspect-ratios and scales. In order to generate rotation boxes, we define the rotation anchors (R-anchors) which contain one more parameter, i.e., the skew angle. According to the characteristics of ships, we set the aspect ratios of R-anchors to be $1:3$, $1:5$, $1:7$ and $1:9$, and the scales as 4, 8, 16, and 32. The angles are divided as same as the literature [26]: $-\frac{\pi}{6}$, 0, $\frac{\pi}{6}$, $\frac{\pi}{3}$, $\frac{\pi}{2}$ and $\frac{2\pi}{3}$ (Figure 5). Similar with the rotation bounding boxes, each R-anchor is indicated by 5 variables (x, y, w, h, θ). Therefore, in the RRPN architecture, each pixel in the feature map corresponds to 96 R-anchors ($4 \times 4 \times 6$), and the depth of classification layer and regression layer will be 192 (2×96) and 480 (5×96), respectively.



(a) aspect ratios (b) scales (c) angles

Figure 5. Anchor strategy used in our RRPN.

3.2.3. Skew Intersection over Union and Non-Maximum Suppression

RPN networks usually generate a lot of bounding boxes, most of which contain no targets. In order to improve detection efficiency, selection of boxes with higher confidence before classification and regression is natural and effective. This process is often realized by Non-Maximum Suppression (NMS), which sorts region proposals according to their classification confidence coefficients, and reserves the one with highest confidence and deletes all the others overlapping more than a certain threshold with it. In the computation of NMS and the evaluation of detection accuracy, Intersection over Union (IoU) is a very important concept. Original IoU measures the degree of overlap between two rectangular bounding boxes A and B. The IoU is defined as the ratio of overlapping area to the area of A and B unions:

$$IoU = (A \cap B)/(A \cup B) \tag{7}$$

In the RRPN scheme, skew IoU and skew NMS take the place of original methods to deal with rotation bounding box issues. The algorithm of skew IoU is listed in Algorithm 1. As shown in Figure 6, we first find intersection points of two boxes, and then find the vertices inside the other [26]. These two kinds of points constitute the polygon of overlapping, so the area can be computed by sorting its vertices and triangulating the polygon. Non-maximum suppression for rotation regions proposals is the same as the original one except for the computation of IoU. Redundant candidate bounding boxes with lower confidences are removed if they have large enough skew IoU with high confidence.

Algorithm 1 Calculate skew IoU

```
Input: Rotation rectangles $R_1, R_2, \ldots, R_N$
Output: $IoU[1,N][1,N]$
Initialize $IoU[1,N][1,N]$ with 0
for each pair of $\langle R_i, R_j \rangle$ $(i < j)$ do
    Point set $PSet \leftarrow \emptyset$
    Add intersection points of $R_i$ and $R_j$ to $PSet$
    Add the vertices of $R_i$ inside $R_j$ to $PSet$
    Add the vertices of $R_j$ inside $R_i$ to $PSet$
    Sort $PSet$ into anticlockwise order
    Compute intersection $I$ of $PSet$ by triangulation
    $IoU(i,j) = Area(I)/\left(Area(R_i) + Area(R_j) - Area(I)\right)$
end for
return $IoU$
```

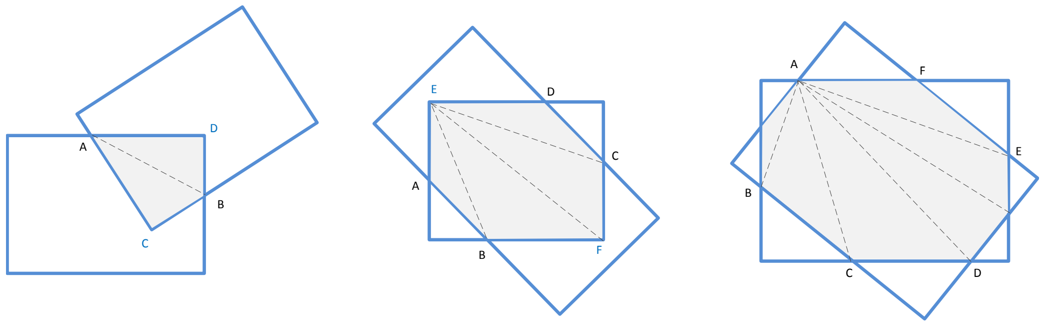


Figure 6. Examples of compute skew IoU. Intersection points are marked in black, and vertices inside the other rectangle are marked in dark blue.

3.2.4. Loss Function for RRPN Training

In the training of RRPN, the sampling strategy for R-anchors needs to be developed. In our application, the positive R-anchors are defined as: (a) the highest skew IoU, or (b) skew IoU larger than 0.5 and the intersection angle with respect to the ground truth of less than $\frac{\pi}{12}$. Negative R-anchors are characterized as: (a) skew IoU lower than 0.2, or (b) skew IoU larger than 0.5 but the intersection angle with the ground truth of larger than $\frac{\pi}{12}$. The loss function for the proposals takes the form of multi-task loss, which is defined as:

$$L(p, l, v^*, v) = L_{cls}(p, l) + \lambda l L_{reg}(v^*, v) \tag{8}$$

where l is the indicator of the class label ($l = 1$ for target and $l = 0$ for background), $p = [p_0, p_1]$ is the probability belonging to background or target. λ is a balancing parameter to trade off two terms of classification and regression. $v = (v_x, v_y, v_w, v_h, v_\theta)$ represents the predicted rotation bounding box, and $v^* = (v^*_x, v^*_y, v^*_w, v^*_h, v^*_\theta)$ denotes the ground truth.

The classification loss for class l is defined as:

$$L_{cls}(p, l) = -\log p_l \tag{9}$$

And for bounding box location regression, the loss function is:

$$L_{reg}(v^*, v) = \sum_{i \in \{x,y,h,w,\theta\}} \text{smooth}_{L_1}(v^*_i - v_i) \tag{10}$$

$$\text{smooth}_{L_1}(x) = \begin{cases} 0.5x^2 & \text{if } |x| < 1 \\ |x| - 0.5 & \text{otherwise} \end{cases}$$

v and v^* are computed by:

$$v_x = \frac{x - x_a}{w_a}, \quad v_y = \frac{y - y_a}{h_a}, \quad v_h = \log\frac{h}{h_a}, \quad v_w = \log\frac{w}{w_a}, \quad v_\theta = \theta \ominus \theta_a \tag{11}$$

$$v^*_x = \frac{x^* - x_a}{w_a}, \quad v^*_y = \frac{y^* - y_a}{h_a}, \quad v^*_h = \log\frac{h^*}{h_a}, \quad v^*_w = \log\frac{w^*}{w_a}, \quad v^*_\theta = \theta^* \ominus \theta_a \tag{12}$$

where x, x_a and x^* denote the predicted bounding box, anchor and the ground truth box, and the same for y, h, w and θ. The operation $a \ominus b = a - b + k\pi$, where $k \in \mathbb{Z}$ to ensure that $a \ominus b \in [-\frac{\pi}{4}, \frac{3\pi}{4})$.

3.3. Contextual RRoI pooling

Due to the fully connected layers for classification or regression, features from the feature map for each proposal should be pooled into a fixed length one. The same as described in Faster R-CNN, our classification and regression network includes two fully connected layers and two *cls* and *reg* multi-task branches (see Figure 2). Therefore, pooling of features is quite necessary in our pipeline. Since our region proposals are of arbitrary orientations, if a common pooling strategy is employed, there will be a lot of redundant information pooled in the horizontal boxes. So the pooling of a rotation Region-of-Interest (RoI) is developed in this subsection. Moreover, with the consideration of our application that background usually contains areas similar to ship targets in terms of structure and texture, if only the features of targets are considered, it is easy to increase false alarms. Therefore, we propose to bring in contextual information of targets when pooling the features.

3.3.1. Pooling of Rotation RoI

According to the fixed size of input of fully connected layers, RoI pooling is employed to convert feature map areas of different sizes (corresponding to different region proposals) into a vector of the same length. For rotation RoI, the pooling process is indicated in Figure 7. We divide the inclined feature map into $H_r \times W_r$ subregions along the long and short sides of the bounding box, and then carry out max pooling within each subregion to generate the RROI pooling results.

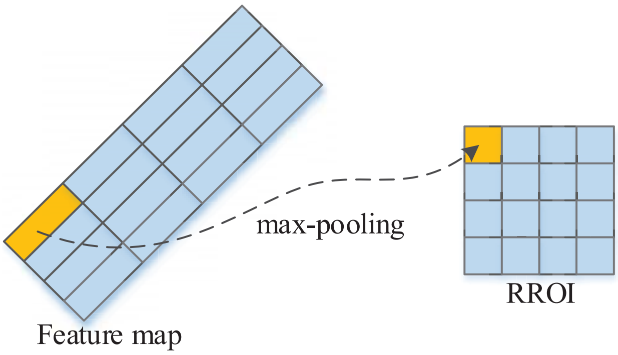


Figure 7. Rotation RoI pooling.

3.3.2. Contextual Information

There are different ways to bring in contextual information. The most common one is the horizontal bounding rectangle, which is used as "ROI Align" in [6]. However, as we have observed from port images, many ships are lying at docks one by one. Under this circumstance, we believe an inclined context is more adequate and reasonable than a horizontal one.

As shown in Figure 8, we extend the rotation bounding box produced by RRPN and obtain a box with context. Let $(x_p, y_p, w_p, h_p, \theta_p)$ denote the rotation box and $(x_c, y_c, w_c, h_p, \theta_c)$ denote the box with context, the relationship between them can be represent as:

$$x_c = x_p, \quad y_c = y_p, \quad h_c = \alpha h_p, \quad w_c = \beta w_p, \quad \theta_c = \theta_p \tag{13}$$

We implement RRoI pooling on feature map areas corresponding to a rotation bounding box and a box with context, respectively, and then concatenate the pooling results to obtain a feature vector that describes the target itself as well as its contextual information. Finally, this feature will be sent to the classification and regression networks.

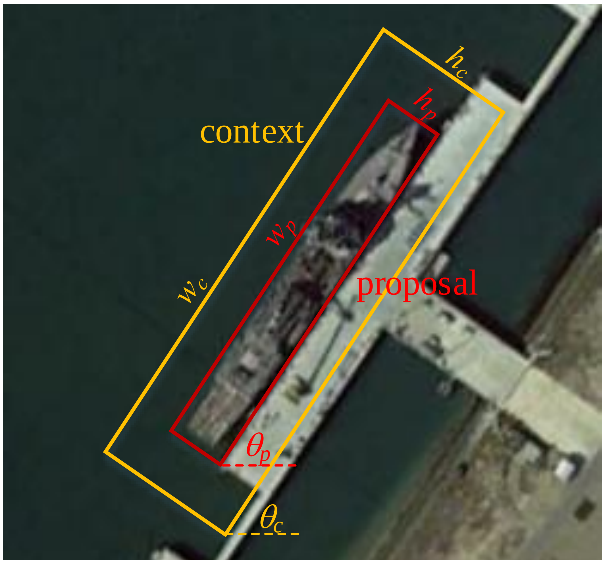


Figure 8. Contextual information of a region proposal.

4. Experimental Results and Discussions

4.1. Data Sets and Evaluation Criteria

Two datasets are employed in our experiments. The first dataset is collected from Google Earth, which covers various ports all over the world and originally contains 278 color images with a resolution of approximate 1 m. The length (width) of each image varies from 1000 to 3000 pixels. This dataset is provided with both semantic labels and rotated bounding box labels, which can be used for port image segmentation and inshore ship detection. We have made it available on: https://github.com/lllltdaf2/Ship-Seg-Detect-data. We augment the original set to 400 images by rotation, where 100 images are selected as the test set and the remaining are used as training ones. Ship targets are manually marked with rotation bounding boxes and very tiny yachts are ignored. For scale evaluation, targets are divided into ones of large sizes and ones of small sizes according to their length, where 80 pixels (approximate 80 m in real world) are set as the threshold of division. Sample images of this dataset are shown in Figure 9. For the images used for training, we crop image blocks of 1000 × 1000 pixels with a stride of 600 on the original images to generate training samples. In the training process, these samples are augmented with random flips.

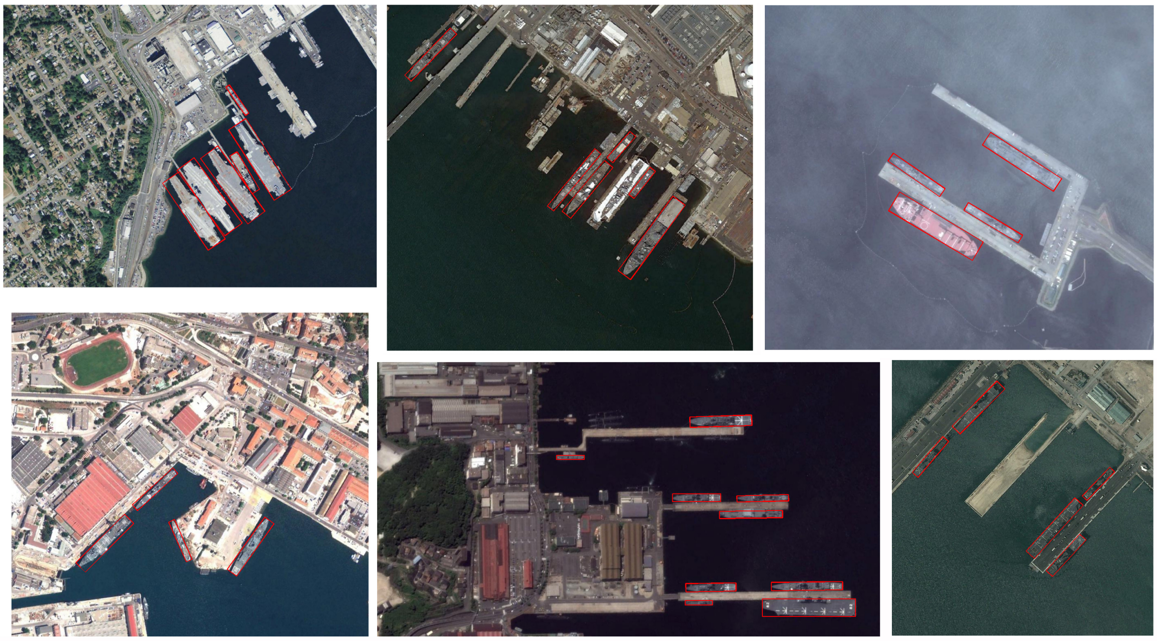


Figure 9. Sample images and manually marked ground truth of the port image dataset collected from Google Earth.

The other dataset is the public HRSC2016 ship dataset [27,39], which contains 1061 images with resolutions between 0.4m and 2m. Images in this dataset range from 300 × 300 to 1500 × 900, and the number of training, validation and test images are 436, 181 and 444, respectively. The HRSC2016 is a multi-applicable dataset, which includes instance labels of three levels. The first level is ship class, and L2 and L3 are ship types of finer categories. Since our method is a detection approach but not a classification one, evaluation on Level 1 is enough to show the capability of ship detection.

Precision and recall are employed as evaluation criteria, which are defined as the ratio of correct number of detected targets to detected number of targets, and the ratio of correct number of detected targets to actual number of targets. The false alarm, which means the number of wrong detected targets, can also reflect the effects of a model. For HRSC2016 dataset, most literatures use Mean Average Precision (mAP) as a criterion. Since mAP is popular on public datasets, we employ it as well for convenient comparison.

In the test process, an IoU threshold of 0.5 is adopted [25,26], which is also convenient for fair comparison with other methods. Actually, as the literature [29] has pointed out, the difficulty of a same threshold of rotated bounding box is not comparable with a horizontal bounding box. As shown in Figure 10, since ships are usually in narrow-rectangle shapes, the rotated IoU overlap between the detected bounding box and the ground truth box is very sensitive to the angle variation, which makes the rotated box more difficult to meet the detected IoU threshold than a horizontal box under the same threshold value. Consequently, methods producing horizontal boxes are listed for reference, and direct comparison on measure values between different bounding boxes does not make much sense.

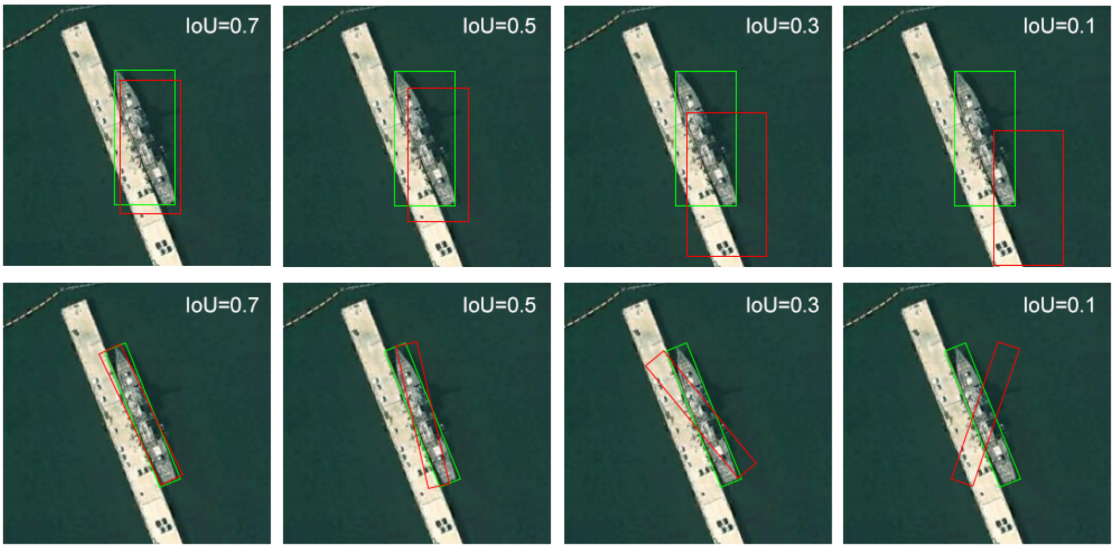


Figure 10. Rotated bounding box is very sensitive to angle variation, as a result, it is more difficult to be determined as "detected" than horizontal box under the same value of IoU threshold.

4.2. Experimental Settings

The experiments were carried out on Ubuntu 16.04, NVIDIA 1080Ti GPU, with Caffe and TensorFlow deep learning framework. Since we have employed the VGG-16 model as the backbone structure in the multi-scale feature extraction module, we use its parameters pre-trained on ImageNet which come from the Caffe Model Zoo (https://github.com/BVLC/caffe/wiki/Model-Zoo) [40] and then fine-tune the model in the end-to-end training. The model is trained end-to-end with gradually decreased learning rates (10^{-3} for the first 40,000 iterations, 10^{-4} for the next 40,000 iterations, and 10^{-5} for the last 20,000), weight decay of 5×10^{-4} and momentum of 0.9. The parameters of contextual expansion boxes are: $\alpha = 2.2$ and $\beta = 1.4$.

4.3. Model Analysis

Because multi-scale features, rotation bounding boxes and contextual RRoI pooling are all employed in our method, experiments are designed to validate the benefits of each approach. In our own dataset, we classify ships into large ships and small ships according to their length (threshold length of 80 pixels, approximate 80m in reality), and the results of the model analysis are given in Table 1.

It can be seen from Table 1 that MSF (multi-scale features), RRPN (rotation region proposal network) and C-RRoI (pooling of contextual rotation RoI) are all beneficial to enhance the precision and recall of ship detection. The employment of a rotation bounding box ensures a more accurate representation of target location, and the skew non-maximum suppression avoids target loss of NMS when ships are densely docked. As a result, Model_1 for analysis shows better recall rates than Faster R-CNN. Precision of Model_2 for analysis is further enhanced compared to Model_1 and Faster R-CNN due to the use of contextual RRoI pooling. When ships are inclined, a rotation bounding box includes less surrounding background information than a horizontal one, which may be insufficient for accurate target description. Since the RRoI is expanded with more contexts, the effect is validated by Model_2. The proposed method includes all three key structures as described in Section 3, where multi-scale features are fused to better describe targets of different sizes. Therefore, results on both small and large ships are superior to the former models.

Table 1. Results of model analysis on our own dataset. MSF means multi-scale features, RRPN respresents rotation region proposal network, C-RRoI indicates pooling of contextual rotation RoI. "✓" means adopted while "×" means not. If one approach is not adopted, that part of structure is employed according to Faster R-CNN [19].

Method	MSF	RRPN	C-RRoI	Target Type	Actual Number	Correct Detected Number	Recall(%)	False Alarm	Precision(%)
Faster R-CNN	×	×	×	Large Small	394 92	327 68	83.0 73.9	37	91.4
Model_1	×	✓	×	Large Small	394 92	334 70	84.8 76.1	39	91.2
Model_2	×	✓	✓	Large Small	394 92	357 79	90.6 85.9	37	92.2
Proposed Method	✓	✓	✓	Large Small	394 92	365 82	92.6 89.1	22	95.3

4.4. Contrast Experiments

In order to evaluate the effectiveness of the proposed method, we further compare it with the following methods on our own dataset: Faster R-CNN [19], SSD [21] and RRPN [26]. The quantitative results are listed in Table 2.

Table 2. Comparative results with other methods on our dataset.

Method	Target Type	Actual Number	Correct Detected Number	Recall(%)	False Alarm	Precision(%)
Faster R-CNN	Large Small	394 92	327 68	83.0 73.9	37	91.4
SSD	Large Small	394 92	293 22	74.4 23.9	95	76.8
RRPN	Large Small	394 92	352 77	89.3 83.7	55	88.6
Proposed Method	Large Small	394 92	365 82	92.6 89.1	22	95.3

Table 2 illustrates that the proposed method outperforms the other contrast approaches on both recall and precision rates. Although the model of RRPN adopts rotation RPN and pooling as well, it ignores the multi-scale features and target context information. Because the port usually contains a lot of interference objects similar to ships, the recall and precision of RRPN model suffer from these problems. Table 3 shows the computational consumption of the above methods. Our proposed method enhances the detection results with the price of a reasonable increase of computation power and running time.

Table 3. Computational consumption of each tested model on our dataset.

Method	Faster R-CNN	RRPN	Proposed Method
Memory occupied	2010 M	2124 M	2546 M
Average time	0.11 s	0.16 s	0.25 s

Table 4 shows the detection results of state-of-the-art methods on HRSC2016 dataset. Baseline methods include 1, 2, 5, 5 and 7 [27,39] which are proposed by the research group that released HRSC2016 data. These methods adopt an SRBBS candidate proposal approach and they are not end-to-end frameworks. Since this dataset provides ground truth of both horizontal and rotated bounding boxes, several typical methods generating horizontal boxes are also listed for reference, including Faster R-CNN [19] retested by us. It should be pointed out that ground truth of horizontal bounding boxes and rotated ones are different. Some results are cited from state-of-the-art methods including RDFPN [6] and RRPN+RROI Pooling [37]. Most of the involved methods adopt VGG-16 as the backbone feature extraction network. It is seen from the table that our method has achieved state-of-the-art performance.

Table 4. Detection results of state-of-the-art methods on HRSC2016 dataset.

No.	Method	End-to-End Model	Bbox Type	mAP
1	SRBBS (NRER-REG-BB) [39]	×	horizontal	55.7
2	SRBBS (NBEB-REG-BB) [39]	×	horizontal	79.7
3	SHD-HBB [41]	✓	horizontal	69.5
4	Faster R-CNN [19]	✓	horizontal	84.0
5	SRBBS (NRER-REG-RBB) [39]	×	rotated	69.6
6	SRBBS (NREB-REG-RBB) [39]	×	rotated	79.0
7	RR-CNN [27]	×	rotated	75.7
8	RDFPN [6]	✓	rotated	76.2
9	RRPN+RROI Pooling [37]	✓	rotated	79.6
10	Proposed Method	✓	rotated	80.8

4.5. Visualized Results

Figure 11 shows the visual effects of contextual RRoI pooling. The top row exhibits some inaccurate results produced by RRPN, including improper bounding boxes and undetected ships, and the bottom row indicates the correction effects of contextual RRoI. By introducing contextual information around the targets, insufficient contextual features of RRoI are supplemented for a better detection precision.

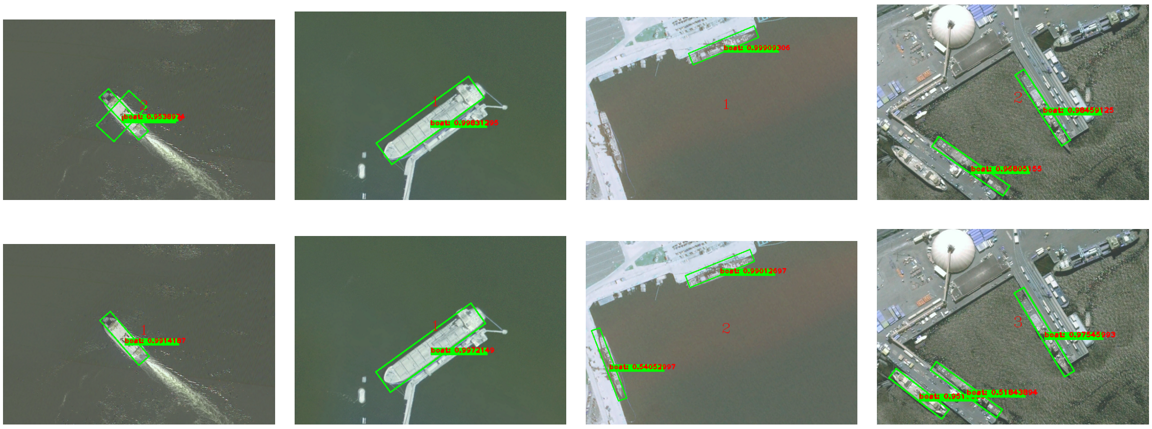


Figure 11. Visual effects of contextual RRoI pooling.

More visualized results are provided in Figure 12 and Figure 13 to display detection effect of the proposed method. The figures show that it can detect inshore ships of different types and various sizes, as well as ships close to each other.

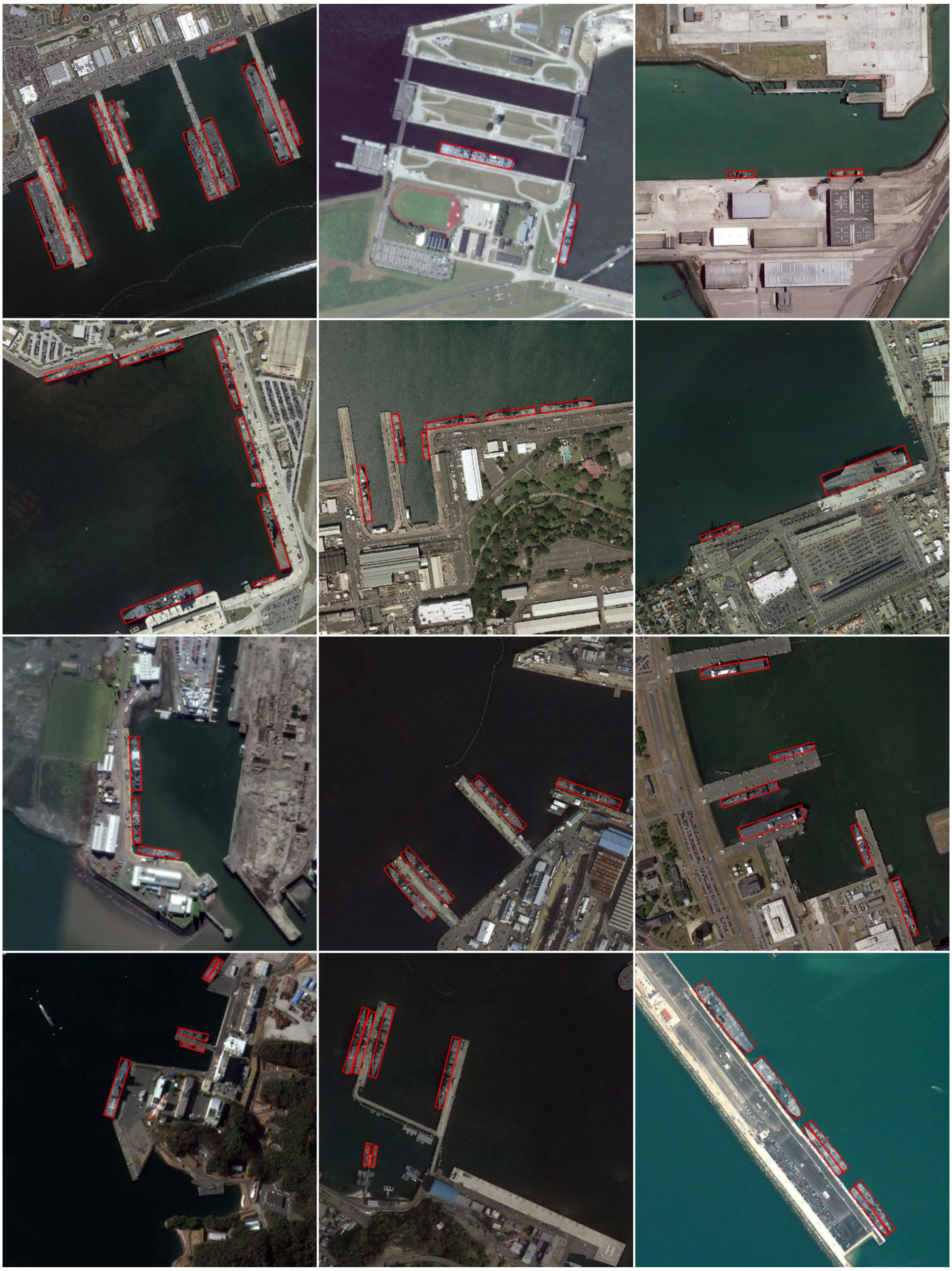


Figure 12. Visualization of detected ships with the proposed method on our dataset.

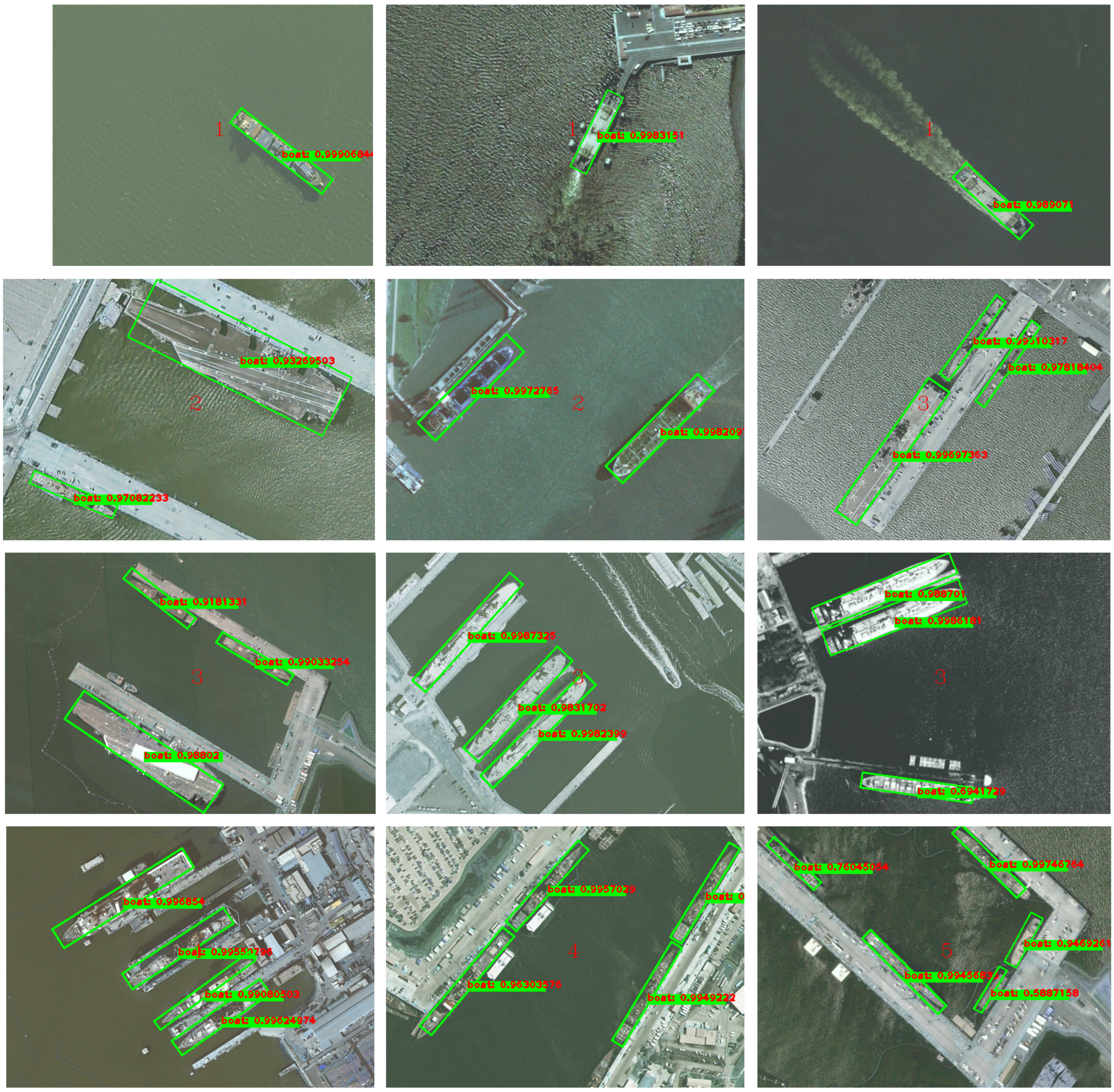


Figure 13. Visualization of detected ships with the proposed method on HRSC2016 dataset.

5. Conclusions

In this paper, we propose an inshore ship detection framework based on a multi-scale feature pyramid fusion network, rotation region proposal network and contextual rotation region of interest pooling. In consideration of various sizes of inshore ships, the proposed method fuses multi-scale features from a pyramid of a backbone feature extraction network to better describe targets of different sizes. In order to locate inshore ships more accurately and distinguish targets that are densely arranged, a rotation region proposal network with skew non-maximum suppression is introduced to generate region proposals. Moreover, to offset the context loss impact of inclined region proposal, a contextual region alongside the rotation RoI is added to supplement effective information of inshore ships. Experiments of model analysis validate that each step mentioned above contributes to the final results of detection, and the proposed model with all three steps achieves state-of-the-art performance compared to other methods.

In the future work, the following aspects are worth further investigation and should be improved for the detection of inshore ships: (1) Backbone detection framework. Currently, Faster R-CNN is employed as the most popular backbone detection network, and improved efficient and concise models may be used as the backbone in the future. (2) A Method of multi-scale feature extraction and fusion.

Many approaches aim to extract multi-scale features, whereas more accurate and faster methods are still in demand. (3) An Approach to produce region proposals. How to generate effective region proposals of reasonable quantity is a key step in object detection, which will affect the efficiency and performance of the whole model. (4) The selection of pooling method, loss function and context addition method. All of these details will influence the model performance especially under conditions of imbalanced samples and complex interferences. (5) Detection and other methods in combination, for example, with semantic segmentation and visual saliency may all contribute to the detection of ships.

Author Contributions: T.T. proposed the idea and wrote the manuscript. Z.P. designed the method pipeline and implemented the validation. X.T. collected experimental data and helped with validation. Z.C. solved problems about software and helped with validation and visualization. All authors have read and agreed to the published version of the manuscript.

Funding: This work is supported by the National Natural Science Foundation of China under Grant No. 41701417 and No. U1711266, the Opening Fund of Ministry of Education Key Laboratory of Geological Survey and Evaluation (Grant No. GLAB2019ZR08), and the Fundamental Research Funds for the Central Universities.

Acknowledgments: The authors would like to thank the colleagues and students in National Key Lab of Multispectral Information Processing Technology, HUST, for their work of dataset collection and assistance of experiments. The authors would also like to thank the anonymous reviewers for their helpful comments and suggestions to improve the manuscript.

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

SAR	Synthetic Aperture Radar
CNNs	Convolutional Neural Networks
R-CNN	Region proposal Convolutional Neural Network
RPN	Region Proposal Network
NMS	Non-Maximum Suppression
SSP-Net	Spatial Pyramid Pooling Network
YOLO	You Only Look Once
SSD	Single Shot Multibox Detector
Bbox	Bounding box
R^{2}CNN	Rotational Region CNN
RRPN	Rotation Region Proposal Network
RRoI	Rotation Region of Interest
MSF	Multi-Scale Feature
RoI	Region of Interest
BN	Batch Normalization
IoU	Intersection over Union

References

1. Kang, M.; Ji, K.; Leng, X.; Lin, Z. Contextual region-based convolutional neural network with multilayer fusion for SAR ship detection. *Remote Sens.* **2017**, *9*, 860. [CrossRef]
2. Chen, W.; Li, X.; He, H.; Wang, L. A review of fine-scale land use and land cover classification in open-pit mining areas by remote sensing techniques. *Remote Sens.* **2018**, *10*, 15. [CrossRef]
3. Li, X.; Tang, Z.; Chen, W.; Wang, L. Multimodal and multi-model deep fusion for fine classification of regional complex landscape areas using ZiYuan-3 imagery. *Remote Sens.* **2019**, *11*, 2716. [CrossRef]
4. Ma, J.; Zhou, H.; Zhao, J.; Gao, Y.; Jiang, J.; Tian, J. Robust feature matching for remote sensing image registration via locally linear transforming. *IEEE Trans. Geosci. Remote Sens.* **2015**, *53*, 6469–6481. [CrossRef]
5. Ma, J.; Jiang, J.; Zhou, H.; Zhao, J.; Guo, X. Guided locality preserving feature matching for remote sensing image registration. *IEEE Trans. Geosci. Remote Sens.* **2018**, *56*, 4435–4447. [CrossRef]

6. Yang, X.; Sun, H.; Fu, K.; Yang, J.; Sun, X.; Yan, M.; Guo, Z. Automatic ship detection in remote sensing images from google earth of complex scenes based on multiscale rotation dense feature pyramid networks. *Remote Sens.* **2018**, *10*, 132. [CrossRef]
7. Wang, L.; Zhang, J.; Liu, P.; Choo, K.K.R.; Huang, F. Spectral–spatial multi-feature-based deep learning for hyperspectral remote sensing image classification. *Soft Comput.* **2017**, *21*, 213–221. [CrossRef]
8. Bi, F.; Chen, J.; Zhuang, Y.; Bian, M.; Zhang, Q. A decision mixture model-based method for inshore ship detection using high-resolution remote sensing images. *Sensors* **2017**, *17*, 1470. [CrossRef]
9. Ma, J.; Wang, X.; Jiang, J. Image super-resolution via dense discriminative network. *IEEE Trans. Ind. Electron.* **2019**. [CrossRef]
10. Zhu, C.; Zhou, H.; Wang, R.; Guo, J. A novel hierarchical method of ship detection from spaceborne optical image based on shape and texture features. *IEEE Trans. Geosci. Remote Sens.* **2010**, *48*, 3446–3456. [CrossRef]
11. Liu, Z.; Wang, H.; Weng, L.; Yang, Y. Ship rotated bounding box space for ship extraction from high-resolution optical satellite images with complex backgrounds. *IEEE Geosci. Remote Sens. Lett.* **2016**, *13*, 1074–1078. [CrossRef]
12. He, H.; Lin, Y.; Chen, F.; Tai, H.M.; Yin, Z. Inshore ship detection in remote sensing images via weighted pose voting. *IEEE Trans. Geosci. Remote Sens.* **2017**, *55*, 3091–3107. [CrossRef]
13. Liu, G.; Zhang, Y.; Zheng, X.; Sun, X.; Fu, K.; Wang, H. A new method on inshore ship detection in high-resolution satellite images using shape and context information. *IEEE Geosci. Remote Sens. Lett.* **2014**, *11*, 617–621. [CrossRef]
14. Ma, L.; Crawford, M.M.; Zhu, L.; Liu, Y. Centroid and covariance alignment-based domain adaptation for unsupervised classification of remote sensing images. *IEEE Trans. Geosci. Remote Sens.* **2018**, *57*, 2305–2323. [CrossRef]
15. Dong, C.; Liu, J.; Xu, F. Ship detection in optical remote sensing images based on saliency and a rotation-invariant descriptor. *Remote Sens.* **2018**, *10*, 400. [CrossRef]
16. Girshick, R.; Donahue, J.; Darrell, T.; Malik, J. Region-based convolutional networks for accurate object detection and segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.* **2015**, *38*, 142–158. [CrossRef]
17. He, K.; Zhang, X.; Ren, S.; Sun, J. Spatial pyramid pooling in deep convolutional networks for visual recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **2015**, *37*, 1904–1916. [CrossRef]
18. Girshick, R. Fast R-CNN. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 11–18 December 2015; pp. 1440–1448.
19. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster R-CNN: Towards real-time object detection with region proposal networks. In Proceedings of the Advances in Neural Information Processing Systems, Montreal, QC, Canada, 7–12 December 2015; pp. 91–99.
20. Redmon, J.; Divvala, S.; Girshick, R.; Farhadi, A. You only look once: Unified, real-time object detection. In Proceedings of the IEEE conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 26 June–1 July 2016; pp. 779–788.
21. Liu, W.; Anguelov, D.; Erhan, D.; Szegedy, C.; Reed, S.; Fu, C.Y.; Berg, A.C. SSD: Single shot multibox detector. In *European Conference on Computer Vision*; Springer: Amsterdam, The Netherlands, 8–16 October 2016; pp. 21–37.
22. Zhang, S.; Wu, R.; Xu, K.; Wang, J.; Sun, W. R-CNN-based ship detection from high resolution remote sensing imagery. *Remote Sens.* **2019**, *11*, 631. [CrossRef]
23. Wu, F.; Zhou, Z.; Wang, B.; Ma, J. Inshore ship detection based on convolutional neural network in optical satellite images. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2018**, *11*, 4005–4015. [CrossRef]
24. Ma, J.; Yu, W.; Liang, P.; Li, C.; Jiang, J. FusionGAN: A generative adversarial network for infrared and visible image fusion. *Inf. Fusion* **2019**, *48*, 11–26. [CrossRef]
25. Jiang, Y.; Zhu, X.; Wang, X.; Yang, S.; Li, W.; Wang, H.; Fu, P.; Luo, Z. R2CNN: Rotational region CNN for orientation robust scene text detection. *arXiv* **2017**, arXiv:1706.09579.
26. Ma, J.; Shao, W.; Ye, H.; Wang, L.; Wang, H.; Zheng, Y.; Xue, X. Arbitrary-oriented scene text detection via rotation proposals. *IEEE Trans. Multimed.* **2018**, *20*, 3111–3122. [CrossRef]
27. Liu, Z.; Hu, J.; Weng, L.; Yang, Y. Rotated region based CNN for ship detection. In Proceedings of the 2017 IEEE International Conference on Image Processing (ICIP), Beijing, China, 17–20 Septmber 2017; pp. 900–904.
28. Xiao, X.; Zhou, Z.; Wang, B.; Li, L.; Miao, L. Ship detection under complex backgrounds based on accurate rotated anchor boxes from paired semantic segmentation. *Remote Sens.* **2019**, *11*, 2506. [CrossRef]

29. Ma, J.; Zhou, Z.; Wang, B.; Zong, H.; Wu, F. Ship detection in optical satellite images via directional bounding boxes based on ship center and orientation prediction. *Remote Sens.* **2019**, *11*, 2173. [CrossRef]
30. Qi, S.; Ma, J.; Lin, J.; Li, Y.; Tian, J. Unsupervised ship detection based on saliency and S-HOG descriptor from optical satellite images. *IEEE Geosci. Remote Sens. Lett.* **2015**, *12*, 1451–1455.
31. Yang, F.; Xu, Q.; Li, B. Ship detection from optical satellite images based on saliency segmentation and structure-LBP feature. *IEEE Geosci. Remote Sens. Lett.* **2017**, *14*, 602–606. [CrossRef]
32. Ma, J.; Zhao, J.; Jiang, J.; Zhou, H.; Guo, X. Locality preserving matching. *Int. J. Comput. Vis.* **2019**, *127*, 512–531. [CrossRef]
33. Han, X.; Zhong, Y.; Zhang, L. An efficient and robust integrated geospatial object detection framework for high spatial resolution remote sensing imagery. *Remote Sens.* **2017**, *9*, 666. [CrossRef]
34. Ren, Y.; Zhu, C.; Xiao, S. Deformable Faster R-CNN with aggregating multi-layer features for partially occluded object detection in optical remote sensing images. *Remote Sens.* **2018**, *10*, 1470. [CrossRef]
35. Lin, H.; Shi, Z.; Zou, Z. Fully convolutional network with task partitioning for inshore ship detection in optical remote sensing images. *IEEE Geosci. Remote Sens. Lett.* **2017**, *14*, 1665–1669. [CrossRef]
36. Li, S.; Zhang, Z.; Li, B.; Li, C. Multiscale rotated bounding Box-based deep learning method for detecting ship targets in remote sensing images. *Sensors* **2018**, *18*, 2702. [CrossRef] [PubMed]
37. Zhang, Z.; Guo, W.; Zhu, S.; Yu, W. Toward arbitrary-oriented ship detection with rotated region proposal and discrimination networks. *IEEE Geosci. Remote Sens. Lett.* **2018**, *15*, 1745–1749. [CrossRef]
38. Lin, T.Y.; Dollár, P.; Girshick, R.; He, K.; Hariharan, B.; Belongie, S. Feature pyramid networks for object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 2117–2125.
39. Liu, Z.; Yuan, L.; Weng, L.; Yang, Y. A high resolution optical satellite image dataset for ship recognition and some new baselines. In Proceedings of the 6th International Conference on Pattern Recognition Application and Methods (ICPRAM 2017). Porto, Portugal, 24–26 February 2017; pp. 324–331.
40. Simon, M.; Rodner, E.; Denzler, J. ImageNet pre-trained models with batch normalization. *arXiv* **2016**, arXiv:1612.01452.
41. Feng, Y.; Diao, W.; Sun, X.; Yan, M.; Gao, X. Towards Automated Ship Detection and Category Recognition from High-Resolution Aerial Images. *Remote Sens.* **2019**, *11*, 1901. [CrossRef]