IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received August 7, 2019, accepted August 31, 2019, date of publication September 9, 2019, date of current version September 20, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2940102

Nearshore Ship Detection on High-Resolution Remote Sensing Image via Scene-Mask R-CNN

YANAN YOU, (Member, IEEE), JINGYI CAO, YANKANG ZHANG, FANG LIU, AND WENLI ZHOU

Beijing Key Laboratory of Network System Architecture and Convergence, Beijing University of Posts and Telecommunications, Beijing 100876, China

Corresponding author: Jingyi Cao (caojingyibupt@gmail.com)

This work was supported by the Fundamental Research Funds for the Central Universities under Grant 2018RC09.

ABSTRACT Deep convolutional neural network (DCNN) can achieve ship detection mission on the high-resolution remote sensing images. However, the false alarms caused by the onshore ship-like objects may decrease the accuracy and feasibility of these DCNN-based detection frameworks. In our work, an end-to-end method, named as Scene Mask R-CNN, is proposed to reduce the onshore false alarms. The scene mask extraction network (SMEN), as a network branch for scene segmentation, is innovatively introduced into the detection framework. The non-target area is marked out by an inferenced scene mask which is used to assist the ship detection. Combining the feature map originated from feature extraction network (FEN) with the inferenced scene mask by using the edge probability weighted (EPW) merging method, the false candidate targets in the non-target area are excluded. This novel mechanism of DCNN-based ship detection not only maintains the detection accuracy, but also effectively suppresses the false alarms in the non-target area. Finally, the validity and accuracy of this method are verified on a ship dataset generated by the high-resolution optical remote sensing images.

INDEX TERMS Ship detection, false alarm suppression, scene mask, convolutional neural network.

I. INTRODUCTION

The mission of ship detection, for searching a target of interest and obtaining its spatial location automatically, is one of the important applications of high-resolution remote sensing (RS) images. In recent researches, the ship detection method based on deep learning provides a feasible solution with high efficiency and intelligence. This innovative framework combines the ship detection method with deep convolutional neural network (DCNN) [1], which conducts a new detection paradigm based on supervised learning. Mass target samples are used to train DCNN and make it have a strong generalization ability. Considering the synchronization of classification and regression processes, DCNN-based target detection has two implementation schemes. One is the two-stage method represented by Faster R-CNN [2], and the other is the one-stage method such as SSD [3] and YOLO [4].

Recently, the two-stage DCNN-based ship detection methods [5]–[8] are widely concerned. They generally outperform the one-stage ship detection methods [9] in detection accuracy and algorithm scalability. Li *et al.* [10] proposes hierarchical selective filtering (HSF) layers to map features with different scales into the same dimensional space, and these layers are applied on Faster R-CNN to detect multi-scale targets effectively. Zhang *et al.* [11] focuses on the ships with multi-direction and designs a rotated region proposal network (RPN), which obtains multi-oriented proposals with ship orientated angle and extracts the discriminative features from the inclined candidate regions to classify the target.

Although the above two-stage methods of DCNN-based ship detection have a good performance, the suppression of onshore false alarms is neglected for most nearshore detection tasks. More specifically, some onshore ship-like objects, such as dock, roof, and road, are interpreted as targets of interest with high probability, even if the trained network has a high detection accuracy.

Generally, as a network training trick, the way of additional negative samples may be adopted to achieve the feature-based suppression of onshore false alarms [12], [13]. Providing sufficient negative samples, the network deliberately learns these representative scenes which are difficult to determine, and some frequent objects which are likely to be misjudged as certain targets, which improves the robustness of the trained network. Wu *et al.* [14] detects the ship head and extracts the

The associate editor coordinating the review of this manuscript and approving it for publication was Muhammad Imran.

surroundings of ship head as region proposals, meanwhile utilizing plenty of hard negative samples, even the onshore false alarms. However, these methods rely on the completeness of negative samples heavily. Considering the coverage of the RS imageries and diversity of objects, it is difficult to gather all unpredictable negative samples for training.

Compared with the feature-based suppression, it is more pertinent to detect targets after excluding the non-target area by using scene information. Therefore, it is preferable to build a framework to achieve scene-level suppression, rather than enthusiastic about some training tricks. In fact, DCNN-based semantic segmentation methods, designed to extract the object mask, such as SNFCN and SDFCN [15], SegNet [16], Unet [17], and DeepUNet [18], are competent to distinguish the target area and the non-target area, and they have potential to facilitate the ship detection mission. Some researches [19]–[21] focus on the segmentation between ship and background by using DCNN-based semantic segmentation in detection tasks. However, few studies have addressed how to combine the detection with the segmentation task to form an end-to-end system, not to mention the research on suppressing onshore false alarm by using scene mask. Supposing that DCNN-based semantic segmentation is imported into the two-step detection process, separating the target and the non-target areas adaptively, quantities of false alarms in the non-target area can be reduced before the location determination of proposal targets.

In order to suppress the false alarms in the non-target area in the nearshore ship detection task, a novel method, named as Scene Mask R-CNN, is proposed in this paper. This end-to-end system contains four sub-networks with different functions. The feature map of the input image is obtained by the feature extraction network (FEN) first, and then the scene mask of target and non-target area is extracted by the scene mask extraction network (SMEN). With the feature combination between the output of FEN and the estimated scene mask, the false-alarm targets existing in non-target area are eliminated entirely. Then region proposal network (RPN) uses the combined feature map to generate the proposed bounding boxes, and these region proposals are imported into the classification and regression network (CRN) to obtain the final detection results. In theory, Scene Mask R-CNN is a feasible re-examination mechanism applied for target detection, and it provides satisfactory detection results meanwhile reducing false alarms in the non-target area.

To verify the feasibility and practicability of the network model,the Scene Mask R-CNN framework is constructed and its performance is evaluated through the relevant experiments. Comparing with the classical Faster R-CNN, our method reaches the accuracy of state-of-art baseline [2], and the onshore false alarms are obviously suppressed due to the restraint of the proposal bounding boxes in the non-target area. Therefore, the main contributions of this paper are summarized: (1) Aiming at the suppression of false alarm in the non-target area, a novel method for DCNN-based ship detection task, called Scene Mask R-CNN, is proposed to estimate the scene mask and detect the targets with an end-to-end process. (2) The scene segmentation and target detection are training synchronously, relying on a multi-task loss function. The target detection result and the segmentation result of the complex scene are produced simultaneously. (3) An edge probability weighted (EPW) merging method is used to optimize the inference process.

The rest of the article is organized as follows. Section II introduces the problem of onshore false alarms in ship detection task. In Section III, the principle of Scene Mask R-CNN is described in detail. In Section IV, some experiments are demonstrated between Scene Mask R-CNN and Faster R-CNN method with different training conditions. Besides, the time consumption and the selection of weighted parameters in loss function are discussed in this section. The relevant conclusions are found in the final part.

II. ONSHORE FALSE ALARMS IN DCNN-BASED SHIP DETECTION TASK

A. BASIC PRINCIPLE OF DCNN-BASED SHIP DETECTION

The two-stage DCNN-based ship detection methods mainly contain two network branches. One is called RPN which generates plenty of region proposals, and each proposal is trained to gain the positive sample and its position information. The other is a classification and regression network to distinguish and locate targets contained in the region proposals.

The RPN [2] is a target location extraction mechanism based on the fully convolutional network [22], it obtains plenty of region proposals from the input image firstly, and then selects the positive samples from the region proposals. Compared with the selective search method [23] used in previous DCNN-based target detection methods, the RPN utilizes convolutional layers to extract the region proposals instead of traversing the input with the exhaustive selective method, greatly reducing the computational overhead.

Specifically, the sliding window mechanism is used to generate target proposals centered at each sliding window, called anchors, containing the location and category information with predefined area size. In order to accommodate the multi-scale targets, the anchors can be set into different scales s with the pyramid structure, and each anchor is transformed through the setting of aspect ratios r (i.e., as the dash boxes shown in Fig. 1). Therefore, various anchor boxes of $W_R \cdot H_R \cdot r \cdot s$ are generated on the input feature map of RPN with the size of $W_R \times H_R$. These boxes represent the initialized region proposals of targets. However, the mechanism that anchors cover the entire input image makes any area to be the potential target area, even containing some non-target areas. Then, the positive and negative samples in the region proposals are confirmed and their position information is estimated through some convolutional layers. It inevitably imports the unnecessary samples in non-target areas, as the red boxes in the gray area shown in Fig. 1, maybe, causing false alarm in the detection result. Therefore, if a mechanism is designed to assist RPN to restrict region proposals only in

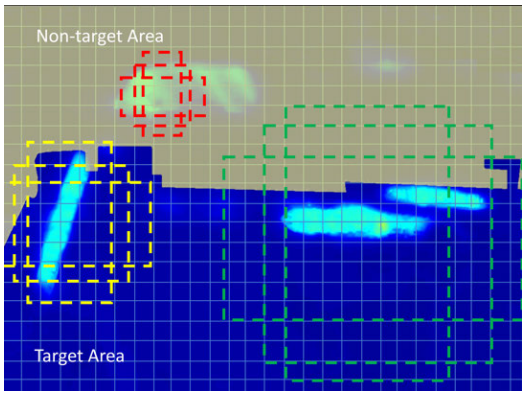


FIGURE 1. The RPN mechanism leads to false alarms in the non-target area. The red boxes represent the anchors that cover the region of false alarms in the non-target area, the other boxes represent the potential targets in the target area, and the blue area and yellow area represent the target area and the non-target area respectively.

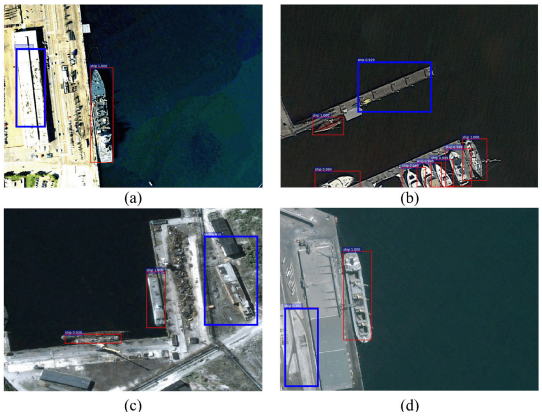


FIGURE 2. Examples of the onshore false alarms in the conventional DCNN method (faster R-CNN). The red boxes represent the correct detection results, and the blue boxes represent the false alarm results, owing to the shape similar to ships.

the target areas, the false alarms in the non-target area can be mostly suppressed.

B. ONSHORE FALSE ALARMS INDUCED BY RPN

With the drawback of RPN mechanism, the false alarm caused by artificial facilities and buildings onshore is an unavoidable problem. In reality, many coastal port areas, several artificial facilities, and buildings onshore have streamlined contours, which are very similar to the shape of ships.

As shown in Fig. 2, the red boxes represent the correct detection results, and the blue boxes represent the false alarm targets owing to the ship-like shapes. In fact, current DCNN-based methods are not competent enough for distinguishing these similar objects. However, the targets we focus on only exist in specific target areas (sea). Obviously, there is a great variation between the background area of ships (sea) and the non-target area (land) in the optical RS image.

Therefore, if the DCNN-based ship detection method can distinguish target area and non-target area, the feature originated from the non-target area will not be imported in the initialized region proposals, and it is possible to suppress the false alarms in the non-target area.

III. SCENE MASK R-CNN

The purpose of our work is to obtain the scene mask of input RS image during the ship detection task, and to reduce the onshore false alarms by the estimated mask, which improves the robustness of region proposals in the detection process. Therefore, the ship detection framework essentially contains two phases, the first phase is to obtain a scene mask via the scene mask extraction network (SMEN), and the second one is to detect the ships in the target area by using RPN, aided by the estimated scene mask.

A. MODEL

The network structure of Scene Mask R-CNN is shown in Fig. 3. It has four sub-networks: FEN, SMEN, RPN and CRN, respectively. The multi-level feature of the input imagery is extracted through FEN built by a series of ResNet blocks (i.e., as the blue dotted box shown in Fig. 3). The predicted mask is generated from the output feature map of FEN by SMEN with sequential convolutional layers and deconvolutional layers (i.e., as the red dotted box shown in Fig. 3). In addition, each feature map obtained by ResNet block connects with its corresponding feature map in SMEN by skip connection method. The scene mask loss is obtained by the predicted mask and its mask label. Simultaneously, the output of FEN is imported into RPN with convolutional layers and Softmax layer to obtain the location and classification of the region proposals (i.e., as the purple dotted box shown in Fig. 3). The region proposals with high scores are imported into CRN built by the ROI pooling layer, fully connection layers, and Softmax layer sequentially. The CRN obtains the target classification and refines target location (i.e., as the green dotted box shown in Fig. 3). The location loss and the classification loss are calculated by the target label and the predicted target. Finally, the total loss is the combination of scene mask loss, location loss, and classification loss.

1) FEATURE EXTRACTION NETWORK

Deeply extracting various features of RS images is crucial to the DCNN-based ship detection method. Multi-level feature maps about the targets and background are extracted through FEN. The residual network (ResNet) [24] establishes some shortcut connections by mapping low-level feature to high-level feature in each residual block. Compared to the classical networks, for example, the VGG net [25], the ResNet has fewer filters and lower complexity. Fewer parameters of ResNet are conducive to the training of DCNN, and it effectively improves the convergence speed.

As a matter of fact, the deeper network induces more abstract features, which benefits the extraction of the multi-level feature maps, especially for better implementing both semantic segmentation and target detection tasks. It is worthy to note that the 101-layer ResNet has more advantages in network depth and feature extraction ability than other structures

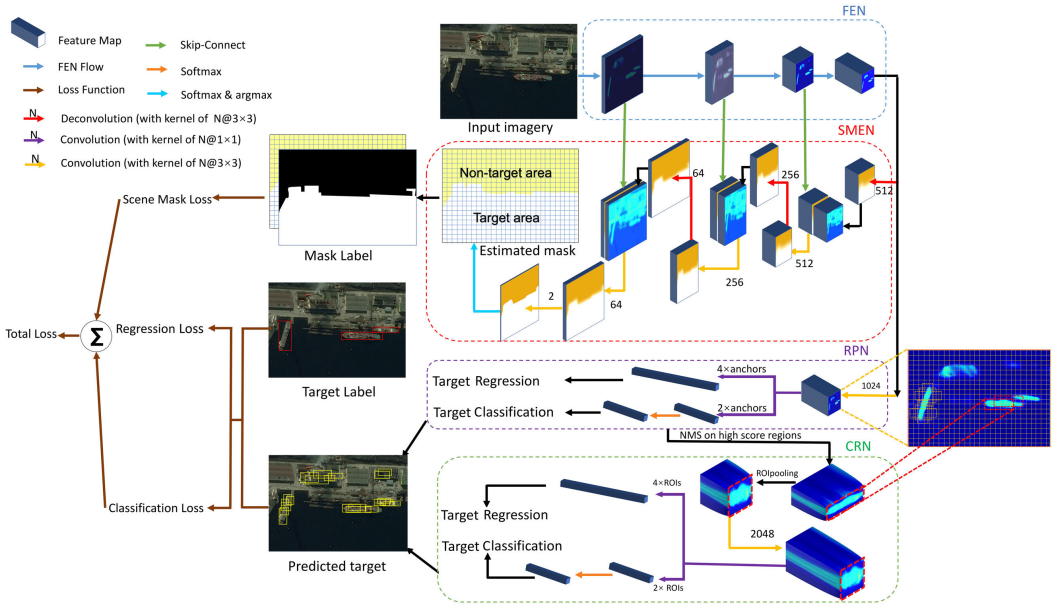


FIGURE 3. Framework of scene mask R-CNN. It contains four sub-networks: Feature Extraction Network (FEN) is used to extract the feature of input image; Scene Mask Extraction Network (SMEN) is used to estimate the scene mask; the location and classification of target are confirmed by Region Proposal Network (RPN) and Classification and Regression Network (CRN) respectively.

TABLE 1. Configuration of the feature extraction network.

Block	Layers	Parameter	Output size
ResNet Block1	Convolution	$7\times7, 64, \text{stride}=2$	$64@m\times n$
	Max_pooling	$3\times3, \text{stride}=2$	$64@m/2\times n/2$
ResNet Block2	Convolution	$3@\begin{bmatrix}1\times1,64\\3\times3,64\\1\times1,256\end{bmatrix}$	$256@m/4\times n/4$
ResNet Block3	Convolution	$4@\begin{bmatrix}1\times1,128\\3\times3,128\\1\times1,512\end{bmatrix}$	$512@m/8\times n/8$
ResNet Block4	Convolution	$23@\begin{bmatrix}1\times1,256\\3\times3,256\\1\times1,1024\end{bmatrix}$	$1024@m/16\times n/16$

TABLE 2. Configuration of the scene mask extraction network.

Layers	Parameter of kernels	Output Size
Deconv1	$512@3\times3\times1024, \text{stride}=2$	$512@m/8\times n/8$
Conv1	$512@3\times3\times1024, \text{stride}=1$	$512@m/8\times n/8$
Deconv2	$256@3\times3\times512, \text{stride}=2$	$256@m/4\times n/4$
Conv2	$256@3\times3\times512, \text{stride}=1$	$256@m/4\times n/4$
Deconv3	$128@3\times3\times256, \text{stride}=2$	$128@m/2\times n/2$
Conv3	$128@3\times3\times256, \text{stride}=1$	$128@m/2\times n/2$
Deconv4	$64@3\times3\times128, \text{stride}=2$	$64@m\times n$
Conv4	$64@3\times3\times128, \text{stride}=1$	$64@m\times n$
Conv5	$2@3\times3\times64, \text{stride}=1$	$2@m\times n$
Softmax & Argmax		$1@m\times n$

with shallower layers. Therefore, ResNet-101 is selected as the implementation of FEN, and the output of each ResNet block contained in the ResNet-101 is connected with SMEN. To this end, the configuration of FEN in our method is defined as collected in Table 1.

2) SCENE MASK EXTRACTION NETWORK

SMEN is utilized to mark out the target area and the non-target area by estimating the scene mask of the input image. In order to extract the precise scene mask, it is necessary to classify each pixel within an input image. Deconvolution [26] is a treatment in DCNN to obtain the classification result of each pixel through the superposition and up-sampling of multi-level feature maps, as shown in Fig. 3. Therefore, SMEN is established on the deconvolutional mechanism. Connecting with the output of FEN, SMEN has four deconvolutional layers and a Softmax layer to estimate scene mask. The configuration of SMEN is shown in Table 2. It is worth noting that the network training requires the target label as well as the scene label of each image in the dataset.

Furthermore, a skip-connection mechanism [24] is applied to optimize the segmentation accuracy for the input scene. As shown in Fig. 3, the low-level feature maps, generated by each ResNet block in FEN, are fused with the high-level feature maps produced by each deconvolutional layer in SMEN. More specifically, each low-level feature map is concatenated with the corresponding high-level one with the same scale through feature maps stacking. Each concatenated feature map is then imported into a convolutional layer, the convolutional layer is used to further extract the feature

of the input image and adjust the depth of the feature maps by using less convolution kernels.

3) REGION PROPOSAL NETWORK

The RPN starts with the output feature map of FEN. Firstly, a series of anchors, generated by the sliding window mechanism on the entire image, yield N initialized anchors with different scales and aspect ratios. Multi-scale feature information is obtained by the pyramid of anchors. Afterwards, the input feature map goes through a 3×3 convolutional layer, and then it enters two branches. In the first branch, the feature map passes a 1×1 convolutional layer and a Softmax layer sequentially to get the classification score of each region proposal. The score represents the probability that the target is a positive sample or a negative sample. In the other branch, the feature map goes through a 1×1 convolutional layer, called bounding box regression [27], to form a $1 \times 1 \times 4N$ feature map that contains the 4 parameters in the coordinate vector of each region proposal. Finally, the region proposals of targets with high scores are screened out for further classification by a reasonable threshold, which controls the probability of targets that would be judged as positive samples.

4) CLASSIFICATION AND REGRESSION NETWORK

CRN imports the region proposals obtained by RPN to classify and regress the targets. These region proposals go through a ROI pooling layer (shown in Fig. 3), which is used to normalize these region proposals to the same size by the spatial pyramid pooling method [28]. Then, the convolutional operation is carried out on these normalized region proposals to obtain the feature map with more channels, and then the extended feature map enters a fully connected layer to form a feature vector. Afterwards, the classification result of the target is obtained by a Softmax layer with the input of feature vector, and the location of target is further optimized by the bounding box regression method. Finally, overlapping bounding boxes are merged by non-maximum suppress (NMS) [29] method.

B. TRAINING

The training process contains forward propagation and back propagation. Forward propagation starts with FEN and then has three branches, which calculates the scene mask loss, the classification loss, and the location regression loss, respectively. Backpropagation [30] updates the parameters of each layer in the network and minimizes the loss function by momentum optimization algorithm [31].

Obviously, Scene Mask R-CNN needs a multi-task loss function [32]. The loss of SMEN (L_{mask}) is the cross-entropy form between the estimated scene mask and its ground truth, as defined in (1),

$$L_{mask} = -\frac{1}{W \times H} \sum_i \sum_j M_t(i,j) \cdot \log(M(i,j)) + (1 - M_t(i,j)) \cdot \log(1 - M(i,j)) \quad (1)$$

where $W \times H$ is the size of the scene mask. $\boldsymbol{M_t}$ is the binary ground-truth, and $\boldsymbol{M}$ is the estimated scene mask.

The loss of classification (L_{cls}) is the cross entropy between the detection results and their ground-truth, as defined in (2),

$$L_{cls} = -\frac{1}{N_{cls}} \sum_k \log\left[p_k p_k^* + (1 - p_k)(1 - p_k^*)\right] \quad (2)$$

where N_{cls} is the number of boxes. p is the predicted probability being a target of box k. p^*, given by the ground-truth label, equals 1 when the label is positive, and is 0 when the label is negative.

The loss of the location regression (L_{reg}) is the $smooth_{L1}$ loss between the vector $\boldsymbol{b}$ containing 4 coordinates of the predicted bounding box and the vector $\boldsymbol{b}^*$ generated from the ground-truth box associated with a positive box, as defined in (3) and (4), respectively.

$$L_{reg} = \frac{\beta}{N_{reg}} \sum_k p_k^* smooth_{L1}\left(\boldsymbol{b}_k - \boldsymbol{b}_k^*\right) \quad (3)$$

$$smooth_{L1}(x) = \begin{cases} 0.5x^2 & \text{if } |x| < 1 \\ |x| - 0.5 & \text{otherwise} \end{cases} \quad (4)$$

where N_{reg} is the normalized parameter, i.e. the number of box locations. β is used to balance the process of classification and regression. The negative boxes are filtered by the effect of the p_k^*.

Besides, the coordinate vector $\boldsymbol{b}_k = [b_x, b_y, b_w, b_h]$ and $\boldsymbol{b}_k^* = [b_x^*, b_y^*, b_w^*, b_h^*]$ are calculated in (5),

$$\begin{aligned} b_x &= (x - x_a)/w_a, b_x^* = (x^* - x_a)/w_a, \\ b_y &= (y - y_a)/h_a, b_y^* = (y^* - y_a)/h_a, \\ b_w &= \log(w/w_a), b_w^* = \log(w^*/w_a), \\ b_h &= \log(h/h_a), b_h^* = \log(h^*/h_a), \end{aligned} \quad (5)$$

where (x, y), w, and h represent center coordinates, width and height of boxes respectively, and the x (or y, w, h), x_a (or y_a, w_a, h_a), and x^* (or y^*, w^*, h^*) are about the predicted box, anchor box, and ground-truth box, respectively.

The total loss is the weighted summation of three losses, as defined in (6),

$$Loss = \lambda(L_{cls} + L_{reg}) + \mu L_{mask} \quad (6)$$

where λ and μ are the weighted parameters that can be adjusted to various training requirements, revealing the emphasis between the target detection task and mask segmentation task in the current model.

C. INFERENCE

In the inference process, the estimated scene mask of the input image is used to reduce the onshore false alarms, as shown in Fig. 4. Firstly, the input image goes through FEN and an output feature map ($C_f @ W_f \times H_f$) is obtained. The width W_f and height H_f of the feature map are smaller than the input image and the channel C_f=1024. The inferenced scene mask, obtained from SMEN, is merged with the feature map originated from FEN. In the merging process, the inferenced

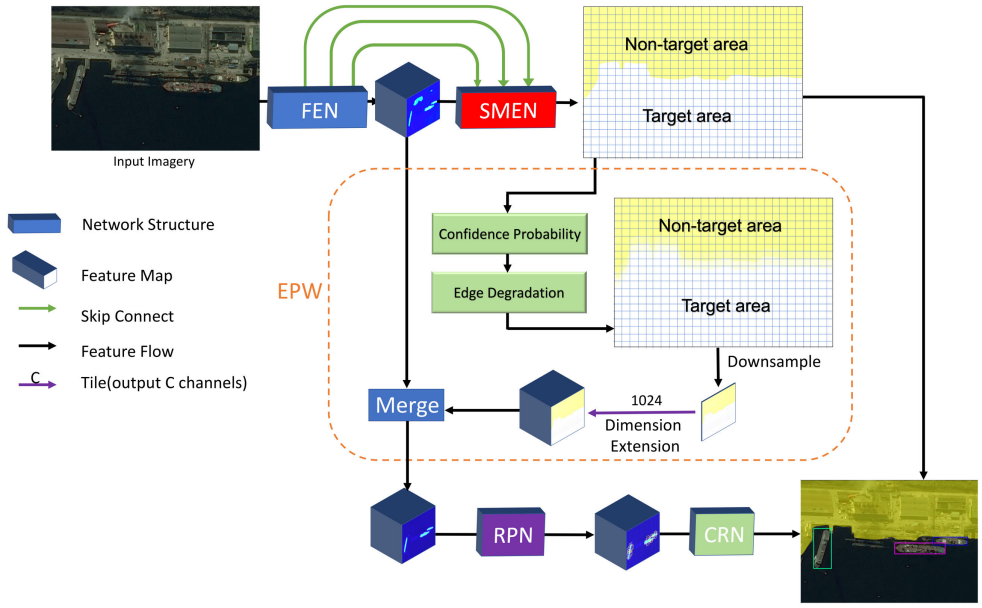


FIGURE 4. Inference process of scene mask R-CNN. The EPW merging operates on the inferenced scene mask, and the feature map is originated from the output of FEN.

scene mask is downsampled to the same size as the output of FEN. Then the merged feature map is introduced into RPN, and the region proposals without false alarms in the non-target area are provided to classify and regress the target.

D. THE EDGE PROBABILITY WEIGHTED MERGING

However, some misjudged pixels of edge in the inferenced mask are inevitable, which conducts the ships very close to the shore to be contained into the land area, even to be omitted during the merging process. Therefore, the edge probability weighted (EPW) merging method, avoiding the omission caused by the misjudged edge, is proposed to fuse the inferenced scene mask and the feature map originated from FEN.

In fact, the accurate segmentation of land and sea is beneficial for nearshore ship detection. Although some existing segmentation methods can improve the accuracy of the edge to assist in subsequent detection tasks, over-segmentation is a common phenomenon. As shown in Fig. 5, whatever SeNet [33], Unet or our method without EPW, the estimated scene mask always covers the targets owing to the rough marking or unclear boundary.

In EPW, the confidence probability $\boldsymbol{P}$ is applied to the inferenced scene mask, and it represents the confidence degree of scene attributes for each pixel. Here, $\boldsymbol{P}$ depends on the spatial relation of observation pixel (i, j) and its nearest pixel (i_B, j_B) belonging to the sea-land boundary in the inferenced scene mask.

The confidence probability $\boldsymbol{P}$ is calculated as (7),

$$P(i,j) = \begin{cases} 1 & (i,j) \in \boldsymbol{\Omega_t} \\ \min\left(\dfrac{\sqrt{(i_B - i)^2 + (j_B - j)^2}}{T}, 1\right) & (i,j) \notin \boldsymbol{\Omega_t} \end{cases} \tag{7}$$

where $\boldsymbol{\Omega_t}$ is the set of pixels in the target area, and T is a constant representing the distance threshold, and its value depends on both the size of targets appearing on imageries and the aspect ratio of real ships. For instance, in our experiment, most of the targets in our dataset with the length of around 150 pixels, and the aspect ratio of ships is close to 5:1. Thus, the constant T is set to 30, representing the maximum distance of the estimated scene mask covering the targets.

To the end, by calculating the ratio of spatial distance (i.e., the Euclidean distance between the pixel (i, j) and (i_B, j_B)) with the threshold T, the confidence probability $\boldsymbol{P}$ of the observation pixels within the threshold distance in non-target area is obtained. Owing to the scope of the probability, the value of $P(i, j)$ is limited in $[0, 1]$. Therefore, for the other pixels in non-target which beyond the threshold distance or the pixels in the target area, their probabilities are set to 1, consistently.

It is deduced that if $P(i, j) = 1$, the corresponding pixel in the estimated scene mask is credible; If $P(i, j) = 0$, it proves that the corresponding pixel in this mask is undependable.

In fact, the inferenced scene mask $\boldsymbol{S}$ is a 0-1 value matrix (i.e., target area is 1 and non-target area is 0). For the pixels around the sea-land boundary, their values of the inferenced scene mask are updated according to $\boldsymbol{P}$. Thus, the updated scene mask $\boldsymbol{M}$ is calculated as (8),

$$\boldsymbol{M} = \mathbf{1} - \boldsymbol{P} \odot (\mathbf{1} - \boldsymbol{S}) \tag{8}$$

where $\mathbf{1}$ refers to the all-1 matrix. According to (8), the boundary of the target and non-target areas in the estimated scene mask is degenerated successfully. In order to demonstrate the modification of scene mask clearly, the calculation process is shown in Fig. 6. $\boldsymbol{D}$ is the degradation matrix (i.e., $\boldsymbol{P} \odot (\mathbf{1} - \boldsymbol{S})$), revealing the degradation of the

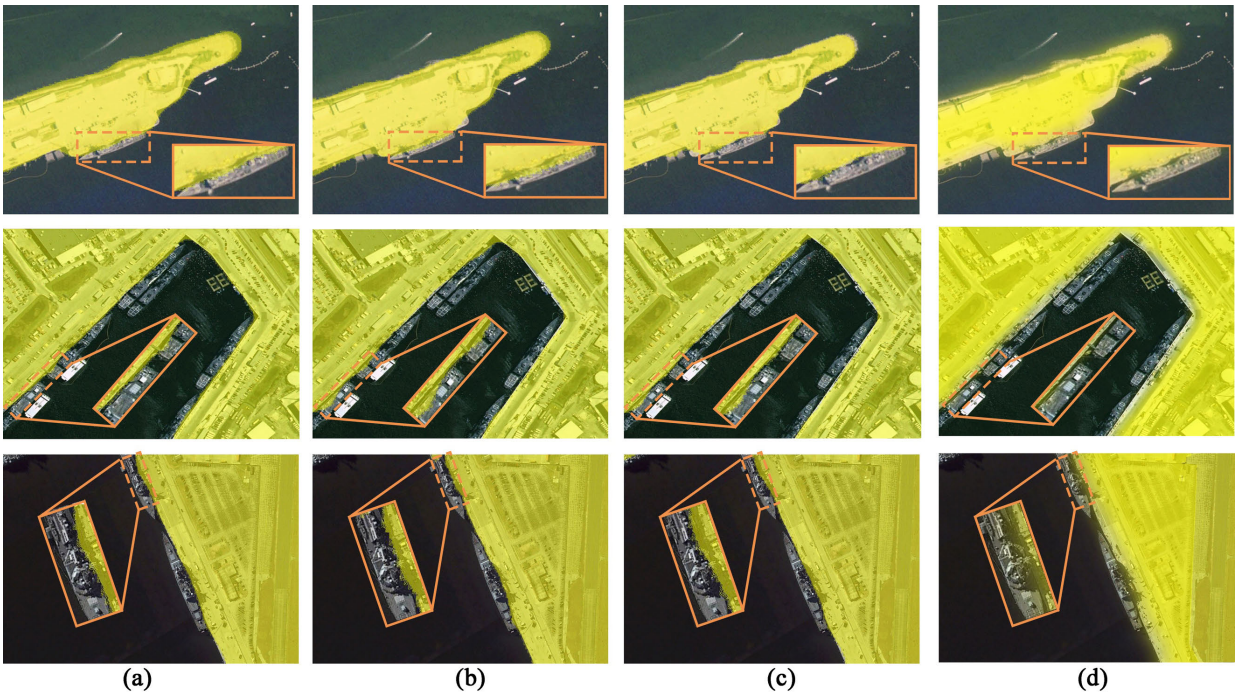


FIGURE 5. Over-segmentation in different segmentation methods and the application of the EPW in our method. (a) SeNet. (b) Unet. (c) Our method without the EPW merging mechanism. (d) Our method with the EPW merging mechanism.

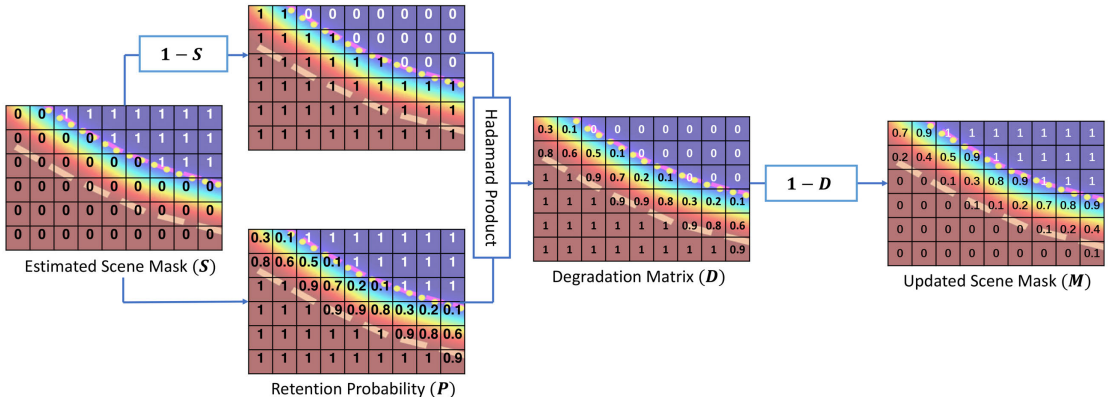


FIGURE 6. The calculation process of the updated scene mask. Each background of the pixelated imagery is obtained from the screenshot of Fig.7. The blue area refers to the target area, and the red area refers to the non-target area, which is consistent with Fig.7. Besides, the purple dash line refers to the sea-land boundary of the estimated scene mask, the yellow dot line refers to the sea-land boundary of ground-truth, the orange dash line refers to the sea-land boundary of the degenerated scene mask after EPW merging method.

non-target area. Finally, the merged feature map $\mathbf{R}$ is defined as (9),

$$\boldsymbol{R} = \boldsymbol{M} \odot \boldsymbol{I} \tag{9}$$

where $\boldsymbol{I}$ is the output of FEN. After EPW merging, $\boldsymbol{R}$ is imported into RPN for further detection.

Based on the above operation, the scene mask can be updated to achieve edge degradation. A simulated scene mask demonstrates the uncertainty of the boundary in the estimated scene mask, as shown in Fig. 7. The purple dash line which crosses the ship refers to the sea-land boundary of the estimated scene mask, the yellow dot line refers to the real sea-land boundary, the orange dash line refers to the sea-land boundary of the degenerated scene mask after EPW merging method, the double-headed arrow refers to the threshold distance T. During the shrinking process from the purple dash line to the orange dash line, the ship feature covered with the error estimated scene mask is partially suppressed according to the confidence probability matrix $\boldsymbol{P}$, instead of completely eliminated.

Benefited from the EPW merging process, even though the ship is partially covered by the estimated scene mask, the updated scene mask has little impact on the true target, making the target not be eliminated completely.

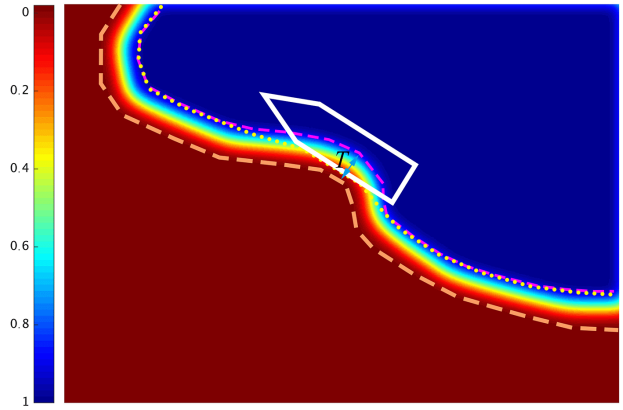


FIGURE 7. Simulated matrix diagram of the scene mask. The purple dash line refers to the sea-land boundary of the estimated scene mask, the yellow dot line refers to the real sea-land boundary, the orange dash line refers to the sea-land boundary of the degenerated scene mask after EPW merging method, the double-headed arrow refers to the threshold distance *T*.

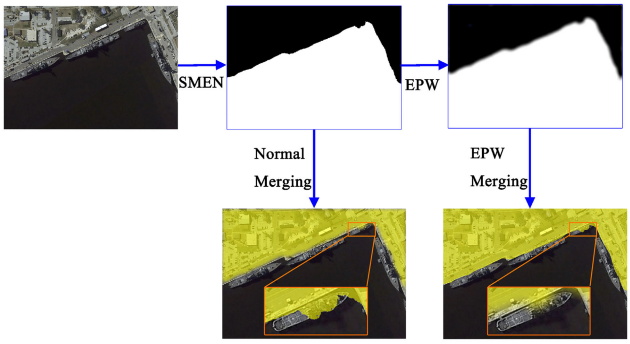


FIGURE 8. Workflow of SMEN and EPW for improving the estimated scene mask.

The workflow of SMEN and EPW is shown in Fig. 8. The results in column (d) of Fig. 5 and Fig. 8 show that EPW is an effectual mechanism to improve the scene mask.

IV. EXPERIMENTS

The high-resolution optical RS imagery dataset is utilized to evaluate the effectiveness of the proposed method. Faster R-CNN is selected as the comparison method to demonstrate the superiority of Scene Mask R-CNN. After that, some details of the experiments are discussed.

A. DATASET

HRSC2016 [34] and other 215 high-resolution optical RS imageries in DOTA [35] are adopted in the experiments. The HRSC2016 dataset consists of several RS images of six famous ports, including Murmansk, Everett, Newport-Rhode Island, Mayport Naval Base, Norfolk Naval Base, and San Diego Naval Base. The resolution of images is between 0.4 m and 2.0 m, and the size of images varies from 300×300 to $1,500 \times 900$ pixels. In addition, 215 high-resolution RS imageries with the size of around $4,096 \times 4,096$ and the same resolution as HRSC2016 are collected. These images contain a large amount of nearshore scenes. Besides, the location information of the targets and segmentation information

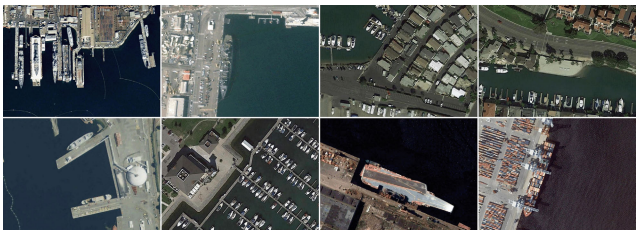


FIGURE 9. Samples of the augmented images in the dataset.

of scene are labeled by manually annotation for complementary images.

Statistically, the dataset contains various types of ships (e.g., cargo ships, cruise ships, yachts, aircraft carriers), and it has abundant multi-scale objects, exhibiting a wide variety of shapes and orientations. Moreover, due to the great period, the RS images with different illumination conditions and the cloud occlusion are contained in the dataset.

Data augmentation, for example, hue variation, brightness adjustment, contrast adjustment, noise addition, cutting, and flipping operations, is applied to enlarge the training samples. The random coefficient between 0.6 and 1.8 is selected for color-shifting, the coefficient between 0.8 and 1.2 is selected to change the brightness, and the coefficient between 0.5 and 1.8 is used to adjust the saturation. In addition, Gaussian noise and salt and pepper noise are added into the input images. Finally, the images in our dataset are rescaled into the same size of $1,200 \times 850$ with the overlap coefficient of 0.3. The statistical result shows that the scale of the labeled ships varies from 64 to 800 pixels, and most of them are around 150 pixels. After augmentation, a total of 11,540 images are divided into the train set, validation set, and test set with the ratio of 8:1:1 respectively. Samples of augmented images are shown in Fig. 9.

B. PROCEDURE AND RESULTS

In order to suppress the onshore false alarms in nearshore ship detection, the network structure is established on Scene Mask R-CNN. TensorFlow [36] is adopted as the deep learning framework. The training process is executed on the GeForce GTX 1080 GPU.

During the training process, consistent with the RPN in Faster R-CNN, the region proposals are obtained based on the shared feature map from the output of FEN. Anchors with multi-scale (2, 4, 8, 16, 32) and different aspect ratios (1:1, 1:2, 2:1) are selected according to the average size and shape of ships in our dataset. For SMEN, the initialization of each convolution kernel obeys the normal distribution with a mean value of 0, and variance which value is inversely proportional to the number of channels in the current feature map and its kernels size. Above settings of parameters are beneficial for increasing convergent speed, and preventing the gradient vanishing [17].

It is crucial to set the proper hyperparameters. In the training process, the momentum gradient descent is adopted with a momentum value of 0.9 to accelerate the convergence of

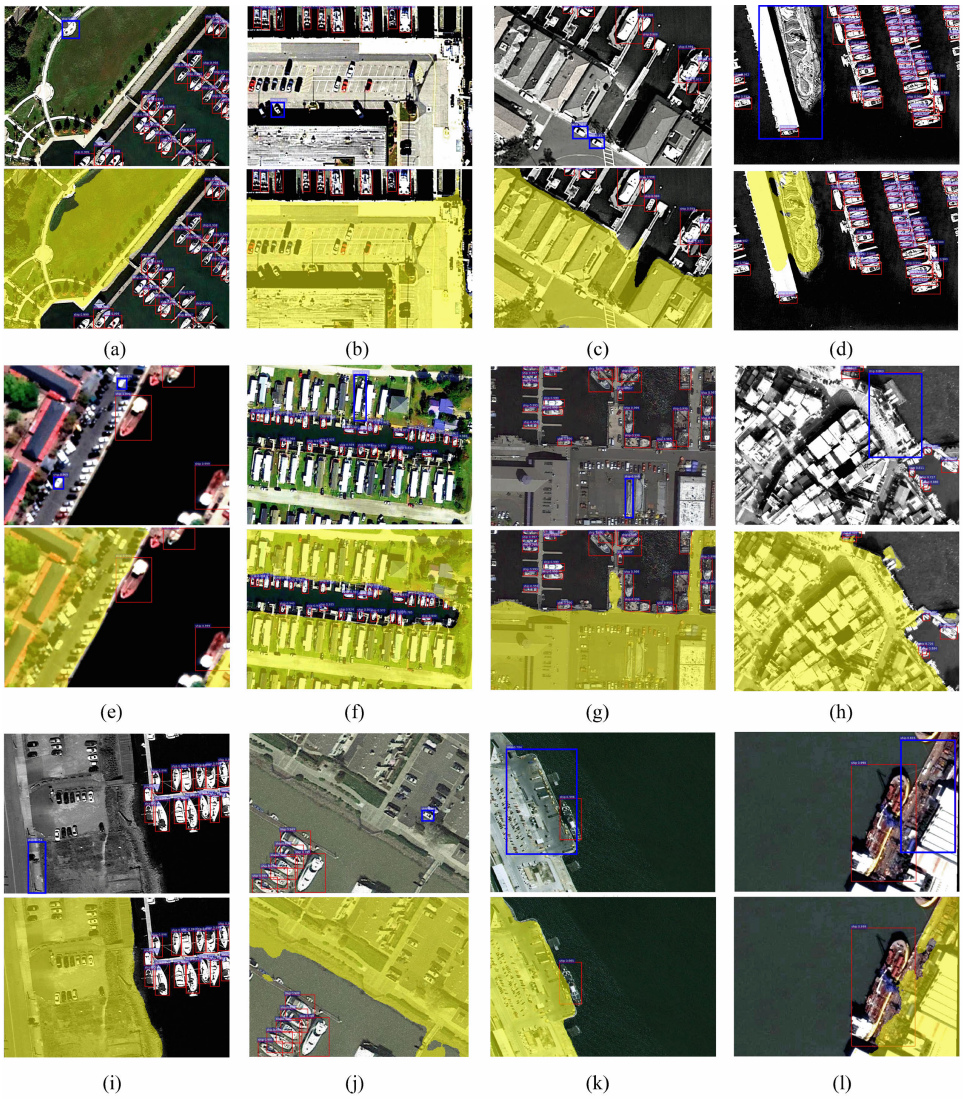


FIGURE 10. Results of the Faster R-CNN and Scene Mask R-CNN. The top rows in (a)-(l) show the results detected by the Faster R-CNN, and the bottom rows in (a)-(l) show the results detected by the Scene Mask R-CNN. The areas covered with yellow color refer to the scene masks, the blue boxes refer to the false alarms appearing on shore, and the red boxes refer to the correct detection results.

training. The number of training steps are 70,000 in total. The learning rate of the first 50,000 steps is 0.001, and it is defined as the basic training. Afterward, the sequential 20,000 steps with a learning rate of 0.0001 are defined as fine-tuning.

Furthermore, Dropout [37] and regularization method [38] are adopted to avoid over-fitting. The dropout rate is set to 0.8. In order to solve the problem of gradient vanishing and gradient exploding during the segmentation of scene mask, the ELU [39] activation function is applied to obtain the mean value that closes to zero. It is proved that the above training tricks make our model have faster convergence speed and more robust with a variety of images.

To analyze the impact of weighted parameter λ and μ in the training process, we conduct three groups of experiments. In the training process, the relationship between λ and μ is configured as follows: $\lambda > \mu$, $\lambda < \mu$, and $\lambda = \mu$. Under each group of experiments, different ratios of the weighted parameters are carried out for comparison. The loss functions with three ratios (i.e., λ:μ = 1:1, λ:μ = 1:5, λ:μ = 5:1) in each training process are depicted in Fig. 11. As it can be seen, the convergence speed of loss function is relatively fast under the condition of λ:μ = 1:1 during iterations, and its loss has more potential to be a stable state.

It is significant to evaluate the weighted parameters with the detection accuracy, as they have similar downward trends. Therefore, the precision and recall after the training of 70,000 steps are applied to measure the detection capability of our method with different ratios of weighted parameters. As shown in Fig. 12, five sets of different parameter ratios are selected for performance evaluation.

Obviously, when the proportion of SMEN loss is large ($\mu > \lambda$), the performance of target detection is reduced,

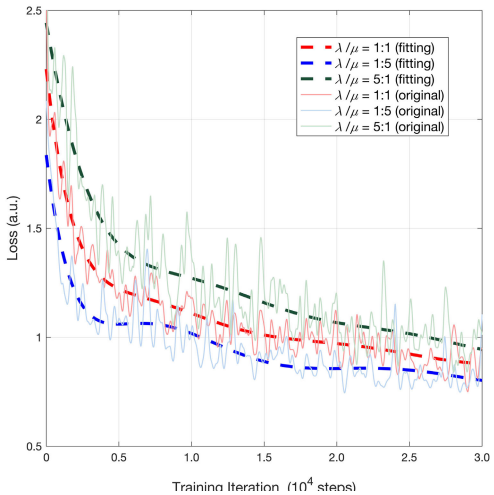


FIGURE 11. Comparison of loss functions on different weighted parameters in the training process. The dash lines refer to the fitting curves of loss functions, and the solid lines refer to the original lines. The red, blue, and green line represent the ratios of weighted parameters of $\lambda{:}\mu = 1{:}1$, $\lambda{:}\mu = 1{:}5$, and $\lambda{:}\mu = 5{:}1$, respectively.

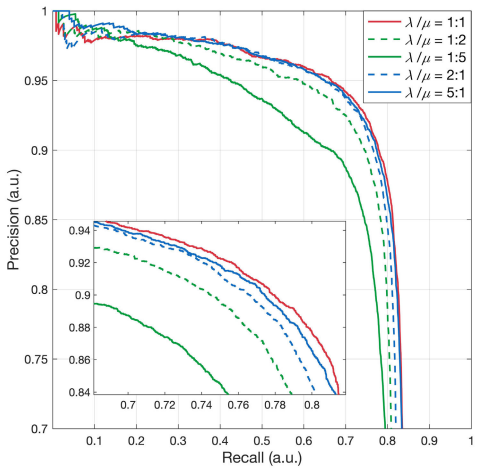


FIGURE 12. Precision-Recall (P-R) curves of ship detection for different ratios of weighted parameters. The red solid line indicates the result of $\lambda{:}\mu = 1{:}1$, the green dash line is $\lambda{:}\mu = 1{:}2$, the green solid line is $\lambda{:}\mu = 1{:}5$, the blue dash line is $\lambda{:}\mu = 2{:}1$, and the blue solid line is $\lambda{:}\mu = 5{:}1$.

owing to the inadequacy of the feature learning for detection and the over-study of the segmentation. Although, when $\mu < \lambda$, the loss of classification and regression is emphasized, there is a little drop in the ability of target detection according to the P-R curves in Fig. 12, which is contrary to the intention to improve the detection with a larger λ. The consistent conclusion is obtained from the result of the mean Average Precision (mAP) in Table 3. Therefore, the ratio of the weighted parameters λ and μ of the multi-loss task is suitable to set to 1:1 .

In the inference process, the EPW merging is applied to combine the estimated scene mask with the output of FEN, and it solves the problem of missed detection of coastal ships caused by an over-segmentation scene mask. Experimentally, the threshold T is set to 30 pixels based on the dataset and the aspect ratio of ships. According to the distance between the

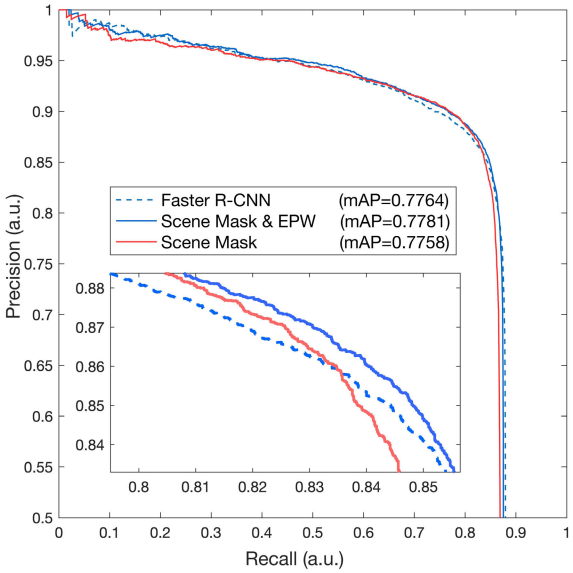


FIGURE 13. Precision and recall (P-R) curves for different detection methods via the basic training process. The blue dotted line indicates the result of Faster R-CNN. The red solid line indicates the result of Scene Mask R-CNN. The blue solid line indicates the result of Scene Mask R-CNN with the EPW merging method.

TABLE 3. Performance evaluation with different ratios of weighted parameters.

Ratios of weighted parameters	mAP
$\lambda:\mu=1:1$	0.7860
$\lambda:\mu=1:2$	0.7739
$\lambda:\mu=2:1$	0.7842
$\lambda:\mu=1:5$	0.7487
$\lambda:\mu=5:1$	0.7856

updated pixel and its nearest pixel belonging to the sea-land boundary of the inferenced scene mask, different merging intensities operate on the inferenced scene mask.

Different from the Faster R-CNN, Scene Mask R-CNN can suppress the false alarms which exist in non-target area (i.e., the land in ship detection task). As shown in Fig. 10, the areas covered with yellow color refer to the inferenced scene masks, the blue boxes refer to the false alarms appearing on shore, and the red boxes refer to the correct detection results. Obviously, our method successfully eliminates the onshore false alarms.

C. PERFORMANCE ANALYSIS

Precision and recall are important indicators for evaluating the performance of target detection. mAP is used to evaluate the performance of Scene Mask R-CNN and Faster R-CNN. The comparison results of mAP are collected in Table 4. As we conclude from the results of mAP, Scene Mask R-CNN reaches the detection ability of Faster R-CNN. Moreover, compared with the detection results in existing literature [34], [35] of HRSC2016 and DOTA, our detection accuracy is acceptable.

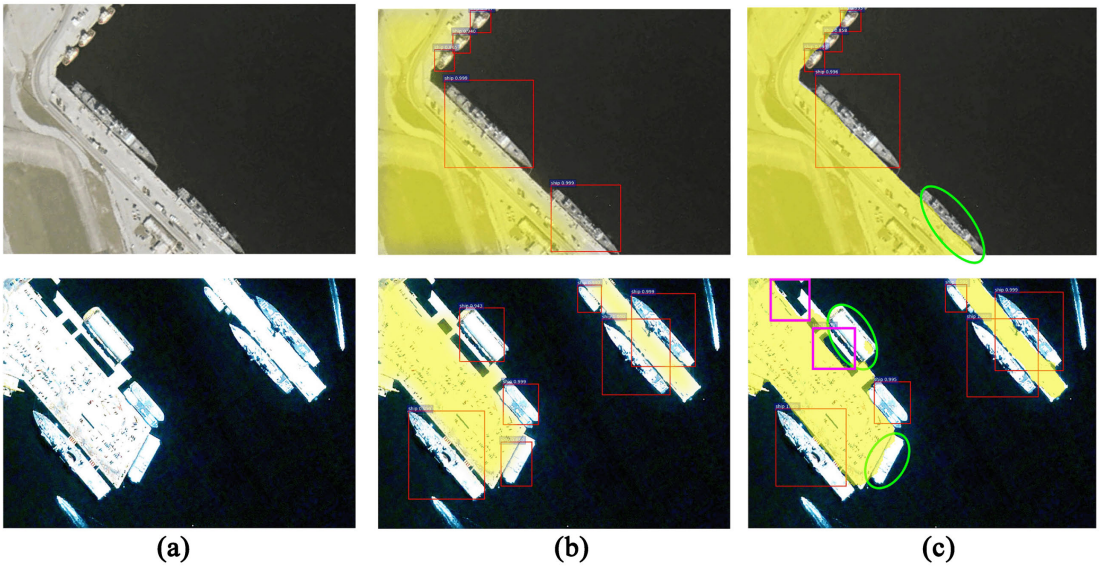


FIGURE 14. Comparison of ship detection results with different merging methods on basic training. (a) The original input images. (b) The result of Scene Mask R-CNN with EPW merging. (c) The result without EPW merging. The red boxes refer to the correct results, the green ellipses refer to the omitted targets, and the pink boxes refer to the misjudged results.

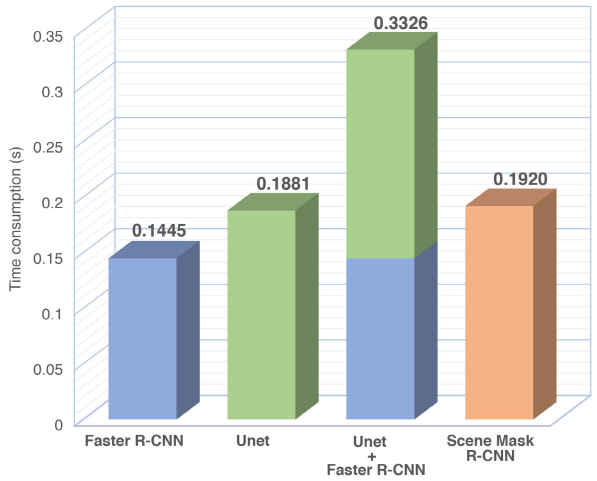


FIGURE 15. Histogram comparison of the time consumption on different networks.

For evaluating the performance of our method intuitively, relevant precision and recall curves of various training methods based on the basic training are shown in Fig. 13, and the area under each curve is equal to the value of the corresponding mAP. Obviously, in the same recall rate, the precision of Scene Mask R-CNN, even without the EPW mechanism, is superior to Faster R-CNN, which indicates our method successfully reduces false alarms in nearshore ship detection task meanwhile maintaining the detection accuracy. Note that the mAP of our method with EPW is slightly higher than both Faster R-CNN and our proposed method without EPW, indicating that EPW contributes to precision and recall.

Concerning the false alarms in the non-target area, the rate of false alarm image (FAI) is an important evaluation indicator. In Table 4, the improvement in mAP value is not significant, because of the low proportion of onshore false alarms in our dataset. However, a great merit is shown in the rate of FAI. Compared with Faster R-CNN, the rate of FAI of Scene Mask R-CNN is obviously reduced with the same basic training process.

As collected in Table 4, the number of images containing false alarms reduces from 183 to 4, with the decreasing FAI rate from 1.58% to 0.03%. It confirms that almost all the onshore false alarms are suppressed through Scene Mask R-CNN with EPW merging.

Besides, the mean accuracy of the estimated scene mask (SMA) of our model with the fine-tuning process is collected in Table 4. The SMA of our method is 0.983. It demonstrates the segmentation ability of our method, which is conducive to enhance the capability of false alarm suppression.

It has been verified the EPW is effective to merge the inferenced scene mask and the output of FEN. As shown in Fig. 14, the ships close to the sea-land boundary may be lost in Scene Mask R-CNN without EPW, which causes the missed detection. (e.g., the green ellipses in column (c) of Fig. 14). Some false alarms may appear due to the misjudgment of sea-land boundary, as the pink boxes shown in column (c) of Fig. 14. However, EPW merging not only suppresses onshore false alarms successfully, but also avoids target omissions, as shown in column (b) of Fig. 14.

High execution efficiency is also the advantage of our DCNN-based method. Considering the fairness of time consumption analysis, we compare it from the perspective of "Segmentation+Detection". As shown in Table 5, we summarize the time consumption of Faster R-CNN, Unet, and our method for the prediction process per image.

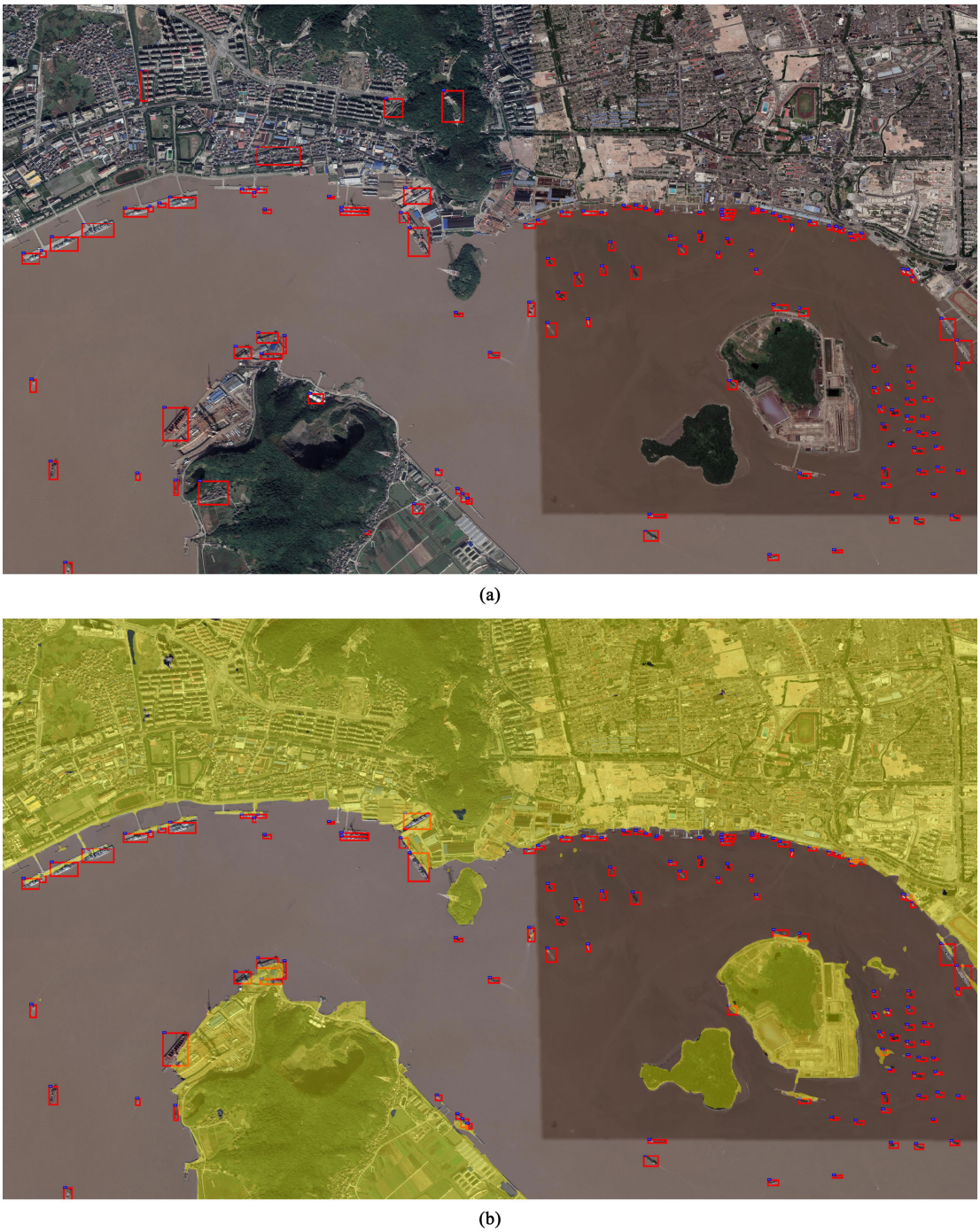


FIGURE 16. The ship detection results for a complex harbor scene. (a) Detection result of faster R-CNN with a lot of false alarms in the non-target area. (b) Detection result obtained from Scene Mask R-CNN. The yellow area in (b) represents the estimated mask of the non-target area (land or island), the red boxes refer to the predicted targets, and the false alarms in the non-target area are suppressed.

For the sake of the intuitiveness of the results, the histogram is depicted to reveal the difference of time consumption. As shown in Fig. 15, it can be concluded that the prediction time of our method is slightly larger than the single Faster R-CNN or Unet structure, limited by the synchronicity of the target detection and the semantic segmentation tasks. However, our network has a great advantage compared with the total time consumption of "Unet+Faster R-CNN", evidently, improving efficiency by 173%. In addition, "Unet+Faster R-CNN", as a non-end-to-end structure, performs lower on suppression of false alarm and detection accuracy, even if SMAs are similar, as shown in Table 4.

The high-resolution image of Zhoushan port in China obtained from GoogleEarth, with the size of 10000 × 6000,

TABLE 4. Performance evaluation with different methods.

Network structure	EPW merging	Basic training	Fine-tuning	mAP	SMA	Rate of FAI
Faster R-CNN		√		0.7764	-	2.82% (327/11540)
Scene Mask R-CNN		√		0.7758	0.981	0.91% (105/11540)
Scene Mask R-CNN	√	√		0.7781	0. 981	0.58% (68/11540)
Faster R-CNN		√	√	0.7828	-	**1.58% (183/11540)**
Unet+Faster R-CNN*		√	√	0.7834	0.980	0.61% (70/11540)
Scene Mask R-CNN	√	√	√	**0.7860**	**0. 983**	**0.03% (4/11540)**

* non-end-to-end system

TABLE 5. Time consumption of prediction on different models.

Model	Faster R-CNN	Unet	Unet+Faster R-CNN	Scene Mask R-CNN
Inference Time Consumption	0.1445 s	0.1881 s	0.3326 s	0.1920 s

is applied to test the performance under the complex scene. The detection result of Scene Mask R-CNN is shown in Fig. 16 (b), the red boxes refer to the predicted targets, and the yellow area represents the estimated mask of the non-target area. Compared with the result of conventional Faster R-CNN in Fig. 16 (a), our method can accurately detect targets in sea, and the eight false alarms located in the non-target area (land or island) are completely reduced, which proves the effectiveness of our network dramatically.

V. CONCLUSION

In the nearshore ship detection mission, the existing DCNN-based detection methods (e.g., Faster R-CNN) pay little attention to the suppression of onshore false alarms. In practical application, some onshore ship-like objects, such as dock, roof, and road, are interpreted as targets of interest with high probability, even if the trained network has a high detection accuracy. In fact, it is more suitable to detect targets after excluding the non-target area by using the scene information.

In this paper, an effective DCNN-based ship detection method, named as Scene Mask R-CNN, is proposed to reduce the onshore false alarms. Scene Mask R-CNN is an end-to-end system and has four sub-networks with different functions. They are feature extraction network (FEN), scene mask extraction network (SMEN), region proposal network (RPN), and classification and regression network (CRN) respectively. Based on the sub-network for scene segmentation, the scene information is utilized to assist the detection process with the non-target area suppression.

This method excludes the region proposals which exist in the non-target area during RPN. In the training process, the multi-task loss is used to optimize the network model; For the inference process, the inferenced scene mask is merged with the feature map originated from the output of FEN by using the EPW merging. The validity and accuracy of Scene Mask R-CNN is verified by the experiments on the high-resolution optical RS imageries dataset. Compared with Faster R-CNN, our method successfully suppresses the onshore false alarms, and achieves a slightly higher detection accuracy than Faster R-CNN in the nearshore ship detection mission.

ACKNOWLEDGMENT

This work is conducted on the platform of Center for Data Science of Beijing University of Posts and Telecommunications.

REFERENCES

[1] G. E. Hinton, S. Osindero, and Y.-W. Teh, "A fast learning algorithm for deep belief nets," *Neural Comput.*, vol. 18, no. 7, pp. 1527–1554, Jul. 2006.

[2] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 6, pp. 1137–1149, Jun. 2017.

[3] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, and S. Reed, "SSD: Single shot multibox detector," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, in Lecture Notes in Computer Science, vol. 9905. Amsterdam, The Netherlands: Springer, Oct. 2016, pp. 21–37.

[4] Y.-L. Chang, A. Anagaw, L. Chang, Y. C. Wang, C.-Y. Hsiao, and W.-H. Lee, "Ship detection based on YOLOv2 for SAR imagery," *Remote Sens.*, vol. 11, no. 7, p. 789, Apr. 2019.

[5] Z. Lin, K. Ji, X. Leng, and G. Kuang, "Squeeze and excitation rank faster R-CNN for ship detection in SAR images," *IEEE Geosci. Remote Sens. Lett.*, vol. 16, no. 5, pp. 751–755, May 2019.

[6] C. Dong, J. Liu, and F. Xu, "Ship detection in optical remote sensing images based on saliency and a rotation-invariant descriptor," *Remote Sens.*, vol. 10, no. 3, p. 400, Mar. 2018.

[7] T. Zhang and X. Zhang, "High-speed ship detection in SAR images based on a grid convolutional neural network," *Remote Sens.*, vol. 11, no. 10, p. 1206, May 2019.

[8] X. Yang, J. Yang, J. Yan, Y. Zhang, T. Zhang, Z. Guo, S. Xian, and K. Fu, "SCRDet: Towards more robust detection for small, cluttered and rotated objects," Aug. 2019, *arXiv:1811.07126*. [Online]. Available: https://arxiv.org/abs/1811.07126v4

[9] L. Qu, S. Wang, N. Yang, L. Chen, L. Liu, X. Zhang, F. Gao, and J. Dong, "Improving object detection accuracy with region and regression based deep CNNs," in *Proc. Int. Conf. Secur., Pattern Anal., Cybern. (SPAC)*, Dec. 2017, pp. 318–323.

[10] Q. Li, L. Mou, Q. Liu, Y. Wang, and X. X. Zhu, "HSF-Net: Multiscale deep feature embedding for ship detection in optical remote sensing imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 12, pp. 7147–7161, Dec. 2018.

[11] Z. Zhang, W. Guo, S. Zhu, and W. Yu, "Toward arbitrary-oriented ship detection with rotated region proposal and discrimination networks," *IEEE Geosci. Remote Sens. Lett.*, vol. 15, no. 11, pp. 1745–1749, Nov. 2018.

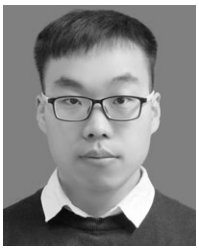
[12] L. Gao, Y. He, X. Sun, X. Jia, and B. Zhang, "Incorporating negative sample training for ship detection based on deep learning," *Sensors*, vol. 19, no. 3, p. 684, Feb. 2019.
[13] X. Li, S. Wang, B. Jiang, and X. Chan, "Inshore ship detection in remote sensing images based on deep features," in *Proc. IEEE Int. Conf. Signal Process. Commun. Comput. (ICSPCC)*, Oct. 2017, pp. 1–5.
[14] F. Wu, Z. Zhou, B. Wang, and J. Ma, "Inshore ship detection based on convolutional neural network in optical satellite images," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 11, no. 11, pp. 4005–4015, Nov. 2018.
[15] G. Chen, X. Zhang, Q. Wang, F. Dai, Y. Gong, and K. Zhu, "Symmetrical dense-shortcut deep fully convolutional networks for semantic segmentation of very-high-resolution remote sensing images," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 11, no. 5, pp. 1633–1644, May 2018.
[16] V. Badrinarayanan, A. Kendall, and R. Cipolla, "SegNet: A deep convolutional encoder-decoder architecture for image segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 12, pp. 2481–2495, Dec. 2017.
[17] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Medical Image Computing and Computer-Assisted Intervention* (Lecture Notes in Computer Science), vol. 9351. Cham, Switzerland: Springer, 2015, pp. 234–241.
[18] R. Li, W. Liu, L. Yang, S. Sun, W. Hu, F. Zhang, and W. Li, "DeepUNet: A deep fully convolutional network for pixel-level sea-land segmentation," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 11, no. 11, pp. 3954–3962, Nov. 2018.
[19] S. Sun, Z. Lu, W. Liu, W. Hu, and R. Li, "SHip net for semantic segmentation on VHR maritime imagery," in *Proc. Int. Geosci. Remote Sens. Symp. (IGARSS)*, Jul. 2018, pp. 6911–6914.
[20] H. Lin, Z. Shi, and Z. Zou, "Maritime semantic labeling of optical remote sensing images with multi-scale fully convolutional network," *Remote Sens.*, vol. 9, no. 5, p. 480, May 2017.
[21] S. Nie, Z. Jiang, H. Zhang, B. Cai, and Y. Yao, "Inshore ship detection based on mask R-CNN," in *Proc. Int. Geosci. Remote Sens. Symp.*, Jul. 2018, pp. 693–696.
[22] E. Shelhamer, J. Long, and T. Darrell, "Fully convolutional networks for semantic segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 4, pp. 640–651, Apr. 2017.
[23] J. R. R. Uijlings, K. E. A. van de Sande, T. Gevers, and A. W. M. Smeulders, "Selective search for object recognition," *Int. J. Comput. Vis.*, vol. 104, no. 2, pp. 154–171, Apr. 2013.
[24] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 770–778.
[25] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," Sep. 2014, *arXiv:1409.1556*. [Online]. Available: https://arxiv.org/abs/1409.1556
[26] M. D. Zeiler, G. W. Taylor, and R. Fergus, "Adaptive deconvolutional networks for mid and high level feature learning," in *Proc. IEEE Int. Conf. Comput. Vis.*, Nov. 2011, pp. 2018–2025.
[27] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 580–587.
[28] K. He, X. Zhang, S. Ren, and J. Sun, "Spatial pyramid pooling in deep convolutional networks for visual recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 9, pp. 346–361, Jun. 2014.
[29] A. Neubeck and L. Van Gool, "Efficient non-maximum suppression," in *Proc. Int. Conf. Pattern Recognit.*, vol. 3. Aug. 2006, pp. 850–855.
[30] R. J. Williams, D. E. Rumelhart, and G. E. Hinton, "Learning representations by back-propagating errors," *Nature*, vol. 2, pp. 533–536, Oct. 1986.
[31] I. Sutskever, J. Martens, G. Dahl, and G. Hinton, "On the importance of initialization and momentum in deep learning," in *Proc. 30th Int. Conf. Mach. Learn. (ICML)*, vol. 3, Feb. 2013, pp. 2176–2184.
[32] R. Girshick, "Fast R-CNN," in *Proc. ICCV*, Dec. 2015, pp. 1440–1448.
[33] D. Cheng, G. Meng, G. Cheng, and C. Pan, "SeNet: Structured edge network for sea–land segmentation," *IEEE Geosci. Remote Sens. Lett.*, vol. 14, no. 2, pp. 247–251, Feb. 2017.
[34] Z. Liu, L. Yuan, L. Weng, and Y. Yang, "A high resolution optical satellite image dataset for ship recognition and some new baselines," in *Proc. 6th Int. Conf. Pattern Recognit. Appl. Methods*, 2017, pp. 324–331.
[35] G.-S. Xia, X. Bai, J. Ding, Z. Zhu, S. Belongie, J. Luo, M. Datcu, M. Pelillo, and L. Zhang, "DOTA: A large-scale dataset for object detection in aerial images," in *Proc. IEEE Comput. Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 3974–3983.
[36] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Trans. Image Process.*, vol. 13, no. 4, pp. 600–612, Apr. 2004.
[37] G. E. Hinton, N. Srivastava, A. Krizhevsky, I. Sutskever, and R. R. Salakhutdinov, "Improving neural networks by preventing co-adaptation of feature detectors," Jul. 2012, *arXiv:1207.0580*. [Online]. Available: https://arxiv.org/abs/1207.0580
[38] Z. Ben Xu, Z. Hai, W. Yao, X. Y. Chang, and L. Yong, "$L_{1/2}$ regularization," *Sci. China Inf. Sci.*, vol. 53, no. 6, pp. 1159–1169, Jun. 2010.
[39] Y. Xie, H. W. Kim, H. J. Kim, and T. L. Song, "Reduction of computational load for implementing iJIPDA filter," in *Proc. 20th Int. Conf. Inf. Fusion*, Jul. 2017, pp. 1–6.

YANAN YOU received the Ph.D. degree from the School of Electronic and Information Engineering, Beihang University, China, in 2015, where he held a postdoctoral position, from 2015 to 2017.

Since September 2017, he has been a Lecturer with the School of Information and Communication Engineering, Beijing University of Posts and Telecommunications, China. His research interests include remote sensing image processing, SAR interferometry processing, deep learning, and big data technology.

JINGYI CAO received the B.E. degree from the Beijing University of Posts and Telecommunications, Beijing, China, in 2019, where she is currently pursuing the M.S. degree with the School of Information and Communication Engineering.

Her research interests include computer vision, pattern recognition, and remote sensing image processing, especially on object detection and segmentation.



YANKANG ZHANG received the B.E. degree from the China University of Petroleum, Qingdao, China, in 2017, where he is currently pursuing the M.S. degree with the School of Information and Communication Engineering, Beijing University of Posts and Telecommunications, Beijing, China.

His research interests include computer vision and remote sensing image processing, especially on object detection and image fusion.

FANG LIU received the Ph.D. degree from Nankai University, Tianjin, China, in 1997.

She is currently an Associate Professor with the School of Information and Communication Engineering, Beijing University of Posts and Telecommunications, Beijing, China. Her research interests include broadband IP networks, network traffic monitoring, machine learning, and data mining.

WENLI ZHOU received the Ph.D. degree in engineering in signal and information processing.

She is currently an Associate Professor with the School of Information and Communication Engineering, Beijing University of Posts and Telecommunications, Beijing, China. Her research interests include network traffic monitoring, user behavior analysis, telecommunications and Internet big data processing, and other research.