IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received December 5, 2020, accepted February 2, 2021, date of publication February 12, 2021, date of current version February 24, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3059033

Soft Thresholding Attention Network for Adaptive Feature Denoising in SAR Ship Detection

RUI WANG[1], SIHAN SHAO[1], MENGYU AN[1], JIAYI LI[1], SHIFENG WANG[1,2], AND XIPING XU[1]

[1]School of Optoelectronic Engineering, Changchun University of Science and Technology, Changchun 130022, China
[2]Key Laboratory of Optoelectronic Measurement and Optical Information Transmission Technology, Ministry of Education, School of Optoelectronic Engineering, Changchun University of Science and Technology, Changchun 130022, China

Corresponding authors: Xiping Xu (peter_opt@sina.com) and Shifeng Wang (sf.wang@cust.edu.cn)

This work was supported in part by the Natural Science Fund of Jilin Province under Grant 20150101047JC, in part by the National Scholarship for Study abroad under Grant 201908220241, and in part by the 111 Project of China under Grant D17017.

ABSTRACT Recently, synthetic aperture radar (SAR) ship detection is used in many applications within the marine field, such as fishery management, traffic control, and urgent rescue operations. Meanwhile, deep learning-based methods have bought new capabilities for ship detection in SAR images on account of high accuracy and robustness. However, several challenges remain to be addressed: 1) the shapes of the ships in SAR images have a relatively extreme aspect ratio comparing to the target objects in the optical images, and 2) complex background and clutter noise result in adverse effects for the network to extract prototypical SAR target features, which limit the ship detection performance. To address these issues, this paper proposes two effective approaches to augment the feature extraction ability of the network. Firstly, IOU (Intersection over Union) K-means is carried out to settle the extreme aspect ratio problem. The IOU K-means, as a preprocessing step, clusters a set of aspect ratios from datasets that are suitable for ship detection. Secondly, we embed a soft thresholding attention module (STA) in the network to suppress the impact of noise and complex background. The comparison results with several state-of-the-art object detection algorithms confirm the efficiency and feasibility of proposed approaches.

INDEX TERMS Ship detection, soft thresholding attention module, feature denoise, deep learning.

I. INTRODUCTION

As one of the important applications of remote sensing, ship detection in synthetic aperture radar images has attracted significant attention in recent years [1]–[8]. Ship detection is also applicable to tactical deployments and ocean defense early warning systems. In terms of civil applications, it is also beneficial to fishery [9], traffic control [10] and maritime surveillance [11], [12]. Despite the wide practical value in these fields, until now, SAR images detection technology still lags behind the optical images because of their dissimilar mechanisms [13].

For this reason, many innovative methods are proposed to solve this problem, which promotes the accuracy and robustness of ship detection techniques. According to our investigations, various types of ship detection methods can be divided into two main categories: (1) Traditional feature extraction techniques, and (2) modern deep learning-base techniques.

The traditional feature extraction techniques for ship detection from the SAR images include methods such as statistical distribution-based [14]–[16], multiple-scale-based [17], template matching [18], and multiple polarization-based [19], [20]. These methods are highly dependent on manual feature extraction and availability of prior knowledge such as predefined thresholding and the distributions of sea clutters [21]–[24]. Generally, most traditional ship detection systems consist of four steps: land masking, preprocessing, prescreening, and discrimination [25].

Land masking is a pre-required stage for most traditional ship detection system. Registering the SAR images with the existing geographic maps is a common means of landing masking [26], yet the easiest methodologies have many shortages. For instance, tidal ranges are self-evident and some minuscule islands and rocks are easily been overlooked, such issues may happen with registration errors.

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

Manual registration can reduce errors, nonetheless, it is not intelligent and cost-effective. Other methodologies to territory masking are to automatically recognize the coastline utilizing particular algorithms. The vast majority of shoreline extraction algorithms are based on the light of these image handling steps, where the speckle noise is removed and using an edge operator. The edge map is then expanded using a mean filter followed by thresholding to distinguish between land and water [5], [27]–[30]. Although the current coastline extraction algorithms overperform registration, these methods are unable to detect in real-time detection in complex sea conditions and their performance is significantly deteriorated.

The objective of the preprocessing stage is to improve the accuracy of the next detection stage. The prescreening stage seeks out the probable ship regions throughout the image. Defining a global threshold of the image is a simple way to proclaim any pixel value beyond the threshold as an anticipative ship pixel. To overcome the speckle noise and region inhomogeneity of the SAR images, constant false alarm (CFAR) is introduced. CFAR calculates the thresholding adaptively. Based on the CFAR, a series of new algorithms have been stated, such as cell-average CFAR, two-parameter CFAR [31], bilateral CFAR [32], etc.

These CFAR-based techniques usually have two drawbacks hindering their development and applications. Firstly, a set of guard windows should be set corresponding to the size of the detected ships. If different sizes of ships are densely clustered, the settled guard windows result in missing detection. Secondly, the CFAR methods rely on the sliding window techniques which are time-consuming. The following stage is the discrimination phase. To discriminate the true target from the false alarms, discrimination algorithms need to extract prototypical (e.g., area, aspect ratio, orientation and wake) and handcrafted features for the target discrimination. These approaches also require expert knowledge to choose advisable features.

The most noteworthy peculiarities of the modern deep learning-based (DL) techniques are their ability of automatic feature extraction, good feature expression level and high recognition accuracy [33]. Upon availability of a labelled dataset, DL trains and learns under the supervision of data to accomplish accurate object detection tasks. State-of-the-art deep learning-based object detection methods are either two-stage or one-stage detectors [34].

Two-stage methods are also called region proposal-based methods, which divide the framework of detection in two stages. The first stage engenders a set of candidate proposals and categorizes them as foreground or background. The second stage then classifies the specific categories and regresses the coordinates of anchors. The highest accuracy with low-efficiency object detectors is achieved by the two-stage methods, such as R-CNN [35], Fast R-CNN [36], Faster R-CNN [37], MSCNN [38], and Cascade R-CNN [39]. In contrast, one-stage detectors achieve the predicting classes and bounding boxes directly based on the regression, such as YOLO [40]–[42], SSD [43], and RetinaNet [44]. These detectors have been optimized for higher speed but their accuracy is often lower than that of the region proposal-based methods. Deep learning-based methods have made significant advancement in the field of optical image detection. Nevertheless, the discrimination between nature scene images and SAR images such as imaging mechanisms and imaging objects result in performance degradation. Therefore, several challenges still exist: 1) the aspect ratio of the ships in SAR images is relatively high, so hand-picked anchor ratios set by the natural scene images in SAR images are not suitable. The proposals generated from the feature maps are also difficult to regress, hence cause inaccurate positioning; 2) ships in the SAR images are minuscule and densely clustered which are easily submerged in noise and complex background (e.g. offshore, inshore, in the inland rivers, and around cays which are visually similar to the ship). Such useless information in images may harm extracting the feature of ships.

To address these problems, in our paper, two novel and effective methods are proposed to build a deep ship detector. Firstly, we design IOU k-means as data preprocessing to cluster on training set bounding boxes to automatically obtain good prior anchors instead of choosing priors by hand[41]. We do not use regular k-means with Euclidean distance because larger boxes result in larger errors than that of the small boxes. Therefore, IOU k-means which is independent of the size of the box leads to higher IOU scores and improves the performance. This method attempts to verify the impact of the model performance of anchors with different scales and aspects. The network can also easily learn to predict good detections if we pick better priors that are adaptive for SAR image ships. Secondly, we design an attention mechanism called soft thresholding attention (STA) block that embed in the network to perform denoising in the feature-level. The STA block, as a feature-level denoising method, can adaptively learn a set of thresholding according to the global information of the features to suppress noise. Comparing with the image-level denoising methods, it can automatically reduce the loss of useful information and suppress complex background and noise such as land speckle and sidelobe effect.

The main contributions of this paper are listed in the following:

1) We propose IOU k-means as data preprocessing, to cluster good prior anchors that are suitable for ship detection to improve the regression ability.
2) We propose SAT block as a feature-level denoising method for the first time in the field of SAR ship detection.

The rest of this paper is organized as follows. Section 2 presents the proposed approaches. Section 3 the experiments are presented, including the dataset and experiment analysis. Section 4 is a discussion, and Section 5 provides our conclusions.

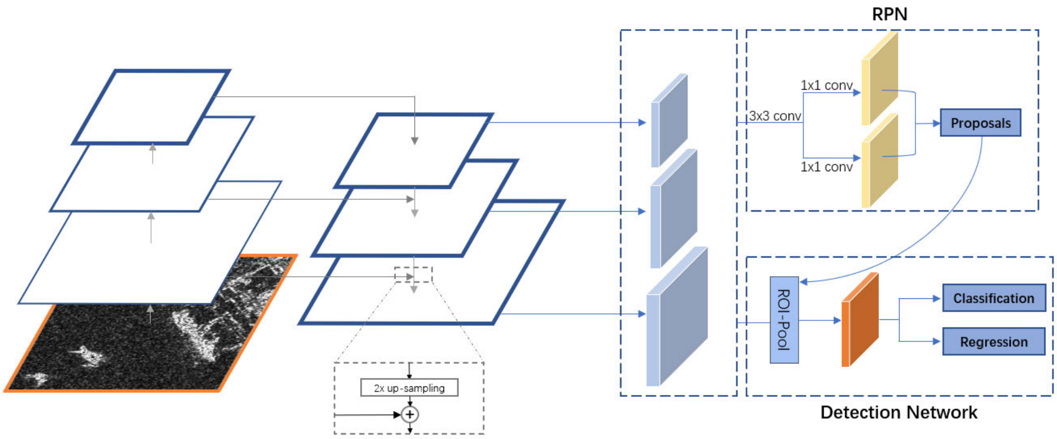


FIGURE 1. The architecture of Faster R-CNN.

II. METHODOLOGY

In this section, we elaborate on the details of the proposed approach.

A. BACKGROUND ON FASTER R-CNN

Faster R-CNN is a two-stage detector and mainly contains three main structures: the convolution networks (ConvNets) as the backbone to extract feature maps, and the region proposal network (RPN) to generate the region proposals. These proposals are then utilized for object classification and bounding box regression in a subnetwork. Images are fed to the ConvNets to obtain their feature maps, and then the RPN is applied to collect a set of the rectangular object proposals and their corresponding foreground and background scores. Region of interest (ROI) aligns these object proposals to sub-network. Finally, these transformed object proposals are given to the subnet for predicting the bounding boxes and classification of the targets included ship. The network structure is illustrated in Figure 1.

B. CONSTRUCTING THE FEATURE EXTRACTION NETWORK

A backbone in Figure 1 shows the bottom-up route for feature extraction and the top-down route for feature fusion. The bottom-up pathway regularly includes CNNs to obtain image's stratified features. With the decrease of the spatial resolution from the bottom to top, the semantic information in the feature maps is then reinforced as the network is getting deeper. The top-down pathway feature pyramid network (FPN) mainly solves the multi-scale problem with the fusion of high-resolution low-level features and low- resolution high-level features. The top-down pathway hallucinates higher resolution features by up-sampling spatially coarser and the lateral connection merges feature maps of the same spatial size from the bottom-up pathway and the top-down pathway.

In our experiment ResNet (see, Table 1) is shown as the bottom-up pathway, we use the feature activations output by each stage's last residual block. We indicate the output of these last residual blocks as $\{C1, C2, C3, C4\}$ for the 2nd, 3rd, 4th and 5th stage outputs. Starting from the 2nd stage, we use a stack of 3 layers included 1×1, 3×3, and 1×1 convolutions, where the 1×1 layers aim to reduce and then increase the dimensions, leaving the 3×3 layer a bottleneck with smaller input /output dimensions. The 2nd, 3rd, 4th and 5th stages are composed of 3, 4, 6 and 3 residual blocks, respectively. The widths (the number of channels) of the convolutions of four stages are $\{C_w, 2C_w, 4C_w, 8C_w\}$ respectively and they have strides of $\{4, 8, 16, 32\}$ pixels with the respect to the input image. To substitute our top-down feature maps, with a coarser-resolution feature map, we upsample the spatial resolution by a factor of 2, using the nearest neighbour upsampling technique. After that, the lateral connection applies 1×1 convolution to reduce the channel dimensions to 256 and then merge the upsampled map and corresponding bottom-up map by element-wise addition. Finally, we append

TABLE 1. Two columns refer to ResNet-50 and STANet-50. Inside the brackets are shapes and operations with specific parameter settings of a residual building. The number of the stacked blocks in each stage is presented outside the brackets. The inner brackets following by the STA module indicate the output of thresholding.

Output	ResNet-50	STANet-50
112×112	conv, 7×7, stride 2	
	max pool, 3×3, stride 2	
56×56	$\begin{bmatrix} conv\ 1\times1, & 64 \\ conv\ 3\times3, & 64 \\ conv\ 1\times1, & 64 \end{bmatrix} \times 3$	$\begin{bmatrix} conv\ 1\times1, & 64 \\ conv\ 3\times3, & 64 \\ conv\ 1\times1, & 64 \\ STA(r=16), & 256 \end{bmatrix} \times 3$
28×28	$\begin{bmatrix} conv\ 1\times1, & 128 \\ conv\ 3\times3, & 128 \\ conv\ 1\times1, & 512 \end{bmatrix} \times 4$	$\begin{bmatrix} conv\ 1\times1, & 128 \\ conv\ 3\times3, & 128 \\ conv\ 1\times1, & 512 \\ STA(r=16), & 512 \end{bmatrix} \times 4$
14×14	$\begin{bmatrix} conv\ 1\times1, & 512 \\ conv\ 3\times3, & 512 \\ conv\ 1\times1, & 1024 \end{bmatrix} \times 6$	$\begin{bmatrix} conv\ 1\times1, & 512 \\ conv\ 3\times3, & 512 \\ conv\ 1\times1, & 1024 \\ STA(r=16), & 1024 \end{bmatrix} \times 6$
14×14	$\begin{bmatrix} conv\ 1\times1, & 1024 \\ conv\ 3\times3, & 1024 \\ conv\ 1\times1, & 2048 \end{bmatrix} \times 3$	$\begin{bmatrix} conv\ 1\times1, & 1024 \\ conv\ 3\times3, & 1024 \\ conv\ 1\times1, & 2048 \\ STA(r=16), & 2048 \end{bmatrix} \times 3$
1×1	Global average pool, 1000-d fc, softmax	

3×3 convolution on each merged map to generate final feature maps. The final maps set is denoted by$\{P2, P3, P4, P5\}$, corresponding to $\{C2, C3, C4, C5\}$ that are of the same spatial sizes.

C. REGION PROPOSAL NETWORK(RPN)

As shown in Figure 2, the RPN includes a 3×3 convolution layer as a sliding window and two 1×1 convolution layers to bring about the region proposals for classification and regression. To generate the region proposals, we slide a 3×3 convolution layer over each position of the convolution feature map output by the backbone. Then each sliding window is mapped to a lower-dimensional feature and at each sliding-window location. We simultaneously predict the multiple region proposals. Generally, each position at the centre of the sliding window over the feature map is associated with k different scales and aspect anchors. In the end, the regression layer has $4k$ output encoding the coordinates of k bounding boxes, and the classification layer outputs $2k$ scores that estimate the probability of ship or background for each proposal. In other words, due to fusion with low-resolution feature and high-resolution feature by feature pyramid network, the anchors can be assigned at different resolution stages $4^2, 8^2, 16^2, 32^2$ toP_2, P_3, P_4, P_5. Considering the diverse scale of ships, we use IOU K-means in Section D to obtain efficient anchors to adopt in each stage.

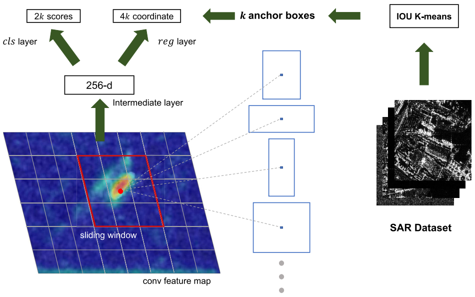


FIGURE 2. The architecture of Region Proposal Network.

Depending on their intersection over union (IOU) ratios with the ground-truth bounding boxes, a positive label is then assigned to an anchor if its IOU is over 0.7 with any ground truth box. A negative label is assigned if its IOU is lower than 0.3 for all ground truth box. Consequently, the 2000 region of interests (ROIs) are collected for each image by top-N and Soft-NMS operation on all proposals.

D. DETECTION NETWORK AND LOSS FUNCTION

For RPN, a binary class label (of being an object or not) is assigned to each anchor and roughly regress predicted anchors. In the detection network, we firstly obtain the mapping relationship between the original image and the feature map in the anchor region. Then the ROI Align [45] is adopted to generate a fixed size of 7×7 features for the feature maps of different sizes. Finally, all the 7×7 features are flattened and fed to the fully connected layers for higher quality detection refinement.

We minimize an objective function following the multi-task loss and the overall loss function is as the following:

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

where the λ is the parameter to balance the loss of classification and regression, i is the index of an anchor in a mini-batch, and p_i is the predicted probability of anchor i being an object. For a positive anchor, the ground truth label p_i^* is 1, and is 0 otherwise; t_i is a vector representing the 4 parameterized coordinates of the predicting bounding boxes, and t_i^* is that of the ground-truth box associated with a positive anchor.

For bounding box regression, (x, y, w, h), (x_a, y_a, w_a, h_a) and (x^*, y^*, w^*, h^*) can represent the predicted box, anchor box and ground-truth box, respectively. We parameterize the 4 coordinates as following:

$$\begin{aligned} t_x &= \frac{x - x_a}{w_a}, t_y = \frac{y - y_a}{h_a} \\ t_w &= log(\frac{w}{w_a}), t_h = log(\frac{h}{h_a}) \\ t_x^* &= \frac{x^* - x_a}{w_a}, t_y^* = \frac{y^* - y_a}{h_a} \\ t_w^* &= log(\frac{w^*}{w_a}), t_h^* = log(\frac{h^*}{h_a}) \end{aligned} \tag{2}$$

To make the predicted boxes close to the bounding boxes, we utilize the smooth L_1 loss function to minimize the error for obtaining a good regressor:

$$L_r eg(t_i, t_i^*) = smooth_{L_1}(t_i - t_i^*) \tag{3}$$

where

$$smooth_{L_1}(x) = \begin{cases} 0.5x^2, |x| < 1 \\ |x| - 0.5, otherwise \end{cases} \tag{4}$$

For category classification, we use the cross-entropy loss:

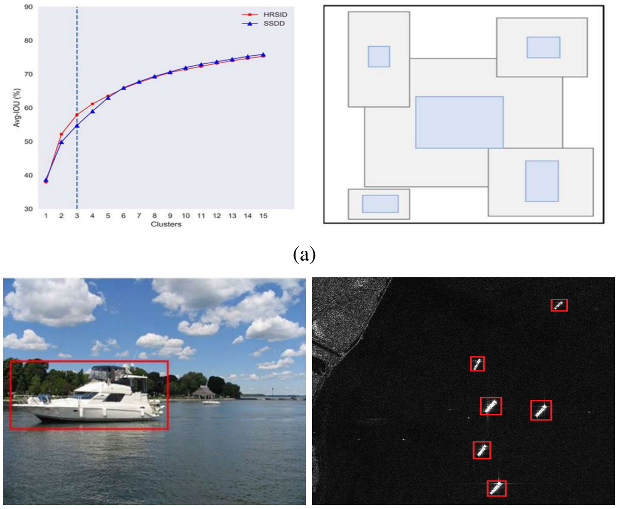
$$L_{cls}(p_i, p_i^*) = -\sum p_i^* log(p_i) \tag{5}$$

E. ANCHOR PRE-DESIGN:IOU K-MEANS

Ship detection techniques based on deep learning usually generate the predicted region proposals of the ship at each position on the feature map. Hence, we may encounter an issue with the anchors using deep learning methods. The aspect and size of anchors are hand-picked which are not suitable for the SAR ship detection. This is because most of the deep learning models are designed for optical scene images. Although the network may learn to adjust the anchors appropriately, if we choose better priors for the model, we can prompt it easier for the model to learn and predict good detections. Here we aim to pre-design better anchors on training set bounding boxes which lead to higher IOU scores and become independent of the box size. Here we use IOU K-means and for our distance metric we use:

$$d(box, centroid) = 1 - IOU(box, centroid) \tag{6}$$

We run IOU K-means for different values of k and plot average the IOU with the closest centroid on three SAR ship detection datasets in [46], [47] and the optical scene images dataset called VOC2007, see Figure 3. As it is seen in Figure 3, we choose $k = 3$ as a good tradeoff between model complexity and high recall. The cluster centroids and the size of anchors in SAR images are significantly different than that of in optical scene images. There are smaller anchors relative to the size of images and the aspect ratio of anchors is also more obvious.



(a)

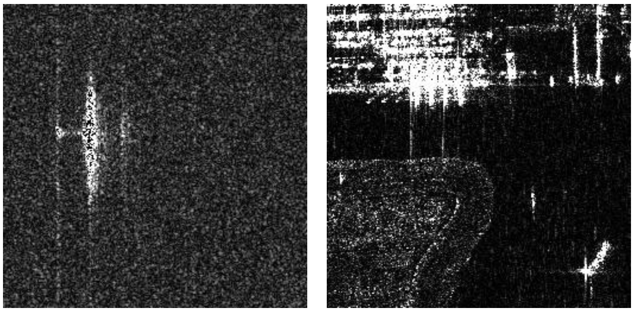
(b)

FIGURE 3. **Pre-designing anchor on two SAR datasets and an optical images dataset. We run IOU K-means clustering on these datasets to obtain good priors for our model. The left shows the average IOU score we get with different *k*. We set *k* = 3 as a tradeoff for recall and complexity of the model. The right image illustrates the aspect ratio and size of the anchors for the SAR and the optical image datasets. Objects in the SAR images are smaller than that of the optical images.**

F. SOFT THRESHOLDING ATTENTION BLOCK (STA)

For satellite SAR system, the radar receives echo signals including ground-based clutter and detection target from the ground. As a coherent imaging system, SAR images inevitably generate speckle. The speckled background in Figure 4(a) acts as the noise on a single detected SAR image. This is because it hides much information of the observed scene which is crucial for ship detection. Besides, considering the metal materials and the superstructure of the ship, the ships have strong backscatters effect as shown in Figure 4(b).

Therefore, in the multi-targets SAR images, the sidelobes of the strong scattering point obscure the main adjacent weak targets, leading to the missed detection of the weak targets. The speckled background and the scattering sidelobes reshape the ship appearances in the SAR images and interfere with the detection process. We need to suppress these noise and complex background to improve the target detection ability.



(a)

(b)

FIGURE 4. **Noise and complex background are harmful to ship detection. (a) SAR image with the speckled background; (b) Ships in the SAR image with scattering sidelobes.**

Soft thresholding function is often utilized as a major step in many signal denoising approaches [48], [49]. Generally, the raw signal is in the processing of domain transformation in which the near-zero numbers are insignificant. The soft thresholding function is then applied to transform the nearzero to zero for signal reconstruction. However, designing such thresholding requires high expertise in signal processing and is often a challenging issue. Attention mechanisms in deep learning [50]–[52] are inspired by the biological visual system which can concentrate on the region of objects and ignore some less important information. We combine soft thresholding function and the attention mechanism to automatically learn a set of thresholding using a gradient descent algorithm. As a result, the integration of the soft thresholding and attention mechanism in deep learning can be a promising method to eliminate the noise-related and complex background information and to construct highly discriminative features.

In this paper, we perform a simple and lightweight architecture unit named soft thresholding attention (STA) block embedded in the backbone, such as ResNet-50, to perform denoising in the feature-level. Compared with the traditional SAR denoise methods in the image-level, it can achieve a performance improvement in the speckled background and scattering sidelobes suppression. This is an additional advantage of a lower computational complexity and information loss. A feature map of the SAR image obtained by the radar system is:

$$Y = X + N \tag{7}$$

where X includes the considered feature maps, Y is a complex matrix with the same size as X which denotes the difference between the reconstructed feature maps and the real scene that includes noise and complex background. In consideration of the sparsity of the feature maps, we can recover the considered features by solving the following optimization problem:

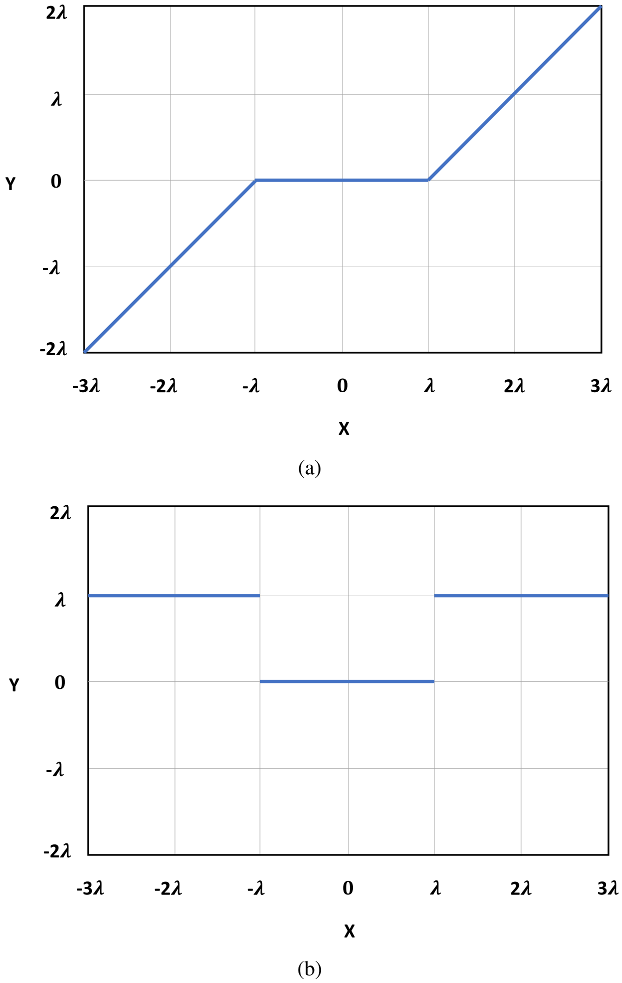
$$\hat{X} = \min_{X}\{\|Y - X\|_2^2 + \lambda\|X\|_1\} \tag{8}$$

the above function is continuous and strictly convex. It is however not differentiable at $x = 0$. The final optimized form of this function is:

$$\hat{X} = \min_{X}\{\|Y - X\|_2^2 + \lambda\|X\|_1\} = \begin{cases} Y - \lambda, Y > \lambda \\ 0, |Y| < \lambda \\ Y + \lambda, Y < \lambda \end{cases} \tag{9}$$

λ is the regularization parameter in (8) and soft thresholding in (9), which controls the reconstructed precision and the sparsity of the estimated scene.

Instead of setting the negative feature to zero in the ReLU activation function, soft thresholding function sets the nearzero features to zero, so that useful negative features can be preserved. The process of soft thresholding and ReLU functions are shown in Figure 5(a). Meanwhile, it can be seen observed that the derivative of output on soft thresholding is either one or zero which is similar to ReLU function. Both of them are effective in preventing gradient vanishing and exploding problems, as shown in Figure 5(b).



(a)

(b)

FIGURE 5. **Illustration of (a) soft thresholding function and (b) its derivative.**

Here we assume that

1) The value of λ is related to the global information of the feature maps. A set of feature maps in each layer also corresponds to a set of λ.
2) In the process of reconstruction, decreasing $\hat{X}$ is positively correlated with the decline of the loss function during the training process. Since the complex background and noise are suppressed in the feature maps, the network can construct highly discriminative feature. The more efficient the detection, the smaller the value of the loss function.

The STA block is built upon a transformation mapping between the input feature maps $X \in R^{H\times W\times C}$ and the reconstruction feature maps $U \in R^{H\times W\times C}$. Since the feature maps in each channel can be associated with a suitable λ to be reconstructed and suppress useless information, we expect to exploit the global information of the feature map in each channel to learn a set of thresholding. A diagram illustrates the structure of an STA block is shown in Figure 6.

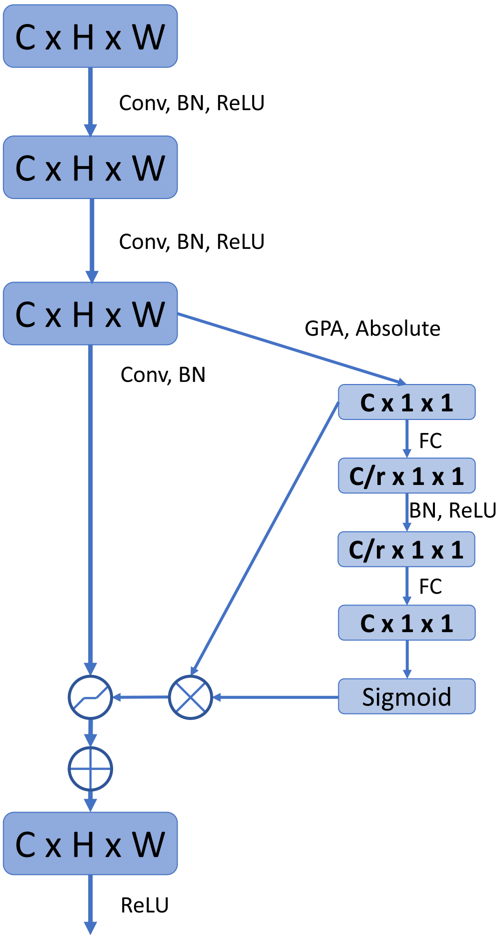


FIGURE 6. **The structure of Soft thresholding Attention Block.**

To choose a series of parameters by exploiting the information of each channel, we consider the signal to each channel in the output feature. Normal convolution kernel is unable to exploit the contextual information in a global region due to its local operation. To address this issue, we introduce extract global spatial information into a channel descriptor which achieved by using global average pooling (GAP) to generate channel-wise statistic. Formally, a statistic $z \in R^{C\times 1\times 1}$ is produced by the squeezing input feature maps X through its spatial dimensions $H \times W$, such as that the c^{th} element of $\mathbf{z}$ is:

$$z_c = \left|\frac{1}{W \times H}\sum_{i=1}^{H}\sum_{i=1}^{W} x_c(i,j)\right| \tag{10}$$

To build up the mapping relation between the context information of feature maps and the regularization parameters, λ, we choose to apply a simple gating mechanism with sigmoid activation to obtain the scaling parameters as:

$$s = \sigma(g(\mathbf{z}, \mathbf{W})) = \sigma(\mathbf{W_2}\delta(\mathbf{W_1 z})) \tag{11}$$

where δ refers to the ReLU function, σ refers to the sigmoid function, $W_1 \in R^{C/r\times C}$, and $W_2 \in R^{C\times C/r}$. To reduce the model complexity, we parameterize the gate mechanism with two fully-connection layers around the non-linearity, i.e., a dimensional-reduction layers with ratio r, a ReLU and then a dimensional-increasing layer returning to the channel dimension of input feature maps. A sigmoid function is also applied at the end of the gating mechanism so that the scaling parameters are scaled to the range of (0, 1) as:

$$\sigma = \frac{1}{1+e^{-z}} \tag{12}$$

The scaling parameter s is then multiplied by the statistic global information vectors z to obtain the thresholding. This arrangement is spurred by the fact that the thresholding for soft thresholding function should be neither negative nor excessively large. If the thresholding is larger than the largest absolute value of feature maps, the output of soft thresholding function is set to zero. Therefore, the thresholding is expressed as:

$$\lambda = \mathbf{s} \cdot \mathbf{z} \tag{13}$$

During multiple iterations of training, irrelevant speckled background and scattering noise of feature maps are set to a near-zero value to suppress the noise.

III. EXPERIMENTS AND RESULTS

In this section, we evaluate the performance of our proposed method for ship detection in the SAR images. Experiments are implemented based on mmdetection [53] which is a well-known open-source deep learning framework and performed on a PC with Intel single Core i7 CPU, NVIDIA GTX-1080TI GPUs, and 64GB RAM. The PC operating system is 64-bit Ubuntu 16.04.

A. DATASET DESCRIPTION

To verify the validity and robustness of our proposed methods, we choose two SAR datasets for ship detection including SSDD [47] and HRSID [46]. SSDD includes 1160 images and 2456 ships, with an average of 2.12 ships in one image. The SAR images in this dataset possess different satellite sensors including Sentinel-1, RadarSat-2, TerraSAR-X. The resolution of these images is 1 10m with Strip-Map (UFS), Fine Strip-Map 1 (FSI), Full Polarization 1 (QPSI), Full Polarization 2 (QPSII) and Fine Strip-Map 2 (FSII) imaging mode. HRSID consists of 5604 cropped SAR images with an overlapped ratio of 25%. There is a total of 16951 ships in HRSID with 3.02 ships per image. The original SAR imageries for constructing HRSID are Sentinel-1B imageries, TerraSAR-X, and TanDEM-X. Compared with SSDD, the SAR images in HRSID dataset have a higher resolution (i.e., under 3m) and contain detailed and accurately represented features of the ships. Under different imaging modes of radar sensors, the ships appear in different forms. A more detailed comparison is shown in Table 2.

B. EVALUATION METRICS

To quantitatively appraise the performance and robustness of our proposed methods, we adopted three popularly used criteria metrics including intersection over union (IoU), precision, recall, and mean average precision (mAP). For single class object detection, mean average precision (mAP) is defined by:

$$mAP = \int_0^1 P(r)dr \tag{14}$$

where r represents recall and $P(r)$ denotes the precision value that $recall = r$ is corresponded to.

$$Precision = \frac{TP}{TP+FP} \tag{15}$$

where TP is the number of the True Positives which in the number of cases where the real ships are correctly detected, and FP is the number of the False Positives, i.e., missed detections. Recall is defined as:

$$Recall = \frac{TP}{TP+FN} \tag{16}$$

where FN is the number of False Negatives, i.e., false alarms. The value of mAP is obtained by the integral of the precision over the interval from $recall = 0$ to $recall = 1$, i.e., the area under the precision-recall (PR) curve. For the ship detection, the larger the value of mAP, the higher the ship detection performance. Nevertheless, mAP does not fully reflect the performance of the object detection framework. Here we divide mAP to calculate evaluation metrics AP, AP_{50} and AP_{75} for more accurate bounding box regression, and AP_L, AP_M, AP_S for large, medium, and small objects, respectively. We then convert all three SAR ship detection datasets to Microsoft Common Object in Context (COCO) metrics as above which are objective and comprehensive metrics for measuring the performance of object detection tasks. The generic mAP metric mentioned above is the same as AP_{50} metric in COCO. For AP_{50}, when the IoU of the ground-truth and the predicted box is greater than 0.5, the test case is

TABLE 2. The characteristics of SSDD and HRSID.

Datasets	Images(num)	Size of ships (num)			Size of Images (pixels)		Resolution (m)
		Small	Medium	Large	Height	Width	
SSDD	1160	1529	935	76	190 $\sim$ 526	214-668	1 $\sim$ 10
HRSID	5604	5604	9242	321	800	800	0.5 $\sim$ 1

TABLE 3. Statistics of SSDD, HRSID, VOC2007 datasets.

Datasets	The average area of images(pixels)	Information of anchor (k=3)		Area ratio
		Area (height × width)	Aspect ratio	
SSDD	416 × 416	12 × 20	0.61	0.0014
		18 × 45	0.40	0.0052
		58 × 27	0.81	0.0241
HRSID	800 × 800	14 × 15	0.93	0.0003
		28 × 44	0.64	0.0019
		62 × 30	2.07	0.0029
VOC2007	416 × 416	26 × 45	0.87	0.0067
		78 × 129	0.60	0.0581
		243 × 281	0.59	0.3946

predicted as a ship, and the intersection over union (IoU) is defined as:

$$IoU(B_p, B_g) = \frac{B_p \cap B_g}{B_p \cup B_g} \tag{17}$$

where B_p and B_g are the area of predicted, and ground-truth boxes, respectively.

For the SAR ship detection, to quantitatively evaluate the performance of the proposed framework, we utilize the standard COCO metrics consisting of AP, AP_{50}, AP_{75}, AP_S, AP_M, AP_L which are defined in Table 4. Therefore, with the rising of IoU threshold, the bounding box regression is better and the ship is well-covered by the predicted bounding box.

C. EXPERIMENT OF IOU K-MEANS

To prove the extreme differences of the object target between the optical scene images and the SAR images, we run IOU K-means to cluster the aspect ratio and calculate the area of the pre-designed anchors of the SAR and optical scene image datasets. We then obtain the ratio of the anchor size to the image size for these datasets as discussed in Section D.

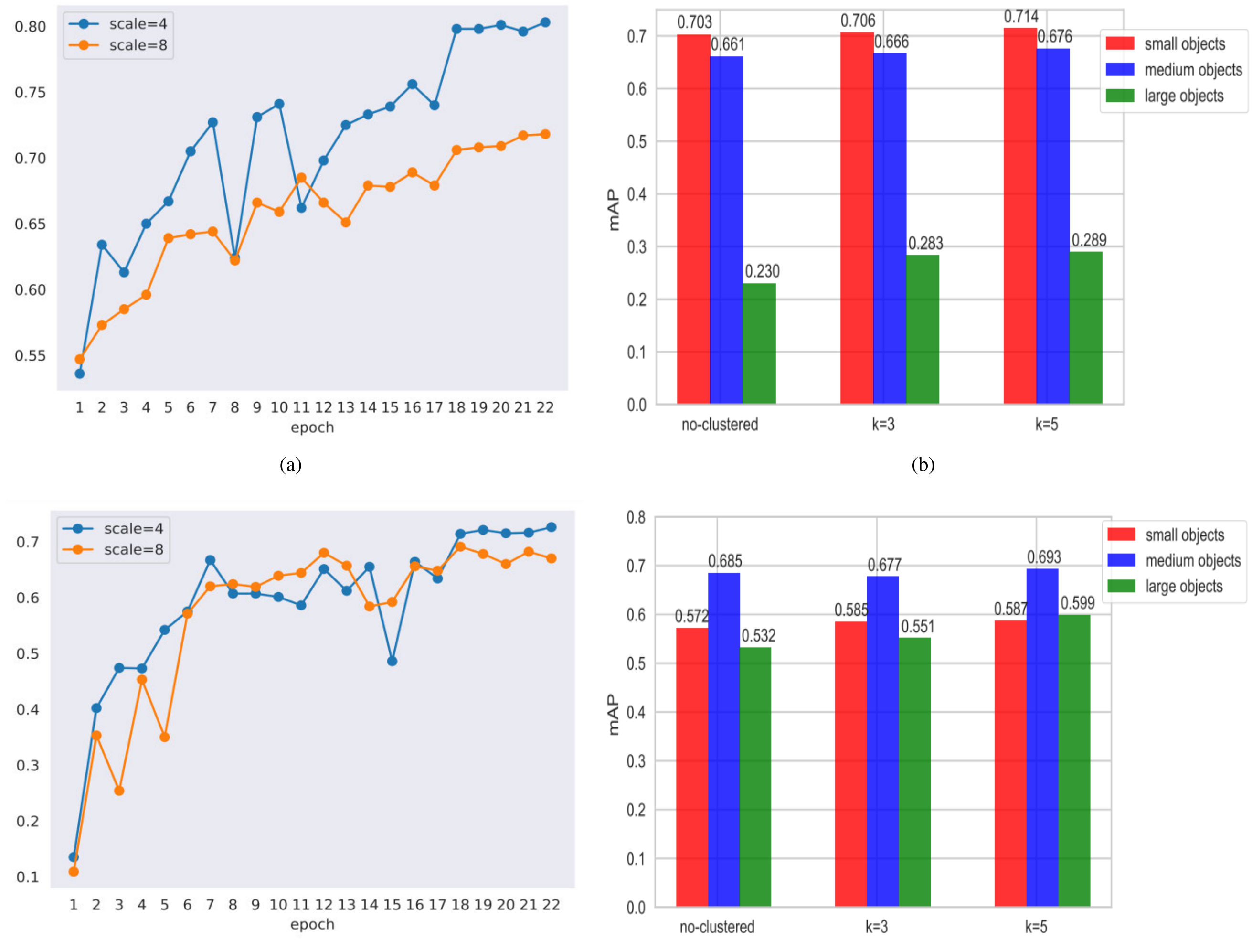
In Table 4, we present the statistics of the information about images and anchors in the three considered data sets. According to the classification indicator of large, medium and small objects in the COCO format, the anchors clustered in SSDD and HRSID datasets mainly correspond to small and medium objects. Furthermore, the anchors in VOC2007 dataset correspond to small, medium and large objects. Subsequently, we analyze the variance of the aspect ratio for these three data sets. It is seen that the variance values in the ascending order are of VOC2007, SSDD and HRSID, which are 0.01682, 0.02802, and 0.38096, respectively. It is also seen that the SAR dataset has a more extreme aspect ratio than the optical image dataset.

TABLE 4. The COCO form Object Detection Evaluation Metrics.

Metrics	Metrics Meaning
AP	AP at IoU=0.50: 0.05: 0.95
AP_{50}	AP at IoU=0.50
AP_{75}	AP at IoU=0.75
AP_S	AP for small objects: are $a < 32^2$
AP_M	AP for medium objects: 32^2< area <96^2
AP_h	AP for large objects:area >96^2

Finally, we analyze the ratio of the anchor area to the image area and find that the objects in SSDD and HRSID datasets occupy small areas of the image, whereas the objects in VOC2007 account for relatively large areas. Therefore, we need to modify the scale and aspect of the anchors before training our model so it learns to efficiently detect ships in the SAR images.

To further demonstrate the impact of the initial anchors of different scales and the aspect ratios on the model performance, we set 4 and 8 as original anchors' scales respectively. In this setting, $scale = 8$ is a general hyperparameter in the optical scene image detection and we perform the IOU K-means with $k = 3$, and $k = 5$ to cluster suitable anchor aspect ratios. In Figure 6, we compare the results of different scales and aspects on the Faster R-CNN. For HRSID and SSDD datasets, the AP_{75} values of $scale = 4$ are around 10.8% and 6.7% larger than those of $scale = 8$, respectively. For different aspects ratios of the anchors that are calculated by the IOU K-means, this method achieves nearly 1.1%, 1.5% and 5.7% performance gains in HRSID dataset, and 1.4%, 0.9% and 6.7% improvement in SSDD datasets in terms of AP_S, AP_M and AP_L, respectively. Note that the performance improvement of detecting the large ships is the highest. Therefore, this method seeks more reasonable initial scales and aspect ratios of the anchors to reduce the difficulty of the model learning and obtain more accurate prediction results.



(a)

(b)

(c)

(d)

FIGURE 7. The detection results with different scales and aspects on HRSID and SSDD datasets respectively: (a), (c) The impact of different scales of initial anchors on the network performance; (b), (d) The impact of different aspects of the initial anchors and their corresponding results with small, medium, and large objects.

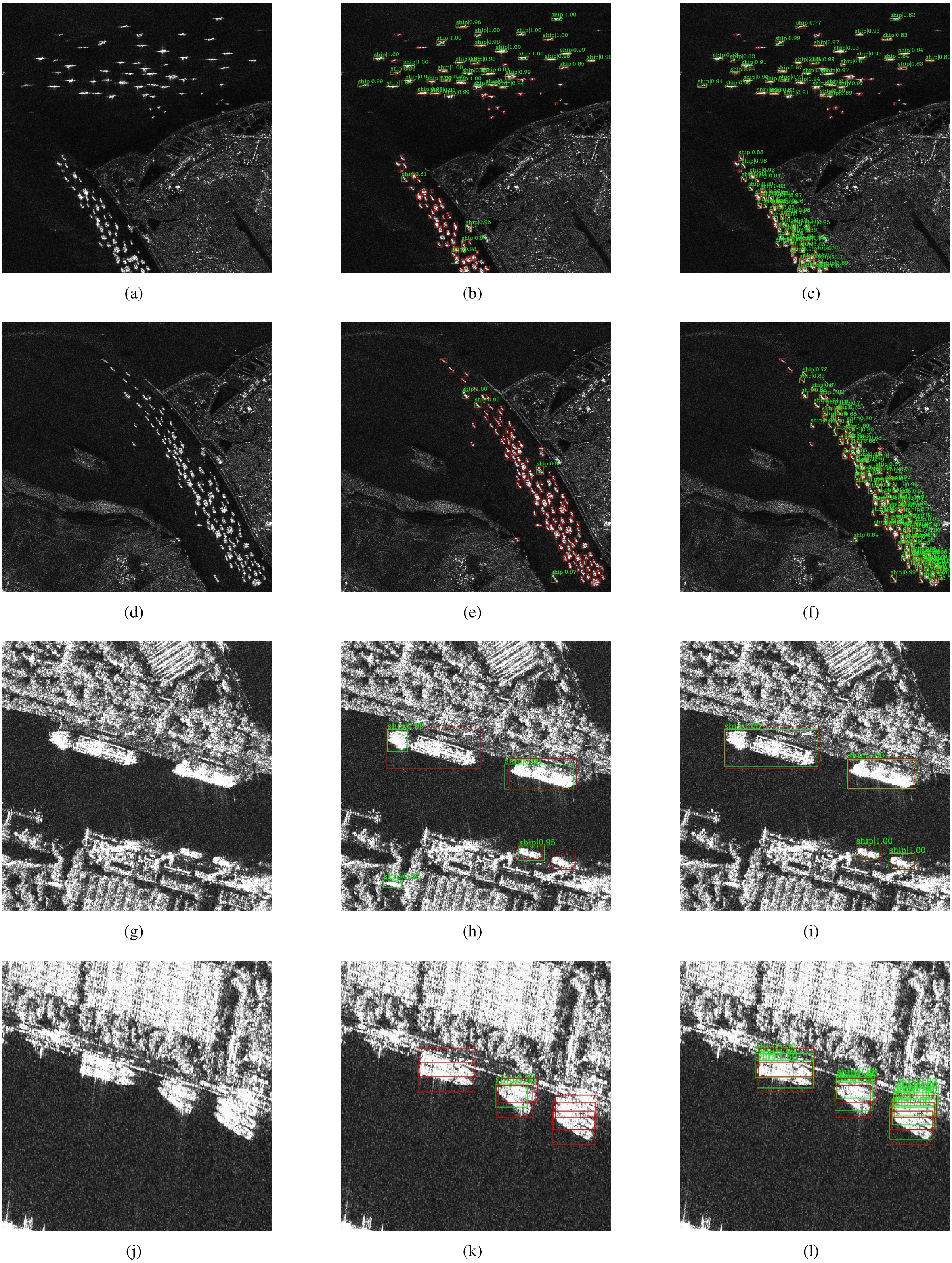
TABLE 5. Quantitative results of the ship detection in SSDD and HRSID datasets based on Faster R-CNN detector. STANet50 (ResNet50 with STA model) overperforms the benchmarking method.

Datasets	Detector	Backbone	mAP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
SSDD	Faster R-CNN	ResNet-50+FPN	62.1	95.4	71.6	57.6	69.4	50.7
		STANet-50+FPN	**63.4**	**95.7**	**76.8**	**59.1**	**70.1**	**55.0**
HRSID	Faster R-CNN	ResNet-50+FPN	68.6	91.9	79.3	70.2	68.6	37.3
		STANet-50+FPN	**69.5**	**92.4**	**81.1**	**70.9**	**68.6**	**37.8**

D. EXPERIMENT RESULT AND ANALYSIS OF STANET

In this subsection, we examine the effectiveness of our proposed method. We evaluate the STA model on SSDD and HRSID datasets with Faster R-CNN detector and follow the standard set of evaluating object detection by the standard mean Average-Precision (AP) scores at the object scales or different box IoUs.

The result is presented in Table 5, where the proposed method, based on Faster R-CNN detector and ResNet50 backbone, demonstrates significant improvement on two SAR datasets. For SSDD dataset, STANet50+FPN achieves 1.3% and 0.9% improvements in terms of mAP for SSDD, and HRSID dataset, respectively. The results also confirm that our method enhances ship detection performance and obtains higher spatial accuracy. Moreover, the value of AP_{75} for STANet50 on SSDD and HRSID is 76.8% and 81.1%, respectively, which suggests improvements equivalent to 5.2% and 1.8%, respectively. The value of AP_{50} is also improved. The results show that the bounding box regression is more efficient and the ship is well covered by



(a) (b) (c)

(d) (e) (f)

(g) (h) (i)

(j) (k) (l)

FIGURE 8. Detection results in four images, shown by the green rectangles. The red rectangles are the ground truth. The first and last two columns of images belong to the HRISD and SSDD dataset, respectively. (a), (d), (g) and (j) demonstrate the original images and (b), (e), (h) and (k) show detection results by Faster R-CNN with ResNet50, respectively. (c), (f), (i) and (l) display results by Faster R-CNN with STA-ResNet50.

TABLE 6. Quantitative results of the ship detection in SSDD and HRSID datasets based on Faster R-CNN detector. STANet50 (ResNet50 with STA model) overperforms the benchmarking method.

Datasets	Detector	Backbone	GFLOPs	Param.	mAP	AP_{50}	AP_{75}
SSDD	RetinaNet	ResNet-50+FPN	36.33	127.82	57.9	93.6	63.1
		STANet-50+FPN	38.85	127.89	**58.3**	**93.9**	**64.4**
	Cascade R-CNN	ResNet-50+FPN	69.15	162.18	63.5	95.5	72.7
		STANet-50+FPN	71.67	162.26	**63.8**	**96.3**	**75.3**
	Faster R-CNN	ResNet-50+FPN	41.35	134.52	62.1	95.4	71.6
		STANet-50+FPN	43.87	134.61	**63.4**	**95.7**	**76.8**
HRSID	RetinaNet	ResNet-50+FPN	36.33	127.82	61.9	85.7	69.6
		STANet-50+FPN	38.85	127.89	**62.6**	**86.2**	**69.9**
	Cascade R-CNN	ResNet-50+FPN	69.15	74.28	70.6	91.3	80.8
		STANet-50+FPN	71.67	74.31	**70.8**	**91.4**	**81.6**
	Faster R-CNN	ResNet-50+FPN	41.35	46.49	68.6	91.9	79.3
		STANet-50+FPN	43.87	46.51	**69.5**	**92.4**	**81.1**

the predicted bounding box. For AP_S, AP_M and AP_L, they are also significantly improved. For SSDD, our STANet50+FPN provides 1.5%, 0.7% and 4.3% gain in terms of AP_S, AP_M and AP_L, respectively. However, it achieves 0.7% and 0.5% improvement in terms of AP_S and AP_L, respectively for HRSID. Compared with the evaluation of SSDD and HRSID, it is seen that the STA model achieves a higher gain for SSDD. This suggests that SAR images in SSDD may include more noise such as complex background and speckles whereas the images in HRSID are less noisy because of their high resolution so the values of AP can be greatly improved due to the characteristic of STA model. Therefore, the STA model plays a significant role in improving detection performance, especially satisfying the complex scene SAR ship detection.

Here we take three typical SAR scenes as the input of our baseline and the proposed method and illustrate the detection results in Figure 8. The figures in the first row in Figure 8, represent the three typical scenes of sparsely distributed small targets, densely clustered small targets, targets in complex background, respectively. As it is seen in Figure 8 (a)-(c), Faster R-CNN with STA-ResNet50 detects more offshore ships with sparse distribution, and accurately detects more inshore targets in a densely clustered ships background. From the detection results shown in Figure 8 (d)-(f), it can further be seen that STA module can improve the detection performance of densely clustered targets. In Figure 7 (g)-(l), the ship detection in pictures with a complex background results in high false alarm rate and inaccurate detection. The STA module addresses these issues through enhancing the extraction of ship features and suppression of irrelevant background as explained in Section 4.A. Comparing the four detection results in the second and third rows of Figure 8, it can be seen that the detector with STA modules can accurately locate the target in various scenes.

E. COMPARISON WITH OTHER METHODS

We embed the STA modules into the popular detector frameworks separately to check if denoising the feature map facilitates ship detection. We select two popular two-stage detection frameworks and a one-stage detection framework, including Faster R-CNN [37], Cascade R-CNN [39] and RetinaNet [44] which use FPN [54] in the backbone. For a fair comparison, we only replace the pretrained backbone model on ImageNet while keeping the other component in the entire detector intact.

Table 6 presented the performance of embedding the backbone with the STA module on three state-of-the-art detectors and two SAR datasets. We find that the STA module introduces few additional parameters and extra calculations, nevertheless the gain of detection performance is significant with more than 1% AP point in terms of AP_{75}. For ship detection using the SSDD dataset, STANet-50 outperforms ResNet-50 by 1.3%, 2.6% and 5.2% on COCO's standard AP_{75} metric for RetinaNet, Cascade R-CNN and Faster R-CNN, respectively. On the other dataset HRISD, STANet-50 achieves nearly 0.3%, 0.8% and 2.2% performance gains comparing with ResNet-50 in terms of AP_{75} for these detectors. For mAP and AP_{50} values, there are slight improvement gains compared to AP_{75}, however there is a significant improvement comparing with the baseline. This corroborates the generalization performance of the STA module for SAR ship detection.

F. COMPARISON WITH OTHER ATTENTION MECHANISM

Next, we choose a representative two-stage detection framework Faster R-CNN to compare STA with several competitive state-of-the-art attention modules, especially for objects with mAP, AP_{50} and AP_{75} metrics. The original backbones are replaced with the corresponding attention embedded ResNet-50, which are pretrained on ImageNet, for a reasonable comparison. The results presented in Tables 7 and 8 show that the STA greatly improves the value of AP_{75} such as 5.2% improvement in SSDD dataset and 2.2% improvement in HRSID dataset. These indicate that the STA module is capable of retaining the feature representation of the precise spatial area and denoise the irrelative areas such as complex background and speckle noise. Therefore it improved the robustness of SAR ship detection.

Meanwhile, the SE/ECA module, as a representative of the channel attention module obtains an improvement that

TABLE 7. Comparison results with other attention models. The performance of the system which is based on Faster R-CNN detector for objects of three scales on SSDD dataset. The size of the image inputs is 416 × 416. The best and the second best are marked as bold and blue, respectively.

Backbone	GFLOPs	Param.	mAP	AP_{50}	AP_{75}
ResNet-50	41.35	46.49	62.1	95.4	71.6
SE+ResNet-50	43.88	**46.51**	$62.8_{(+0.7)}$	$\mathbf{96.1_{(+0.7)}}$	$70.8_{(-0.8)}$
ECA+ResNet-50	**41.35**	**46.50**	$\mathbf{63.0_{(+0.9)}}$	$\mathbf{96.2_{(+0.8)}}$	$\mathbf{73.7_{(+2.1)}}$
CBAM+ResNet-50	43.88	46.49	$61.3_{(-0.8)}$	$95.1_{(-0.3)}$	$71.9_{(+0.3)}$
NL+ResNet-50	55.00	51.46	$62.2_{(+0.1)}$	$95.4_{(+0.0)}$	$72.7_{(+1.1)}$
STA+ResNet50	**43.87**	46.51	$\mathbf{63.4_{(+1.0)}}$	$95.7_{(+0.3)}$	$\mathbf{76.8_{(+5.2)}}$

TABLE 8. The performance of the system which is based on Faster R-CNN detector for the objects of three scales on HRSID dataset. The size of the input images is 800 × 800. The notations are the same as in Table 7.

Backbone	GFLOPs	Param.	mAP	AP_{50}	AP_{75}
ResNet-50	41.35	134.52	**68.6**	91.9	79.3
SE+ResNet-50	43.88	**135.46**	$68.3_{(-0.3)}$	$92.0_{(+0.1)}$	$79.4_{(+0.1)}$
ECA+ResNet-50	**41.35**	135.45	$68.5_{(-0.1)}$	$\mathbf{92.1_{(+0.2)}}$	$\mathbf{79.5_{(+0.2)}}$
CBAM+ResNet-50	43.88	**135.41**	$68.3_{(-0.3)}$	$\mathbf{92.1_{(+0.2)}}$	$\mathbf{79.5_{(+0.2)}}$
NL+ResNet-50	55.00	152.78	$68.0_{(-0.6)}$	$91.8_{(-0.1)}$	$79.3_{(+0.0)}$
STA+ResNet50	**43.87**	**135.46**	$\mathbf{69.5_{(+0.9)}}$	$\mathbf{92.4_{(+0.5)}}$	$\mathbf{81.5_{(+2.2)}}$

falls between CBAM and NL (Non-Local attention module). It indicates that recalibrating channel-wise feature responses is helpful for the representation of model both in optical images and SAR images. Comparing CBAM and STA attention mechanism, CBAM is a method of fusing spatial attention and channel attention, whereas the STA module can be considered as a special spatial attention mechanism for denoising feature maps. The performance of CBAM is dropped compared to baseline and this suggests that using the spatial attention mechanism for the SAR image should be done carefully otherwise it may cause irrelevant spatial information to be falsely enhanced and submerge the features of the ships because of the characteristic of the SAR images. The NL module is a method that integrates non-local mean [55] operation and attention mechanism. We compare it with STA and find that the NL module has a feature denoising function in theory, but with a slight improvement and in some cases the achieved performance maybe even worse than the baseline for some metrics. In summary, the attention module proposed based on the optical images is not necessarily suitable for the SAR ship detection images, and STA module can reconstruct the representative feature maps and suppress noise in feature maps.

IV. DISCUSSION

In this Section, we use Faster R-CNN with ResNet50 as our baseline model to further explore the advantages of the STA module.

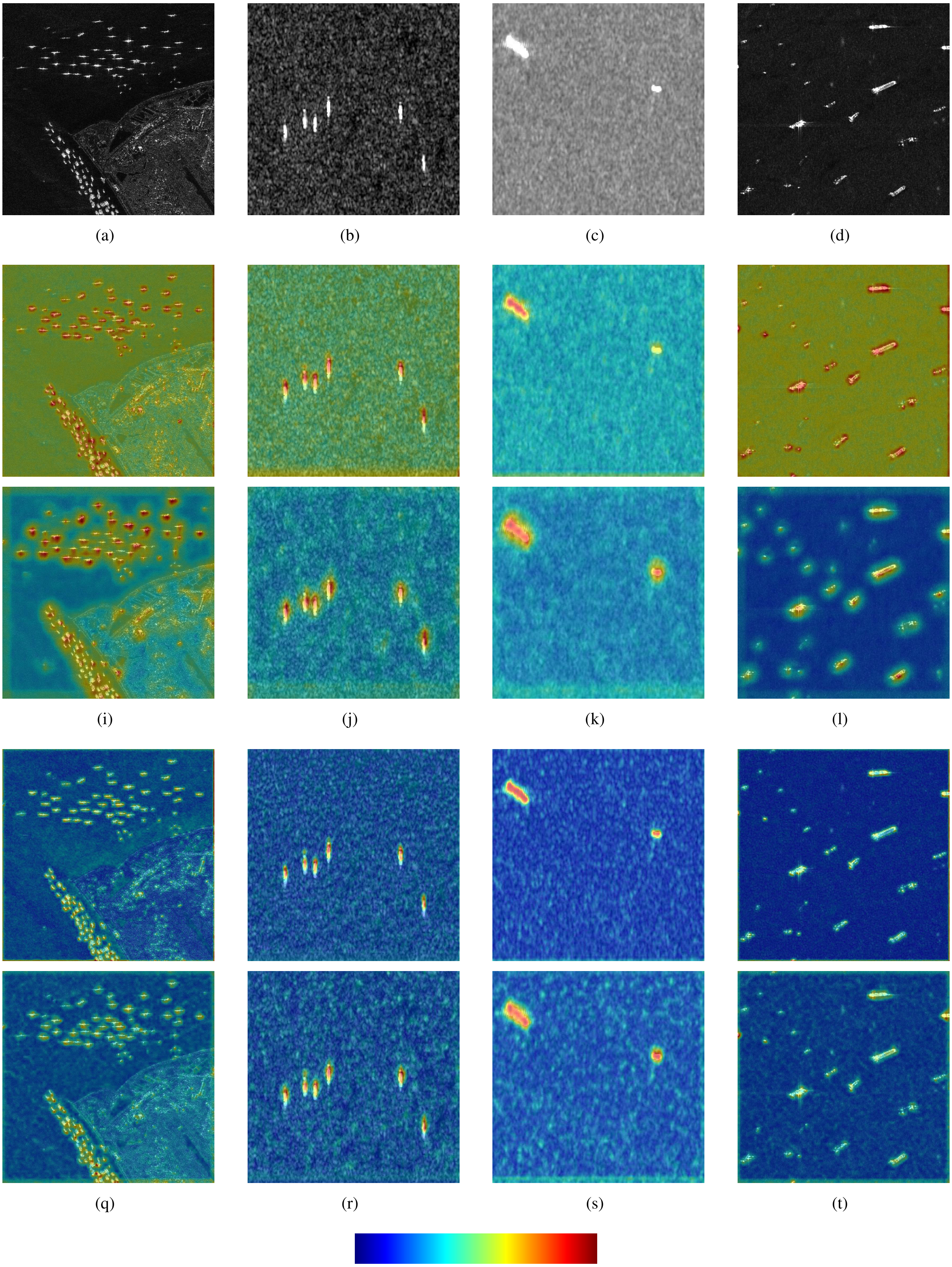
A. VISUAL ANALYSIS OF THE FEATURE MAPS

In general, as the convolution layers deepen, the size of the feature maps is decreased and only more abstract semantic meanings are preserved. If we visualize the feature maps on the deep layer, the important spatial information can be observed. Therefore, the feature map visualization verifies the effectiveness of the proposed method. In this section, we utilize the heatmap to visualize the spatial response of different stages' response where the blue colour indicates low spatial response, and the red presents a high response. To better understand the relationship between the feature map and the original image, here we resize the feature map to the same size as the images and then superimpose them with a certain coefficient ($c = 0.2$). The results are shown in Figure 8.

As it is seen in Figure 8, compared with our baseline, the model's response to the background becomes very low because of the STA module. Furthermore, the detailed features of the target are highlighted. By comparing the results in Figure 8 (a), (e) and (i), it is also seen that the background features are not suppressed in the first stage for our baseline but the STA module enables suppressing the response of the background in the feature map at the first stage.

Furthermore, we can assume that the model has limited ability to suppress irrelevant information as the network is deeper during the training process while the STA module can improve this ability. The STA module results in a more discriminative response of the target while suppressing irrelevant background information, see, Figure 9(e) and (i). This is the reason why it has good performance on densely clustered ship detection in Figure 8(a). It is seen in Figure 9(b), (f), (j), (k), (g) and (k), that the speckled background in the SAR image significantly interferes with the feature extraction of the model, STA also decreases the response of the speckled noise in the feature map to focus on the target feature. For the offshore clustered ships with less complex background and noise in Figure 9(d), (h) and (i), STA obtains more discriminative feature maps faster.

By visualizing the intermediate feature map of our baseline and the proposed method, it is verified that CNN has a limited ability to denoise and eliminate irrelevant information in

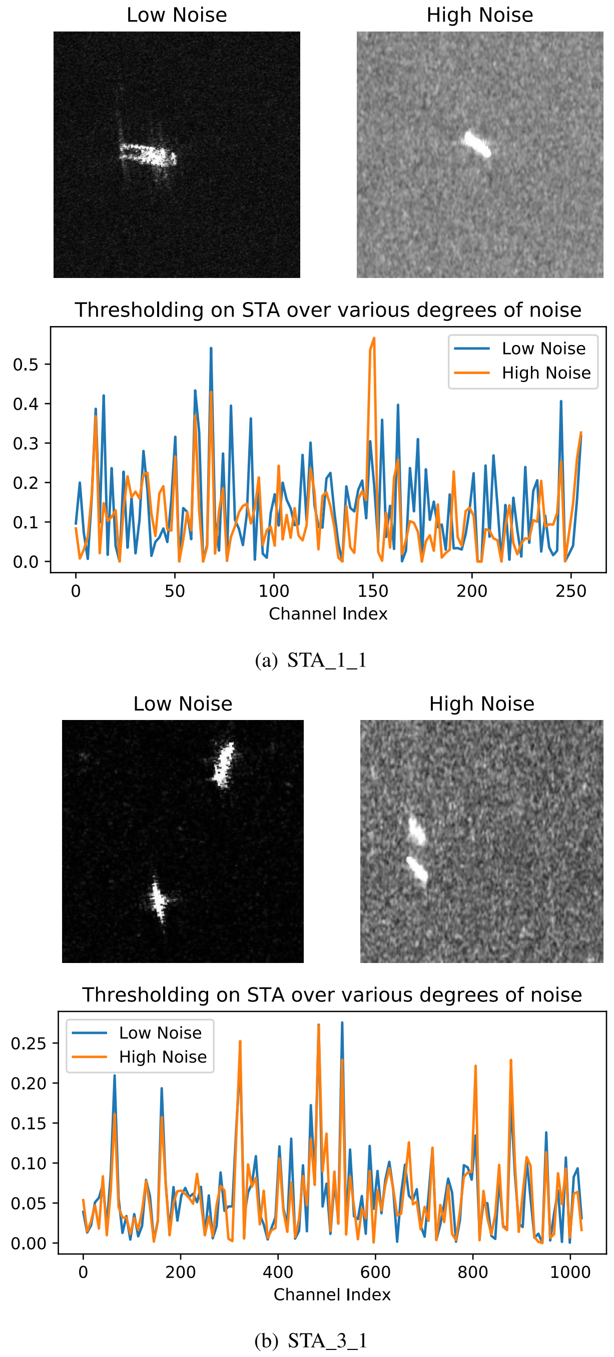


(a) (b) (c) (d)

(i) (j) (k) (l)

(q) (r) (s) (t)

FIGURE 9. Visualization of the feature maps: (a), (b), (c) and (d) denote the original image patches from HRSID and SSDD datasets; (e), (f), (g) and (h) display the feature map in the first two stages of our baseline (ResNet50), respectively; (i), (j), (k) and (l) exhibit the feature map of the first and second stages of our method (STA-ResNet50), respectively. The response intensity gradually decreases from red to blue.



(a) STA_1_1

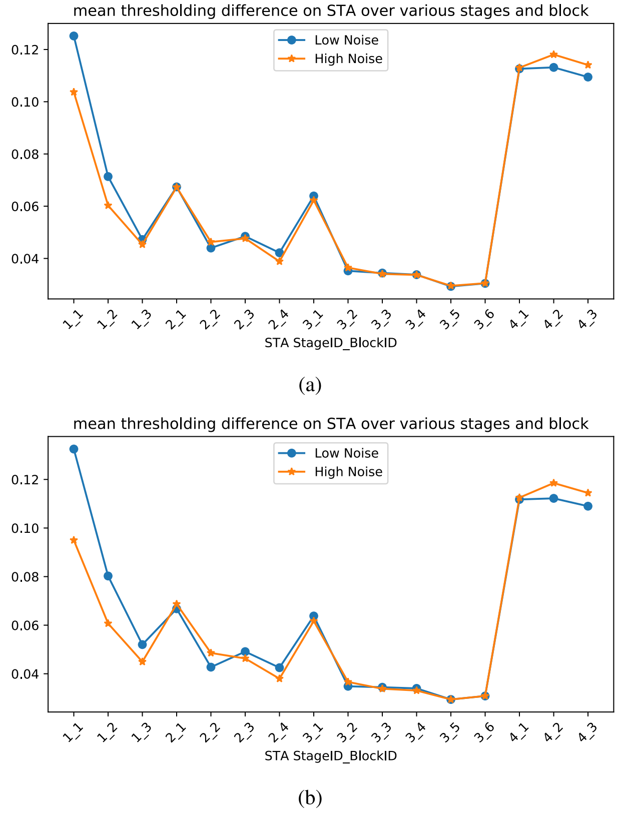
(b) STA_3_1

FIGURE 10. Thresholding caused by the STA module at different depth in the STA-ResNet-50 on SSDD. Each set of thresholding is named according to the following scheme: STA_stageID_blockID.

feature extraction. By using STA, the denoising and suppression ability of the feature map is improved, hence that the detection accuracy is increased.

B. THRESHOLDING ANALYSIS OF STA MODULE

To provide a clearer picture of the thresholding function of the STA module, in this section we study example values from the STA-ResNet-50 model and examine their distribution



(a)

(b)

FIGURE 11. Averaged thresholding values at different depth in the STA-ResNet-50.

for different noise degrees at various depths in the network. In particular, we would like to understand how thresholding varies in different stages across images of different noise degrees.

First, we calculate the thresholding values for different depth in each STA unit. Figure 10 show the thresholding values in all channels in the shallow stage and the deeper stage for two images with different degree noise. It is seen that in the shallow stage, for high background noise, the thresholding values are generally lower than that of in the low-noise images. It suggests that due to the limited representation capability of the shallow stage, the noise slightly affects the discrimination thresholding values between the target and the background. However, given the action of multiple block STAs in previous stages, the thresholding values of the target and the background are fully distinguished in the deeper stage. The thresholding values of low-noise and high-noise images tend to be similar.

We make the following observation about the averaged thresholding values in each stage and blocks for images with different noise degree. Figure 11 shows the results for two random samples at different STA units. A surprising pattern is observed about the role of thresholding values across various depth: the deeper the network, the more similar thresholding values in the low and middle-level stages (e.g., STA_2_3, STA_1_3). However, at much higher layers (e.g., STA_5_2), the information expressed by the feature map is highly abstract, so the pattern is disappeared.

V. CONCLUSION

In this paper, we proposed two methods, the IOU k-means, and the STA module. The IOU k-means is a pre-design anchor technique that results in performance enhancement. The STA module is an architectural block designed to enhance the representational power of a network by empowering it to perform dynamic feature denoising and recalibration. Through extensive experimental studies, we showed the effectiveness of IOU k-means and STA module, which achieves significant improvement across multiple datasets in SAR image ship detection. Besides, the experiments on IOU k-means illustrated the distinction between optical and SAR images. The STA module shed light on the limitations of the previous architectures to adequately enhance the spatial response of the targets' feature and suppress irrelevant information. The proposed method in this paper improves ship detection in SAR images with complex background and noise.

ACKNOWLEDGMENT

(Rui Wang and Sihan Shao contributed equally to this work.)

REFERENCES

[1] H. Dai, L. Du, Y. Wang, and Z. Wang, "A modified CFAR algorithm based on object proposals for ship target detection in SAR images," *IEEE Geosci. Remote Sens. Lett.*, vol. 13, no. 12, pp. 1925–1929, Dec. 2016.

[2] T. Li, Z. Liu, R. Xie, and L. Ran, "An improved superpixel-level CFAR detection method for ship targets in high-resolution SAR images," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 11, no. 1, pp. 184–194, Jan. 2018.

[3] T. Tang, D. Xiang, and H. Xie, "Multiscale salient region detection and salient map generation for synthetic aperture radar image," *J. Appl. Remote Sens.*, vol. 8, no. 1, Dec. 2014, Art. no. 083501.

[4] C. Wang, F. Bi, W. Zhang, and L. Chen, "An intensity-space domain CFAR method for ship detection in HR SAR images," *IEEE Geosci. Remote Sens. Lett.*, vol. 14, no. 4, pp. 529–533, Apr. 2017.

[5] S. Wang, M. Wang, S. Yang, and L. Jiao, "New hierarchical saliency filtering for fast ship detection in high-resolution SAR images," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 1, pp. 351–362, Jan. 2017.

[6] X. Wang and C. Chen, "Ship detection for complex background SAR images based on a multiscale variance weighted image entropy method," *IEEE Geosci. Remote Sens. Lett.*, vol. 14, no. 2, pp. 184–187, Feb. 2017.

[7] D. Xiang, T. Tang, Y. Ban, and Y. Su, "Man-made target detection from polarimetric SAR data via nonstationarity and asymmetry," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 9, no. 4, pp. 1459–1469, Apr. 2016.

[8] D. Xiang, T. Tang, W. Ni, H. Zhang, and W. Lei, "Saliency map generation for SAR images with Bayes theory and heterogeneous clutter model," *Remote Sens.*, vol. 9, no. 12, p. 1290, Dec. 2017.

[9] M. Petit, J.-M. Stretta, H. Farrugio, and A. Wadsworth, "Synthetic aperture radar imaging of sea surface life and fishing activities," *IEEE Trans. Geosci. Remote Sens.*, vol. 30, no. 5, pp. 1085–1089, 1992.

[10] F. Gao, F. Ma, Y. Zhang, J. Wang, J. Sun, E. Yang, and A. Hussain, "Biologically inspired progressive enhancement target detection from heavy cluttered SAR images," *Cognit. Comput.*, vol. 8, no. 5, pp. 955–966, Oct. 2016.

[11] Z. Zhao, K. Ji, X. Xing, H. Zou, and S. Zhou, "Ship surveillance by integration of space-borne SAR and AIS-review of current research," *J. Navigat.*, vol. 67, no. 1, p. 177, 2014.

[12] S. Brusch, S. Lehner, T. Fritz, M. Soccorsi, A. Soloviev, and B. van Schie, "Ship surveillance with TerraSAR-X," *IEEE Trans. Geosci. Remote Sens.*, vol. 49, no. 3, pp. 1092–1103, Mar. 2011.

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

[14] M. E. Smith and P. K. Varshney, "VI-CFAR: A novel CFAR algorithm based on data variability," in *Proc. IEEE Nat. Radar Conf.*, May 1997, pp. 263–268.

[15] G. Gao, L. Liu, L. Zhao, G. Shi, and G. Kuang, "An adaptive and fast CFAR algorithm based on automatic censoring for target detection in high-resolution SAR images," *IEEE Trans. Geosci. Remote Sens.*, vol. 47, no. 6, pp. 1685–1697, Jun. 2009.

[16] A. Farrouki and M. Barkat, "Automatic censoring CFAR detector based on ordered data variability for nonhomogeneous environments," *IEE Proc.-Radar, Sonar Navigat.*, vol. 152, no. 1, pp. 43–51, Feb. 2005.

[17] D. Pastina, F. Fico, and P. Lombardo, "Detection of ship targets in COSMO-SkyMed SAR images," in *Proc. IEEE RadarCon (RADAR)*, May 2011, pp. 928–933.

[18] Q. V. Le, J. Ngiam, A. Coates, A. Lahiri, B. Prochnow, and A. Y. Ng, "On optimization methods for deep learning," in *Proc. 28th Int. Conf. Int. Conf. Mach. Learn. (ICML)*. Madison, WI, USA: Omnipress, 2011, pp. 265–272.

[19] M. Messina, M. Greco, L. Fabbrini, and G. Pinelli, "Modified Otsu's algorithm: A new computationally efficient ship detection algorithm for SAR images," in *Proc. Tyrrhenian Workshop Adv. Radar Remote Sens. (TyWRRS)*, Sep. 2012, pp. 262–266.

[20] Q. V. Le, W. Y. Zou, S. Y. Yeung, and A. Y. Ng, "Learning hierarchical invariant spatio-temporal features for action recognition with independent subspace analysis," in *Proc. CVPR*, Jun. 2011, pp. 3361–3368.

[21] Q. Fan, F. Chen, M. Cheng, S. Lou, R. Xiao, B. Zhang, C. Wang, and J. Li, "Ship detection using a fully convolutional network with compact polarimetric SAR images," *Remote Sens.*, vol. 11, no. 18, p. 2171, Sep. 2019.

[22] 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, Aug. 2018.

[23] X. Huang, W. Yang, H. Zhang, and G.-S. Xia, "Automatic ship detection in SAR images using multi-scale heterogeneities and an a contrario decision," *Remote Sens.*, vol. 7, no. 6, pp. 7695–7711, Jun. 2015.

[24] K. El-Darymli, P. McGuire, D. Power, and C. 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.

[25] D. J. Crisp, "The state-of-the-art in ship detection in synthetic aperture radar imagery," Dept. Defence, Austral. Government, Tech. Rep., 2004, p. 115.

[26] C. C. Wackerman, K. S. Friedman, W. G. Pichel, P. Clemente-Colón, and X. Li, "Automatic detection of ships in RADARSAT-1 SAR imagery," *Can. J. Remote Sens.*, vol. 27, no. 5, pp. 568–577, Oct. 2001.

[27] Z. Liu, F. Li, N. Li, R. Wang, and H. Zhang, "A novel region-merging approach for coastline extraction from Sentinel-1A IW mode SAR imagery," *IEEE Geosci. Remote Sens. Lett.*, vol. 13, no. 3, pp. 324–328, Mar. 2016.

[28] C. Liu, J. Yang, J. Yin, and W. An, "Coastline detection in SAR images using a hierarchical level set segmentation," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 9, no. 11, pp. 4908–4920, Nov. 2016.

[29] C. Liu, Y. Xiao, and J. Yang, "A coastline detection method in polarimetric SAR images mixing the region-based and edge-based active contour models," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 7, pp. 3735–3747, Jul. 2017.

[30] R. Touzi, A. Lopes, and P. Bousquet, "A statistical and geometrical edge detector for SAR images," *IEEE Trans. Geosci. Remote Sens.*, vol. 26, no. 6, pp. 764–773, Nov. 1988.

[31] M. Weiss, "Analysis of some modified cell-averaging CFAR processors in multiple-target situations," *IEEE Trans. Aerosp. Electron. Syst.*, vol. AES-18, no. 1, pp. 102–114, Jan. 1982.

[32] X. Leng, K. Ji, K. Yang, and H. Zou, "A bilateral CFAR algorithm for ship detection in SAR images," *IEEE Geosci. Remote Sens. Lett.*, vol. 12, no. 7, pp. 1536–1540, Jul. 2015.

[33] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, pp. 436–444, May 2015.

[34] J. Han, D. Zhang, G. Cheng, N. Liu, and D. Xu, "Advanced deep-learning techniques for salient and category-specific object detection: A survey," *IEEE Signal Process. Mag.*, vol. 35, no. 1, pp. 84–100, Jan. 2018.

[35] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Region-based convolutional networks for accurate object detection and segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 1, pp. 142–158, Jan. 2016.

[36] R. Girshick, "Fast R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 1440–1448.

[37] 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.

[38] Z. Cai, Q. Fan, R. S. Feris, and N. Vasconcelos, "A unified multi-scale deep convolutional neural network for fast object detection," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2016, pp. 354–370.

[39] Z. Cai and N. Vasconcelos, "Cascade R-CNN: Delving into high quality object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 6154–6162.

[40] 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. (CVPR)*, Jun. 2016, pp. 779–788.

[41] J. Redmon and A. Farhadi, "YOLO9000: Better, faster, stronger," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 7263–7271.

[42] J. Redmon and A. Farhadi, "YOLOv3: An incremental improvement," 2018, *arXiv:1804.02767*. [Online]. Available: https://arxiv.org/abs/1804.02767

[43] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg, "SSD: Single shot multibox detector," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2016, pp. 21–37.

[44] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollar, "Focal loss for dense object detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 2, pp. 318–327, Feb. 2020.

[45] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 2961–2969.

[46] S. Wei, X. Zeng, Q. Qu, M. Wang, H. Su, and J. Shi, "HRSID: A high-resolution SAR images dataset for ship detection and instance segmentation," *IEEE Access*, vol. 8, pp. 120234–120254, 2020.

[47] J. Li, C. Qu, and J. Shao, "Ship detection in SAR images based on an improved faster R-CNN," in *Proc. SAR Big Data Era, Models, Methods Appl. (BIGSARDATA)*, Nov. 2017, pp. 1–6.

[48] D. L. Donoho, "De-noising by soft-thresholding," *IEEE Trans. Inf. Theory*, vol. 41, no. 3, pp. 613–627, May 1995.

[49] K. Isogawa, T. Ida, T. Shiodera, and T. Takeguchi, "Deep shrinkage convolutional neural network for adaptive noise reduction," *IEEE Signal Process. Lett.*, vol. 25, no. 2, pp. 224–228, Feb. 2018.

[50] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7132–7141.

[51] X. Li, W. Wang, X. Hu, and J. Yang, "Selective kernel networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 510–519.

[52] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "CBAM: Convolutional block attention module," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 3–19.

[53] K. Chen *et al.*, "MMDetection: Open MMLab detection toolbox and benchmark," 2019, *arXiv:1906.07155*. [Online]. Available: http://arxiv.org/abs/1906.07155

[54] T.-Y. Lin, P. Dollar, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 2117–2125.

[55] A. Buades, B. Coll, and J.-M. Morel, "A non-local algorithm for image denoising," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, vol. 2, Jun. 2005, pp. 60–65.

RUI WANG received the B.S. degree in measurement-control technology and instrument and the M.S. degree in instrument science and technology from the Changchun University of Science and Technology, in 2016 and 2018, respectively, where he is currently pursuing the Ph.D. degree in intelligent test technology and system. His research interests include laser measurement technology, modern metrology testing technology, and imaging processing technology.

SIHAN SHAO is currently pursuing the B.S. degree in opto-electronics information science and engineering with the Changchun University of Science and Technology. He currently plans to applicate the joint master's and Ph.D. degrees in instrument science and technology with the Changchun University of Science and Technology. His research interests include pattern recognition and computational imaging.

MENGYU AN is currently pursuing the B.S. degree in opto-electronics information science and engineering with the Changchun University of Science and Technology. He currently plans to pursue the Ph.D. degree in optical engineer. He is also the President of the Artificial Intelligence and Photoelectric Detection Association, Changchun University of Science and Technology. His research interests include laser measurement technology and environmental perception technology.

JIAYI LI is currently pursuing the B.S. degree in opto-electronics information science and engineering with the Changchun University of Science and Technology. She currently plans to applicate the master's degree with the Changchun University of Science and Technology. Her research interests include pattern recognition and laser measurement.

SHIFENG WANG received the B.S. degree in mechatronic engineering, the M.S. degree in measuring and testing technology and instruments, and the Ph.D. degree in instruments science and technology from the Changchun University of Science and Technology, in 2002, 2005, and 2008, respectively, and the Ph.D. degree from the Faculty of Engineering and Information Technology, University of Technology Sydney, in 2013. He has been a Professor with the Department of Detection and Information Engineering. His research interests include robot environment perception, machine vision, and photoelectric target detection and tracking.

XIPING XU received the B.S. degree in electronic engineering, the M.S. degree in precision machinery engineering, and the Ph.D. degree in optical engineering from the Changchun University of Science and Technology, in 1993, 1999, and 2004, respectively. Since 2009, he has been a Professor with the Instruments Science and Technology, Changchun University of Science and Technology. His research interests include intelligent instrument systems and photoelectric detection technology.