Shape Similarity Intersection-Over-Union Loss Hybrid Model for Detection of Synthetic Aperture Radar Small Ship Objects in Complex Scenes

Peng Chen, Hui Zhou, Ying Li, Bingxin Liu, and Peng Liu

***Abstract*—With the continuous development and utilization of marine environments, the demand for accurate identification of ship targets at sea is increasing in both military and civilian fields. Synthetic aperture radar (SAR) is used to detect ship targets at sea and can provide 24-h detection under any weather conditions. Deep-learning models enable the effective detection of ship targets using SAR images; however, the recognition accuracy may be low or false positives may occur in complex scenarios wherein it is difficult to detect the ship targets. Current target-detection tasks include target classification and positioning through bounding-box regression. Herein, a regression loss function is derived to calculate the position of the bounding box, and intersection over union (IoU) is applied to estimate the positioning accuracy. As a result, a gap exists between the commonly used positioning losses for regressing the parameters of a bounding box and the optimization of these metric values. Therefore, the proposed hybrid model combines classification, localization, and segmentation with a novel multitask loss function for boundary-box localization based on the improved IoU. This solves the problem of inconsistency between training and evaluation and improves the positioning accuracy. Experiments were conducted using the SAR dataset for ship detection; the dataset was labeled by SAR experts and included multiscale ship chips with a resolution of 256 pixels in both range and azimuth. In summary, the experimental results indicate that the proposed hybrid model could improve the detection accuracy in complex scenarios, and its false-positive rate is significantly lower than those of the other models.**

***Index Terms*—Complex scene, image segmentation, multitarget ship detection, multitask loss function, synthetic aperture radar (SAR) image.**

I. Introduction

DETECTION of ship targets at sea and in ports is used in various maritime activities, such as illegal fishing, oil spill monitoring, and marine traffic management [1], [2]. Ship detection using synthetic aperture radar (SAR) systems has received considerable attention in recent years because of the wide-area coverage and 24-h imaging capabilities of the SAR systems [3]. The resolution of the systems is constant when they are located far from the observed target. Therefore, using SAR systems has become crucial to enable ship-target detection at sea [4], [5]. In general, offshore ship detection experiences low background noise, making foreground target extraction reasonably easy. However, for inshore ship detection, the background noise is high, which is accompanied by the effects of different land types. This is because harbors, where ships dock, generally occur along the coastlines; hence, there is a constant concentration of ships at near-shore areas, which presents shore image contrasts, making it difficult to extract and identify the target.

Unlike optical imaging, the imaging system of SAR is based on the coherent principle, and the gray value of adjacent pixels would generate some random changes when signals are returned. This process would bring speckle noise, which makes it more difficult to detect meaningful changes. Objects with a large backscattering coefficient have greater brightness on SAR images. Differences in brightness constitute a single-channel grayscale image [6]. In recent decades, feature-based ship-detection methods have been used to perform accurate ship detection [7]. These methods deliver high performance in open sea areas. For example, in 2010, Zhu *et al.* [8] proposed a method that combines a histogram of oriented gradients (HOG) with the features and characteristics of spatial and rotational deformation. Candidates were extracted based on thresholds, and objects were detected using the hierarchical classification method. Shi *et al.* [9] combined the characteristics of the circular frequency with HOG features and applied the AdaBoost classifier to ship detection. Zhang *et al.* [10] generated new HOG features by normalizing the polar angle as well as detecting ships and other objects at sea using the support vector machine classifier. However, this type of feature cannot effectively adapt to complex background environments in SAR images.

In comparison, deep-learning models can automatically learn distinctive features; hence, they are frequently used for ship detection in SAR imaging [11]. Liu *et al.* [12] utilized sea–land segmentation to obtain the proposed location of the ship target and distinguish it from other objects using a convolutional neural network (CNN). Kang *et al.* [13] proposed a region-based CNN

Manuscript received May 21, 2021; revised July 25, 2021 and August 23, 2021; accepted September 4, 2021. Date of publication September 14, 2021; date of current version October 4, 2021. This work was supported by the National Natural Science Foundation of China under Grant 51609032. *(Corresponding author: Peng Chen.)*

Peng Chen, Bingxin Liu, and Peng Liu are with the Navigation College, Dalian Maritime University, Dalian 116026, China (e-mail: chenpeng@dlmu.edu.cn; liubingxin@dlmu.edu.cn; liupeng@dlmu.edu.cn).

Hui Zhou is with the College of Computer and Software, Dalian Neusoft University of Information, Dalian 116023, China (e-mail: zhouhui@neusoft.edu.cn).

Ying Li is with the Environmental Information Institute, Dalian Maritime University, Dalian 116026, China (e-mail: yldmu@126.com).

Digital Object Identifier 10.1109/JSTARS.2021.3112469

with contextual information and multilayer features for ship detection. They improved the detection accuracy by combining high-resolution graphic and semantic features and eliminated detection errors based on contextual information. Wang *et al.* [14] detected ships within complex backgrounds in SAR images using a single-shot multibox detector (SSD) model and transfer learning to improve the detection accuracy. Cui *et al.* [15] proposed a dense attention pyramid network-based method that can detect multiscale ships in different scenes of SAR images with high accuracy. Li *et al.* [16] proposed a model that combines the generative adversarial network and Fast R-CNN. The model delivered higher detection accuracy but low efficiency. Wei *et al.* [17] proposed a SAR ship-detection algorithm based on the improved Faster R-CNN, which was trained to initially identify small targets and initialize the parameters of the detection model. Training and tests were performed on Sentinel-1 SAR images for ship detection. The experimental results indicate that this method performed well in detecting small or dim ship targets in SAR images. Over the past few years, deep-learning networks have generated image-segmentation models with excellent performance and considerably improved the accuracy. These networks can also effectively deepen the understanding of images based on improvements in detection and classification models. Nie *et al.* [18] proposed a ship-detection and segmentation method based on the Mask R-CNN model, which can accurately detect and segment ships at the pixel level.

Several CNN-based ship detectors have been applied to improve the ship-detection performance. However, these detectors may have some drawbacks. According to SAR images, when buildings, islands, or harbors have double backscattering reflections, they are likely to have the same backscatter coefficient value as that of ships [19]. This may lead to false positives in complex scenarios, resulting in low ship-detection accuracies, as shown in Fig. 1. CNN-based ship detectors rely on backbone CNN architectures that are pretrained on image classification tasks to extract the feature maps of input images and use the last layer of feature vectors for object localization and classification. Different feature layers of a CNN have different spatial resolution and semantic information. For example, the lower layers utilize lower level semantic features in comparison to those utilized by the final layers. As a result, the object position in the lower layer is more accurate, which enhances the detection of small objects, whereas, in the final layers, only large objects are detected because smaller objects lose significant signals during downsampling in the pooling areas. The feature pyramid network (FPN) integrates multilayer features to improve the detection of small objects by replacing the feature extractor in the detector (such as Faster R-CNN) and generating feature maps. In this manner, PANet [20], NAS-FPN [21], and other networks based on FPN solved the problem of multiscale target detection through the structure of cross-scale connections.

In addition, the mean average precision (mAP) is calculated based on the intersection-over-union (IoU) threshold, which is the most common evaluation metric used for object localization. However, in object detectors, the optimization under a regression loss function limits the consistency of optimization and evaluation. During the training of ship detectors, when calculating the regions of interest (ROIs) of different ship targets in the process of localization and segmentation, each pixel was first established as the center of the feature maps. Anchor boxes with varying pixel areas and different ratios were assigned, generating a large number of proposals. Subsequently, the IoUs of the proposal and the ground truth were calculated. A proposal with an IoU greater than the threshold was retained as the ROI. These proposals enable objects with various scales and ratios to be equally represented in training a detector. The optimization of localization and segmentation is based on a regression loss function of the ROI bounding box. A gap exists between the optimization of regression losses and IoU values [22]–[24]. IoU is a common evaluation metric used for comparing the similarity between two arbitrary shapes. IoU encodes the shape properties of the objects under comparison, such as the widths, heights, and locations of two bounding boxes, and then calculates a normalized measure that focuses on these areas.

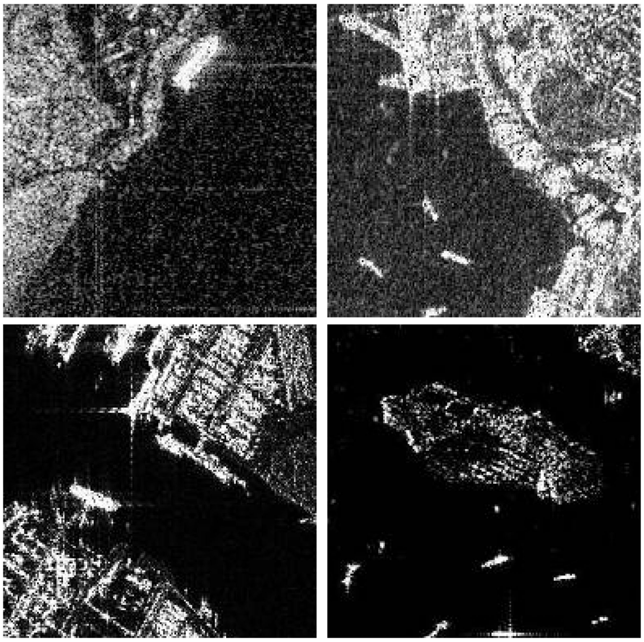


Fig. 1. Dataset of ship targets in complex scenarios (e.g., harbor, offshore, island, and other backgrounds).

Hence, this article proposes a hybrid model that simultaneously performs localization, classification, and segmentation. FPN was employed as a bone network for target detection, and the image-segmentation model with a deconvolution algorithm was included for target separation. Using this hybrid model, each pixel of the detected ships was extracted to achieve accurate segmentation. Meanwhile, a hybrid model on FPN with a new multitask loss function, an anchor box loss function, and an improved IoU loss was directly applied instead of using the original bounding-box regression loss function. The experimental results indicate that the proposed model increased the ship-recognition accuracy and significantly reduced the false-positive rate in complex scenarios.

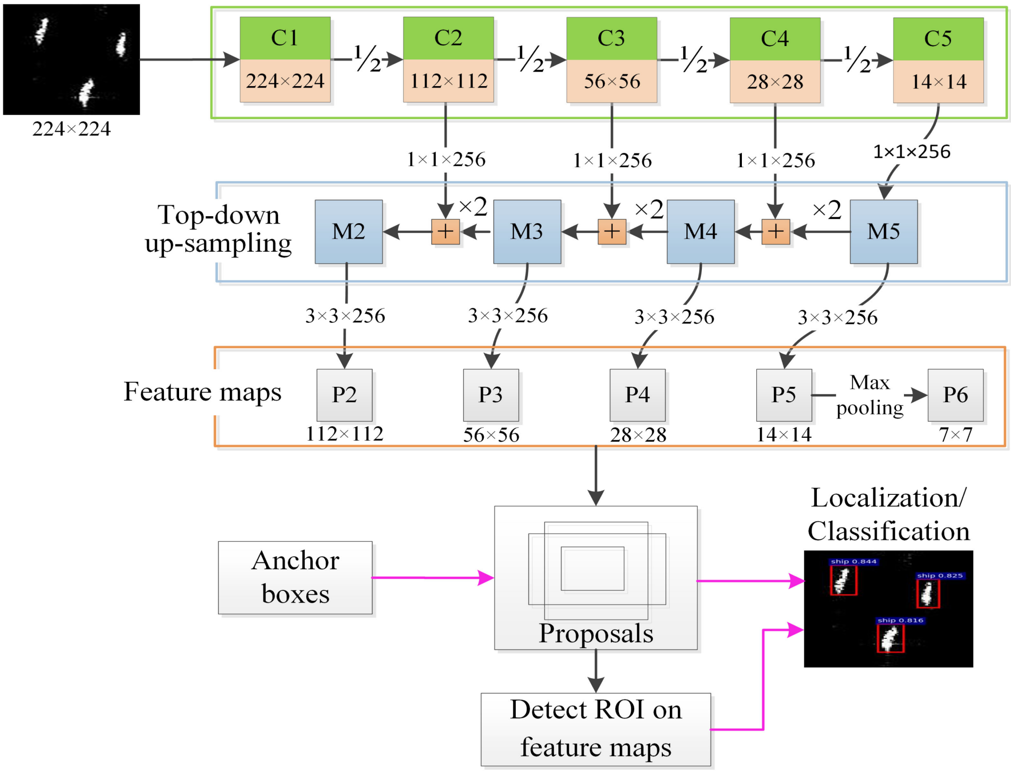


Fig. 2. Localization and classification modules. The green frame is the convolutional layer, the blue frame is the merging of each layer in addition to the upsampling result, and the orange frame is the feature map. Finally, the pink arrows indicate the classification and positioning results.

II. Materials and Methods

A. Revisiting FPN

The emergence of FPN [25] established the dominant positions of multiscale and small-scale detectors. FPN realizes classification and localization to detect objects. In this architecture, either VGG [26] or Resnet [27] can be used as the backbone network. An image was introduced into the pretrained backbone network, and the convolution layer then used bottom–up feature mapping to form layers {C1, C2, C3, C4, and C5}, as shown in Fig. 2. Next, {C1, C2, C3, C4, and C5} were laterally connected with the upsampling results through a 1 × 1 convolution kernel (256 channels) to form new feature maps {M2, M3, M4, and M5}. Finally, another 3 × 3 convolution was performed on M2–M5 to eliminate the aliasing effect caused by upsampling. The feature maps {P2, P3, P4, and P5} were then obtained, and the max pooling was realized for P5 to obtain the last layer of P6, which finally forms feature maps {P2, P3, P4, P5, and P6}.

When calculating the ROIs of different ship targets, each pixel was first established as the center of the feature maps {P2, P3, P4, P5, and P6}. Proposals were generated using the anchor boxes with varying pixel areas and length-to-width ratios. They were then retrained as ROIs using the threshold of their IoUs with ground-truth information. Most SAR ships in complex scenarios are small targets, which increase the difficulty of target extraction. Therefore, the quality of candidate areas generated using an IoU can be improved during target detection, which is conducive to achieving high detection accuracy.

In addition, different ground truths were assigned to different feature layers according to the length and width of each ground truth

$$k = \left\lfloor k_0 + \log_2 \frac{\sqrt{w \times h}}{224} \right\rfloor \tag{1}$$

where w and h refer to the width and height of the ground truth, respectively; k is the level an ROI is assigned to the feature map level P_k; and k_0 represents the lowest level mapped feature when $w, h = [224, 224]$. In other words, targets of different sizes were mapped to different feature pyramid levels to ensure that small targets were mapped to low feature levels that retained a considerable amount of location information. ROI and ground truth were selected for regression to achieve localization and classification.

B. Hybrid of FPN and Segmentation

The instance-segmentation method aims to achieve accurate detection of ship targets in SAR images and precise segmentation of targets from backgrounds. In complex scenarios, ships are presented as relatively small targets. If the ocean is calm, its scattering mechanism is a single reflection. If not, volume scattering is observed, which weakens the contrast between the ocean and the ship. Close proximity to islands, harbors, and offshore areas can result in inaccurate or missed detection due to the backscattering of SAR images. First, feature map layers were constructed through the FPN, where each convolution layer involved convolution and pooling operations during localization,

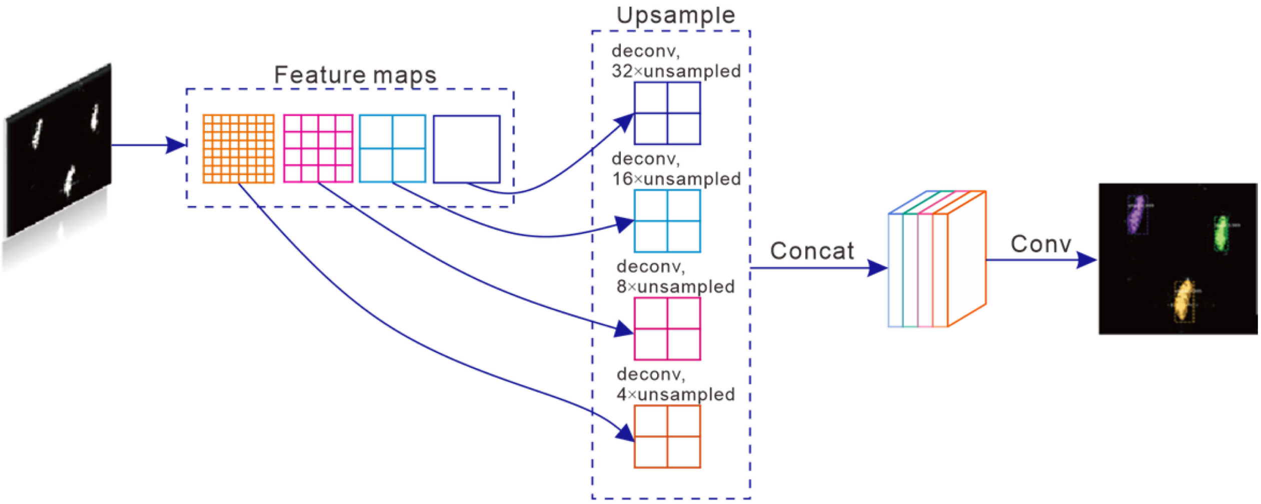


Fig. 3. Image-segmentation module. First, extract the feature maps of the image. Feature maps are embedding spaces with reduced size, which help to integrate and reduce computational complexity. Second, upsample and concatenate the feature maps, and use four different deconvolutions to extract features in parallel, which can extract multiscale semantic information on multilayer feature maps. Finally, perform a convolution on the extracted features to obtain a segmentation map.

classification, and recognition (i.e., downsampling). A decrease in the information acquired from each pixel was observed, allowing object features to be extracted. This was conducive to achieving object recognition. However, less pixel information can also lead to inaccurate localization of bounding boxes for object detection in complex scenarios, such as backgrounds with false targets or obscure nearby targets. Object segmentation can restore the downsampled image to its original size after it has been located and recognized. The output image is the same size as the original, with information annotation specifying the potential classification of each pixel. In contrast to the bounding box used in object detection, the supplementary segmentation model can accurately cut around the edges of the ship.

The ROI, selected in the feature maps, such as FPN, restored the pixel size of the original image through transposed convolution. The category of each pixel was subsequently detected. This process is illustrated in Fig. 3, and the specific steps are listed as follows.

Step 1: The convolutional network was shared with the models for localization and classification, and the feature map layers of the image were extracted.

Step 2: The feature maps that correspond to ROIs were calculated.

Step 3: Upsampling was performed on the output features *inconv* (*m*, *n*) of the ROI through transposed convolution; the output matrix *deconv* (*m'*, *n'*) was obtained.

Upsampling was performed 2^S times from *inconv* (*m*, *n*) to *deconv* (*m'*, *n'*), where *S* is the number of mapped feature layers of the ROI. Consequently

$$\begin{aligned} m' &= (m-1)\text{stride} + \text{kernelsize} - 2\text{padding} \\ n' &= (n-1)\text{stride} + \text{kernelsize} - 2\text{padding} \end{aligned} \tag{2}$$

where kernelsize is the size of the kernel, the zero-padding parameter padding is set at 1, and the parameter for stride length is stride $= 2^S$.

Step 4: Deconv (*m'*, *n'*) was obtained from different feature layers using transposed convolution and upsampling. It is different from the original image as it has a size of 224 × 224 pixels. The feature subgraph of ROI mapped to the bottom layer P6 is an example. Following FPN convolution, a feature map with a size of 7 × 7 × 256 was formed. Transposed convolution and upsampling are presented in the blue frame of Fig. 3. Here, let $S = 5$. After transposed convolution calculation, 0.5 was used as a threshold for binarization to generate a mask for the segmentation of the background and foreground.

On each layer of the hybrid model, the ROI was defined using the IoU greater than the threshold, as shown in Fig. 4. The output matrix was $m \times n$, and the number of channels was 256. Subsequently, two steps were performed simultaneously on the ROI feature vectors. One step involved classification and localization, and another step used transposed convolution for object segmentation, performed alongside localization and classification.

C. *IoU_SS Optimization Multitask Loss Function*

Ships in complex scenarios in SAR images are predominantly small targets, which are difficult to extract. The quality of ROI generated from an IoU during target detection determines the quality of detection. In this case, we directly used the optimization function as a loss, which improved the detection accuracy.

The hybrid model included classification, localization, and segmentation. Therefore, its multitask loss function *L* includes these three tasks

$$\begin{aligned} L = &\sum_i Lcls(pi, ui) \\ &+ \lambda 1 \cdot \frac{1}{N_p} \sum_i L_{\text{mask}}(pi, yi) + \lambda 2 \cdot L_{\text{IoU\_SS}} \end{aligned} \tag{3}$$

where λ_1 and λ_2 are the normalization coefficients, $L_{\text{cls}}(p_i, u_i) = -\log p_i u_i$ is the classification loss function, and

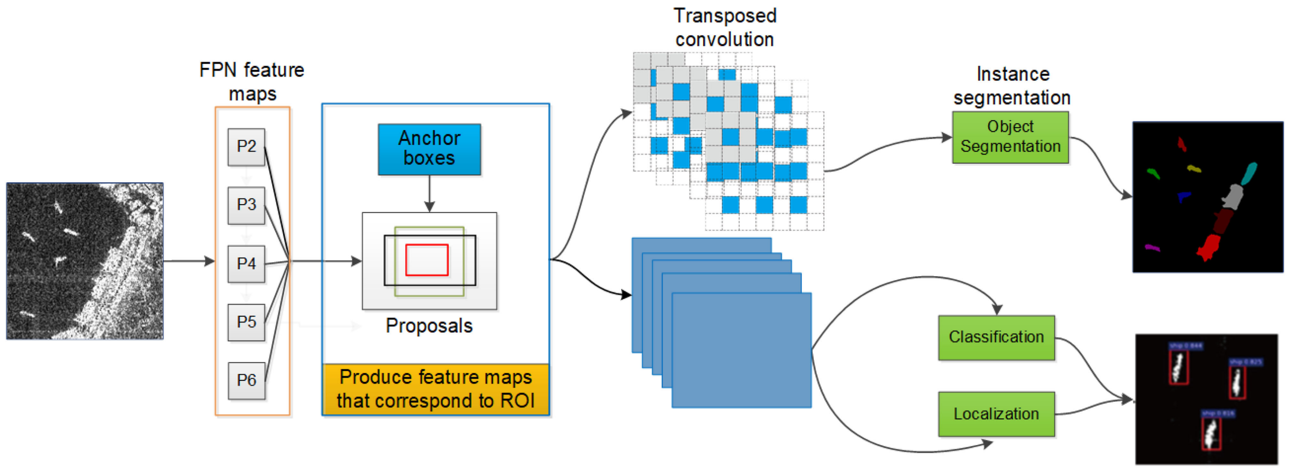


Fig. 4. Architectural diagram for the hybrid model.

the $p_i = (p_0, p_1, \ldots, p_k)$ is the probability distribution of each selected ROI. For the category of target k (whether it is a ship), if the calculated proposal is assigned with a positive label, then $u_i = 1$; if it is assigned with a negative label, then $u_i = 0$.

L_{mask} is the loss function for semantic segmentation and a binary cross-entropy classification loss function based on pixel calculations, determining the foreground and background. Each mask contains N_{pixel} pixels. L_{mask} is the mean of the binary cross-entropy loss for all masked pixels in a selected ROI

$$L_{\text{mask}} = \frac{1}{N_{\text{pixel}}} \sum_{j=1}^{N_{\text{pixel}}} (y_j \times \log p_j + (1 - y_j) \times \log p_j) \quad (4)$$

where y_j represents the foreground or background, and p_j is the corresponding probability.

$L_{\text{IoU\_SS}}$ is the novel localization loss function for the confidence score of ROI, which is applied directly instead of the original bounding-box regression loss. This ensures the consistency of the evaluation and optimization schemes

$$\begin{aligned} L_{\text{IoU\_SS}} = & \sum_{i}^{N_p} p_i [1 - \text{IoU\_SS}]^2 \\ & + \lambda_s \sum_{i}^{N_p} (1 - p_i)[1 - \text{IoU\_SS}]^2 \end{aligned} \quad (5)$$

where λ_s is the penalty factor for the detected ship, and p_i is the probability of true positives.

IoU is the intersection-over-union ratio of the predicted bounding box and the actual boundary box (ground truth). ***pb*** and ***gt*** correspond to the predicted bounding box and ground-truth coordinates, respectively

$$\begin{aligned} \boldsymbol{pb} &= (x^p_{\min}, x^p_{\max}, y^p_{\min}, y^p_{\max}) \\ \boldsymbol{gt} &= (x^g_{\min}, x^g_{\max}, y^g_{\min}, y^g_{\max}) . \end{aligned} \quad (6)$$

Then, we obtain

$$\begin{aligned} Ap &= (x^p_{\max} - x^p_{\min}) \times (y^p_{\max} - y^p_{\min}) \\ Ag &= (x^g_{\max} - x^g_{\min}) \times (y^g_{\max} - y^g_{\min}) \end{aligned} \quad (7)$$

$$\begin{aligned} AI &= \begin{cases} (x^I_2 - x^I_1) \times (y^I_2 - y^I_1), & \text{if } x^I_2 > x^I_1, y^I_2 > y^I_1 \\ 0, & \text{otherwise} \end{cases} \\ AU &= Ap + Ag - AI \end{aligned} \quad (8)$$

where $x^I_1 = \max(x^p_{\min}, x^g_{\min})$, $x^I_2 = \min(x^p_{\max}, x^g_{\max})$, $y^I_1 = \max(y^p_{\min}, y^g_{\min})$, $y^I_2 = \min(y^p_{\max}, y^g_{\max})$, A_I is the intersection of the predicted bounding box and ground truth, and A_U is the union of the predicted bounding box and ground truth

$$\text{IoU} = \frac{AI}{AU}. \quad (9)$$

This article proposes an IoU called IoU_SS (IoU with shape similarity) to calculate the shape similarity of the predicted bounding box and ground truth based on the absolute sum of the differences (ASD) and the sum of absolute differences (SAD) [28], and it can be defined as follows:

$$\text{IoU\_SS} = \text{IoU} - \left(\cos(\frac{d_{\text{ASD}}}{d_{\text{SAD}}} \times \frac{\pi}{2}) \right) \quad (10)$$

$$d_{\text{ASD}}(\boldsymbol{pb}, \boldsymbol{gt}) = \left| \sum_{k=1}^{n} (\boldsymbol{pbk} - \boldsymbol{gtk}) \right| \quad (11)$$

$$d_{\text{SAD}}(\boldsymbol{pb}, \boldsymbol{gt}) = \sum_{k=1}^{n} |\boldsymbol{pbk} - \boldsymbol{gtk}|. \quad (12)$$

In the IoU_SS loss, $\frac{d_{\text{ASD}}}{d_{\text{SAD}}}$ is the reflect shape similarity, and the larger the value, the higher the similarity; otherwise, the similarity is considered to be low. As shown in Fig. 5, the value range of $\cos(\frac{d_{\text{ASD}}}{d_{\text{SAD}}} \times \frac{\pi}{2})$ is [0, 1], as shown in Fig. 5. When $\frac{d_{\text{ASD}}}{d_{\text{SAD}}} = 1$, $\cos(\frac{d_{\text{ASD}}}{d_{\text{SAD}}} \times \frac{\pi}{2}) = 0$. This explains that the predicted bounding box and ground truth are the same or similar. Otherwise, the value is (0, 1]; in this case, $-1 \leq \text{IoU\_SS} \leq 1$.

IoU reflects the degree of coincidence between the proposals and the ground truth; i.e., the greater the degree of coincidence, the greater the value. Thus, IoU_SS has the same characteristics while considering the similarity of the shape of the prediction frame and target object.

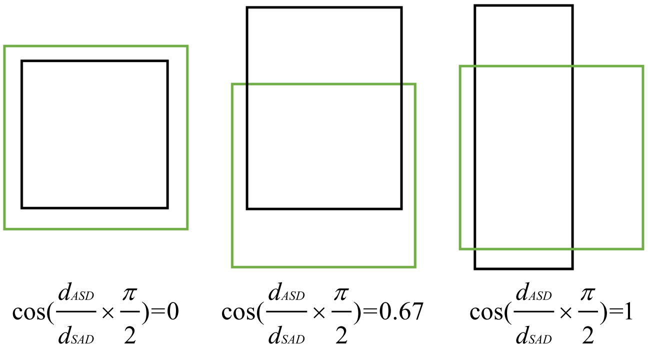

$\cos(\frac{d_{ASD}}{d_{SAD}} \times \frac{\pi}{2})=0$
$\cos(\frac{d_{ASD}}{d_{SAD}} \times \frac{\pi}{2})=0.67$
$\cos(\frac{d_{ASD}}{d_{SAD}} \times \frac{\pi}{2})=1$

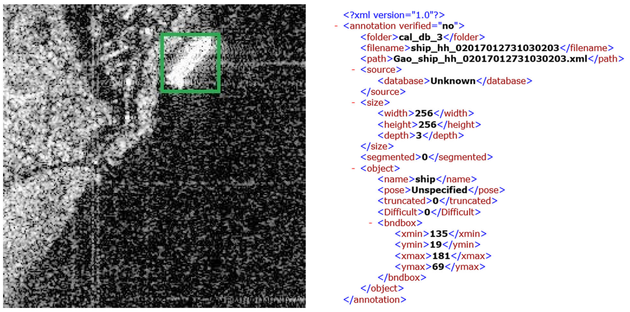
Fig. 5. These examples with the bounding box (black box) and ground truth (green box) represented by coordinates. For all three cases in the set, their shape similar distance between two rectangles are very different.

TABLE I
DETAIL INFORMATION OF THE DATASET FOR ORIGINAL SAR IMAGERY

Sensor	Imaging Mode	Resolution Rg.xAz.(m)	Incident Angle(°)	Polarization
GF-3	UFS	3×3	20~50	Single
	FSI	5×5	19~50	Dual
	QPSI	8×8	20~41	Full
	FSII	10×10	19~50	Dual
	QPSII	25×25	20~38	Full
Sentinel-1 SLC	SM	1.7×4.3~ 3.6×4.9	20~45	Dual
Sentinel-1 GRD	IW	20×22	29~46	Dual

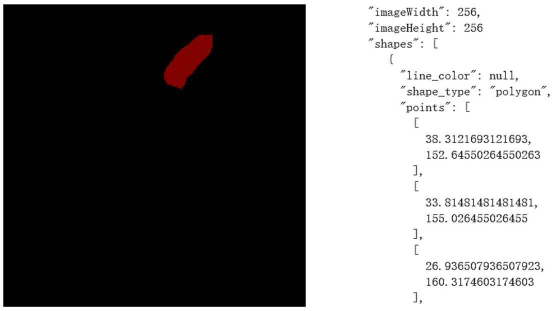
TABLE II
OVERVIEW OF THE DATASET

DataType	Image	Island	Harbor	Offshore	Sea area
Training	5355	1070	1070	1070	2145
Verification	1530	305	305	305	615
Testing	765	153	153	153	306



```
<?xml version="1.0"?>
- <annotation verified="no">
    <folder>cal_db_3</folder>
    <filename>ship_hh_02017012731030203</filename>
    <path>Gao_ship_hh_02017012731030203.xml</path>
  - <source>
      <database>Unknown</database>
    </source>
  - <size>
      <width>256</width>
      <height>256</height>
      <depth>3</depth>
    </size>
    <segmented>0</segmented>
  - <object>
      <name>ship</name>
      <pose>Unspecified</pose>
      <truncated>0</truncated>
      <Difficult>0</Difficult>
    - <bndbox>
        <xmin>135</xmin>
        <ymin>19</ymin>
        <xmax>181</xmax>
        <ymax>69</ymax>
      </bndbox>
    </object>
</annotation>
```

Fig. 6. Detailed information of a labeled ship chip. The green rectangles indicate the ground truth, and the label <bndbox> in the right image shows the location information of the ground truth.



```
"imageWidth": 256,
"imageHeight": 256
"shapes": [
  {
    "line_color": null,
    "shape_type": "polygon",
    "points": [
      [
        38.3121693121693,
        152.64550264550263
      ],
      [
        33.81481481481481,
        155.026455026455
      ],
      [
        26.936507936507923,
        160.3174603174603
      ],
```

Fig. 7. Result of object segmentation. The edge is made up of points, and connecting these points together is the edge polygon of the target.

TABLE III
STEP-BY-STEP EXPERIMENTAL PROCEDURE ON GAOFEN (GF)-3 AND SENTINEL-1 SAR SETS

Step	Methods	Island	Harbor	Offshore	mAP(%)
A	FPN-Resnet	93.72%	77.93%	70.32%	80.66%
B	A+Segmentation	95.36%	80.68%	77.69%	84.57%
C	B+IoU Loss	96.30%	88.05%	86.91%	90.42%
D	C+IoU_SS Loss	96.30%	94.44%	87.17%	92.63%

III. RESULTS AND DISCUSSION

A. *Implementation Details*

1) Dataset: The dataset, which entails 102 Chinese Gaofen-3 images and 108 Sentinel-1 images, was constructed by Wang *et al.* [29] and labeled by SAR experts. It comprises 43 819 ship chips with a resolution of 256 pixels in both range and azimuth. These ships have distinct scales and backgrounds. Furthermore, some of the ships were captured in complex scenes, such as islands, harbors, and offshore, as shown in Fig. 1.

For Gaofen-3, the image modes included ultrafine strip map (UFS), fine strip map 1 (FSI), full polarization 1 (QPSI), full polarization 2 (QPSII), and fine strip map 2 (FSII). The resolution of the SAR images in the dataset ranged from 3 to 10 m. For Sentinel-1, the imaging modes were S3 strip map (SM), S6 SM, and IW mode. The details of these images, including resolution, incidence angle, and polarization, are summarized in Table I.

In addition to noninterference sea-area scenes, the dataset included complex scenes, which are divided into three categories—offshore, island, and harbor—as shown in Fig. 1. An overview of the dataset is presented in Table II. The training, verification, and testing sets constitute 70%, 20%, and 10% of the dataset, respectively. Furthermore, the Labelme software was used to label the ship locations. Each ship chip corresponded to an extensible markup language file similar to that in the PASCAL VOC detection dataset, indicating the ship location, ship chip name, and <bndbox> label that defines the ground truth, as shown in Fig. 6. Each pixel of images in training sets was marked as a ship object or background pixel-by-pixel and simultaneously, as shown in Fig. 7.

2) Evaluation Metric: The results of the comparison in terms of detection precision P, recall R, and miss rate P_m are defined as

$$P = N_{\mathrm{TD}}/N$$
$$R = N_{\mathrm{TD}}/N_{\mathrm{GT}}$$
$$P_m = N_{\mathrm{FN}}/N_{\mathrm{GT}} \tag{13}$$

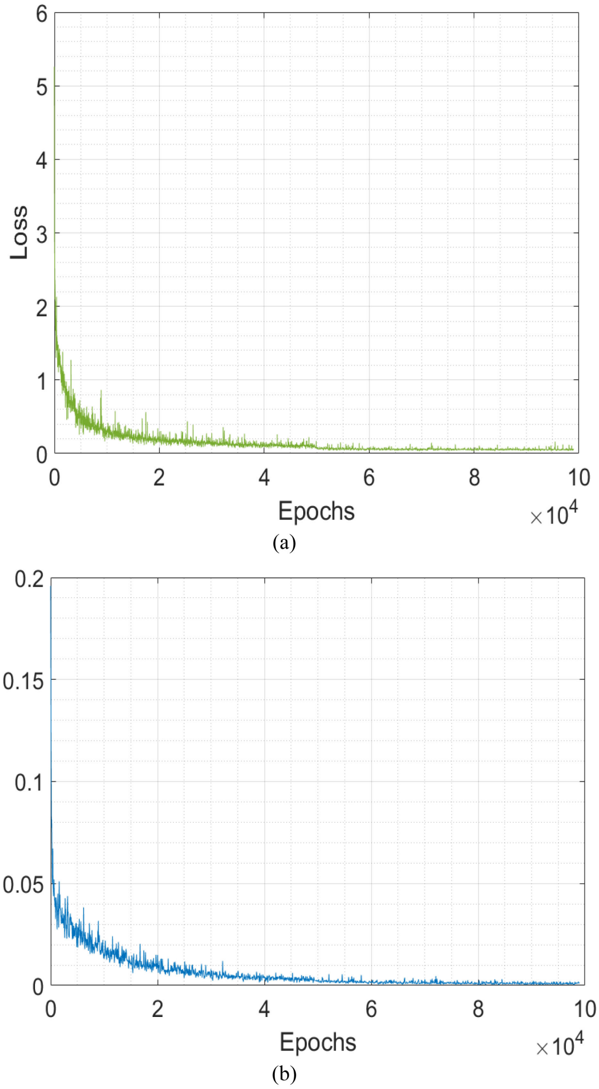


Fig. 8. Convergence of the loss function for the confidence score. (a) Traditional proposals regression loss function. (b) IoU_SS loss function. After 10^5 iterations, (b) converges faster than (a) and the effect is improved.

where N_{TD} is the number of ship targets detected correctly (true detection, TD), N_{GT} is the actual number of ship targets (ground truth, GT), N_{FD} is the number of incorrectly detected targets (false detection, FD), N_{FN} is the number of missed detected targets (false negative, FN), and N is the total number of all ship targets detected. The average precision (AP) is defined as

$$\mathrm{AP} = \int_0^1 P(R)dR \tag{14}$$

where P represents the precision, R represents the recall, and P is a function that takes R as a parameter, which is equal to taking the area under the curve. Different IoU thresholds can calculate the different numbers of ROIs and subsequently detect different N_{TD}.

For the calculation of mAP, each IoU threshold corresponded to an AP value. mAP denotes the mean of AP values, which helps assess the detection effect of the model. n denotes the number of scene categories

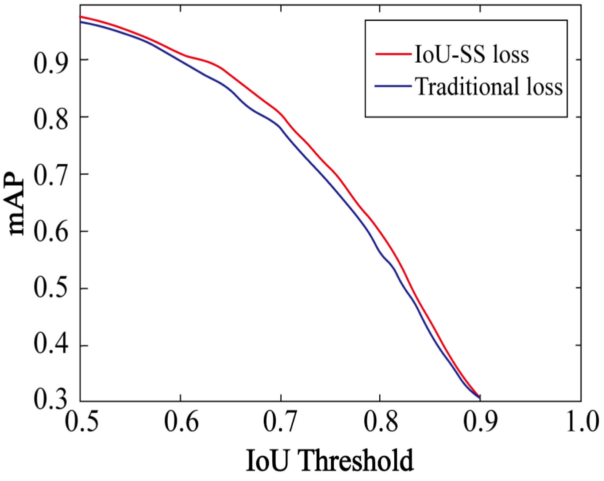


Fig. 9. mAP value against different IoU thresholds, for the proposed model trained using IoU_SS loss function (red) and the traditional regression loss function (blue).

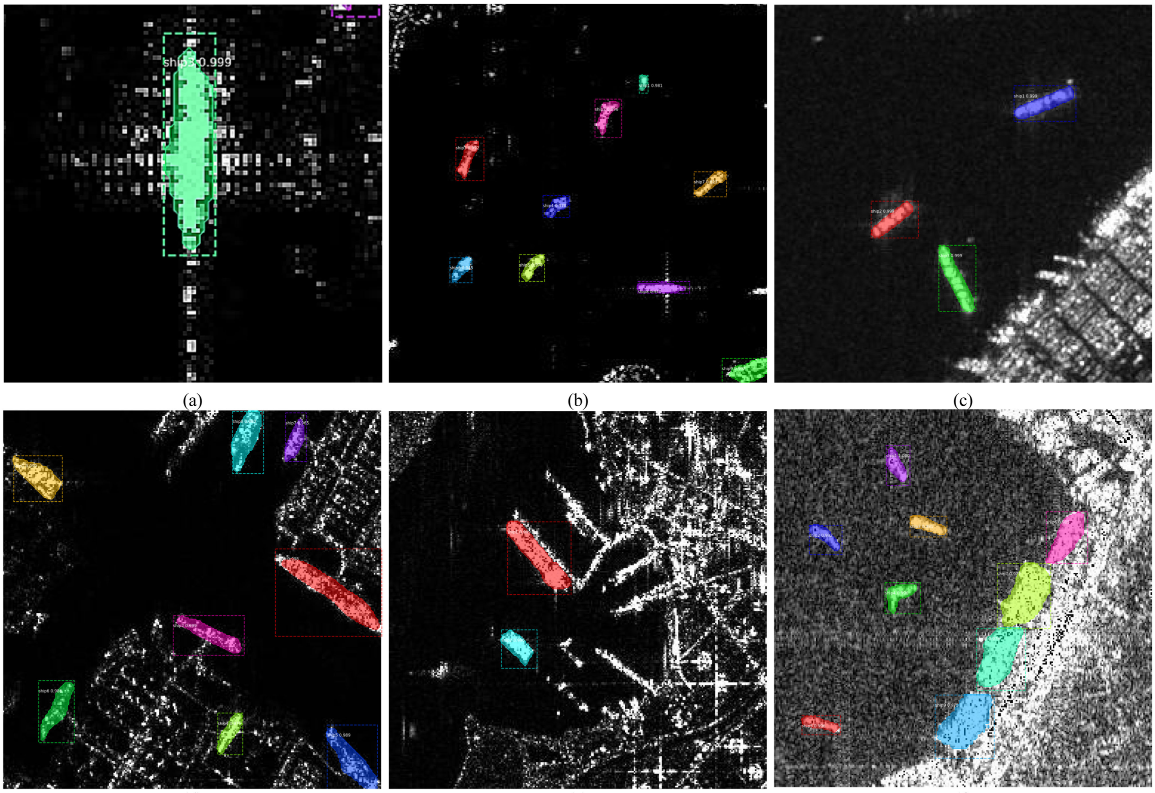
$$\mathrm{mAP} = \frac{\sum_{i=1}^{n} \mathrm{AP}i}{n}. \tag{15}$$

3) Implementation Details: Ubuntu 14.0 was used to conduct the experiments, the GPU was NVIDIA Tesla V100, and the development platform was Keras. During the hybrid model experiments, ResNet-101 was used as the backbone network. To verify the effectiveness of the model, IoU was introduced into the loss function to replace the traditional bounding-box regression loss function. The training schedule was the same as that of the FPN. Under a larger batch size setting, the Adam optimizer was adopted to optimize the gradient descent. The values of the empirical learning rate, batch size, moment, and momentum were 0.0001, 64, 0.99, and 0.0001, respectively.

B. Experimental Procedure

In this section, we present the effectiveness of each module, as given in Table III.

A+Segmentation (Step B) outperformed the baseline FPN-Resnet (Step A). The mAP in step B increased by 3.91%, and for scenes of the island, harbor, and offshore, the AP increased by 1.64%, 2.75%, and 7.37%, respectively. This is because segmentation was used to predict whether each pixel in the image belonged to the ship. Compared with results obtained in step B, the introduction of IoU loss in step C increased the mAP by 5.85%, and the AP increased significantly for the harbor and offshore scenarios by 7.37% (80.68–88.05%) and 9.22% (77.69–86.91%), respectively. In step D, we replaced the IoU loss with IoU_SS loss and incorporated the shape similarity in the loss function. As a result, the mAP increased to 92.63%. In summary, when using the proposed method (Steps A–D), the mAP increased from 80.66% to 92.63%. The convergence of different loss functions is shown in Fig. 8, and the number of iterations is 10^5. The hybrid model with IoU_SS can effectively distinguish ships from their complex backgrounds, which proves that the proposed method performs robustly even in complex



(a) (b) (c)

(d) (e) (f)

Fig. 10. Ship-detection results of the proposed hybrid model with IoU_SS for certain typical scenes. (a) Shows a single ship; the rectangular frame is the calculated positioning frame, the confidence rate is 0.999, and the ship object divided by pixels is marked in green. (b) Shows the result for open sea water. (c)–(f) Depict complex scenes, where (c) represents offshore and (d)–(f) contain building, dock, and harbor scenes.

scenes. Fig. 9 shows a comparison of the IoU_SS loss function and the traditional regression loss function. We can observe that the varying mAP values correspond to the IoU threshold; i.e., $0.5 \leq \text{IoU} \leq 0.95$. Furthermore, it is evident that under different thresholds, the IoU_SS loss function yields a higher mAP than that yielded by the traditional loss function throughout the training process.

The results presented in Fig. 10 indicate that the hybrid model can detect multiscale ship objects under various scenarios. As shown in Fig. 10(a), each ship detected by the hybrid model is marked pixel-by-pixel and equipped with a detection frame. For a low-interference scenario [such as the open sea, as shown in Fig. 10(b)], the model could produce accurate detection results even with the presence of multiple dense ship targets. Moreover, when a ship was located in a complex scenario, such as land, harbor, or island, as shown in Fig. 10(c)–(f), the hybrid model with the novel loss function could detect the ship. These results suggest that the model has high accuracy and can be directly used to detect ship targets through SAR images without requiring sea–land segmentation.

As the backbone network model, we used FPN, segmentation (FCN), and the hybrid model proposed in this article as a comparison and followed their training protocol using the reported default parameters and the number of iterations on each benchmark. Fig. 11 shows the precision–recall curves of FPN, segmentation (FCN), and hybrid model on the SAR dataset.

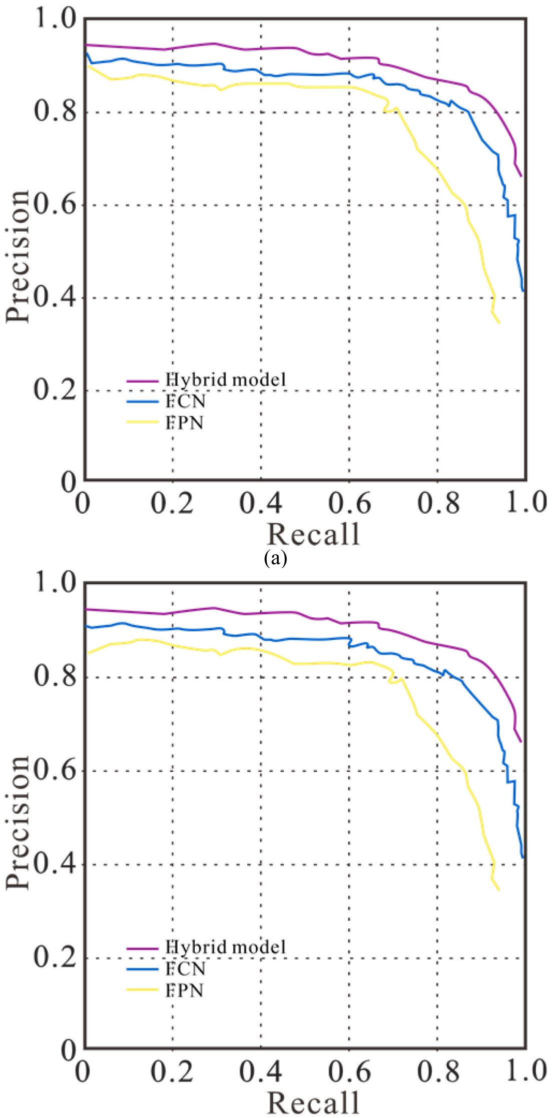
Fig. 11 shows that the precision and recall of the hybrid model are higher than those of FPN and FCN, respectively, whether in open sea water (a) or in complex scenarios (b).

C. Comparison With IoU_SS Loss and Other Loss Functions

We implemented Faster R-CNN [30], FPN [15], and Mask R-CNN [31] in Keras and trained these using regression loss to obtain the baseline results (trained using regression loss). To train Faster R-CNN, FPN, and Mask R-CNN using IoU and IoU_SS losses, we replaced their original loss in the final bounding-box refinement stage with L_{IoU} and $L_{\text{IoU\_SS}}$ losses, as shown in (3). Similar to the experiment on the models, we regularized the new regression loss against the other losses, such as classification and segmentation losses. The final results on the SAR ship dataset have been presented in Tables IV–VI. The mAP of the hybrid model proposed herein is increased by 6.78%, 3.46%, and 2.34% as compared with Faster R-CNN, FPN, and Mask R-CNN, respectively.

D. Comparison With Existing Detection Models

The proposed hybrid model was compared with SSD [14], Faster R-CNN [30], FPN [15], and Mask R-CNN [31] using the Keras platform. The learning rate for SSD was set at an experimental value of 0.00001. The batch size was chosen as 18 for SSD, the moment was set at 0.99, the momentum was set at 0.0005, and the number of iterations was 100 000. The



(a)

(b)

Fig. 11. PR curves of FPN, segmentation (FCN), and hybrid model on the different datasets. (a) PR curves on SAR ships. (b) PR curves on SAR ships in complex scenes.

TABLE IV
COMPARISON BETWEEN THE PERFORMANCES OF FAST R-CNN TRAINED USING ITS OWN LOSS (REGRESSION LOSS) AS WELL AS L_{IoU} AND $L_{\mathrm{IoU\_SS}}$ LOSSES

Loss/Evalutation	mAP(%)
Regression Loss	84.42%
IoU loss	87.95%
IoU_SS loss	91.20%

Results are reported on the set of SAR ship images.

TABLE V
COMPARISON BETWEEN THE PERFORMANCES OF FPN TRAINED USING ITS OWN LOSS (REGRESSION LOSS) AS WELL AS L_{IoU} AND $L_{\mathrm{IoU\_SS}}$ LOSSES

Loss/Evalutation	mAP(%)
Regression Loss	89.86%
IoU loss	91.93%
IoU_SS loss	93.32%

Results are reported on the set of SAR ship images.

TABLE VI
COMPARISON BETWEEN THE PERFORMANCES OF MASK R-CNN TRAINED USING ITS OWN LOSS (REGRESSION LOSS) AS WELL AS L_{IoU} AND $L_{\mathrm{IoU\_SS}}$ LOSSES

Loss/Evalutation	mAP(%)
Regression Loss	91.38%
IoU loss	91.90%
IoU_SS loss	93.72%

Results are reported on the set of SAR ship images.

TABLE VII
AP (%) OF DETECTION UNDER DIFFERENT SCENARIOS ENTRIES WITH THE BEST APS FOR EACH SCENE ARE BOLDFACED

Model	Island	Harbor	Offshore	mAP (%)
SSD	85.93	70.81	67.16	74.63
Faster R-CNN	89.95	76.69	71.33	79.32
FPN	93.72	77.93	70.32	80.66
Mask R-CNN	95.36	80.68	77.69	84.57
Hybrid model with IoU_SS	96.30	94.44	87.17	92.63

parameters for Faster R-CNN, FPN, and Mask R-CNN were as follows: learning rate = 0.0001, batch size = 64, moment = 0.99, momentum = 0.0001, and number of iterations = 100 000. The backbone in SSD was VGG19. The Faster R-CNN, FPN, Mask R-CNN, and hybrid models used ResNet-101 as the backbone.

As given in Table VII, in terms of the mAP, the proposed hybrid model with IoU_SS clearly outperformed SSD by 18.00% (92.63% versus 74.63%) and Faster R-CNN by 13.31% (92.63% versus 79.32%). Moreover, for complex scenarios, the mAP of the proposed model significantly increased by 11.97% and 8.06%, compared with FPN and Mask R-CNN, respectively. Of these scenarios, the most significant improvement was noted for the harbor scene, with the AP of FPN increasing by 16.51% (94.44% versus 77.93%) and that of Mask R-CNN increasing by 13.76% (94.44% versus 80.68%). The interferences in the harbor scene, such as docks and buildings, resulted in a lower accuracy for the traditional method. Furthermore, the proposed model was more effective in identifying ships under extremely complex scenes.

As given in Table VIII, the missed detection rate of the proposed hybrid model was reduced by 2.03%, 4.93%, 4.32%, and 1.28%, compared with that of SSD, Faster R-CNN, FPN, and Mask R-CNN, respectively. This shows that under complex scenarios, the hybrid model could detect ships more effectively than the other models could. Although Mask R-CNN and the proposed model exhibited lower missed detection rates than those of the other models, the missed detection rate of the proposed hybrid model was lower for scenes involving more

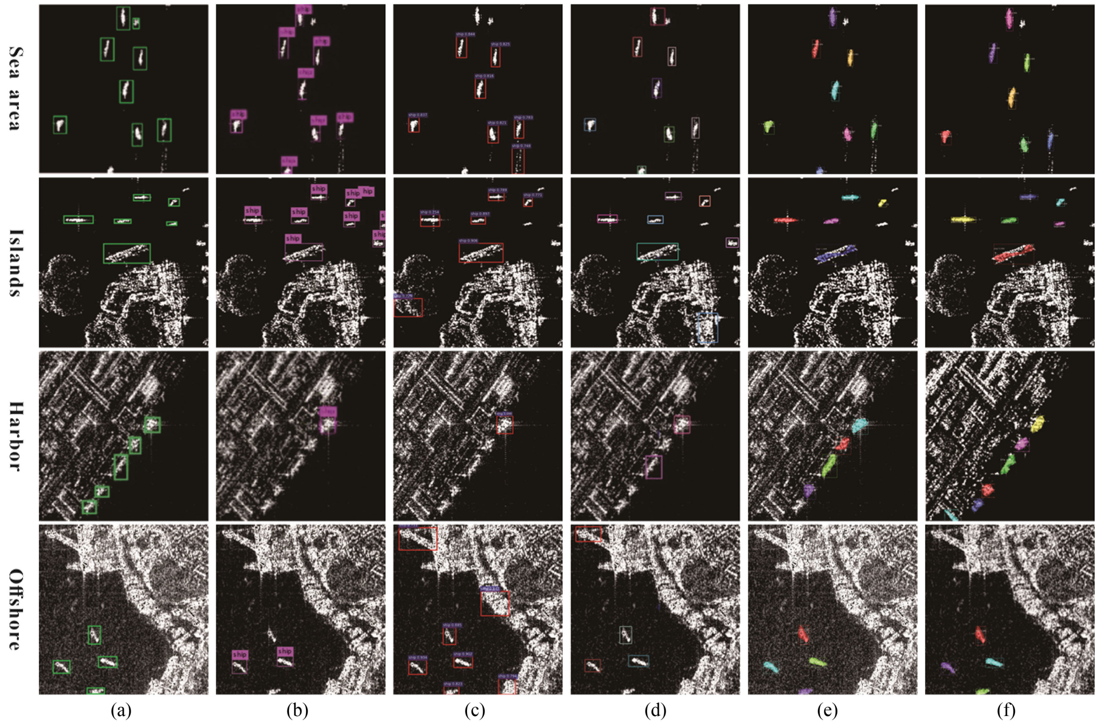


Fig. 12. Comparison of detection results for different models. The green bounding boxes indicate the actual location of ships in the images, and the remaining bounding boxes represent the detected ships from four different models.

TABLE VIII
MISSED DETECTION RATE P_M (%) UNDER DIFFERENT SCENARIOS

Model	Island	Harbor	Offshore	Average
SSD	7.60	11.7	11.8	10.4
Faster R-CNN	7.32	15.45	17.3	13.3
FPN	7.29	14.28	16.5	12.69
Mask R-CNN	7.25	10.0	11.7	9.65
Hybrid model with IoU_SS	7.25	8.51	9.34	8.37

complex backgrounds (such as the harbor and offshore scenes). Furthermore, the hybrid model outperformed the Mask R-CNN.

Fig. 12 presents the detection results for SSD, Faster R-CNN, FPN, Mask R-CNN, and the proposed method. The first column, as shown in Fig. 10(a), depicts the ground truth. Ships appear as bright spots in SAR images. In complex scenes, objects, such as buildings and docks, often backscatter and appear as bright spots, similar to that of the ship target. This causes several false alarms leading to false positives. This is a huge setback in the development of ship-target detection. SSD achieved a high missed detection rate for small ships and berthing ships, resulting in a low recognition rate. In addition, SSD exhibited a high false alarm rate when the target was close to or near the edge of the image. Faster R-CNN exhibited better recognition effect for ships of different scales. However, in some complex scenes, owing to significant speckle noise and interference from numerous objects, objects similar to ships were misidentified as ships. The detection rate of FPN was slightly higher than those of the first two methods because the pyramid feature map was added in the FPN. However, for the harbor and offshore scenes, the recognition rate of FPN decreased when the ships were at the edges or when multiple objects were significantly close to each other. Mask R-CNN exhibited a high recognition rate for different scenes; however, some targets at the edge of the image were not detected. These four detection models delivered lower detection accuracies than that of the proposed hybrid model. Additionally, they exhibited high rates of missed or false detections under the complex scenarios involving high interference from harbors and offshore. Under the harbor scenario, SSD, Faster R-CNN, and FPN achieved a detection precision of less than 81%, whereas the hybrid model successfully detected all ships in the image, achieving a detection precision of 94.44%. For the hybrid model, significant speckle noise and high interference from numerous objects similar to ships were noted in the offshore scenario, and although a false positive was produced, a higher detection accuracy than those of the other detection models was retained. Furthermore, in complex scenes, where other objects interfered with ship targets, the proposed hybrid model could effectively distinguish the ship targets from the complex background, proving that the hybrid model with SS achieves robust performance even under complex scenes.

IV. CONCLUSION

Herein, we proposed a model that combines object detection and segmentation with a novel multitask loss function. The hybrid model was designed primarily for ship-target detection in SAR images under complex scenarios. The results revealed that the hybrid model could effectively extract ship targets under complex scenarios involving high background interference. The experimental results indicate that the object-segmentation algorithm can be integrated into the proposed hybrid model to extract the boundary information of targets. Furthermore, the IoU_SS multitask loss function with the shape similarity loss afforded better convergence than the traditional loss function while ensuring consistency in optimization and evaluation. Compared with the other detection models employed for performance verification of the proposed model, the proposed hybrid model achieved superior detection accuracy and significantly reduced rates of false positives during ship-target detection under complex scenarios, such as islands, harbors, and offshore areas. Although the proposed low false positives, we intend to combine the characteristics of SAR images with improved models to further reduce these false negatives in future research.

References

[1] K. Eldhuset, “An automatic ship and ship wake detection system for spaceborne SAR images in coastal regions,” *IEEE Trans. Geosci. Remote Sens.*, vol. 34, no. 4, pp. 1010–1019, Jul. 1996.

[2] N. Yokoya and A. Iwasaki, “Object detection based on sparse representation and Hough voting for optical remote sensing imagery,” *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 8, no. 5, pp. 2053–2062, May 2015.

[3] B. Tings, C. Bentes, D. Velotto, and S. Voinov, “Modelling ship detectability depending on TerraSAR-X-derived metocean parameters,” *CEAS Space J.*, vol. 11, no. 1, pp. 81–94, 2019.

[4] C. Brekke and A. H. S. Solberg, “Oil spill detection by satellite remote sensing,” *Remote Sens. Environ.*, vol. 95, no. 1, pp. 1–13, 2005.

[5] Y. Wang, C. Wang, H. Zhang, Y. Dong, and S. Wei, “Automatic ship detection based on retinanet using multi-resolution Gaofen-3 imagery,” *Remote Sens.*, vol. 11, no. 5, 2019, Art. no. 531.

[6] C. Wang, H. Zhang, F. Wu, S. Jiang, B. Zhang, and Y. Tang, “A novel hierarchical ship classifier for COSMO-SkyMed SAR data,” *IEEE Geosci. Remote Sens. Lett.*, vol. 11, no. 2, pp. 484–488, Feb. 2014.

[7] G. Cheng, P. Zhou, and J. Han, “Learning rotation-invariant convolutional neural networks for object detection in VHR optical remote sensing images,” *IEEE Trans. Geosci. Remote Sens.*, vol. 54, no. 12, pp. 7405–7415, Dec. 2016.

[8] C. Zhu, H. Zhou, R. Wang, and J. Guo, “A novel hierarchical method of ship detection from spaceborne optical image based on shape and texture features,” *IEEE Trans. Geosci. Remote Sens.*, vol. 48, no. 9, pp. 3446–3456, Sep. 2010.

[9] Z. Shi, X. Yu, Z. Jiang, and B. Li, “Ship detection in high-resolution optical imagery based on anomaly detector and local shape feature,” *IEEE Trans. Geosci. Remote Sens.*, vol. 52, no. 8, pp. 4511–4523, Aug. 2014.

[10] W. Zhang, X. Sun, H. Wang, and K. Fu, “A generic discriminative part-based model for geospatial object detection in optical remote sensing images,” *ISPRS J. Photogrammetry Remote Sens.*, vol. 99, pp. 30–44, 2015.

[11] Z. Huang, Z. Pan, and B. Lei, “Transfer learning with deep convolutional neural network for SAR target classification with limited labeled data,” *Remote Sens.*, vol. 9, no. 9, 2017, Art. no. 907.

[12] Y. Liu, M.-H. Zhang, P. Xu, and Z.-W. Guo, “SAR ship detection using sea-land segmentation-based convolutional neural network,” in *Proc. Int. Workshop Remote Sens. Intell. Process.*, 2017, pp. 1–4.

[13] M. Kang, X. Leng, Z. Lin, and K. Ji, “A modified faster R-CNN based on CFAR algorithm for SAR ship detection,” in *Proc. Int. Workshop Remote Sens. Intell. Process.*, 2017, pp. 1–4.

[14] Y. Wang, C. Wang, and H. Zhang, “Combining a single shot multibox detector with transfer learning for ship detection using Sentinel-1 SAR images,” *Remote Sens. Lett.*, vol. 9, no. 8, pp. 780–788, 2018.

[15] Z. Cui, Q. Li, Z. Cao, and N. Liu, “Dense attention pyramid networks for multi-scale ship detection in SAR images,” *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 11, pp. 8983–8997, Nov. 2019.

[16] J. Li, C. Qu, S. Peng, and Y. Jiang, “Ship detection in SAR images based on generative adversarial network and online hard examples mining,” *J. Electron. Inf. Technol.*, vol. 41, pp. 143–149, 2019.

[17] S. Wei, P. Jiang, Q. Yuan, and M. Liu, “Detection and recognition of SAR small ship objects using deep neural network,” *Xibei Gongye Daxue Xuebao/J. Northwestern Polytechnical Univ.*, vol. 37, no. 3, pp. 587–593, 2019.

[18] X. Nie, M. Duan, H. Ding, B. Hu, and E. K. Wong, “Attention Mask R-CNN for ship detection and segmentation from remote sensing images,” *IEEE Access*, vol. 8, pp. 9325–9334, 2020.

[19] Y. Sun, W. Lei, and X. Ren, “Remote sensing image ship target detection method based on visual attention model,” *Proc. SPIE*, vol. 10605, 2017, Art. no. 106053Z.

[20] Q. Zhao *et al.*, “M2det: A single-shot object detector based on multilevel feature pyramid network,” in *Proc. AAAI Conf. Artif. Intell.*, 2019, pp. 9259–9266.

[21] G. Ghiasi, T.-Y. Lin, and Q. V. Le, “NAS-FPN: Learning scalable feature pyramid architecture for object detection,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 7036–7045.

[22] T. Zhang, X. Zhang, J. Shi, and S. Wei, “HyperLi-Net: A hyper-light deep learning network for high-accurate and high-speed ship detection from synthetic aperture radar imagery,” *ISPRS J. Photogrammetry Remote Sens.*, vol. 167, pp. 123–153, 2020.

[23] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, “You only look once: Unified, real-time object detection,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 779–788.

[24] J. Ding, B. Chen, H. Liu, and M. Huang, “Convolutional neural network with data augmentation for SAR target recognition,” *IEEE Geosci. Remote Sens. Lett.*, vol. 13, no. 3, pp. 364–368, Mar. 2016.

[25] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, “Feature pyramid networks for object detection,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 2117–2125.

[26] W. Ao, F. Xu, Y. Qian, and Q. Guo, “Feature clustering based discrimination of ship targets for SAR images,” *J. Eng.*, vol. 2019, no. 20, pp. 6920–6922, 2019.

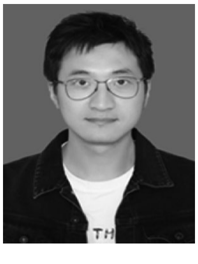
[27] F. Zhang, Y. Wang, J. Ni, Y. Zhou, and W. Hu, “SAR target small sample recognition based on CNN cascaded features and AdaBoost rotation forest,” *IEEE Geosci. Remote Sens. Lett.*, vol. 17, no. 6, pp. 1008–1012, Jun. 2020.

[28] H. Ling and D. W. Jacobs, “Shape classification using the inner-distance,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 29, no. 2, pp. 286–299, Feb. 2007.

[29] Y. Wang, C. Wang, H. Zhang, Y. Dong, and S. Wei, “A SAR dataset of ship detection for deep learning under complex backgrounds,” *Remote Sens.*, vol. 11, no. 7, 2019, Art. no. 765.

[30] 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.

[31] S. Wei *et al.*, “Precise and robust ship detection for high-resolution SAR imagery based on HR-SDNet,” *Remote Sens.*, vol. 12, no. 1, 2020, Art. no. 167.



Peng Chen was born in Dalian, Liaoning, China, in 1982. He received the B.S. degree in geographical information science from the China University of Geosciences, Wuhan, China, in 2005, and the Ph.D. degree in environmental science from Dalian Maritime University, Dalian, China, in 2012.

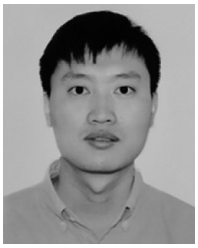
Since 2012, he has been with the Faculty of Navigation College, Dalian Maritime University, where he is currently an Associate Professor of traffic engineering. His research interests include machine learning, remote sensing image processing, and modeling and simulation of complex systems.

Hui Zhou was born in Hubei, China, in 1983. She received the B.Sc. and M.Sc. degrees in geographic information systems from the China University of Geosciences, Wuhan, China, in 2005 and 2008, respectively.

Her research interests include machine learning, data mining, pattern recognition, and object detection.

Ying Li was born in Liaoning, China, in 1968. She received the Ph.D. degree in geosciences from Tohoku University, Sendai, Japan, in 1996.

She is currently a Professor with Navigation College, Dalian Maritime University, Dalian, China. She has authored more than 100 refereed journals and conference papers. Her research interests include geographic information science, marine environment information technology, and ship pollution detection.



Bingxin Liu received the Ph.D. degree in environmental science from Dalian Maritime University, Dalian, China, in 2013.

He is currently an Associate Professor with Navigation College, Dalian Maritime University. His research interests include remote sensing mechanisms and deep learning.

Peng Liu received the Ph.D. degree in marine science from Kobe University, Kobe, Japan, in 2015.

He is currently an Associate Professor with Navigation College, Dalian Maritime University, Dalian, China. His research interests include image processing, data mining, and pattern recognition.