

# A Ship Target Location and Mask Generation Algorithms Base on Mask RCNN

Lin Shaodan<sup>1,\*</sup>, Feng Chen<sup>2</sup>, Chen Zhide<sup>2</sup>

<sup>1</sup>Department of Information Engineering, Fujian Chuanzheng Communications College, Fuzhou 350007, China

<sup>2</sup>College of Mathematics and Informatics, Fujian Normal University, Fuzhou 350007, China

## ARTICLE INFO

### Article History

Received 03 Jun 2019

Accepted 04 Oct 2019

### Keywords

Mask RCNN

Mask

Region proposal network

Upsample

ROI align

## ABSTRACT

Ship detection is a canonical problem in computer vision. Motivated by the observation that the major bottleneck of ship detection lies on the different scales of ship instances in images, we focus on improving the detection rate, especially for the small-sized ships which are relatively far from the camera. We use the Smooth function combined with L1 and L2 norm to optimize the region proposal network (RPN) loss function and reduce the deviation between the prediction frame and the actual target to ensure the accurate location of the ship target. With the Two-Way sampling combined with the shared weight to generate the mask, we solve the problems of inaccurate segmentation, target loss and small interference when Mask Region Convolution Neural Network (RCNN) is used to segment an instance. We create the experimental data sets from the deep learning annotation tool—Labelme. Experiments show that the improved Mask-RCNN model has a confidence rate of 82.17%. Serving as the basic network, the test accuracy rate of ResNetXt-101 is 3.3% higher than that of the original ResNet-101, which can better realize the function of ship target location and mask generation.

© 2019 The Authors. Published by Atlantis Press SARL.

This is an open access article distributed under the CC BY-NC 4.0 license (<http://creativecommons.org/licenses/by-nc/4.0/>).

## 1. INTRODUCTION

With the development of social economy, it is an important task for every coastal country to monitor and identify the Marine vessels, for it plays an important role in protecting Marine resources, monitoring and managing fishing boats, fighting against smuggling and so on. Exact detection and identification of ships can help to better monitor the water transport. An optimized ship identification system can effectively improve the shipping environment, maintain the maritime traffic order, fully enhance the operational capacity of the port and reduce the traffic accidents. In addition, ship identification technology has military strategic significance for accurately tracking the ship targets and realizing the precision guidance [1].

Until recent years, ship identification technology has not yet got popularized. The common method involves preprocessing, feature extraction and classifier. The main defect of this method is that it needs complex image processing algorithm but cannot get a satisfactory experimental result. In addition, convolutional neural network (CNN) is also well applied to the ship identification. However, the latest data shows that the identification accuracy is not ideal [2], hence it is extremely urgent to improve the identification accuracy. Ship identification is also one of the research hotspots in the field of pattern recognition. In 2014, Jiang Shao Feng *et al.* proposed Synthetic Aperture Radar (SAR) commercial ship classification algorithm based on structural characteristics [3], which can

classify bulk cargo ships, container ships and fishing boats. With the rise of method of neural network, Liang Jin Xiong used BP neural network to identify the infrared images of aircraft carriers, destroyers, frigates, passenger ships, container ships and oil ships in 2015 [4]. In 2016, Zhao Liang with his research team designed CNN for ship image classification, and acquired a better classification effect [5].

With the breakthrough of Deep CNN in image classification task, Jia Yangqing, as well as other scholars and researchers, has generated strong interest and devotion in it. Studies have shown that Deep CNN model which has been well systematically trained with ImageNet is used to extract deep convolution activation feature (DeCAF) [6]. When combined with different classifiers, it is able to obtain great results in identifying various scenes and birds. The experiments show that the sixth layer, the fully connected layer (DeCAF6) based on AlexNet [7] network, is superior to other layers in both bird recognition and target recognition tasks. For this reason, AlexNet network is adopted and DeCAF is introduced to conduct network training for the data set of transport ships with complex background. Then, DeCAF6 features of the network are extracted. Combining with SVM classifier, this paper respectively classifies and identifies three types of transport ships with high similarity, namely container ships, bulk carriers and oil and gas ships. The CNN architecture proposed by Krizhevsky *et al.* [8] has made a great breakthrough in Image-net Large Scale Visual Recognition Challenge (ILSVRC2012) (In the image classification task, the performance is 10% higher than that of the traditional feature architecture). Subsequently, Razavian *et al.* [9] pointed out

\*Corresponding author. Email: [linshaodan66@qq.com](mailto:linshaodan66@qq.com)

that the CNN model obtained through pre-training on the large-scale image library could be used to extract the visual features of images and applied to various image recognition tasks, such as scene classification and multi-tag classification. Ma Xiaofeng *et al.* [10] proposed an attention-based mechanism to extract environmental features by adding a fully convolutional network (FCN) branch to the detection network. Yu Zhang *et al.* [11] proposed a method to adjust the parameters such as the size of region proposal network (RPN) anchor and threshold value of non-maximum suppression (NMS), so as to improve the certainty of image mask extraction in multi-target scenes. However, when these models are applied to ship image recognition in real complex situations, there are still some problems in precision, quality and accuracy, where there is a gap from the idealized requirements of actual application. In the field of pattern recognition, ship image recognition has been one of the main points researchers focused on. By means of Segmentation and Artificial Neural Network, G.K. Yuksel [12] provided a way out to extract ship characteristics and identify ships from three-dimensional (3D) ship silhouette image. K. Rainey *et al.* [13] put forward the CNN for ship identification, and Corbane *et al.* [14] detected small ship targets by wavelet transform and radon transform of satellite images. The experiments on satellite ship images show that the recognition results are good.

Although these existing studies have acquired good recognition effect, they mainly focus on the ship images with large shape differences in a single background. Because of the complex backgrounds, small differences in hull shapes and the shooting angles, a large number of images of transport ships taken in real ports and waterways are of high similarity, making the traditional method of classification and identification for ship images unable to gain good results. How to improve the effect of this kind of ship image recognition is a difficult problem in the research of ship image recognition. In order to make up for the above shortcomings and to solve the related problems, this paper proposes a vessel target positioning and mask generation algorithm based on Mask RCNN [15], which has strong detection performance for Marine vessels and greatly enhances the accuracy and practicability of the algorithm. It has important theoretical and applied value to people's livelihood, economic construction, fishery administration supervision, maritime traffic management and even military field.

In the detection of offshore vessels, the experimental results show that the improved Mask RCNN model achieves better performance and higher precision than the existing deep learning vessel detection technology in a relatively short time.

## 2. RELATED WORK

The main idea of Mask RCNN is to locate multiple feature regions in an image, input each region into CNN for feature extraction [16] and generate a Mask on the feature-extracted region. The biggest feature of Mask RCNN is to separately extract the classified

regression information of the image to be tested (that is, the border information of the target to be tested) and combine it with the pyramid features (the length and width of the image input for pyramid region of interest [ROI] processing, that is, PyramidROIAlign) for mask generation, which is shown in Figure 1.

In the ship target positioning Mask RCNN architecture, ResNet works as the backbone network for feature mapping and FCN works as the backbone network for Mask prediction. According to the principle of network effect, "the deeper, the better," ResNetXt-101 is adopted in this paper as the basis of backbone network, and ROI Align layer is employed to align the extracted features with the pre-input [17]. ROI Align layer can accurately locate the ship targets, which is very important in the Mask RCNN system. In this paper, CNN is connected to the back of the ROI Align layer to extract the input features for classification and mask generation. Ship classification gets the alignment feature map of  $7 \times 7$  from the ROI Align layer, and ship mask generation gets the alignment feature map of  $14 \times 14$  from the ROI Align layer. In addition, CNN is composed of four convolution layers and two fully connected layers, each convolution layer of which is followed by Relu activation and maximum pooling sampling. Finally, the loss function output of Mask RCNN is as follows:

$$L_{Mask-RCnn} = L_{cls} + L_{box} + L_{mask} \quad (1)$$

In Formula (1),  $L_{cls}$ ,  $L_{box}$  and  $L_{mask}$  refer to classified loss function, prediction box loss function and mask loss function respectively.

## 3. REGION PROPOSAL NETWORK (RPN)

In fact, RPN is a core network of Mask RCNN (as shown in Figure 2), which is an important leading step to realize FPN feature layer selection and ROI Align. RPN is the FCN, which can conduct end-to-end training for the mission to generate testing suggestion box. Therefore, it can predict the boundary of the target and the score at the same time, as long as two extra full convolution layers (Cls-Layer and Reg-Layer) are added. Reg-Layer is used to predict the coordinates  $x$ ,  $y$ , width  $w$  and height  $h$  of the feature region corresponding to the center anchor, while Cls-Layer is used to determine whether the feature region is the foreground or the background. The sliding window processing method can ensure that Reg-Layer and Cls-Layer are associated with all feature spaces.

When inputting a convolution layer of  $3 \times 3 \times 512 \times 512$  to the RPN network, we will get an output of  $11 \times 11 \times 512$ , where  $11 \times 11$  represents the size of feature region. It is assumed that there are several  $11 \times 11$  big black boxes in the original image. Each of them has nine little color boxes and  $3 \times 3 = 9$  anchors. The anchors are located in the center of the  $3 \times 3$  sliding window. We can predict multiple feature regions through one sliding window. If there are  $k$  feature regions (that is,  $k$  reference frames), then each reference frame can be confirmed by the size, ratio and uniqueness of the anchor points in the

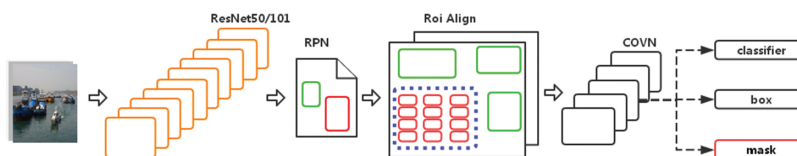


Figure 1 | Mask RCNN model.

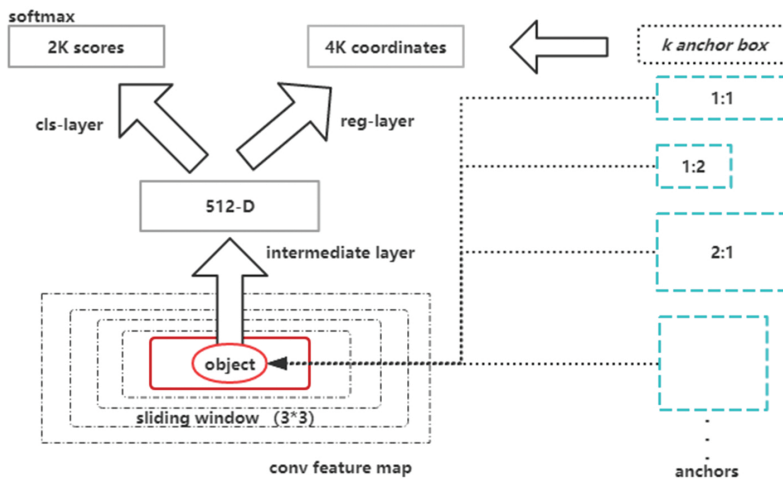


Figure 2 | Region proposal network (RPN) structure.

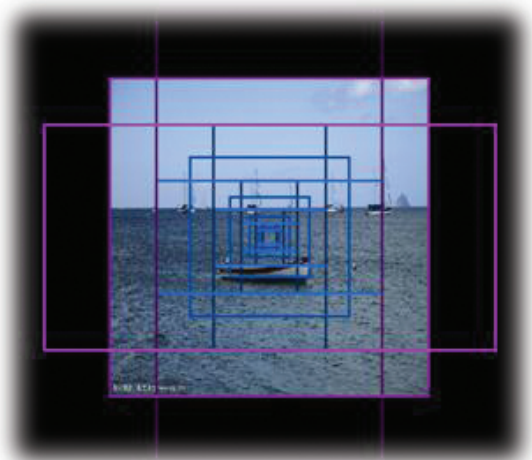


Figure 3 | Generation of nine anchor boxes.

sliding window. An anchor can refer to an anchor box or a reference frame. If  $k = 9$ , it indicates three dimensions ( $128^2$ ,  $256^2$ ,  $512^2$ ) and three ratios (1:1, 1:2, 2:1) [18]. When nine reference frames corresponding to the current sliding window position are identified, the  $W \times H$  feature map and its corresponding  $W \times H \times k$  anchors will bear the scale invariance (as shown in Figure 3), which can produce the boxes with a certain ratio of  $11 \times 11 \times 9$  at different positions. This is enough to pinpoint all target objects. If it is assumed that there is a target object in the original image, it can be located in the frame simply by translating the black box.

#### 4. LOSS FUNCTION

In this paper, the calculation rules of Loss function are set as follows:

1. The positive and negative sample calibration rules of anchors must be set before calculating the loss. Assuming that the intersection over union (IOU) between the reference box and ground truth (GT) of Anchor was greater than 0.7, it is marked as a positive sample; and when the overlap (IOU) between the reference box and GT of Anchor is less than 0.3, it will be marked as a negative sample. As for the remaining samples that

are neither positive nor negative, they will not participate in the final training [19].

2. The loss function of training RPN includes regression calculation loss and classification loss (that is, Softmax loss).

In this paper, the gradient method combining norm  $L_1$  with norm  $L_2$  is used to solve the problem that the gradient value will not be too large when the difference between the prediction box and the GT is too large and can be small enough when the difference is very small. When the input  $x$  is less than 1, norm  $L_1$ . (Formula 2) is used to make the gradient smoother and easier to differentiate. When the input  $x$  is greater than 1, norm  $L_2$  (Formula 3) is used to effectively avoid the gradient explosion and reduce the outliers.

$$smooth_{L_1}(x) = 0.5 * x^2 \{if |x| < 1\} \tag{2}$$

$$smooth_{L_2}(x) = |x| - 0.5 \{if |x| > 1\} \tag{3}$$

The detection block location of loss function is set as  $L_{loc}$ , the predicted translation and scaling parameter is set as  $t^u$ , and the actual translation and scaling parameter is set as  $v$ . Thus the loss rate of detection block location is evaluated by comparing the difference between the two parameters, which is as shown in Formula (4).

$$L_{loc}(t^u, v) = \sum_{i \in \{x, y, w, h\}} smooth_{L_{(1,2)}}(t_i^u - v_i) \tag{4}$$

The loss rate of RPN is obtained by Formula (2), Formula (3) and Formula (4). In Formula (5),  $p_i$  refers to the probability that the  $i$ -th anchor is a target object. When the target is found,  $p_i$  is 1; otherwise,  $p_i$  is 0.  $t_i$  is the four coordinate parameters of the boundary box  $(t_x, t_y, t_w, t_h)$  when predicting RPN, which corresponds to the coordinates of positive anchor point GT.  $L_{cls}$  is the loss function of dichotomy (target/non-target), and  $L_{reg}$  is the loss function of  $smooth_{L_{(1,2)}}$ , which is used to predict boundary box and GT box. The normalized total loss rate of RPN is obtained by the combination of  $N_{cls}$ ,  $N_{reg}$  and  $\lambda$  (balance weight). Through trial and error, we find that setting  $\lambda$  to 10,  $N_{cls}$  to 512 and  $N_{reg}$  to 2500 can keep the balance of loss values of the two parts of RPN. Finally, a series of feature regions are acquired by RPN through training.

$$L(\{p_i\}, \{t_i\}) = \frac{1}{N_{cls}} \sum_i L_{cls}(p_i, p_i^*) + \lambda \frac{1}{N_{reg}} \sum_i p_i^* L_{reg}(t_i, t_i^*) \quad (5)$$

## 5. ROI ALIGN

The feature region obtained by RPN processing is processed by ROI Align, so that it can provide required ROI input for mask generation. ROI Align is the improvement of ROI Pooling of Fast RCNN. ROI Align cancels the quantization operation. For the pixel whose coordinates are floating point numbers in the quantization, bilinear interpolation is used to calculate the pixel value, so as to prevent the loss of some feature point information from later aggregation [20]. Through experiments, we find that ROI Align is obviously effective to enhance the dataset with a large number of close-up large objects. However, as for the the dataset with a larger number of long-range small targets, ROI Align does not work well. It is mainly because small targets are easier to be affected by non-linear problems (for example, for large targets, the deviation of 0.1 in pixel is insignificant, but for small targets, the impact of error is great). For the reasons above, the anchor point cannot locate the distant hull target in the generated top-level anchors, which is shown in Figure 4.

In order to solve the above problem of small targets location, we make RPN generate a total of 261888 anchor points, and set the scale of the feature image as (32, 64, 128, 256, 512) and the proportion as [0.5, 1, 2]. When ResNet generates five different-scale feature maps, the trunk network will be input into the RPN network and respectively generate ROI. Then, the RPN network generates several Anchor-Boxes through these five feature maps of different scales. The number of anchor boxes is shown in Table 1.



Figure 4 | The generated top-level anchors.

Table 1 | Number of anchor boxes generated by five different-scale feature maps.

Level 1	196608	[256, 256]
Level 2	49152	[128, 128]
Level 3	12288	[64, 64]
Level 4	3072	[32, 32]
Level 5	768	[16, 16]

In this paper, nearly 1000 ROIs are retained after the NMS operation (1000 is a variable parameter). Because of the difference in strides, ROI Align operation is performed on the stride corresponding to the Feature maps of five different scales [Level 1, Level 2, Level 3, Level 4, Level 5] respectively. The ROI generated after this operation is the combination of down-sampling and NMS, which solves the problem that small targets cannot be located in ROI Align. Through the above method, the generated ROI includes all target vessels in the figure (as shown in Figure 5), providing the accurate input source for the subsequent mask generation.

## 6. MASK GENERATION

The mask generation in the Mask RCNN is achieved by citing an additional network separately. This mask generation branch is a full-convolutional-neural networks, where the positive region selected by the ROI Align classifier is used as input to generate a mask of 14\*14 low resolution pixels. The size of the mask is such small that it helps to keep the light weight of mask generation branch network. In the training process, we reduce the size of actual mask to 14\*14 to calculate the loss function. In the inference process, we amplify the size of the predicted mask to the border of ROI, so as to get the final result of mask. The procedure of mask generation is shown in Figure 6.

Mask generation is mainly to add a series of convolution operations after ROI Align expands its output dimension, so as to predict the mask more accurately. The operating parameters are shown in Table 1. This process is conducted through the ROI feature map generated by many inputs of the same size in ROI Align. Each output channel has 256 fusion feature layers. The number of the outputs is 256 and the size is 14\*14. As shown in Figure 3, the output is generated by one-way mask through four levels. Each level involves a convolution layer, a normalized layer and an activation function, which is finally connected to a deconvolution layer to acquire the mask.

The mask generated in the original model cannot be dynamically given in accordance with the target window. In the process of distant hull detection, mask generation branches cannot ensure the



Figure 5 | The region of interest (ROI) generated.

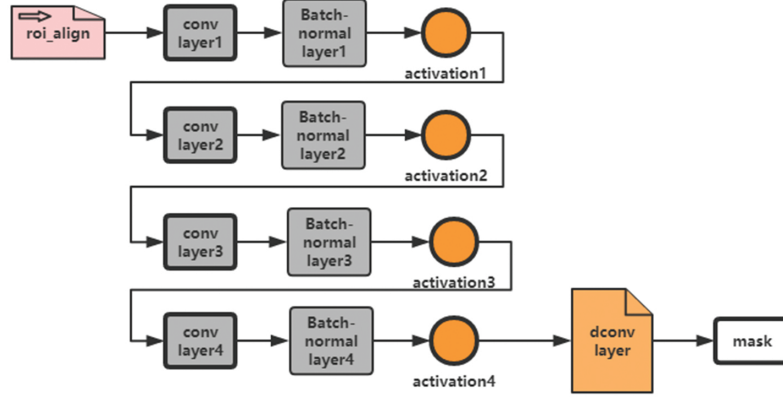


Figure 6 | Process of mask generation.

precise segmentation for instance, resulting in target loss or tiny interferences. To solve these problems, we propose an improved mask generation structure.

## 7. IMPROVED MASK STRUCTURE

In this paper, the mask branch is improved. After three times of convolution and aggregation, each ROI region  $x$  is shown as follows

$$\begin{aligned} x' &= \sum_i w_a^i x_i + b_{1j} \\ x'' &= \sum_i w_b^i x_i + b_{2j} \\ x''' &= \sum_i w_c^i x_i + b_{3j} \end{aligned} \quad (6)$$

Among them,  $w_a^i, w_b^i, w_c^i$  are the weight vectors, and  $b_{1j}, b_{2j}, b_{3j}$  are the offset vectors. In this paper, a random gradient descent algorithm is used to minimize the expected risk. The cost function is as follows.

$$J(w) = \frac{1}{2} (h_w(x^{(i)}) - y^{(i)})^2 \quad (7)$$

Therefore, the rule for updating the parameters of Random gradient descent (SGD) at this time is  $w_j = w_j + \alpha (y^i - h_w(x^i)) x_j^i$ , and the output results are upsampled and restored to get a new  $x', x''$ , that is,

$$x' = \text{upsampling}(x'), x'' = \text{upsampling}(x'') \quad (8)$$

It is the upsampling function. After another convolution and pooling process, the output result  $f_1(x), f_2(x)$  is obtained by activating the linear rectifier (relu) function, which is shown as follows.

$$\begin{aligned} f_1(x) &= \text{relu} \left( \sum_i w_a^i x_i' + b_j \right) \\ f_2(x) &= \text{relu} \left( \sum_i w_c^i x_i' + b_j \right) \end{aligned} \quad (9)$$

Among them,  $w_c^i$  is the weight of the third convolution,  $b_j$  is the bias quantity, and  $\text{relu}$  is the activation function. Finally, in order to obtain the segmentation result, the  $f_1(x)$  and  $f_2(x)$  are merged and multiplied by the shared weight  $\lambda$  obtained by the Time Distribution layer operation.

$$f(x) = \text{mask} (f_1(1) + f_2(2)) \cdot \lambda \quad (10)$$

The structure is shown in Figure 7.

The improved mask generation process mainly better the mask segmentation effect of each ROI. The pooling layer 1 and 3 are connected, and the pooling layer 3, connected to the pooling layer 2, is the convolution of the pooling layer 1. The pooling layer 1 and 3 are connected to the upper sampling layer respectively, so as to constitute a Two-Way sampling. Finally, the two channels are activated and merged into one path and multiplied by a shared weight to generate a mask. With the use of two-way sampling, the feature information of the image increases while the dimension of image description does not increase on the premise that the number of channels remains unchanged. Increasing the amount of information per dimension only is obviously beneficial to the final mask generation.

### 7.1. Shared Weight Generation

Through the Time Distribution layer, the multi-layer convolution features of the RPN network output can be shared. Therefore, the real significance of Time Distribution is to share the weight among the feature maps of different levels. The Time Distribution layer applies mask generation branch to a specific vector and performs an operation on each vector without increasing the final output vector. The Time Distribution layer performs upsampling, convolution, normalization, pooling and other operations on each time slice. Through a series of operations, we get a result, that is, the shared weight. Shared weight generation is shown in Figure 8.

In the process of mask generation improvement, through the operations of Time Distribution\_Layer 1, Time Distribution\_Layer 2 and Time Distribution\_Layer 3, different time slices are defined as  $f_1(x), f_2(x), f_3(x)$  to generate the output weight  $P_1, P_2, P_3$  respectively, and the output weight value is twice as much as its corresponding input weight, which is shown in Formula (11).

$$\begin{aligned} P_1 &= \sum_i 2w_a^i x_i + b_{1j} \\ P_2 &= \sum_i 2w_b^i x_i + b_{2j} \\ P_3 &= \sum_i 2w_c^i x_i + b_{3j} \end{aligned} \quad (11)$$

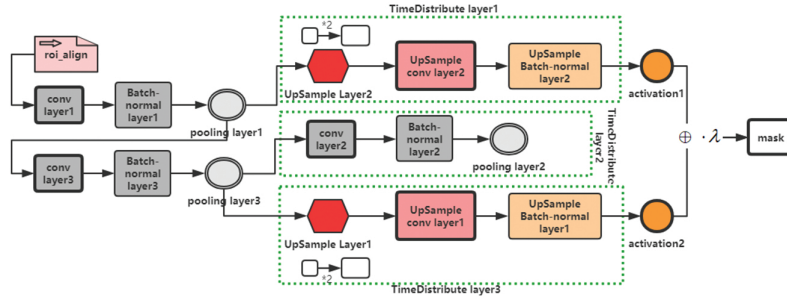


Figure 7 | Process of the improved mask generation.

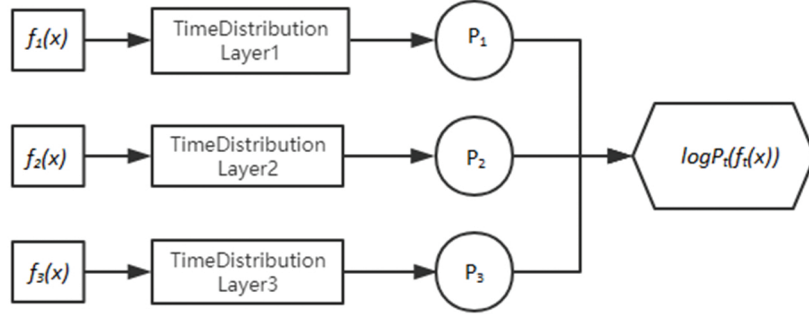


Figure 8 | Shared weight generation.

The Shared Weight  $\lambda$  required in Formula (11) is expressed as  $\log P_t(f_i(x))$ , as shown in the Formula (12)

$$\log P_t(f_i(x)) = \sum_{i=1}^N \log P_t(f_i(x) | f_1(x), f_2(x), f_3(x), \dots, f_n(x)) \quad (12)$$

## 8. EXPERIMENT

### 8.1. Dataset Preparation

The data set used in this experiment is the ship pictures crawled from the network by crawler program. Due to the image size requirement of Mask RCNN, the size of the data set image is 512\*512. LabelMe is used to mark the area where the ship target is located in the image. LabelMe is an open annotation tool created by the Computer Science and Artificial Intelligence Laboratory of the Massachusetts Institute of Technology (MIT CSAIL) [21]. Through LabelMe software, it is convenient to mark and preserve the boundary of all kinds of ship targets in the image.

LabelMe generates label images, mask images and logo images (as shown in Figure 9). We have produced the Mask RCNN standard dataset 3000, which includes three kinds of ocean ship images: ocean-going cargo ship, inshore fishing boat and large warship.

### 8.2. Training

In this paper, the time of training cycle is set as 5, each cycle iterates 100 steps, and the number of training steps usually depends on the size of the dataset. It takes a big amount of memory resources and time for large-scale data set training to use the above model;

therefore, it needs a high-end hardware. The hardware parameters involved in this paper are CPU: Intel (R) Core (TM) i7-6820HQ 8 core, Memory: 32G, and GPU: Nvidia Quadro M2000M; while the software environment is built with Python3, Keras2, TensorFlow1.4. Comparing with the previous Mask RCNN architecture, the improved Mask RCNN network training can optimize the training time, loss rate and so on, which is shown as follows. In this paper, four methods are used for training:

1. RPN<sub>(L1)</sub> with L1 Norm and pre-improved mask (Mask<sub>(O)</sub>) Generation method;
2. RPN with L1 Norm and improved Mask<sub>(N)</sub> Generation method;
3. RPN<sub>(L1, L2)</sub> with L1 and L2 Norm and pre-improved mask (Mask<sub>(O)</sub>) Generation method;
4. RPN<sub>(L1, L2)</sub> with L1 and L2 norm and improved mask (Mask<sub>(N)</sub>) generation method.

Using the above four methods, the gradient loss rates of RPN and Mask are shown in Figure 10, which shows that the gradient trend of RPN (smooth) is basically consistent with that of RPN (loss), Mask (smooth) and Mask (loss).

Table 2 shows the comparison of training parameters for the above four methods.

When the L1 and L2 norms are combined with the improved mask generation method, the accuracy precision (AP) value is the highest. The training time increases with the increase of the computational complexity of the model. In addition, in the training, we substitute the backbone network Resnet-101 with VGG19, Resnet-50, ResNeXt-50 and ResNeXt-101, each of which is combined with



Figure 9 | Ship target marking.

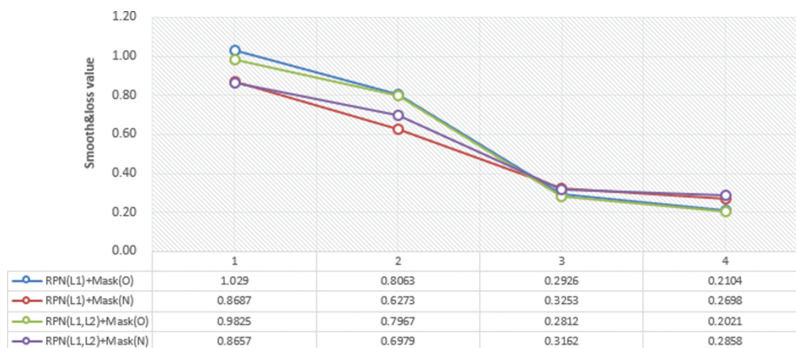


Figure 10 | Comparison of gradient loss rate generated by region proposal network (RPN) and mask.

Table 2 | Comparison of training parameters of mask RCNN model.

Method	RPN-Loss	Mask-Loss	AP	Elapsed Time
RPN (L1) + Mask (O)	Smooth = 1.029 value = 0.8063	Smooth = 0.2926 Value = 0.2104	80.3%	37 m 45 s
RPN (L1) + Mask (N)	Smooth = 0.8687 value = 0.6273	Smooth = 0.3253 Value = 0.2698	80.2%	39 m 59 s
RPN (L1, L2) + Mask (O)	Smooth = 0.9825 value = 0.7967	Smooth = 0.2812 Value = 0.2021	81.1%	37 m 56 s
RPN (L1, L2) + Mask (N)	Smooth = 0.8657 value = 0.6979	Smooth = 0.3162 Value = 0.2858	82.1%	41 m 03 s

RPN, region proposal network.

the improved Mask generation method respectively [22]. The comparison of AP and the training duration is shown in Table 3.

### 8.3. Test

With 1000 ship images as the test set, the test results of the improved method are compared with those of other methods. Through several groups of control experiments, the test model obtains the precision in accordance with Formula (13) and the recall in accordance with Formula (14). This index is used to measure the effect of the model on ship positioning.

$$precision = \frac{TP}{TP + FP} \tag{13}$$

$$recall = \frac{TP}{TP + FN} \tag{14}$$

Table 3 | Comparison of training parameters for different basic backbone networks.

Backbone Networks	Method	AP (%)	Elapsed Time
Resnet-101	RPN (L1) + Mask (N)	80.2	39 m 59 s
	RPN (L1, L2) + Mask (N)	82.1	41 m 03 s
Resnet-50	RPN (L1) + Mask (N)	75.3	35 m 43 s
	RPN (L1, L2) + Mask (N)	77.7	35 m 40 s
VGG19	RPN (L1) + Mask (N)	73.2	33 m 21 s
	RPN (L1, L2) + Mask (N)	73.4	33 m 21 s
ResNeXt-50	RPN (L1) + Mask (N)	77.9	36 m 52 s
	RPN (L1, L2) + Mask (N)	78.4	37 m 19 s
ResNeXt-101	RPN (L1) + Mask (N)	83.7	43 m 05 s
	RPN (L1, L2) + Mask (N)	84.6	44 m 12 s

AP, accuracy precision; RPN, region proposal network.

Among them, TP represents the expected positive samples with the actual identification in line with the expectation, while FP represents the expected positive samples that the actual identification

shows negative samples [23]. It is assumed that TP is the number of images that have accurately located the ship target, FP is the number of images that do not locate the ship target or just locate part of the ship targets and FN does not locate the ship target at all. We usually use the number of images whose test accuracy is lower than a certain low threshold value. The test results are shown in Table 4:

When the threshold value is 0.80, the number of images whose recognition rate is greater than or equal to 0.80 is TP value, the number of images whose recognition rate is less than 0.80 but greater than or equal to 0.70 is FP value, and the number of images whose recognition rate is less than 0.70 is FN value.

When the threshold value is 0.75, the number of images whose recognition rate is greater than or equal to 0.75 is TP value, the number of images whose recognition rate is less than 0.75 but greater than or equal to 0.70 is FP value, and the number of images whose recognition rate is less than 0.70 is FN value.

It shows that the test precision and recall rate of the improved method are higher than those of the original method.

In this paper, Resnet-101, Resnet-50, VGG19 and ResNeXt-50, ResNeXt-101 are used to test the un-improved and improved mask generation models. Figure 11 shows the comparison of test precision of different models.

As you can see from Figure 11, the model testing precision of ResNetXt-101 + RPN (L2) + Mask (N) is as high as 93.60%, which is 3.3% higher than that of ResNet-101 and 33.1% higher than that of VGG19. The data shows that the improvement of the basic backbone network is also a method to improve the model testing precision.

**Table 4** | Test results.

Method	Threshold Value	Precision (%)	Recall (%)
RPN (L1, L2) + Mask (N)	0.80	87	96
RPN (L1, L2) + Mask (N)	0.75	90	96
RPN (L1, L2) + Mask (O)	0.80	85	93
RPN (L1, L2) + Mask (O)	0.75	87	93

RPN, region proposal network.

## 9. THE SOLUTION OF SHIP TARGET LOCATION

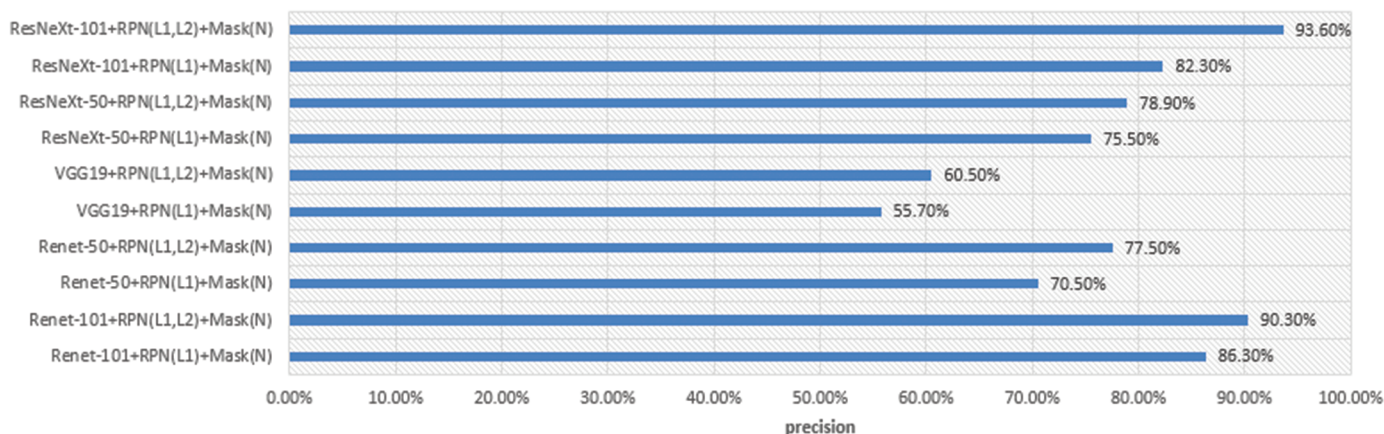
This paper defines the solution of ship target location through six processes:

(1) Image Input; (2) Anchor-Box Setting; (3) Top Anchor-Box Extraction; (4) Anchor-Box Selection and Optimization; (5) ROI Extraction; (6) Target Mask Generation (as shown in Figure 12). In this solution, three different ship types are tested, that is, ocean-going freighters, inshore fishing boats and large warships respectively. The test images are divided into three categories: close range, medium vision and prospect, as shown in Figure 12. Figure 12(a) displays the original image input with the size of 512\*512, Figure 12(b) displays the nine anchor boxes of different colors from dark to light, Figure 12(c) displays the anchor box at the top of the extraction, and Figure 12(d) displays the remaining anchor box after cutting the edges of the image. Figure 12(e) displays the extracted ROI, and Figure 12(f) displays the mask generation for the target image. The testing result shows the processing effect of setting up RPN-Box, extracting ROI and generating mask in the un-improved and improved Mask RCNN network.

## 10. COMPARISON WITH OTHER METHODS

In this section, several models are compared and tested from the perspective of  $F_1$ -score and run time.  $F_1$ -score is used as matrix for the network performance evaluation. We conduct the experiments to compare RCNN, Faster RCNN and Mask RCNN with the improved Mask RCNN. The results are shown in Table 5.

As shown in Table 5, the improved model has better  $F_1$ -score performance. What's more, the performance of this model is 5.3% better than that of the Faster RCNN. This table also presents the running time of different models. As shown in Table 5, the running time of the proposed Mask RCNN is about 1 minute to 2 minutes slower than that of the existing technology. This shows that it will not cost too much to modify the Mask RCNN.



**Figure 11** | Comparison of test precision of different models.



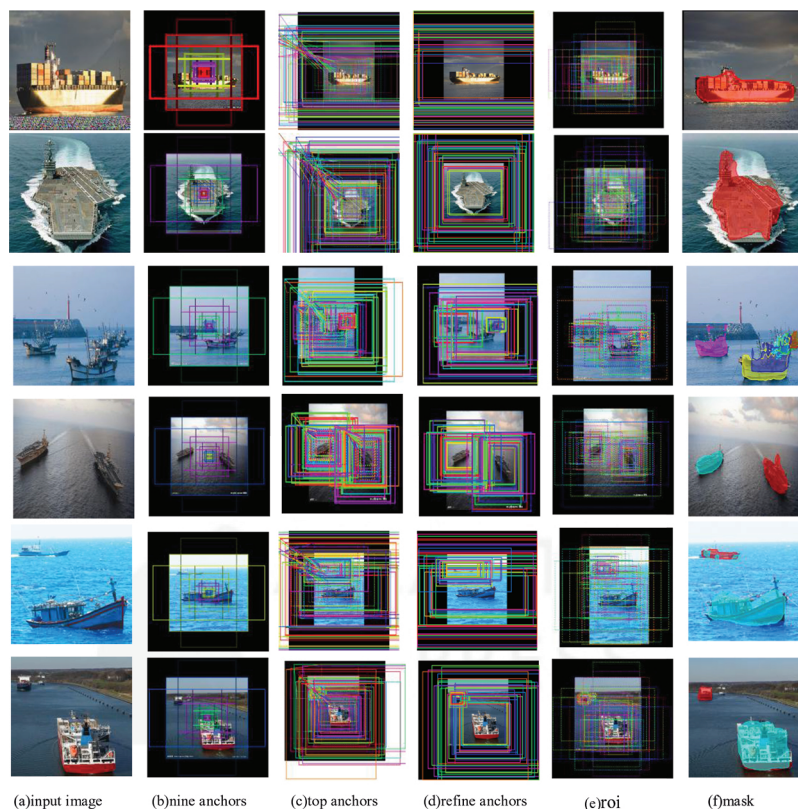


Figure 12 Solution of ship target location.

Table 5 Comparison with other methods.

Method	F <sub>1</sub> -Score	Running Time
Improved Mask RCNN	0.926	44 m 12 s
Faster RCNN	0.873	43 m 10 s
RCNN	0.835	40 m 08 s
CNN	0.751	38 m 19 s

CNN, convolutional neural network.

## 11. CONCLUSION

Aiming at detecting the target of offshore ships, this paper improves the RPN loss function of Mask RCNN and the target mask generation algorithm. In the process of distant hull detection, mask generation branches cannot ensure the precise segmentation for the instance, resulting in target loss or tiny interferences. We proposed an improved mask generation structure to solve these problems.

ResNet-101, ResNet-50, VGG19 and ResNeXt-50 and ResNeXt-101 are used to verify the precision and efficiency of ship target recognition. The test results show that the improved Mask RCNN model can detect and locate the target accurately, and the proposed Mask RCNN is more accurate than the existing ones. Although it will not cost a lot to modify the Mask RCNN, it still has the disadvantage that the mask generation cannot completely cover the edge of the target. In the following work, we will use the edge filtering algorithm combined with a larger sample base to further improve the precision of the model.

## AUTHORS' CONTRIBUTIONS

Lin Shaodan contributed the central idea, analysed most of the data, and wrote the initial draft of the paper. The remaining authors contributed to refining the ideas, carrying out additional analyses and finalizing this paper.

## ACKNOWLEDGMENTS

The authors would like to thank all the participants taken part in the experiments. This work was supported in part by the National Science Foundation of China (Grant No. 61841701) and the Guangdong Natural Science Foundation (Grant No. 2019B010137002) and Fujian Vocational College Intelligent Equipment Application Technology Collaborative Innovation Center Construction Project (Grant No. 2016-7) and the Science and Technology Project from Transportation Department of Fujian Province (Grant No. 201934).

## REFERENCES

- [1] C. Lian, S. Qiang, D. Liang, Z. Yan-Jun, G. Sone, Comparative analysis of the development of research vessel at home and abroad, *Ship Sci. Technol.* 36 (2014), 2–7.
- [2] M.H. Zwemer, R.G.J. Wijnhoven, P.H.N. de With, Ship detection in harbour surveillance based on large-scale data and CNNs, in *International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISIGRAPP), Funchal, 2018, vol. 5, pp. 153–160.*

- [3] J.S. Feng, W. Chao, W. Fan, Z. Bo, Algorithm for merchant ship classification in COSMO-SkyMed images based on structural feature analysis, *Remote Sens. Technol. Appl.* 29 (2014), 607–615.
- [4] L.J. Xiong, W.K. Qi, Ship recognition based on BP network, *Ship Sci. Technol.* 37 (2015), 206–209.
- [5] Z. Liang, W.X. Feng, Y. YI-TAO, Research on ship recognition method based on deep convolutional neural network, *Ship Sci. Technol.* 38 (2016), 119–123.
- [6] J. Donahue, Y. Jia, O. Vinyals, *et al.*, DeCAF: a deep convolutional activation feature for generic visual recognition, *Comput. Sci.* 50 (2013), 815–830.
- [7] A. Krizhevsky, I. Sutskever, G.E. Hinton, ImageNet classification with deep convolutional neural networks, in *International Conference on Neural Information Processing Systems*, Curran Associates Inc., Lake Tahoe, 2012, pp. 1097–1105.
- [8] O. Russakovsky, J. Deng, H. Su, *et al.*, ImageNet large scale visual recognition challenge, *Int. J. Comput. Vision.* 115 (2015), 211–252.
- [9] A.S. Razavian, H. Azizpour, J. Sullivan, *et al.*, CNN features off-the-shelf: an astounding baseline for recognition, in *IEEE Conference on Computer Vision and Pattern Recognition Workshops*, IEEE Computer Society, Columbus, 2014, pp. 512–519.
- [10] M. Xiaofeng, W. Li, Z. Shi, Attention-based convolutional networks for ship detection in high-resolution remote sensing images, in *Chinese Conference on Pattern Recognition and Computer Vision (PRCV)*, Springer, Cham, 2018.
- [11] Y. Zhang, Y. Zhang, S.-X. Li, J.-H. Zhang, Accurate detection of berthing ship target based on mask R-CNN, in *Proceeding of SPIE 10836, 2018 International Conference on Image and Video Processing, and Artificial Intelligence*, Shanghai, 2018.
- [12] G.K. Yüksel, B. Yalituna, Ö.F. Tartar, *et al.*, Ship recognition and classification using silhouettes extracted from optical images, in *24th Signal Processing and Communication Application Conference (SIU)*, IEEE, Zonguldak, 2016, pp. 1617–1620.
- [13] K. Rainey, J.D. Reeder, A.G. Corelli, Convolution neural networks for ship type recognition, in *SPIE Defense + Security*, Baltimore, 2016, vol. 9844.
- [14] C. Corbane, M. Petit, Fully automated procedure for ship detection using optical satellite imagery, in *Proceedings of SPIE - The International Society for Optical Engineering*, Noumea, 2008, vol. 7150.
- [15] K. He, G. Gkioxari, P. Dollar, *et al.*, Mask R-CNN, *IEEE Trans. Pattern Anal. Mach. Intell.* 99 (2017), 1.
- [16] D. Liang, K. Park, P. Krompiec, Facial feature model for a portrait video stylization, *Symmetry.* 10 (2018), 442.
- [17] D. Chen, S. Zhang, W. Ouyang, J. Yang, Y. Tai, Person search via a mask-guided two-stream CNN model, in *European Conference on Computer Vision (ECCV)*, Munich, 2018, pp. 764–781.
- [18] Y. Zhong, J. Wang, J. Peng, L. Zhang, Anchor box optimization for object detection, *CoRR abs/1812.00469*, 2018.
- [19] Z. Lin, K. Ji, X. Leng, G. Kuang, Squeeze and excitation rank faster R-CNN for ship detection in SAR images, *IEEE Geosci. Remote Sens. Lett.* 16 (2019), 751–755.
- [20] D. Schweitzer, R. Agrawal, Multi-class object detection from aerial images using mask R-CNN, in *IEEE International Conference on BigData*, Seattle, 2018, pp. 3470–3477.
- [21] B.C. Russell, A. Torralba, K.P. Murphy, W.T. Freeman, LabelMe: a database and web-based tool for image annotation, *Int. J. Comput. Vision.* 77 (2008), 157–173.
- [22] J. Lei, W. Shi, Z. Lei, F. Li, Efficient power component identification with long short-term memory and deep neural network, *EURASIP J. Image Video Process.* 122 (2018), 122–136.
- [23] R. Huang, P.K. Mungai, J. Ma, K.I-K. Wang, Associative memory and recall model with KID model for human activity recognition, *Future Generation Comp. Syst.* 92 (2019), 312–323.