Target Detection of Marine Ships Based on a Cascade RCNN

Lei Wu, Shicai Chen* and Chengcheng Liang

Key Laboratory of Marine Dynamic Simulation, Dalian Maritime University, Dalian, Liaoning Province, China
*Email: 549175057@qq.com

Abstract. Faster R-CNN is a classic algorithm with high accuracy and a wide range of applications in the field of target detection, and Cascade R-CNN is improved based on Faster R-CNN. The article applies the Cascade R-CNN method to the detection of marine ship targets. It improves the traditional Faster R-CNN algorithm and extracts areas that may contain pedestrians through RPN. In this paper, a multi-layer cascade detector is used to distinguish and classify the target area, and an algorithm is designed to detect and verify the data set. In the end, it is concluded that the Cascade R-CNN algorithm performs better than the traditional algorithm.

1. Introduction
The development and competition of the ocean by mankind runs through the history of the development of human civilization [1]. In today's increasingly deepening globalization, the ocean is an important link for economic and cultural exchanges between countries in the world. As a tool for mankind to develop and utilize the ocean, ships play an irreplaceable role [2]. When the existing intelligent recognition algorithm performs image recognition, the recognition accuracy is often low due to the poor weather conditions of the image to be recognized, the complex shore-based background [3], and the small ship target to be recognized. With the rapid development of various sensors, especially the continuous advancement of imaging technology, ship identification technology has shown a diversified development trend [4]. Recognition based on target image information has gradually become the focus of research in the field of ship target recognition [5]. This paper is based on the Cascade R-CNN algorithm, aiming at the problems that ships usually encounter at sea in harsh environments such as rain, snow, fog, and other interference factors such as reefs. Ship targets are easily blocked, and the targets are small, and it is difficult to identify and locate with high accuracy. A method of multi-target and multi-scale fusion is proposed, and then the feature fusion is down-sampled to extract the features of small targets and occluded targets, and zoom in to make less feature loss. This paper proposes that the algorithm model combined with ships has significant advantages over the traditional Faster R-CNN, which can detect and identify ships and ship categories in complex environments with high precision and accuracy.

2. The Principle of Cascade RCNN
The Cascade RCNN network is mainly divided into two steps. The first is to locate the location of the target, and then to classify the target. First, provide a picture to the network, and obtain the features of the image through the feature extraction network; then, the feature pyramid locates candidate targets on the feature map, and selects different IOU threshold detection models through the cascade detector to train and train the selected samples. return. As can be seen from the figure, the difference between
Cascade RCNN and the traditional Faster RCNN is that it adds several detection models to the latter detection model. Each detection model is trained based on the positive and negative samples of different IOU, and the output of the previous detection model is the input of the latter detection model. The detection models are all focused on detecting the proposal of the IOU in a certain range, thereby improving the accuracy of the detection. The specific structure of Cascade RCNN used in this article is shown in figure 1.

![Cascade RCNN structure diagram.](image)

**Figure 1.** Cascade RCNN structure diagram.

### 3. The Method of Cascade RCNN

#### 3.1. Multi-scale Feature Fusion

After a calibrated picture is randomly selected and imported, and after multiple convolution and pooling of the network, the feature map of each layer is obtained as shown in figure 2.

![Feature map visualization.](image)

**Figure 2.** Feature map visualization.

Looking from left to right in Figure 2, you can see the entire feature extraction process. In (b) to (c), the image of the ship can be clearly seen, and the resolution can be better. In addition to the background is a little fuzzy and unclear, it still contains the basic information and features of the original image. However, in (d) to (f), the image is gradually blurred, and the background, contour, and color difference are gradually extracted; the ship marked by the red frame is gradually blurred until it finally disappears. The features in the original image are also gradually extracted in the step-by-step convolution, so as to obtain the key information of the image, which has more global semantic information. In this way, it is better to distinguish ship targets that are blocked, and the target is small, and it is difficult to be recognized with high precision. For small targets, single-scale training is difficult to detect small targets, and multi-scale training can improve the robustness of the model. Multi-scale is divided into training stage multi-scale and test stage multi-scale, among which the multi-scale training stage is divided into image pyramid and feature pyramid. In this paper, the image pyramid method is used in the training stage to send images of multiple resolutions to the network for recognition. During training, a scale is randomly selected every certain cycle for training, so that the trained model is robust and can accept any size as input. For example, by merging figure 2(e) with figure 2(f), deep features can be integrated into the shallow more and more network, so that the small network retains important information while incorporating more semantic information. First, we perform deep convolution on the input image, and then perform dimensionality reduction operations on the features on P2, and then perform an addition operation to the processed P2 and the processed next [6]. The idea behind it is to obtain a strong semantic information, which can improve the detection performance. Then merge the down-sampled P2 with the Conv3 output feature map of
changing the number of channels. Finally got P3. Since directly superimposing features can easily cause feature discontinuities and lead to feature confusion, a 3×3 convolution is used to convolve the fused feature maps to eliminate the difference in feature distribution between different feature maps and ensure the stability of the features. In the same way, P4 and P5 can be obtained, and the specific fusion formula is:

\[ F_i = conv(I_i) + \text{dim}(I_{i+1}), i \in [2, n] \]  

Among them is the feature map output after the i-th level fusion, representing the feature map output by the ResNet101 network, representing the down-sampling of the i-th layer feature map, and representing the 1×1 convolutional dimensionality reduction operation on the i+1-th layer feature map, n represents the ResNet101 network level. Using the above fusion method is beneficial to the detection of fuzzy targets and small targets.

3.2. Hollow Convolution Downsampling

In the process of the multi-scale feature fusion process described in the previous step, the fuzzy target and the small target features in the photo gradually become less or even lost in this fusion process [7]. Because multi-scale accumulates the processed low-level features and processed high-level features, the purpose of this is because the low-level features can provide more accurate location information, and multiple down-sampling and up-sampling operations make the location information of the deep network There are errors. Therefore, it is proposed to use hole convolution for downsampling to reduce feature loss. Contrasted to the ordinary convolution, the hole convolution has more parameters with the hole rate r, which can expand the receptive field and maintain the image resolution without adding additional calculations. Since targets of different scales correspond to different receptive fields, three kinds of hole convolutions with different hole rates are paralleled for down-sampling, and information of different ranges and sizes around the target can be obtained. At the same time, the convolution ranges of different hole convolutions are different, so that different range features can be retained after convolution, reducing feature loss, and finally the down-sampled feature maps are fused. The specific operation is shown in figure 3:

![Figure 3. Dilated convolution down sampling.](image)

As shown in figure 3, multiple hole convolutions with different hole ratios are connected in parallel to uniformly process the input feature map. The specific method is as follows: down-sampling is performed using three types of 3×3 hole convolutions a step size of 2, where the hole rates are 1, 2, and 3 respectively. Then use the concat fusion method with batch normalization to fuse the convolved feature maps, and finally use 1×1 convolution to perform. This expression can be expressed as:

\[ F = H_{3.1(x)} + H_{3.2(x)} + H_{3.3(x)} \]  

Among them is the hole convolution, c is the size of the convolution kernel, and d is the hole rate, which is the feature after fusion. Compared with single convolutional downsampling, multi-branch hole convolutional downsampling makes the final fused feature map not only retains the features after single convolution downsampling, but also incorporates target surrounding information to enrich target features while reducing features lost.
4. Experimental Results and Experimental Analysis

4.1. Experimental Data
In this experiment, the data set consists of 2960 pictures, including 671 containers, 306 ro-ro ships, 610 bulk carriers, 291 oil tankers, 458 liquefied gas ships, 372 passenger ships, and 252 fishing boats.

4.2. Formatting Author Affiliations
The experimental operating environment is shown in table 1:

| Category                  | Version number                        |
|---------------------------|---------------------------------------|
| Operating system          | Ubuntu 16.04                          |
| CPU                       | XeonSilver 4007 CPU                    |
| GPU                       | NVIDIA GTX 1660Ti                      |
| Volume number             | Bold type                             |
| The version of python     | Python 3.7                             |

4.3. Experimental Data
In this experiment, the data set consists of 2960 pictures, including 671 containers, 306 ro-ro ships, 610 bulk carriers, 291 oil tankers, 458 liquefied gas ships, 372 passenger ships, and 252 fishing boats.

4.4. Evaluation Index
Image detection needs to use a rectangular box to select the target detection object. According to the overlap ratio between the detection result and the target box greater than 0.90, it is regarded as a qualified candidate. The IOU calculation formula between the predicted instance A and the real instance B is:

\[
IOU(A, B) = \frac{|A \cap B|}{|A \cup B|}
\]

In the formula, A is the predicted instance; B is the real instance; IOU is the cross-union ratio.

Determine whether the image content matches according to whether the name of the ship at sea in the picture is consistent with the candidate name. In order to fully evaluate the effectiveness of the model, the recall and precision must be checked at the same time. The calculation formula for the recall and precision of the test results is:

\[
R = \frac{TP}{TP + FN}
\]

\[
P = \frac{TP}{TP + FP}
\]

\( R \) is the recall rate; \( TR \) is a true example, which means that the model correctly predicts the positive category sample as a positive category; \( FN \) is a true negative example, which means that the negative category sample is correctly predicted as a negative category.

\( P \) is the precision rate; \( FP \) is a false positive example, which means that the negative category sample is incorrectly predicted as the positive category.

It is the most ideal situation when the accuracy rate and recall rate evaluation indicators are both optimal; but under normal circumstances, the accuracy rate is high, the recall rate is low; conversely, the recall rate is high, and the accuracy rate is low. Therefore, this article uses \( F_1 \) comprehensive evaluation indicators to comprehensively consider the accuracy rate and recall rate, so as to evaluate the performance of the model more reasonably. The \( F_1 \) calculation formula is:
\[ F_1 = \frac{2PB}{P+R} \] 

(6)

4.5. Training Parameters
The initial learning rate of training is set to 0.01, and it drops to 0.001 after the 4th iteration, and to 0.0001 after the 8th iteration, and then continues to iterate for 4 iterations at a learning rate of 0.0001 to stop. The optimization function uses the stochastic gradient descent method, and the momentum and decay rate are set to 0.9000 and 0.0001 respectively, and experiments are performed on the PYTORCH architecture.

4.6. Data Set Experiment Results
This section will conduct comparative experiments and ablation experiments on the ship data set to verify the effectiveness of the proposed innovations. The IOU and the confidence threshold are set to 0.5, and the detection results on the ship data set are shown in table 2:

| Model category               | P    | FPS  | \( F_1 \) |
|-------------------------------|------|------|---------|
| Faster rcnn_resnet50          | 78.96% | 10.9s | 0.94518 |
| Cascade rcnn_resnet50         | 89.72% | 7.1s  | 0.98124 |

The experimental results show that the experimental results of Cascade RCNN are compared with those of the traditional Faster RCNN as shown in table 2. The accuracy of the improved model based on Cascade RCNN is very high. It is much higher than Faster RCNN to detect ships at sea under more complicated and severe conditions, and the missed detection rate is less. The Cascade RCNN model will perform better, its generalization ability will be better, and its accuracy will be higher.

5. Conclusion
Based on the network idea of Cascade RCNN, this paper designs an algorithm system that is more suitable for maritime ship identification in complex environments. On the verification set, the F1 index achieved a score of 0.98124, which achieved accurate detection and identification of maritime ships, and obtained the following Conclusion:

1) When the positioning accuracy of the detected target is high, the advantage of Cascade RCNN is significantly higher than that of Raster RCNN. Use a Roman e for an exponential e; for example, \( y = e^x \).

2) According to the characteristics of the marine ship type detection data set, the method of multi-scale feature fusion on the picture, combined with the method of cavity convolution down sampling, cascade detector, etc., can significantly enhance the detection effect of the model.

3) When the size of the detected picture far exceeds the size of the detected target, the coarse-to-fine method proposed in this paper can greatly improve the accuracy of detection while saving computing resources.

4) Combining models with similar performance but different network structures can better improve the generalization ability and detection effect of the model.

6. References
[1] Agarwal S, Terrail J O D and Jurie F 2018 J. ArXiv Preprint arXiv 1809 03
[2] Kim K J, Kim P K, Chung Y S and Choi D H 2019 Proc. Int. Conf.on IEEE Access (Shanghai) vol 7 (IEEE) p 78311
[3] Nguyen H 2019 J. Mathematical Problems in Engineering 03 11
[4] Zhou T, Li Z and Zhang C 2019 J. Machine Learning and Cybernetics 10 11
[5] Cai Z and Vasconcelos N 2018 Proc. Int. Conf. on IEEE computer vision and pattern recognition (Salt Lake City) vol 31 (CVPR) p 6154
[6] Kong F F and Song B B 2020 J. Computer Engineering and Applications 56 20
[7] Jin T, Xiaofeng S, Rui R and W 2016 J. Data Acquisition and Processing 01 32