Inshore Dense Ship Detection in SAR Images Based on Edge Semantic Decoupling and Transformer

Yongsheng Zhou, *Member, IEEE*, Feixiang Zhang, Qiang Yin, *Senior Member, IEEE*, Fei Ma, *Member, IEEE*, and Fan Zhang, *Senior Member, IEEE*

***Abstract*—Synthetic aperture radar ship detection has recently received significant attention from scholars. However, accurately distinguishing between ships is challenging due to the significant overlap between inshore ship labels. In addition, some labeled boxes contain interference information, such as land areas, which can cause false alarms and confusion in ship feature learning. To address these challenges, this article creates an edge semantic decoupling (ESD) module, adds semantic segmentation branches, and introduces the edge semantic information of ships into the training process. As a result, the model can accurately distinguish between ship targets even when significant overlap exists between inshore labeled boxes. In addition, considering that transformer has the benefit of capturing global and contextual information, this article introduces it into the detection layer to construct a transformer detection layer (TDL) to limit the interference of land and other regions within the labeled box. Experimental results from the public SAR ship detection dataset show that the proposed ESD module and TDL detection layer effectively distinguish different ship targets in the inshore dense ship area, which is less affected by interference areas, such as land in the labeled box. The average precision improves to 96.72%, and both false alarms and miss detections inshore are reduced.**

***Index Terms*—Edge semantic decoupling (ESD), inshore ship detection, transformer.**

I. Introduction

Synthetic aperture radar (SAR) has the capability to conduct all-day and all-weather observations, allowing for long-term monitoring of the ocean without the interference of weather conditions like cloud cover, fog, etc., [1], [2], [3], [4], [5], [6]. With the increase of public datasets and the development of convolution neural networks (CNN), more and more researchers are utilizing CNNs for SAR ship detection [7], [8], [9], [10], [11], [12], [13].

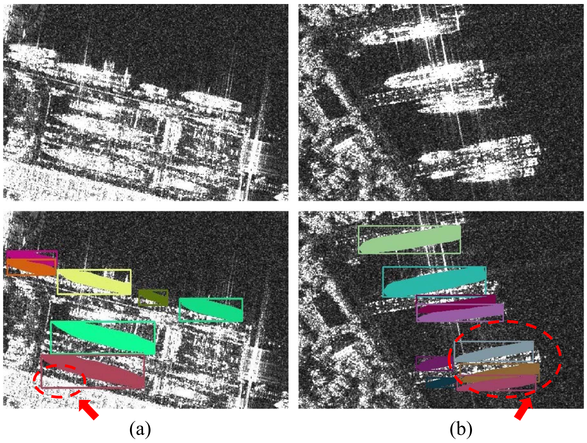


Fig. 1. Issue of labeled boxes for inshore ships. (a) Contain interference information. (b) Labeled boxes overlapped.

Current methods for SAR ship detection using CNNs can be classified into two categories as follows: 1) single-stage and 2) two-stage detection. The two-stage detection method involves a two-step process where the detection box is first coarsely extracted using region proposal network (RPN) [14], followed by regression and classification of the box. Some representative methods using this approach are faster region-CNN (R-CNN) [14] and cascade R-CNN [15]. The single-stage detection method has the advantage of being faster and more efficient, as it can perform regression and classification without the need for coarse extraction. Some representative methods using this approach are YOLOv3 [16] and YOLOv4 [17]. In consideration of practical application requirements, the current trend in SAR ship detection primarily favors the single-stage detection method due to its speed and simplicity.

The detection of ships in SAR images using CNN-based methods requires a large number of labels. However, dense inshore ship labels often overlap, making it challenging to differentiate between different ship targets. As a result, the dense region is prone to miss detection, as illustrated in Fig. 1. Tian et al. [18] attempted to solve this issue by utilizing rotating boxes for detection. Rotating box labeling has been applied to mitigate a significant portion of the inshore interference. However, this approach faces a limitation in its ability to include contextual semantic information, resulting in a higher likelihood of false alarms in inshore scenarios. Conversely, the horizontal box contains richer contextual information; but, its susceptibility to

Manuscript received 1 March 2023; revised 3 April 2023 and 28 April 2023; accepted 9 May 2023. Date of publication 17 May 2023; date of current version 5 June 2023. This work was supported in part by the National Natural Science Foundation of China under Grant 62171016 and Grant 62271034, and in part by the Fundamental Research Funds for the Central Universities under Grant buctrc202001 and Grant XK2020-03. *(Corresponding author: Qiang Yin.)*

Yongsheng Zhou, Feixiang Zhang, Qiang Yin, and Fei Ma are with the College of Information Science and Technology, Beijing University of Chemical Technology, Beijing 100029, China (e-mail: zhyosh@mail.buct.edu.cn; 2019200710@mail.buct.edu.cn; yinq@mail.buct.edu.cn; mafei@mail.buct.edu.cn).

Fan Zhang is with the College of Information Science and Technology, Interdisciplinary Research Center for Artificial Intelligence, Beijing University of Chemical Technology, Beijing 100029, China (e-mail: zhangf@mail.buct.edu.cn).

Digital Object Identifier 10.1109/JSTARS.2023.3277013

higher levels of inshore interference creates a pressing issue for detection of inshore ships. One critical challenge lies in how to suppress the inshore interference when working with a horizontal box containing significant contextual semantic information. Another approach was used by Ma et al. [19] who employed key point estimation to differentiate individual targets in dense inshore ships. However, this method could only highlight the central area of the ship, not the edges, resulting in a biased fit of inshore ship labels in the dense region. Su et al. [20] and Wu et al. [21] attempted to overcome the problem of dense overlap by merging the edge semantic information of ships and employing instance segmentation methods for detection. However, their algorithms were complex and inefficient to implement. Ge et al. [22] demonstrated that decoupling the detection layer can improve the regression and classification performance of the model. Decoupling can be utilized to construct a simple and easy-to-implement module that introduces inshore edge semantic information about ships, thereby addressing the miss detection problem in dense inshore scenarios.

The inshore ship labels often contain interference information, such as land, which can lead to land false alarms and mislead ship feature learning, as illustrated in Fig. 1. Wang et al. [23] and Hou et al. [24] showed that introducing contextual semantic information in combination with the scene effectively reduces land false alarms in the inshore region. Ke et al. [25] expanded the encoding and increased contextual information to obtain feature maps of multiple sensory fields, which was more effective but computationally complex and unsuitable for practical applications. Zhu et al. [26] demonstrated that introducing transformer can capture contextual information more comprehensively, especially for high-density occlusion objects, with minimal computational overhead. The structure of the transformer is composed of an encoder and decoder, which can better obtain the contextual semantic information of the target, improve the feature extraction ability, and better locate the edges of the target in the target detection [27]. Therefore, transformer can be introduced to build a plug-and-play module for learning contextual information to reduce the land false alarms caused by land area interference in ship labels.

In this article, an SAR ship detection method based on edge semantic decoupling and transformer is proposed to address the issue of inshore dense ship detection. To tackle the challenge of miss detection caused by inshore label overlap, a semantic segmentation layer is added by decoupling the detection layer, thereby enhancing the model's ability to differentiate ship edges and reduce miss detection in dense scenes. Furthermore, to mitigate the interference from regions, such as land in the labeled boxes and facilitate feature learning, a transformer detection layer is constructed that leverages the transformer's capacity to capture global and contextual information. This enables the model to better distinguish between inshore false alarm targets and ship targets, leading to a reduction in land false alarms.

In summary, it is worthwhile to note the following contributions of the proposed method.

1) The edge semantic decoupling (ESD) module is introduced to address the challenge of distinguishing between dense inshore ship targets in SAR images. By adding semantic segmentation branches and incorporating edge semantic information, the model is able to accurately discriminate between ships even in regions with significant overlap between inshore labeled boxes.
2) The transformer detection layer (TDL) is introduced to limit interference caused by land areas and other regions within the labeled box. By taking advantage of the transformer's ability to capture global and contextual information, the TDL helps to reduce false alarms and improve the accuracy of ship target detection.

The rest of this article is organized as follows. Section II presents the proposed method. In Section III, the proposed method is validated by comparison to other methods. Section IV presents discussions. Finally, Section V concludes this article.

II. Methodology

Fig. 2 illustrates the overall structure of the proposed method. It consists of the feature extraction and detection layer parts, with the solid orange line representing the improved ESD module and the transformer-based TDL module. In Section II-A, the ESD module designed for SAR images of dense inshore ships is introduced first. Then the design details of the TDL module are presented. Finally, the decoupling loss function is presented.

A. Edge Semantic Decoupling Module

Inshore ships are known to have a dense appearance with significant labeled box overlap, which can negatively impact the model's ability to assess individual ship targets, resulting in the inclusion of several ship targets within a detection box (i.e., missed detection), as illustrated in Fig. 1(b). Conventional single-stage detection algorithms utilize a single branch to simultaneously handle the tasks of classification and detection box coordinate regression. However, the goals of classification and localization is different, as classification is primarily concerned with the texture information of the target, while localization is focused on the edge information of the target. This difference in focus can lead to conflicts between the two tasks as follows.

1) Higher-level convolutional fields have a larger receptive field, allowing them to extract more global information, which is useful for classification. However, the corresponding areas in the original image become larger, which can be detrimental to localization. Therefore, while the information contained in higher level feature maps is advantageous for classification, it is not necessarily helpful for localization.
2) Lower-level convolution and other operations correspond to smaller areas of the original image, making them more accurate for localization. However, they may only contain local information about the object and therefore are not suitable for classification. Consequently, the information contained in the lower-level feature map is suitable for target localization but not for classification.

In order to overcome the limitations of performing the two tasks in one branch, the approach proposed in this article is inspired by [22] and [28], which separates the tasks of classification and detection box coordinate regression into two distinct

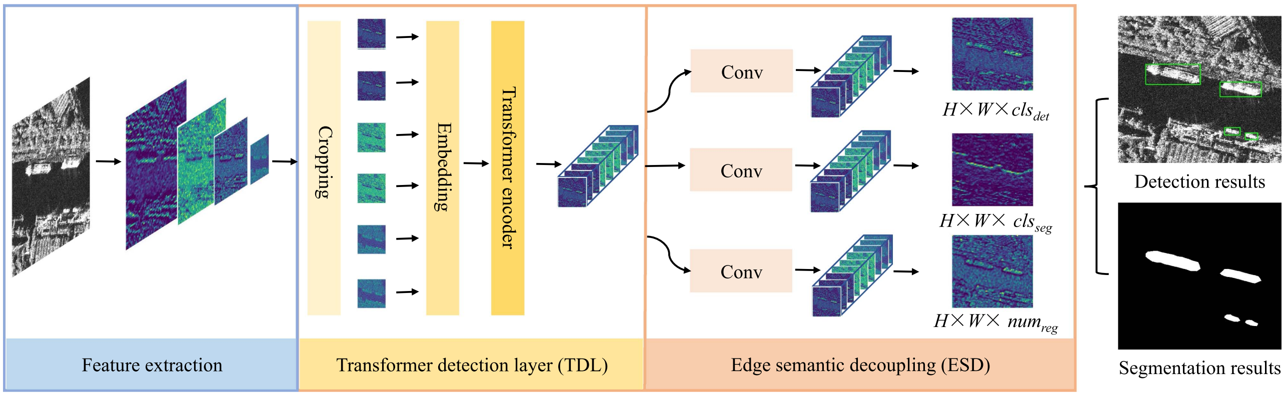


Fig. 2. Overall structure of the method proposed in this article.

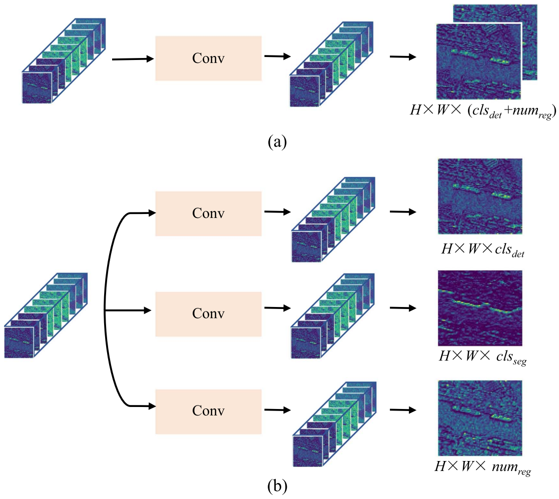


Fig. 3. (a) Single-branch structure. (b) Edge semantic decoupling (ESD) structure.

branches. Specifically, an additional ship semantic segmentation branch is incorporated to capture edge information of ships, enhancing the model's ability to differentiate between different ship targets in dense scenes and reducing the occurrence of missed detections.

The ESD module and the conventional single branch are compared in the middle of Fig. 3. It can be observed that the decoupled detection head performs multiple tasks, obtaining the results of both ship detection and semantic segmentation simultaneously, compared to the conventional single branch. All three branches, regression, classification, and segmentation, are used separately after the backbone network to further specialize in learning features using more convolutional layers, decoupling while making the learned features richer and helping to further define the ship's position and edges. This not only introduces semantic information about the ship's edge but also significantly improves the model's scalability and ability to carry out multiple tasks. An example of conventional single-branch decoding is provided as follows for reference:

$$\text{Infer}_{\text{single}} = \text{Conv}\left(ch, \text{cls}_{\text{det}} + \text{num}_{\text{reg}}\right) \tag{1}$$

where $\text{Infer}_{\text{single}}$ represents the prediction result of the conventional single-branch structure, and Conv represents the 2-D convolution. ch represents the number of the channels of the feature map extracted by the detection layer, which is set to 255 in this article. cls_{det} represents the number of detected target categories, which is set to 1 as there is only one ship category in the detection task of this article. num_{reg} represents the coordinate values of the regression. It is set to 4, which corresponds to the upper left and lower right horizontal and vertical coordinates of the detection box. The edge semantic decoupling is as follows:

$$\begin{aligned}\text{Infer}_{\text{decouple}} &= \text{Conv}\left(ch, \text{cls}_{\text{det}}\right) + \text{Conv}\left(ch, \text{num}_{\text{reg}}\right)\\ &\quad + \text{Conv}\left(ch, \text{cls}_{\text{seg}}\right)\end{aligned} \tag{2}$$

where $\text{Infer}_{\text{decouple}}$ is the prediction result of the edge semantic decoupling structure and cls_{seg} is the number of split categories. Since there is only one ship category in the detection task of this article, the value is set to 1.

B. Transformer Detection Layer Module

Due to the proximity of inshore ships to land, their labeled boxes often include land, as illustrated in Fig. 1(a). This can lead to the model mistakenly identifying certain features of the land as features of the ship in the absence of contextual information, resulting in false alarms.

Compared to CNN, the transformer architecture can efficiently extract contextual information of the target by uniformly cropping the input into multiple patches and utilizing a multiheaded attention mechanism [29]. To mitigate the impact of land areas in the detection labels, this article introduces transformer to construct the TDL module, which incorporates contextual information and enhances the model's ability to distinguish land targets, thus reducing false alarms.

The input of transformer is a 1-D sequence of token embeddings. To handle 2-D feature maps, feature maps $\mathbf{x} \in \mathbb{R}^{h \times w \times c}$ are reshaped into a sequence of flattened 2-D patches $\mathbf{x}_p \in \mathbb{R}^{n \times (p^2 \cdot c)}$, where (h, w) is the resolution of the feature map, c is the number of channels, (p, p) is the resolution of each feature patch, and $n = hw/p^2$ is the resulting number of patches. The process of patch embedding can be described as

$$\text{Output}_{\text{embedding}} = \text{Flatten}(\text{Conv}(\text{Part}(x))), \tag{3}$$

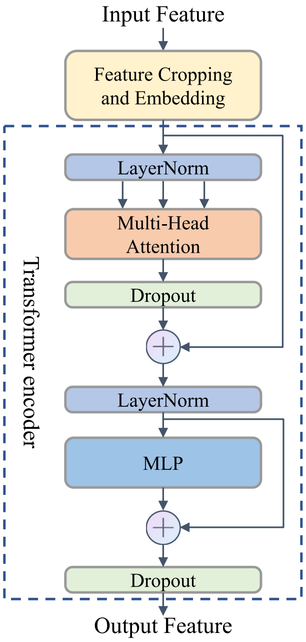


Fig. 4. TDL module structure.

where $\mathrm{Output}_{\mathrm{embedding}}$ means embedded patches, Part represents the chunking operation, which divides the input feature map into patches of a specific size. Conv is used to reduce dimensions, and Flatten is used to construct a 1-D vector by pulling flat.

The design of the transformer detection layer is illustrated in Fig. 4, which mainly consists of a multihead attention module and a feedforward neural network multilayer perceptron (MLP) module.

The multihead attention module aliquots the input $x \in R^{N\times d_{in}}$ in the feature dimension to obtain several copies of $x_i \in R^{N\times d_i}, \quad i = 1, 2, \ldots n$, where N is the sequence length, d denotes the feature dimension and $\sum_{i=1}^{n} d_i = d_{in}$. Each x_i is processed with an attention to obtain n copies of the output, which are then stitched together in the feature dimension to obtain the final result. The calculation of a single attention is as follows:

$$\mathrm{Attention}(Q, K, V) = \mathrm{Softmax}\left(\frac{QK^T}{\sqrt{d}} + B\right)V \quad (4)$$

where Q, K, and V are obtained from the input x_i through the fully connected layer and B is the position information. The components of MLP are shown as follows:

$$\mathrm{MLP} = \mathrm{drop}\left(fc(\mathrm{drop}(\mathrm{act}(fc(x))))\right) \quad (5)$$

where drop means dropout operation, fc means fully connected layer, x is the input, act represents Gaussian error linear unit (GELU) activation function as follows:

$$\begin{aligned}\mathrm{GELU}(x) &= xP(X \le x) = x\Phi(x)\\ &\approx 0.5x\left(1 + \tanh\left[\sqrt{2/\pi}\left(x + 0.044715x^3\right)\right]\right).\end{aligned} \quad (6)$$

The input of transformer encoder are embedded patches, and the added LayerNorm and Dropout layers are used to prevent overfitting. The TDL module mainly replaces one convolutional layer of the detection layer, which enhances the ability to capture diverse contextual information with only a minor increase in computational cost. It also leverages the self-attention mechanism to explore the potential of feature representation.

C. Decoupling Loss Function

The decoupling loss function in this article is designed to optimize both the detection and segmentation tasks in a single network structure. Instead of training and optimizing the two tasks individually, the decoupling loss allows for the inclusion of edge semantic information in the ship detection optimization process by back-propagating the loss after both ship segmentation and detection have been performed.

The designed decoupling loss function has two components, namely, 1) ship detection loss function and 2) ship semantic segmentation loss function.

1) Ship Detection Loss Function: The loss function of the ship detection component is as follows:

$$\mathrm{loss}_{\mathrm{det}} = \mathrm{loss}_{\mathrm{CIoU}} + \mathrm{loss}_{\mathrm{cls}}\,. \quad (7)$$

$\mathrm{loss}_{\mathrm{CIoU}}$ is the detection box regression loss. In order to more effectively filter out the high quality detection results that are closer to the labeled box, this article uses CIoU [30] as the regression loss for ship detection. The CIoU is calculated as

$$\mathrm{loss}_{\mathrm{CIoU}} = 1 - \mathrm{IoU} + \frac{\rho^2\left(B_p, B_g\right)}{c^2} + \alpha \quad (8)$$

where $\rho^2(B_p, B_g)$ represents the Euclidean distance between the center point of the detection box and the labeled box, B_p is the detection box, B_g is the labeled box, c represents the length of the diagonal between the upper left and lower right corners of the smallest outer rectangle of the detection box and the labeled box. α is a parameter to measure the consistency of the aspect ratio and v is a tradeoff parameter

$$\alpha = \frac{v}{1 - \mathrm{IoU} + v} \quad (9)$$

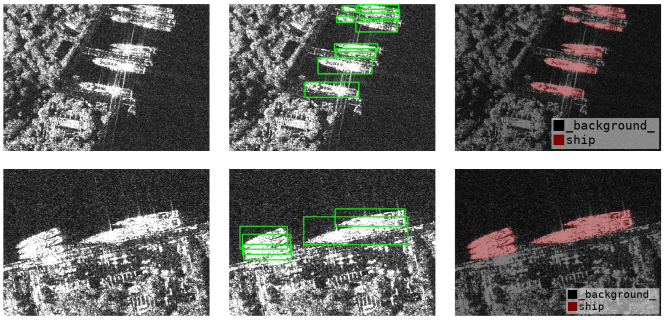
$$v = \frac{4}{\pi^2}\left(\arctan\frac{w^{gt}}{h^{gt}} - \arctan\frac{w}{h}\right)^2 \quad (10)$$

where w and h represent the width and height of the prediction box, respectively; w^{gt} and h^{gt} represent the width and height of the labeled box, respectively. The IoU is defined as

$$\mathrm{IoU}(B_p, B_g) = \frac{|B_p \cap B_g|}{|B_p \cup B_g|}. \quad (11)$$

$\mathrm{loss}_{\mathrm{cls}}$ is the category classification loss for ship detection. In the dataset used in this study, there is only one category, so the foreground and background of the ships need to be separated. The binary cross-entropy function used is shown as follows:

$$\mathrm{loss}_{\mathrm{cls}} = -\frac{1}{n}\sum_{i=1}^{n}\left[I^{\mathrm{obj}}\log d + (1 - I^{\mathrm{obj}})\log(1 - d)\right] \quad (12)$$



Original image Ship detection labels Ship segmentation labels

Fig. 5. SSDD dataset labels.

where I^{obj} is the value of the detection label: 0 for the background and 1 for the ship. d is the output of the detection layer by the Sigmoid function, and n is the number of detected samples.

2) Ship Semantic Segmentation Loss Function: Since the ship semantic segmentation also needs to distinguish between just two types of pixels, i.e., ship and background, the same binary cross-entropy function is used as (12)

$$\text{loss}_{\text{seg}} = -\frac{1}{n}\sum_{i=1}^{n}\left[I^{\text{obj}}\log s + (1 - I^{\text{obj}})\log(1 - s)\right] \tag{13}$$

where I^{obj} is the true value of the semantic segmentation of the ship edges. The value 0 indicates that the pixel belongs to the background, whereas 1 indicates that it belongs to the ship. s is the output of the segmentation layer by the Sigmoid function and n is the number of segmented samples.

The joint loss for detection and segmentation in this article is shown in (14), and the values of the optimization process are involved in the final parameter update.

$$\text{loss}_{\text{total}} = \text{loss}_{\text{det}} + \text{loss}_{\text{seg}}\,. \tag{14}$$

III. Experimental Results and Discussion

The effectiveness of the proposed method is evaluated using the publicly accessible dataset SSDD labeled by Zhang et al. [7]. This section first introduces the dataset and the hyperparameter settings of the experiments, followed by presenting the ablation experiment results of each module. Finally, the proposed method is compared with the current mainstream detection algorithms.

A. Dataset and Experimental Parameter Setting

The SSDD dataset used in this study contains SAR images with resolutions ranging from 1–15 m, sourced from RADARSAT-2, TerraSAR-X, and Sentinel-1. The dataset includes ship detection box labels as well as ship semantic segmentation labels. An example of a labeled dataset can be seen in Fig. 5.

The SSDD dataset was labeled with 1160 images containing 2456 ship targets. Among the 2456 ship targets in the dataset, 928 were used for training and 232 were used for testing. In addition, 46 of the test images were taken from the coast and 186 were taken from the ocean.

Algorithm 1: Update Parameters During Training.

```
Initialization:
accumulate = 0;
batch-size = 4;
while accumulate <= 64 do
    if accumulate<64 then
        ($\text{loss}_{\text{det}}$+$\text{loss}_{\text{seg}}$).backward();
        accumulate = accumulate+batch-size;
    else
        optimizer.step();
        optimizer.zero_grad();
        accmulate = 0;
    end
end
```

The network is implemented using the Pytorch deep learning framework. The optimizer utilized is stochastic gradient descent (SGD) with momentum. A Geforce RTX 2070 GPU is used to train 100 epochs, starting with an initial learning rate of 1e-3, momentum of 0.9, and weight decay of 5e-4. Joint training is necessary to incorporate ship edge semantic information into the optimization of the ship detection model throughout the training phase. Given that semantic segmentation and target detection have different labels, the losses are calculated separately during training. First, the segmentation and detection losses are added together for back-propagation. Then, the gradient is accumulated to a preset value and a parameter update is performed. The training process is shown in Algorithm 1.

Gradient accumulation training offers the advantage of achieving large batches even on machines with limited video memory, thereby mitigating the oscillations lost during training and allowing for faster acquisition of the best model.

B. Evaluation Metric

In this article, the average precision (AP) metric is used to assess the performance of the ship detection model, which is calculated as follows:

$$\text{AP} = \int_0^1 \text{P}(\text{R})d\text{R} \times 100\% \tag{15}$$

where

$$\text{P} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{16}$$

$$\text{R} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{17}$$

where TP, FP, and FN refer to the number of correctly predicted ship targets, the number of incorrectly predicted ship targets, and the number of ship targets judged to be nonship targets, respectively. P represents the accuracy rate, which is the proportion of the number of correct predictions to the total number of predictions among all predictions. R represents the recall rate, which is the proportion of the number of correctly predicted ship targets to the total number of annotations among all annotated ship targets. AP describes the area under the Precision–Recall

TABLE I
COMPARISON OF THE ESD MODULE WITH CONVENTIONAL SINGLE-BRANCH STRUCTURE

Method	P/%	R/%	AP_{50}/%	AP_{50-95}/%
Single branch	87.26	87.79	93.24	58.41
ESD	**91.71**	**90.10**	**96.68**	**59.08**

The bold entities means the best result in the same column.

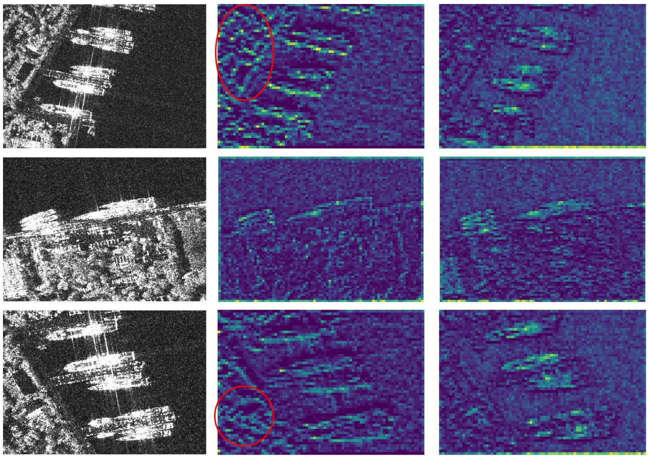


Fig. 6. Feature map visualization.

TABLE II
ABLATION EXPERIMENTS OF ESD AND TDL

ESD	TDL	P/%	R/%	AP_{50}/%	AP_{50-95}/%
–	–	87.26	87.79	93.24	58.41
–	✓	**93.22**	88.09	94.18	52.14
✓	–	91.71	90.10	96.68	59.08
✓	✓	92.44	**92.43**	**96.72**	**64.24**

The bold entities means the best result in the same column.

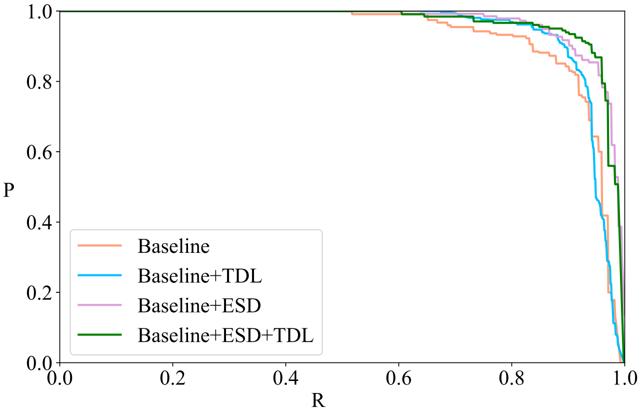


Fig. 7. P-R curve of the ablation experiment of ESD and TDL.

(P-R) curve. It is a compromise between the two metrics and also shows the overall performance of different methods.

C. Effectiveness of ESD

The aim of the ESD module is to reduce the ship's missed detection in inshore dense scenarios. To assess the module's effectiveness, comparative experiments were performed on the publicly available SSDD dataset, and the experimental results are shown in Table I.

In Table I, P, R and AP represent the precision, recall, and average precision in the inshore region, respectively. AP_{50} represents the AP of inshore ship detection calculated with 0.5 as the threshold value. AP_{50-95} means that the threshold value of IOUs is taken from 0.5 to 0.95 in steps of 0.05, and then the average value of APs under these IOUs is calculated. Compared with AP_{50}, the calculation of AP_{50-95} is more rigorous and better reflects the advantages and disadvantages of the model.

As can be seen from Table I, compared to the conventional single-branch detection structure, the P, R, AP_{50}, and AP_{50-95} of the inshore ships are improved after adding the ESD module. The R has increased by 2.31%, indicating that the miss detection of the inshore ships is alleviated. To further analyze the impact of the ESD module, feature visualization is performed in this article. The PLT image processing package is used in the model inference process to save the feature matrices at different scales by channel and colorize them, which makes the visualized feature maps visually better compared with grayscale maps. The feature visualization results in Fig. 6 also demonstrate that the inclusion of the ESD module results in sharper edges of ships in dense areas and clearer distinction between individual ships. The interference in the land area is also effectively suppressed, which helps to decrease the rate of miss detection in the dense inshore scenario.

D. Ablation Experiments of ESD and TDL Module

The proposed transformer-based TDL module is capable of effectively extracting contextual information about the target, which enables it to accurately differentiate between the ship target and land-based false alarms. Compared with the baseline, the TDL module is added, and the number of model parameters only increases by 0.0064%, which is a small increase in computational burden. Ablation experiments were conducted to confirm the effectiveness of this module. According to the experimental results in Table II, the addition of the TDL detection layer improves the P, R, AP_{50}, and AP_{50-95}. Compared to the baseline algorithm, the P improves by 5.96%, and adding TDL to ESD, the P improves by 0.73%, indicating that TDL can effectively combine contextual information to reduce false alarms in the inshore region. The AP_{50-95} is improved by 5.16%, indicating that the model's overall performance has been optimized. The P-R curves in Fig. 7 with the enclosed region of the coordinate axes are the values of AP_{50}. From which, the improvement in accuracy of the proposed method in this article can be seen more intuitively.

To demonstrate the superiority of the proposed method in this article, comparative experiments were conducted with typical two-stage detection algorithms (Faster R-CNN [14], Cascade R-CNN [15]), rotating box algorithm (OSCD-Net [31]) and typical single-stage detection algorithms (YOLOv3 [16],

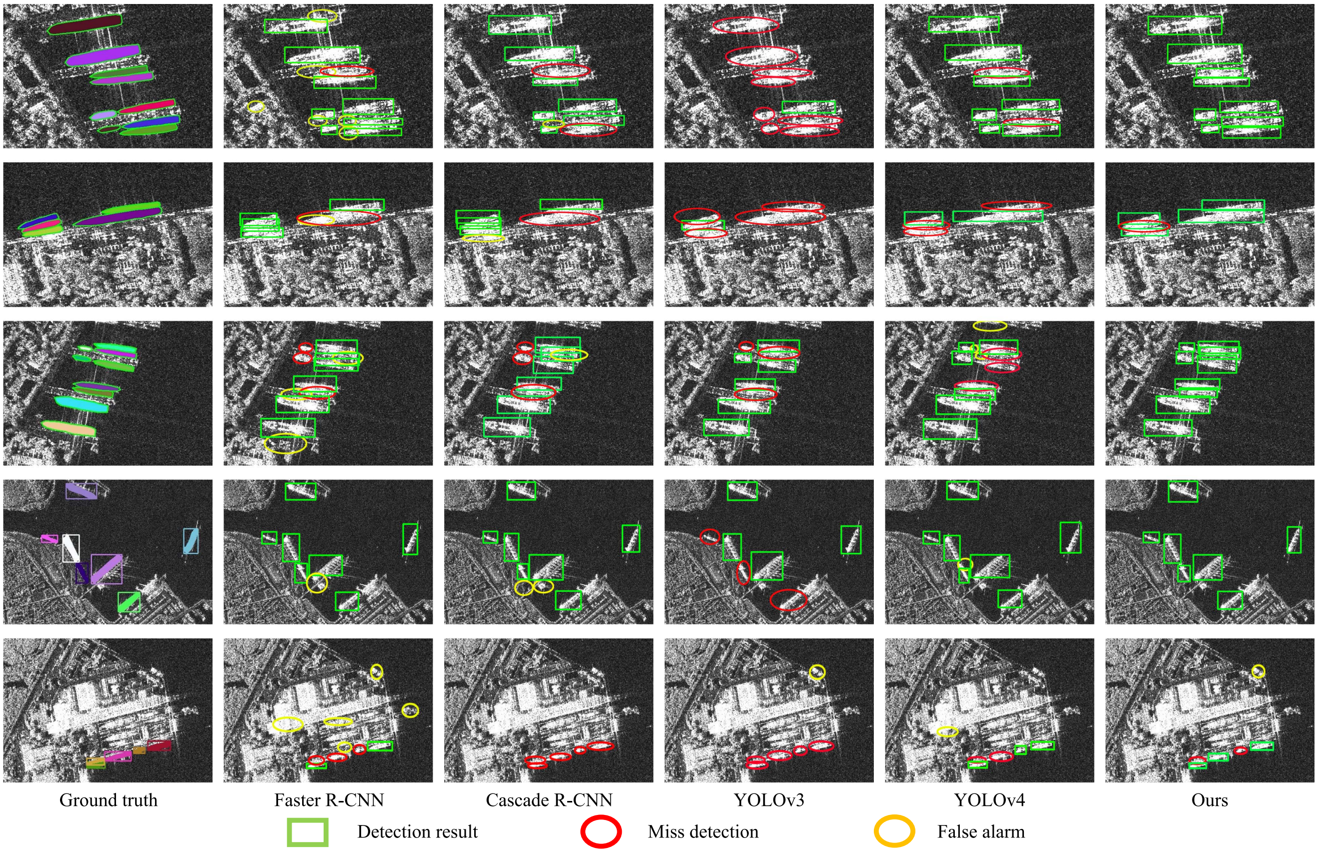


Fig. 8. Comparison with some detection results of relevant benchmark methods.

TABLE III
Comparison With Other Well-Known Methods

Method	P/%	R/%	AP_{50}/%
Faster R-CNN	57.08	91.58	88.57
Cascade R-CNN	86.66	91.58	89.97
YOLOv3	84.09	88.10	80.47
YOLOv4	87.25	75.58	83.98
OSCD-Net [31]	89.25	90.45	89.85
Ours	**92.44**	**92.43**	**96.72**

The bold entities means the best result in the same column.

YOLOv4 [17]). Among the above methods, the backbone network of faster R-CNN, cascade R-CNN, and OSCD-Net is ResNet. YOLOv3, YOLOv4, and the backbone network of the method proposed in this article use DarkNet. The ship detection results are shown in Table III, where the experimental results of OSCD-Net is derived from [31].

The proposed method in this article has been shown to be more effective than several commonly used conventional single-stage, two-stage, and rotate box detection algorithms in detecting SAR inshore ships. The detection results of different comparison methods are shown in Fig. 8. The ground truth images are colorized for different ship targets to facilitate better differentiation of dense adjacent ship targets. The red ellipse in the result comparison graph indicates the miss detection and the yellow ellipse indicates the false alarm. These visualizations provide an intuitive demonstration of the effectiveness of the proposed method in suppressing false alarms and miss detection in the inshore scenario when compared to conventional detection algorithms.

IV. Discussion

The detection of inshore ships presents a greater challenge than that of ships located solely at sea due to the higher rates of false alarms and missed detections.

The higher rates of false alarms are due to that SAR images of inshore ship targets are prone to interference from nonship targets. Therefore, context information is needed to better differentiate between ships and false alarms. In this article, a transformer-based TDL detection layer is introduced to capture global and context information, and comparative experiments have shown that adding TDL can effectively reduce false alarms.

The higher rates of missed detections are due to that inshore ships are densely arranged, making it difficult to distinguish between adjacent targets. To alleviate this issue, ship edge semantic information needs to be introduced to better distinguish adjacent targets. In this article, by decoupling the detection layer and adding a semantic segmentation branch to introduce ship edge information, the ship recall rate was improved and missed detections were reduced. Compared to methods that increase computational complexity, such as dilated convolution or fusion of high-resolution feature layers to add context information, the proposed TDL layer only adds a small number of parameters while achieving high accuracy. Compared to using instance segmentation to introduce ship edge information, the proposed ESD method is simple to implement and does not require complex instance labels, making it an effective and easily implementable module.

However, training the proposed method requires ship semantic segmentation labels, which undoubtedly increases the annotation workload for large datasets. How to reduce the dependence of the decoupled semantic segmentation layer on ship semantic labels is a direction for future algorithm improvements. Compared to SAR images of purely sea scenes, inshore scenes are more complex and ship detection is more difficult. Therefore, how to improve the detection of inshore ships while increasing a minimal or even no computational burden is an important and meaningful research direction.

V. Conclusion

This article presents a novel method for detecting dense inshore ships in SAR images using an ESD module and a transformer-based TDL layer. The ESD module incorporates edge semantic information of inshore ship targets during training, improving the model's ability to distinguish between neighboring ships and reducing miss detections. Meanwhile, the TDL layer utilizes transformer to extract contextual information and reduce false alarms caused by interference from land and other regions in the labeled boxes. The results of comparison experiments with some two-stage and single-stage detection algorithms on the SSDD dataset showed the proposed method achieved the highest AP_{50} of 96.72%, demonstrating its effectiveness in detecting inshore ships. The simple structure of the single-stage detector makes it easier to perform improvements and experiments, so this article performs experimental validation on a single-stage detector. However, the TDL and ESD proposed in this article are both plug-and-play improvement modules, which are less dependent on the overall structure of the detection algorithm. Theoretically, they can be fully ported to two-stage detectors, and whether the porting is effective requires extensive experimental verification. The future work is to explore the potential of combining this method with other two-stage detection frameworks for further optimization and improvement.

Acknowledgment

The authors would like to thank Zhang et al. [7] for providing SSDD dataset used in this article. The authors would also like express their deepest gratitude to the anonymous reviewers, whose careful work and thoughtful suggestions helped to improve this article considerably.

References

[1] K. El-Darymli, P. McGuire, D. Power, and C. R. Moloney, "Target detection in synthetic aperture radar imagery: A state-of-the-art survey," *J. Appl. Remote Sens.*, vol. 7, no. 1, Mar. 2013, Art. no. 071598.

[2] Z. Yue et al., "A novel semi-supervised convolutional neural network method for synthetic aperture radar image recognition," *Cogn. Computation*, vol. 13, no. 4, pp. 795–806, 2021.

[3] H.-C. Li, W.-S. Hu, W. Li, J. Li, Q. Du, and A. Plaza, "A^3CLNN: Spatial, spectral and multiscale attention ConvLSTM neural network for multisource remote sensing data classification," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 33, no. 2, pp. 747–761, Feb. 2022.

[4] J.-Y. Yang, H.-C. Li, W.-S. Hu, L. Pan, and Q. Du, "Adaptive cross-attention-driven spatial–spectral graph convolutional network for hyperspectral image classification," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, 2022, Art. no. 6004705.

[5] F. Ma, F. Zhang, D. Xiang, Q. Yin, and Y. Zhou, "Fast task-specific region merging for SAR image segmentation," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5222316.

[6] F. Ma, F. Zhang, Q. Yin, D. Xiang, and Y. Zhou, "Fast SAR image segmentation with deep task-specific superpixel sampling and soft graph convolution," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5214116.

[7] T. Zhang et al., "SAR ship detection dataset (SSDD): Official release and comprehensive data analysis," *Remote Sens.*, vol. 13, no. 18, pp. 3690–3730, Sep. 2021.

[8] 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, pp. 765–778, Mar. 2019.

[9] X. Sun, Z. Wang, Y. Sun, W. Diao, Y. Zhang, and K. Fu, "AIR-SARShip-1.0: High-resolution SAR ship detection dataset," *J. Radar*, vol. 8, no. 6, pp. 852–862, 2019.

[10] T. Zhang et al., "LS-SSDD-v1.0: A deep learning dataset dedicated to small ship detection from large-scale Sentinel-1 SAR images," *Remote Sens.*, vol. 12, no. 18, pp. 2997–3033, Sep. 2020.

[11] X. Sun et al., "FAIR1M: A benchmark dataset for fine-grained object recognition in high-resolution remote sensing imagery," *ISPRS J. Photogrammetry Remote Sens.*, vol. 184, pp. 116–130, 2022.

[12] Y. Zhou, F. Zhang, F. Ma, D. Xiang, and F. Zhang, "Small vessel detection based on adaptive dual-polarimetric feature fusion and sea-land segmentation in SAR images," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 15, pp. 2519–2534, 2022.

[13] F. Ma, X. Sun, F. Zhang, Y. Zhou, and H.-C. Li, "What catch your attention in SAR images: Saliency detection based on soft-superpixel lacunarity cue," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5200817.

[14] 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.

[15] Z. Cai and N. Vasconcelos, "Cascade R-CNN: Delving into high quality object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Salt Lake City, UT, USA, 2018, pp. 6154–6162.

[16] J. Redmon and A. Farhadi, "YOLOv3: An incremental improvement," 2018, *arXiv:1804.02767*.

[17] A. Bochkovskiy, C.-Y. Wang, and H.-Y. M. Liao, "YOLOv4: Optimal speed and accuracy of object detection," 2020, *arXiv:2004.10934*.

[18] T. Tian, Z. Pan, X. Tan, and Z. Chu, "Arbitrary-oriented inshore ship detection based on multi-scale feature fusion and contextual pooling on rotation region proposals," *Remote Sens.*, vol. 12, no. 2, pp. 339–357, Jan. 2020.

[19] X. Ma, S. Hou, Y. Wang, J. Wang, and H. Wang, "Multiscale and dense ship detection in SAR images based on key-point estimation and attention mechanism," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5221111.

[20] H. Su et al., "HQ-ISNet: High-quality instance segmentation for remote sensing imagery," *Remote Sens.*, vol. 12, no. 6, pp. 989–1012, Mar. 2020.

[21] Z. Wu, B. Hou, B. Ren, Z. Ren, S. Wang, and L. Jiao, "A deep detection network based on interaction of instance segmentation and object detection for SAR images," *Remote Sens.*, vol. 13, no. 13, pp. 2582–2607, Jul. 2021.

[22] Z. Ge, S. Liu, F. Wang, Z. Li, and J. Sun, "YOLOX: Exceeding yolo series in 2021," 2021, *arXiv:2107.08430*.

[23] R. Wang, Y. Huang, Y. Zhang, J. Pei, J. Wu, and J. Yang, "An inshore ship detection method in SAR images based on contextual fluctuation information," in *Proc. 6th Asia-Pacific Conf. Synthetic Aperture Radar*, Xiamen, China, 2019, pp. 1–5.

[24] X. Hou and F. Xu, "Inshore ship detection based on multi-aspect information in high-resolution SAR images," in *Proc. 6th Asia-Pacific Conf. Synthetic Aperture Radar*. Xiamen, China, 2019, pp. 1–4.

[25] X. Ke, X. Zhang, and T. Zhang, "GCBANet: A global context boundary-aware network for SAR ship instance segmentation," *Remote Sens.*, vol. 14, no. 9, pp. 2165–2185, Apr. 2022.

[26] X. Zhu, S. Lyu, X. Wang, and Q. Zhao, "TPH-YOLOv5: Improved YOLOv5 based on transformer prediction head for object detection on drone-captured scenarios," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 2778–2788.

[27] K. Li, M. Zhang, M. Xu, R. Tang, L. Wang, and H. Wang, "Ship detection in SAR images based on feature enhancement swin transformer and adjacent feature fusion," *Remote Sens.*, vol. 14, no. 13, pp. 3186–3208, Jul. 2022.

[28] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 2980–2988.

[29] A. Dosovitskiy et al., "An image is worth 16x16 words: Transformers for image recognition at scale," 2020, *arXiv:2010.11929*.

[30] Z. Zheng et al., "Enhancing geometric factors in model learning and inference for object detection and instance segmentation," *IEEE Trans. Cybern.*, vol. 52, no. 8, pp. 8574–8586, Aug. 2022.

[31] J. Zhang, R. Huang, Y. Li, and B. Pan, "Oriented ship detection based on intersecting circle and deformable ROI in remote sensing images," *Remote Sens.*, vol. 14, no. 19, pp. 4749–4769, Sep. 2022.

Yongsheng Zhou (Member, IEEE) received the B.E. degree in communication engineering from Beijing Information Science and Technology University, Beijing, China, in 2005, and the Ph.D. degree in signal and information processing from the Institute of Electronics, Chinese Academy of Sciences, Beijing, China, in 2010.

He was with the Academy of Opto-Electronics, Chinese Academy of Sciences, during 2010 and 2019. He is currently a Professor of Electronic and Information Engineering with the College of Information Science and Technology, Beijing University of Chemical Technology, Beijing, China. His research interests include target detection and recognition from microwave remotely sensed image, digital signal, and image processing.

Feixiang Zhang received the M.S. degree in information and communication engineering from the College of Information Science and Technology, Beijing University of Chemical Technology, Beijing, China, in 2022.

His research interests include image processing and deep learning-based small target detection.

Qiang Yin (Senior Member, IEEE) received the B.S. degree in electronic and information engineering from Beijing University of Chemical Technology, Beijing, China, in 2004, and the M.S. and Ph.D. degrees in signal and information processing from the Institute of Electronics, Chinese Academy of Science, Beijing, in 2008 and 2016, respectively.

From 2008 to 2013, she was a Research Assistant with the Institute of Electronics, Chinese Academy of Sciences. From 2014 to 2015, she was a Research Fellow with the European Space Agency, Rome, Italy. She is currently an Associate Professor with the College of Information Science and Technology, Beijing University of Chemical Technology. Her research interests include polarimetric/polarimetric interferometric synthetic aperture radar and deep learning.

Fei Ma (Member, IEEE) received the B.S., M.S., and Ph.D. degrees in electronic and information engineering from Beijing University of Aeronautics and Astronautics, Beijing, China, in 2013, 2016, and 2020, respectively.

He is currently with the College of Information Science and Technology, Beijing University of Chemical Technology, Beijing, China, as an Associate Professor. His research interests include radar signal processing, image processing, machine learning, and target detection.

Fan Zhang (Senior Member, IEEE) received the B.E. degree in communication engineering from the Civil Aviation University of China, Tianjin, China, in 2002, the M.S. degree in signal and information processing from Beihang University, Beijing, China, in 2005, and the Ph.D. degree in signal and information processing from Institute of Electronics, Chinese Academy of Sciences, Beijing, in 2008.

He is currently a Full Professor of Electronic and Information Engineering with the Beijing University of Chemical Technology, Beijing, China. His research interests incldue remote sensing image processing, high-performance computing, and artificial intelligence.

Dr. Zhang is also an Associate Editor of IEEE ACCESS and a Reviewer of the IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING, the IEEE JOURNAL OF SELECTED TOPICS IN APPLIED EARTH OBSERVATIONS AND REMOTE SENSING, the IEEE GEOSCIENCE AND REMOTE SENSING LETTERS, and the *Journal of Radars*.