IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received July 23, 2018, accepted August 27, 2018, date of publication September 10, 2018, date of current version October 8, 2018.

Digital Object Identifier 10.1109/ACCESS.2018.2869289

A Cascade Coupled Convolutional Neural Network Guided Visual Attention Method for Ship Detection From SAR Images

JUANPING ZHAO, ZENGHUI ZHANG, (Member, IEEE), WENXIAN YU, AND TRIEU-KIEN TRUONG, (Life Fellow, IEEE)
Shanghai Key Laboratory of Intelligent Sensing and Recognition, Shanghai Jiao Tong University, Shanghai 200240, China

Corresponding author: Zenghui Zhang (zenghui.zhang@sjtu.edu.cn)

This work was supported by the National Natural Science Foundation of China under Grant 61331015.

ABSTRACT Convolutional neural networks (CNNs) have found applications in ship detection from synthetic aperture radar (SAR) images. However, there are some challenges hamper their advance. First, the detected bounding boxes are not very compact. Second, there are quite a few missing detections for small and densely clustered ships. Third, objects with analogical scatterings on land are detected as ships by making mistake. This is due to: 1) the CNN-based SAR ship detectors cannot utilize the spatial information very sufficiently; 2) features learned from CNNs only describe SAR images in space domain while neglecting the information hidden in frequency domain; and 3) information contained in the meta-data file, which may link to other sources, is not taken into account. To overcome these problems, in this paper, a cascade coupled CNN-guided (3C2N-guided) visual attention method for SAR ship detection is proposed. This method considers the newly presented 3C2N model as a qualified ship proposal generator because the images' spatial information is utilized more sufficiently. The 3C2N model, with coupled CNN as the baseline, consists of a sequence of cascade detectors for training. Complementally, a pulse cosine transformation-based visual attention model in frequency domain is operated on the adaptive regions for ship discrimination. This could further refine the proposals' locations and could significantly reduce the missing detections and false alarms. In addition, the digital elevation model data are adopted to remove ship-like targets on land. Experimental evaluations on 25 Sentinel-1 images demonstrate that the proposed method is superior to the previous state-of-the-art methods.

INDEX TERMS Cascade coupled convolutional neural network (3C2N), pulse cosine transformation (PCT), ship detection, synthetic aperture radar (SAR).

I. INTRODUCTION

With a wealth of irreplaceable characteristics, such as day-and-night, all-weather, active imaging and wide-swath, synthetic aperture radar (SAR) shows its unique superiority and has been an important tool for marine surveillance regardless of cloud cover conditions [1]–[7]. In particular, ship detection from SAR images is playing an increasingly essential role both in civil and military regime [8]–[10]. However, it is still a challenging task due to the relatively small size, i.e., a ship in Sentinel-1 images may only account for several pixels in length. In addition, the various complex background conditions further render the method difficult for ship detection, for example, sometimes ships might be located in high clutter ocean environments.

In the past decades, various threads relating to SAR ship detection have been explored by researchers. As mentioned in [11]–[14], a practical architecture of a SAR ship detection system usually consists of four stages: that is, land masking, preprocessing, prescreening, and discrimination. The land masking stage distinguishes ocean area from land and defines a smaller scope to be detected for subsequent stages, trying to eliminate the adverse effect caused by land [15]. The preprocessing stage is intended to transform the original SAR imagery into a new image from which ship detection is more easier and the detection performance is improved. There are various preprocessing ways, such as speckle filtering [16]. In the prescreening stage, some potential ship pixels are detected as candidate ship targets. Among the approaches

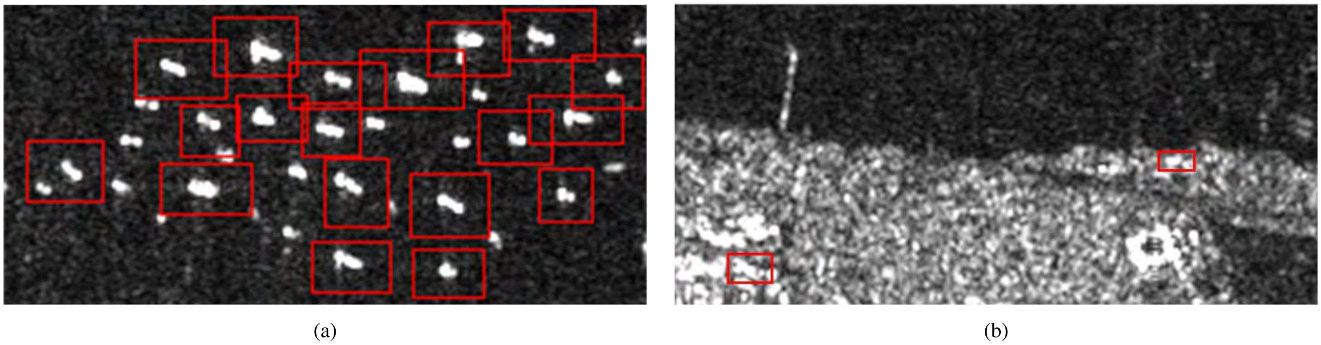


FIGURE 1. SAR ship detection result by the state-of-the-art CCNN method. The red rectangles indicate the detected bounding boxes by CCNN. (a) CCNN detection result in areas, where ships are small in size and are densely clustered. The bounding boxes are very loose and there are quite a few missing detections. (b) CCNN detection result on land area. Some ship-like objects on land are detected by making mistake.

searching for the candidate pixels, the constant false alarm rate (CFAR) [8], [17] and the generalized-likelihood ratio test algorithm [18] account for the most prevailing position. The core of which is based on the sea clutter modeling and parameter estimation to find the optimal pixels. Finally, the ship discrimination stage is intended to exclude the regions containing false alarms, and accept the ones containing real targets [19]. Another stream in traditional SAR ship detection system is the visual attention based methods [9], [20]–[22], which is inspired by human vision system. The visual attention model has also been used in Polarimetric SAR ship detection [23], [24]. By applying this technology into the ship detection system, the detection performance has further improved so that both the missing detections and the false alarms have been decreased. These traditional detection systems share three advantages. Firstly, they make full use of the statistical characteristics of the SAR image. Secondly, the analysis in frequency domain could improve the detection performance further. Thirdly, with the aid of some prior knowledge, such as shape information, the detection performance has been highly improved. However, all these traditional detection systems for SAR images still pose great challenges. On one hand, the hand-crafted features for discrimination have limited representation capability for ship description, thereby leading to a low detection accuracy, especially when they are immersed in complex scenes. On the other hand, the traditional systems show a multi-step operation mode, which is very time-consuming and therefore, it is not suitable for software implementations.

It is well-known that convolutional neural networks (CNNs) [25]–[27] are multi-layer architectures, which enable one to extract multi-level feature representations to depict ship targets. With the end-to-end deep learning framework, CNNs have achieved significant success in object detection from optical remote sensing images. For more details, see [28]–[31]. They also have been verified the powerful capability in SAR ship detection. In order to detect small ship targets automatically, Kang *et al.* [32] firstly designed a CNN-based method for SAR ship detection, which is composed of a region proposal network (RPN) and an object detection network with contextual features. The detection performance shown in this work has been improved for small-sized ships by fusing both the deep semantic and shallow high-resolution features. Recently, a densely connected multi-scale neural network (DCMSNN) is proposed by Jiao *et al.* [33] to detect multi-scale and multi-scene ship targets. Based on Faster R-CNN framework [34], the DCMSNN leverages the densely connected network [35] as its main convolutional trunk. More recently, in 2018, Juanping *et al.* [36] proposed a coupled convolutional neural network (CCNN) for small and densely clustered SAR ship detection. This method is mainly composed of an exhaustive ship proposal network (ESPN) for proposal generation and an accurate ship discriminative network (ASDN) for excluding false alarms. In ESPN, features from different layers are reused and the proposals are predicted from several representative intermediate layers to obtain reliable ship proposals as many as possible. Note that in ASDN, the context information for each proposal is combined with the original deep features in order to rule out false alarms as accurately as possible.

Clearly, the success of CNN-based methods for SAR ship detection can be attributed to the following reasons: Firstly, CNNs could learn image features automatically in the end-to-end deep learning framework. Secondly, in addition to some multi-scale strategies [37]–[40], image information in space domain is utilized in a relatively sufficient mode by a series of convolution, pooling, and other spatial-like operations in CNNs [41]. Even though highly improved detection results have been advanced by CNNs to a great extent in comparison with the traditional ship detection systems, there still exist some challenges for CNN-based methods. First of all, the regressed bounding boxes are not very compact, which is not desirable for practical applications. Meanwhile, the missing detections are very severe for ships in some complex background conditions and areas, where small and densely clustered ship targets are filled. In addition, few false alarms, especially the ones on land, still cannot be effectively removed. These drawbacks can be visualized clearly in Fig. 1 in which the red rectangles indicate the detected bounding boxes. The detection results in Fig. 1(a) give us a view that the bounding boxes are too loose so that it sometimes leads to missing detections for small-sized ships in densely packed areas. In Fig. 1(b), the ship-like objects on land are detected by making mistake. The reasons can be attributed by: 1) The images' spatial information by CNN is not utilized very sufficiently; 2) Information hidden in frequency domain has been neglected by CNNs; 3) Information attached in the meta-data file, which may link to other source data, has not been utilized.

To the best of our knowledge, researches integrating the visual attention model in frequency domain into the CNN-based SAR ship detection framework have never seen in literature. From the perspective of information mining, features learned from CNNs only describe the SAR image in space domain, while information relating to frequency domain has been completely neglected. It is worthwhile to point out that information in frequency domain could provide an extremely important clue for ship detection, due to the fact that ships are always moving-targets floating on the ocean surface.

In order to achieve these goals, a cascade coupled CNN guided (3C2N-guided) visual attention method for SAR ship detection is proposed in this paper. This method firstly generates the rough locations as ship proposals by a pre-trained 3C2N model, which could provide a sufficient way to guide ship detection. With the goal of utilizing the spatial information in CNN more sufficiently, the 3C2N deep learning framework leverages the CCNN architecture as base network, and follows a sequence of discrimination networks, rather than single, to further improve the detection quality. This is motivated by the literature [42]. Then, adaptive proposal regions are generated and they are despeckled by the non-local-mean (NLM) algorithm [43] to make the following discrimination work easier. In order to fully utilize information hidden in the frequency domain and reduce the computation complexity, the pulse cosine transformation (PCT) model [44] is employed on the adaptive regions, rather than the original wide-swath SAR imagery. Thus, the saliency maps are generated in which the pixels belonging to the ship targets are highlighted, while others are weakened. Hence, with the constraints of ship length and digital elevation model (DEM) available, a refined detection result could be achieved. Finally, to avoid repeated detections, the non-maximum suppression (NMS) [45] is adopted to reduce redundancy. The proposed method can be essentially viewed as combining the advantages of spatial superiority by the 3C2N method, including the cascade spatial structure, the ships' structure feature, geometric feature, contextual feature, etc., and the superiority of frequency domain by the PCT-based visual attention method. Experiments on the real Sentinel-1 images demonstrate the superiority of the proposed method for SAR ship detection, achieving the accurate bounding boxes, the lowered missing detections, and the lowered false alarms.

The rest of this paper is organized as follows: Section II illustrates the proposed method in detail, including the 3C2N model for ship proposal generation and the PCT-based visual attention model for ship discrimination. Section III provides the details of the experiments, the detection results, and the corresponding discussions. This paper concludes with a brief summary in Section IV.

II. METHODOLOGY

This section is dedicated to propose a new method, which is composed of two stages: ship proposal generation by using a newly presented 3C2N model and followed by a PCT-based visual attention model to discriminate ship targets. The 3C2N model is trained with a large amount of Sentinel-1 images and could characterize testing images in space domain more sufficiently. Complementally, the PCT-based visual attention model features the testing images in frequency domain. The overall scheme of the proposed method is presented in Fig. 2.

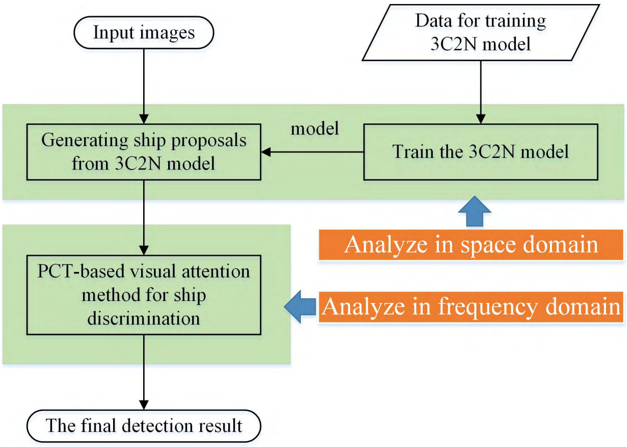

Input images
Data for training 3C2N model
model
Generating ship proposals from 3C2N model
Train the 3C2N model
Analyze in space domain
PCT-based visual attention method for ship discrimination
Analyze in frequency domain
The final detection result

FIGURE 2. **The overall framework of the proposed method.**

A. SHIP PROPOSAL GENERATOR: A CASCADE COUPLED CONVOLUTIONAL NEURAL NETWORK

Zhao *et al.* [36] analyzed many SAR ship detection methods and have concluded that the CCNN method achieves the state-of-the-art performance. Based on the architecture of the CCNN model, the 3C2N model, a cascade structure, is proposed to generate ship proposals. The network architecture is also motivated by the literature [42]. Aiming to improve the detection quality, the 3C2N model is composed of an ESPN and two ASDNs. In what follows, the architecture of the 3C2N model and the mechanism of the way to generate ship proposals are presented.

1) THE 3C2N ARCHITECTURE

The 3C2N framework mainly consists of three significant modules: an ESPN, ASDN-1, and the ASDN-2. Three of them share a CNN trunk for feature learning. Here, the VGG-16 [41] architecture, which is verified to be efficient in this task, is utilized as the shared CNN trunk for feature learning. The 3C2N model is regarded as a ship proposal generator, which is trained with annotated Sentinel-1 images.

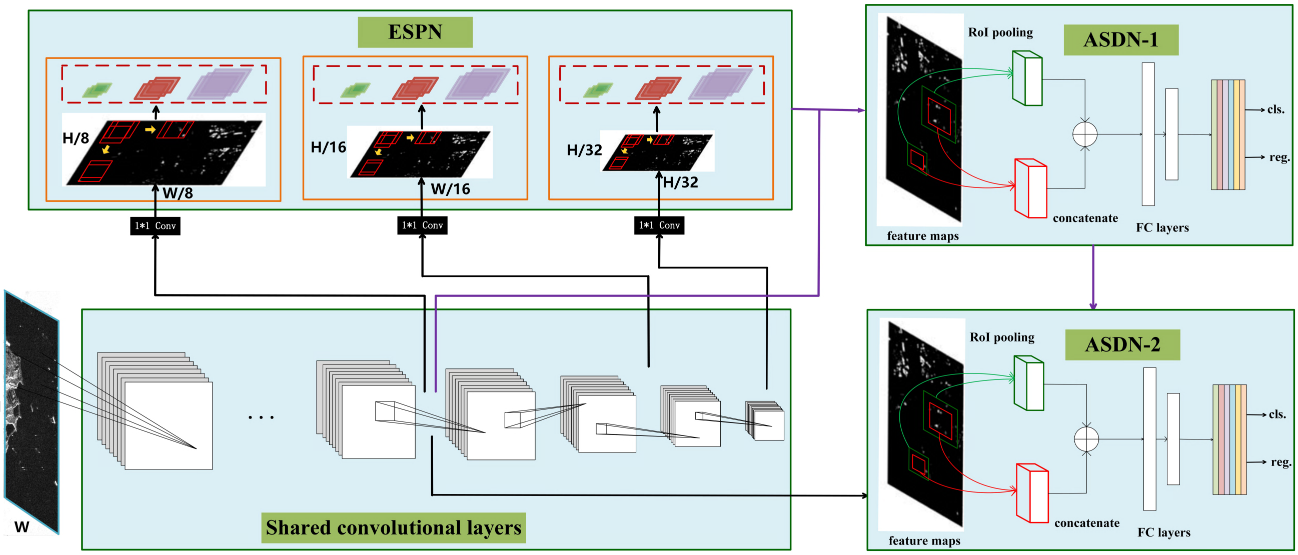


FIGURE 3. The architecture of the 3C2N model, which sufficiently analyzes the SAR image in space domain using the multi-scale technology and the cascade structure. It mainly consists of three significant parts: one ESPN and two cascade ASDNs. The ESPN and ASDN-1 is used for generating ship proposals and ASDN-2 is followed to execute ship discrimination. All of them share some convolutional layers for feature learning. In this figure, "FC" represents fully connected, "cls." and "reg." denote classification (ship-region or non-ship region) and bounding box regression, respectively.

The overall architecture of the 3C2N model are shown in Fig. 3. The first level ship proposals are generated via the ESPN module and the second level ship proposals are generated from the ASDN-1 module. The ship candidate regions are generated by ASDN-2 module. Leveraging the spatial information more sufficiently, this cascade architecture could further improve the detection quality according to [42].

In particular, image features in ESPN are extracted from three different representative layers, i.e., Conv4_3, Conv5_3, and Conv6_1. When assuming the size of the input SAR image to be $H \times W$, the resolution of the feature maps from these three layers are transformed to $\frac{H}{8} \times \frac{W}{8}$, $\frac{H}{16} \times \frac{W}{16}$, and $\frac{H}{32} \times \frac{W}{32}$, respectively. Then, the branches are reused and are extended to three different sub-branches in that each branch is convolved via a set of small convolutional filters, i.e., 3×3, 5×5, 7×7, to perceive different-sized objects. Finally, the first level ship proposals are predicted from each sub-branch. This technique is verified to be effective for perceiving different-sized objects in [36]. The objective function in ESPN can be formulated as

$$L_{\text{ESPN}}(\Theta_{\text{p}}) = \sum_{m=1}^{M} \sum_{t \in S^m} \alpha_m l^m(X_t, Y_t, B_t|\Theta_{\text{p}}), \tag{1}$$

where the number of detection layers, denoted by M, is equal to 9, which stands for three proposal branches with three different sub-branches as detection layers, the training sample of each prediction layer is denoted by S^m, α_m denotes the weight for the m-th detection layer's loss, the variables, X_t, Y_t, and B_t, represent the local features, labels (i.e., ships or non-ships), and the coordinates of the t-th candidate region, respectively, Θ_{p} denotes parameters in this network, and l^m indicates the loss function of the m-th sub-branch, referring to the loss function in [34].

In the ASDN-1 and ASDN-2 modules, image features for each candidate region are composed of two parts: one is the CNN feature learned directly from the candidate regions by the shared convolutional layers and the other is the contextural feature learned from the corresponding context regions, which are set as 1.5 times larger than the proposal ship regions. The size of each context region is determined by means of the grid-searching strategy. Next, both of the two feature sets are pooled to feature maps with the same resolution via region of interest (RoI) pooling, see the red and green cubes in ASDN-1 and ASDN-2 part of Fig. 3. Then they are combined by concatenation for further classification (ship-like region or non-ship like region) and bounding box regression. Both of the ESPN and ASDN-1 modules are used for ship proposal generation and ASDN-2 are adopted for ship discrimination.

The overall loss function of the 3C2N deep learning framework must have the form

$$\begin{aligned} L(\Theta_{\text{p}}, \Theta_{\text{d1}}, \Theta_{\text{d2}}) &= L_{\text{ESPN}}(\Theta_{\text{p}}) \\ &\quad + \alpha_{M+1} \sum_{t \in S^{M+1}} l_{\text{ASDN-1}}(X_t, Y_t, B_t|\Theta_{\text{d1}}) \\ &\quad + \alpha_{M+2} \sum_{t \in S^{M+2}} l_{\text{ASDN-2}}(X_t, Y_t, B_t|\Theta_{\text{d2}}), \end{aligned} \tag{2}$$

where α_{M+1} & α_{M+2} denote the weight of the ASDN-1's loss $l_{\text{ASDN-1}}$ and the ASDN-2's loss $l_{\text{ASDN-2}}$, respectively, see [34]. And Θ_{d1} & Θ_{d2} stand for the added parameter set of fully connected layers in the ASDN-1 module and the ASDN-2 module, respectively.

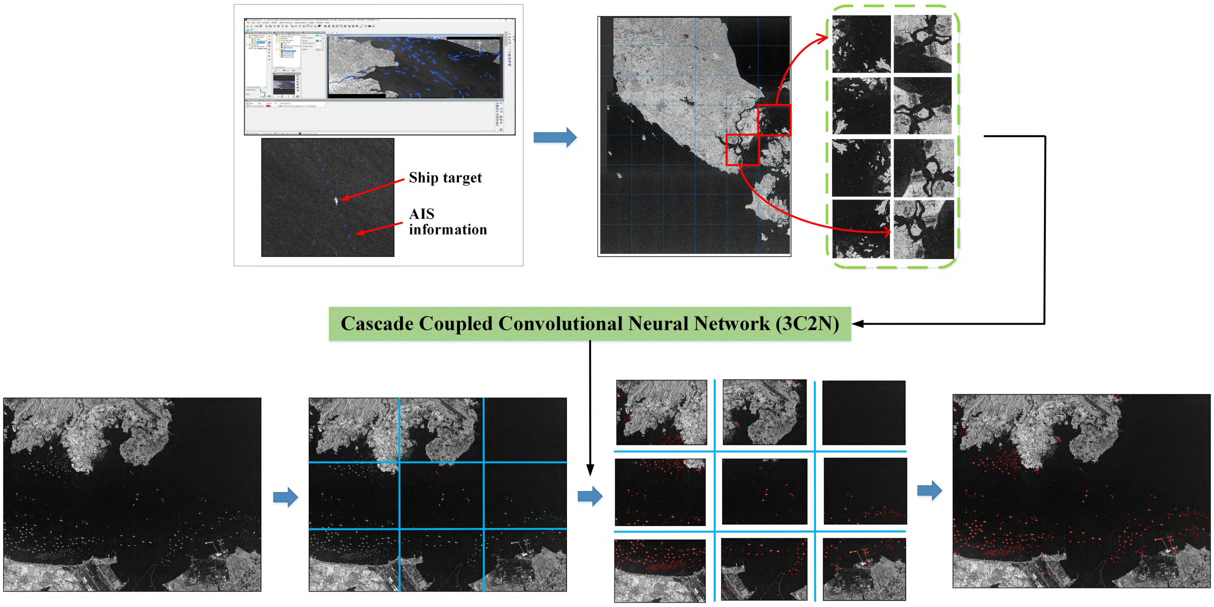


FIGURE 4. Ship proposals are generated by using the 3C2N model.

2) SHIP PROPOSAL GENERATION VIA THE 3C2N MODEL

The 3C2N, implemented with the deep learning framework Caffe [47], popularizes the use of end-to-end CNNs for small and densely clustered ship detection from SAR images. When using this method for ship detection, training is needed to generate a component model for further research. Fig. 4 subtly depicts the process of using 3C2N for ship proposal generation. The upper row of this figure shows the data preparation, i.e., data collection, data annotation, and data augmentation. The image testing for finding the ship proposals is illustrated by the bottom row. The method can be divided into five steps as follows:

(i) Build the training set. Before training the 3C2N model, sixty wide-swath Sentinel-1 images containing small and densely clustered ship targets are collected, all of which are with interferometric wide-swath (IW) mode ground-range detected (GRD) format. Ships in the images are manually annotated by expert inspections on the Sentinel-1 application platform (SNAP) [48] partially with the help of automatic information system (AIS) information.

(ii) Augment and tile the training data set. The training data set is tiled into image blocks of size 1024 × 768 in pixel and the corresponding ground-truth locations and labels are attached in a XML file. Image blocks in the training data set are then augmented eight-fold via rotation from 0° to 315° with an even gap 45°. Finally, they are flipped horizontally and vertically, respectively.

(iii) Train the 3C2N model. Before training the model, the training configurations and the optimization strategy are conducted. The network training is started with an initial learning rate of 0.001 and the learning rate changes in accordance with

$$LR = 0.001 \times 0.1^{iter/5k}, \tag{3}$$

where *iter* represents the current number of iteration. All the parameters in this framework are iteratively updated, totally 35k iterations, by minimizing the loss function formulated as (2). It is worth mentioning that the parameter of intersection over union (IoU) in ASDN-1 and ASDN-2 are set to be 0.5 and 0.6, respectively. The tiled training image blocks and the corresponding labels are successively fed into the 3C2N deep learning framework to train the network and generate the 3C2N model. The entire training procedure proceeds recursively until the overall loss function is converged.

(iv) In the testing stage, ship proposals are generated by using the pre-trained 3C2N model. The original SAR images to be tested S are divided into small image blocks, denoted by S_b, which has nearly the same size as the training image blocks. Notably, the upper-left corner's index in the original imagery is denoted as $[X_{ul}, Y_{ul}]$. Thereafter, the pre-trained 3C2N model takes these image blocks as input and then the model outputs the candidate ship regions $[x_b, y_b, h, w]$, given by

$$[x_b, y_b, h, w] = \text{3C2N}(S_b), \tag{4}$$

where x_b & y_b denote the index of the upper-left corner along the row and column directions in the testing image block S_b, respectively, and h & w indicate the height and width of the proposal region, respectively.

(v) Transform both the locations and labels of all the ship proposals into a unified format. The detection results of the same original imagery are stitched together. Meanwhile, the corresponding locations and labels are transformed into the coordinates in the original imagery S, denoted by $[x_s, y_s, h, w]$, where x_s and y_s represent the upper-left corner's coordinate, i.e., the row and column index, in the original imagery space, respectively. Now, the relationship can be mathematically expressed by

$$[x_s, y_s, h, w] = [x_b, y_b, h, w] + [X_{ul} - 1, Y_{ul} - 1, 0, 0]. \quad (5)$$

Algorithm 1 illustrates a clear road-map for ship proposal generation by using the pre-trained 3C2N model. This algorithm takes the images to be tested and the pre-trained 3C2N model as input. Also, it outputs $ShipProposal = [x_s, y_s, h, w]$ as ship proposals' locations.

Algorithm 1 Ship Proposal Generation by Using the 3C2N Model

```
Input:
    Input imagery $S$ for testing;
    Pre-trained 3C2N model;
Output:
    Ship proposals' locations $ShipProposal$;
Initialize $ShipProposal = zeros()$;
Set an index counter $index = 1$;
Tile the input imagery into $N_b$ small-sized image blocks $S_b$ for testing;
for each $i \in [1, N_b]$ do
    Denote the upper-left corner's index of $S_b^i$ as $[X_{ul}^i, Y_{ul}^i]$ in the original imagery system;
    Get the proposals' locations in the image block coordinate system $Loc$ via (4). Therefore, $\forall 1 \leq j \leq size(Loc, 1), Loc(j, :) = [x_b, y_b, h, w]$;
    for each $j \in [1, size(Loc, 1)]$ do
        Obtain ship proposals' locations $ShipProposal(index, :) = [x_s, y_s, h, w]$ in the input imagery coordinate system by (5);
        $index = index + 1$;
    end for
end for
return $ShipProposal$.
```

B. SHIP DISCRIMINATION: A PCT-BASED VISUAL ATTENTION MODEL

Even though the detection results of the 3C2N model have shown to be superior in SAR ship detection, they display a far cry from meeting commands for practical applications. For example, in small and densely clustered areas, ship targets are usually very close to each other, which leads to big overlaps between the detected regions of adjacent ship targets. This phenomenon is aggravated by the loosely detected bounding boxes. Meanwhile, few false alarms still exists in the land area because the appearance of some scattering objects on land are visually similar to ship targets by SAR. It is because the CNN-based methods only focus on the images' spatial information, while neglecting the information hidden in frequency domain, and other references relating to geographical information are not considered. For these reasons, a PCT-based visual attention model in frequency domain is presented for ship discrimination. This model aims to refine the proposals' bounding boxes so that reducing the missing detections, especially in the small and densely clustered areas. It is worth noting that the DEM data, which could be linked by the Sentinel-1's meta-data file, are adopted for ruling out regions containing analogical scatterings on land. The workflow of this method is shown in Fig. 5.

For each proposal region obtained by the presented 3C2N model, the PCT-based visual attention method for ship discrimination can be implemented by twelve steps as bellow:

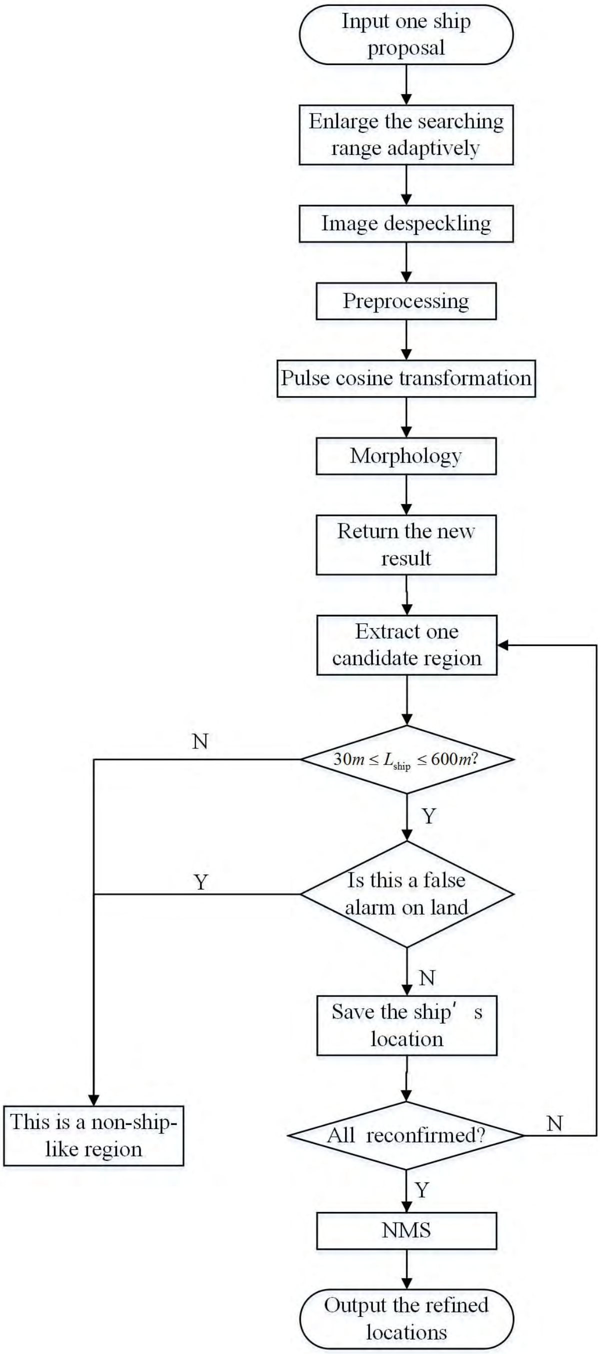

Input one ship proposal
Enlarge the searching range adaptively
Image despeckling
Preprocessing
Pulse cosine transformation
Morphology
Return the new result
Extract one candidate region
30m ≤ L_ship ≤ 600m?
N
Y
Is this a false alarm on land
Y
N
Save the ship' s location
This is a non-ship-like region
All reconfirmed?
N
Y
NMS
Output the refined locations

FIGURE 5. Workflow of the PCT-based visual attention method for ship discrimination.

Step-1, extract one proposal region. The proposal region can be obtained via the pre-trained 3C2N model, i.e., denoted by a rectangular $[x_s, y_s, h, w]$. This is illustrated in Algorithm 1.

Step-2, enlarge the proposal region adaptively. In order to lower the missing detections, the proposal region is enlarged adaptively in the original detection map according to its original size. The detailed mechanism is described in Fig. 6. It follows from this figure that the red rectangle represents the original proposal region and the green one denotes the adaptive region. In particular, we enlarge the proposal region five times both in the row and the column spaces with the center pixel fixed and then obtain the adaptive region to be $R_{ad} = [x_s - 2h + 1, y_s - 2w + 1, 5h, 5w]$. This is because the majority of the missing detections in areas with small-size and densely clustered targets are very close to the detected regions, no more than two-fold of the bounding box's size, according to our observation on the 3C2N detection results. Particularly, if a proposal region's border is beyond the original imagery, the imagery border is contributed to the adaptive region.

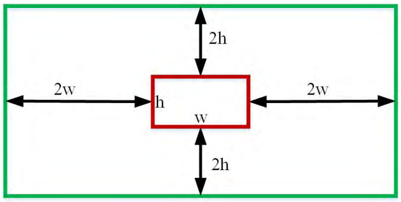


FIGURE 6. **The mechanism of the adaptive region. The red rectangle represents the original proposal region and the green one denotes the adaptive region.**

Step-3, reduce the speckle noise in the adaptive regions. The NLM algorithm mentioned earlier is employed in this paper, which is known to be one of the most effective despeckling method for SAR images. The NLM despeckled image can be written as

$$I_{NL} = NonLocalMean(R_{ad}), \tag{6}$$

where $NonLocalMean(\cdot)$ represents the NLM algorithm which transforms the input image to a despeckled one, denoted by I_{NL}.

Step-4, image preprocessing. In this step, image preprocessing is intended to smooth the sea clutter and the land area and to enhance the real targets. In order to remove land areas and pixels with high intensity, while still maintaining the ships with only a few pixels, the following strategy is adopted. For each pixel in the despeckled adaptive region, i.e., $p \in I_{NL}$, the pixel value in this position is transformed to

$$f(p) = \frac{[\mu_{\text{local}}]^2}{2 \times [\sigma_{\text{adap}}]^2}, \tag{7}$$

where μ_{local} indicates the mean value of pixels in a local region centered at the current location, i.e., a moving window of size 3×3 centered at p, and σ_{adap} is the standard deviation of the despeckled adaptive image block I_{NL}. Then, the preprocessed image block I_{PP} is obtained.

Step-5, the PCT-based visual attention model used for detection. This step is to transform the preprocessed image block I_{PP} in space domain into frequency domain. Then, visually intensive pixels are highlighted in the transformed saliency map $I_{saliency}$ via the PCT method [49]. The details of this step is analyzed in Section II-C. It is worth mentioning that the results detected from one adaptive region may be multiple. In addition, the results detected by the PCT-based visual attention model are the binary maps, denoted by I_{bi}, which serves as the candidate regions for the following steps. In the binary maps, the pixels belonging to the ship targets are set to be 1 and others are set to be 0.

Step-6, image morphology. In this step, we firstly determine the connected components. Then, we compute the area for each component and remove the connected components whose area is less than a certain value. In this paper, the one whose area is less than 10 pixels is removed. This parameter is set empirically.

Step-7, verification. In order to further improve the detection accuracy, the candidate detections are verified via several basic constraints. First of all, considering the limitation on ship size, we empirically remove the ships whose length is larger than 600 m or lower than 30 m. This can be mathematically written as

$$30 < h_r \cdot PS_{az}, w_r \cdot PS_{rg} < 600, \tag{8}$$

where h_r & w_r illustrate the newly refined height and width of the detected region, respectively, and PS_{az} & PS_{rg} represent the pixel spacing in azimuth and ground range directions, respectively.

Then, we adopt the digital elevation model (DEM) [50] as an important reference to remove land areas, where there are still ship-like targets falsely detected by the 3C2N model. The DEM is a complementary information of the SAR image, which could be indexed by the geographical locations in the meta-data file. It was gathered from the shuttle radar topographic mission (SRTM) 3 Arc-Second Global.

Step-8, extract the refined ship targets. The ship targets detected by the PCT-based visual attention method are located at $[x_{rb}, y_{rb}, h_r, w_r]$, which is characterized in the adaptive image block coordinate system. The detected targets may be multiple.

Step-9, coordinate transformation. The refined detections are transformed from the adaptive image block system to the original imagery coordinate system by

$$[x_{rs}, y_{rs}, h_r, w_r] = [x_{rb}, y_{rb}, h_r, w_r] + [x_s - 2h, y_s - 2w, 0, 0]. \tag{9}$$

Step-10, return to Step-8 recursively until all the candidate regions detected by the PCT-based visual attention method have been verified.

Step-11, return to Step-1 to process another proposal region obtained by the 3C2N model until all the objective regions has been discriminated.

Algorithm 2 The 3C2N-Guided Visual Attention Model for Ship Discrimination

```
Input:
    The imagery $S$ to be detected;
    Ship proposals' locations $ShipProposal$ in the imagery.
Output:
    All the refined locations $RefinedShip$ in the input imagery;
Compute the number of ship proposals obtained from 3C2N via $N_p = size(ShipProposal, 1)$;
Define a counter $count = 1$;
Initialize $RefinedShip = zeros()$;
for each $i \in [1, N_p]$ do
    Get the original proposal region $[x_s^i, y_s^i, h^i, w^i] = ShipProposal(i, :)$;
    Obtain the adaptive region $R_{ad} = [x_s^i - 2h^i + 1, y_s^i - 2w^i + 1, 5h^i, 5w^i]$;
    Reduce the speckle noise by using the NLM algorithm and achieve $I_{NL} = NonLocalMean(R_{ad})$;
    Preprocess the despeckled image via (7) and obtain the preprocessed image block $I_{PP}$;
    Obtain the saliency map $I_{saliency}$ of the adaptive region by the PCT-based visual attention method via (10) to (18);
    Compute the binary map $I_{bi}$ by using (19) in which only the pixels belonging to ship targets are highlighted;
    Get the objects' locations of the binary map via morphology and record the number of detected targets $N_m$;
    for each $j \in [1, N_m]$ do
        Get the current object's location $temp = [x_{rb}^{ij}, y_{rb}^{ij}, h_r^{ij}, w_r^{ij}]$ in the coordinate system of the image block;
        while $(30 < h_r^{ij} \cdot PS_{az} < 600) \wedge (30 < w_r^{ij} \cdot PS_{rg} < 600)$ do
            if the object is not located on land then
                Transform the coordinate of the detected bounding boxes back to the original imagery system via (9), denoted by
                $RefinedShip(count, :) = temp + [x_s^i - 2h^i, y_s^i - 2w^i, 0, 0]$;
                $count = count + 1$;
            end if
        end while
    end for
end for
return $RefinedShip$.
```

Step-12, reduce redundancy. The NMS algorithm [51] needed in this method is adopted to avoid repeated detections.

The details for implementing the 3C2N-guided visual attention method are illustrated in Algorithm 2. In this algorithm, the ship proposals' locations *ShipProposal* obtained from the pre-trained 3C2N model are taken as input and the algorithm produces the refined detection result stored in *RefinedShip*. The core of this algorithm is the PCT model, which processes the adaptive image blocks in frequency domain. The referenced DEM data also provides a good way to eliminate false alarms on land.

C. THE PULSE COSINE TRANSFORMATION MODEL ON ADAPTIVE REGIONS

Considering the computational efficiency and the great capability to predict eye fixation, in this paper, the PCT model provides a good way to predict visual attention objects. Further details about this algorithm is described as follows:

Given an image block $I \in \mathbb{R}^{5h \times 5w}$, to remove noisy pixels and enhance the binary image quality, the image is firstly transformed via

$$I_{fl}(x, y) = \begin{cases} I(x, y), & \text{if } I(x, y) \geq T_1 \\ T_1, & \text{otherwise} \end{cases}, \tag{10}$$

where the threshold T_1 can be determined by

$$T_1 = \mu(I_{fl}) + \alpha \cdot \sigma(I_{fl}). \tag{11}$$

Here, $\mu(I_{fl})$ represents the mean value of the enhanced image block I_{fl}, called flooding image [21], $\sigma(I_{fl})$ denotes the standard deviation of the enhanced image block I_{fl}, and α is a constant value, which is empirically set to be 0.6. It is of interest to note that ship targets can be considered as noisy pixels if a greater α is used. Otherwise, the noise of the image can be decreased.

In the next stage, the PCT model is utilized to generate visual attention maps. With the flooding image I_{fl} available, the transformed image in frequency domain will be obtained by means of the two-dimensional discrete cosine transformation (DCT). That is,

$$\begin{aligned} I_c(u, v) &= \mathscr{C}(I_{fl}(x, y)) \\ &= a_u a_v \sum_{x=0}^{5h-1} \sum_{y=0}^{5w-1} I_{fl}(x, y) \\ &\quad \times \cos\frac{(2x+1)u\pi}{2 \cdot 5h} \cdot \cos\frac{(2y+1)v\pi}{2 \cdot 5w}, \end{aligned} \tag{12}$$

where $\mathscr{C}(\cdot)$ indicates the two-dimensional DCT operation, and

$$a_u = \begin{cases} \dfrac{1}{\sqrt{5h}}, & \text{if } u = 0 \\ \sqrt{\dfrac{2}{5h}}, & \text{if } 1 \leq u \leq 5h-1, \end{cases} \tag{13}$$

$$a_v = \begin{cases} \dfrac{1}{\sqrt{5w}}, & \text{if } v = 0 \\ \sqrt{\dfrac{2}{5w}}, & \text{if } 1 \leq v \leq 5w-1, \end{cases} \tag{14}$$

where $0 \leq x \leq 5h-1$, $0 \leq y \leq 5w-1$, $0 \leq u \leq 5h-1$, and $0 \leq v \leq 5w-1$. In what follows, the encoded image block I_{sn} in frequency domain is adopted by

$$I_{sn}(u, v) = sign(I_c(u, v)) = \begin{cases} -1, & \text{if } I_c(u, v) < 0 \\ 0, & \text{if } I_c(u, v) = 0 \\ 1, & \text{if } I_c(u, v) > 0, \end{cases} \tag{15}$$

where the function $sign(\cdot)$ indicates the sign function by which the positive pixel values and the negative values are coded to be 1 and -1, respectively. Otherwise, the pixel values are set to be zero.

Now, $\mathscr{C}^{-1}(\cdot)$ denotes the two-dimensional inverse DCT operation. The input image in frequency domain can be transformed back into space domain by the following manner:

$$\begin{aligned} I_{inv}(x, y) &= abs(\mathscr{C}^{-1}(I_{sn}(u, v))) \\ &= abs(\sum_{u=0}^{5h-1} \sum_{v=0}^{5w-1} a_u a_v I_{sn}(u, v) \\ &\quad \times \cos \frac{(2x+1)u\pi}{2 \cdot 5h} \cdot \cos \frac{(2y+1)v\pi}{2 \cdot 5w}), \end{aligned} \tag{16}$$

where

$$abs(\eta) = \begin{cases} \eta, & \text{if } \eta \geq 0 \\ -\eta, & \text{otherwise}, \end{cases} \tag{17}$$

Then, the saliency map $I_{saliency}$ can be obtained by convolving F_g with the square of I_{inv}, given by

$$I_{saliency} = F_g \otimes [I_{inv}]^2, \tag{18}$$

where F_g denotes the two-dimensional gaussian low pass filter and $\otimes$ represents the two-dimensional convolution process.

The PCT, represented by (12) to (15), retains the sign of the two-dimensional DCT coefficient and neglects the amplitude information. The sign function is utilized in frequency domain and functions on the two-dimensional coefficient because the sign is much important to simulate the activation or restriction of the neurons in human visual system. In other words, the PCT codes the image into -1, 0, and 1. It is the coding operation that simulates the neuron's pulse in human brain. Finally, the saliency map is obtained by using (16) to (18).

Additionally, in order to achieve the detection results more clearly, a binary image I_{bi} is obtained from the saliency map. The mathematical formulation can be expressed by

$$I_{bi}(x, y) = \begin{cases} 1, & \text{if } I_{saliency} \geq T_2 \\ 0, & \text{otherwise}, \end{cases} \tag{19}$$

where $T_2 = \mu(I_{saliency}) + \beta \cdot \sigma(I_{saliency})$ is a threshold. Here, $\mu(I_{saliency})$ and $\sigma(I_{saliency})$ are the average value and the standard deviation of the obtained saliency map, respectively, and β, which is empirically set to be 2.5, is a balancing parameter between $\mu(I_{saliency})$ and $\sigma(I_{saliency})$.

III. EXPERIMENTS AND RESULTS

A. EXPERIMENTAL ENVIRONMENT

The training of the CNN-based methods, implemented by the deep learning framework Caffe [47], are executed on a workstation with two Intel 32 Core i7 CPUs with 64G RAM and four ENVIDIA GTX-1080 GPU with 8GB memory. The testing part of the CCNN method, the 3C2N method, and the 3C2N-guided visual attention method are programmed with MATLAB 2015b. The operating system is Ubuntu 16.04.

B. EXPERIMENTAL DATA

The data set used in this paper is collected from 60 wide-swath Sentinel-1 images [52]. The data is collected from five typical scenes because of their intense marine traffic: Shanghai Port (China), Shenzhen Port (China), Tianjin Port (China), Yokohama Port (Japan), and Singapore Port (Singapore). The Sentinel-1 images are suitable to verify the effectiveness of the proposed method due to three reasons: 1) There are a large amount of multi-scale ship targets in the images; 2) The background conditions of ships are varying from simple to extremely complex; 3) The ships are distributed in different ways, sometimes they are single and sometimes they are densely clustered. Among the images, 52 of them are used for training and the other 8 images are used for testing. It is noted that the validation data set is selected randomly from the training set, 21816 ship samples are included. In this experiment, the data acquired in the interferometric wide-swath (IW) mode, high resolution ground-range detected (GRD) format with C band is provided by the European space agency (ESA). Such level-1 products are generally available for most data users and consist of focused SAR data detected in magnitude with a native range by azimuth resolution estimated as 20m × 22m and a 10m × 10m pixel spacing. The polarization of this data set is selected as VH. This is because the side-lobe effect for VV polarization is much more severe than that for VH polarization, which is may interfere the detection of SAR ships. And the average image size is 25000 × 18000 (rg × az). Here, rg and az represent the ground range and the azimuth directions, respectively. The main information of this data set is listed in Table 1.

The ground-truths are annotated by professional SAR image interpreters partially with the help of AIS information

TABLE 1. Meta-data of wide-swath Sentinel-1 images.

PARAMETER	VALUE
Satellite	Sentinel-1
Imaging Mode	IW
Band	C
Polarization	VH
Product Type	GRD
Resolution (rg × az)(m)	20 × 22
Pixel Spacing (rg × az)(m)	10 × 10
Average Size Per Image (rg × az)(pixel)	25000 × 18000

on the official ESA Sentinel application platform (SNAP) software, see [48]. Specifically, they are marked with rectangles, which are denoted by the upper-left corner's coordinate with the object rectangle's height and width. To better exhibit the detection results of the proposed method, we tile the testing imagery into image blocks of size 200 × 200 (rg × az) without overlap. It is worth mentioning that the radiometric calibration is performed by using SNAP.

C. EVALUATION METRICS

Now, the precision, the recall, and the F_1 score, which are three widely used criterias in object detection, are adopted in order to evaluate the detection performance quantitatively. The precision is defined as

$$precision = \frac{N_{TP}}{N_{TP} + N_{FP}}, \tag{20}$$

where, N_{TP} & N_{FP} are the number of true-positives and the number of false-positives, respectively. True positives and false positives are the correctly detected ships and the falsely detected ships, respectively.

Similarly, when assuming N_{FN} as the number of false negatives, the recall can be formulated more explicitly as

$$recall = \frac{N_{TP}}{N_{TP} + N_{FN}}. \tag{21}$$

Now, let F_1 score be a comprehensive evaluation metric, which can be mathematically given by

$$F_1 = 2 \times \frac{precision \times recall}{precision + recall}. \tag{22}$$

As is widely known that a higher precision and a higher recall rate are both expected. Whereas, in fact, these two evaluation metrics are on the opposite sides. It means that a higher precision corresponds to a lower recall rate and a lower precision corresponds to a higher recall rate. It is worthwhile to point out that a higher F_1 score means a more desirable comprehensive ship detection performance.

D. BASELINE METHODS

To verify the effectiveness of the proposed method, the CCNN method [36] is chosen as a baseline to demonstrate the superiority of the proposed techniques. The newly presented 3C2N deep learning method is also evaluated to verify its superiority over the CCNN. The intermediate features of the PCT model are analyzed in Section III-E.2. Moreover, the influence of the NLM algorithm and the preprocessing strategy are discussed with configurations, namely, "Ours-without-NLM" and "Ours-without-PP", separately. These two configurations represent the proposed method without using the NLM algorithm and the proposed method without preprocessing, respectively.

E. EXPERIMENTAL RESULTS

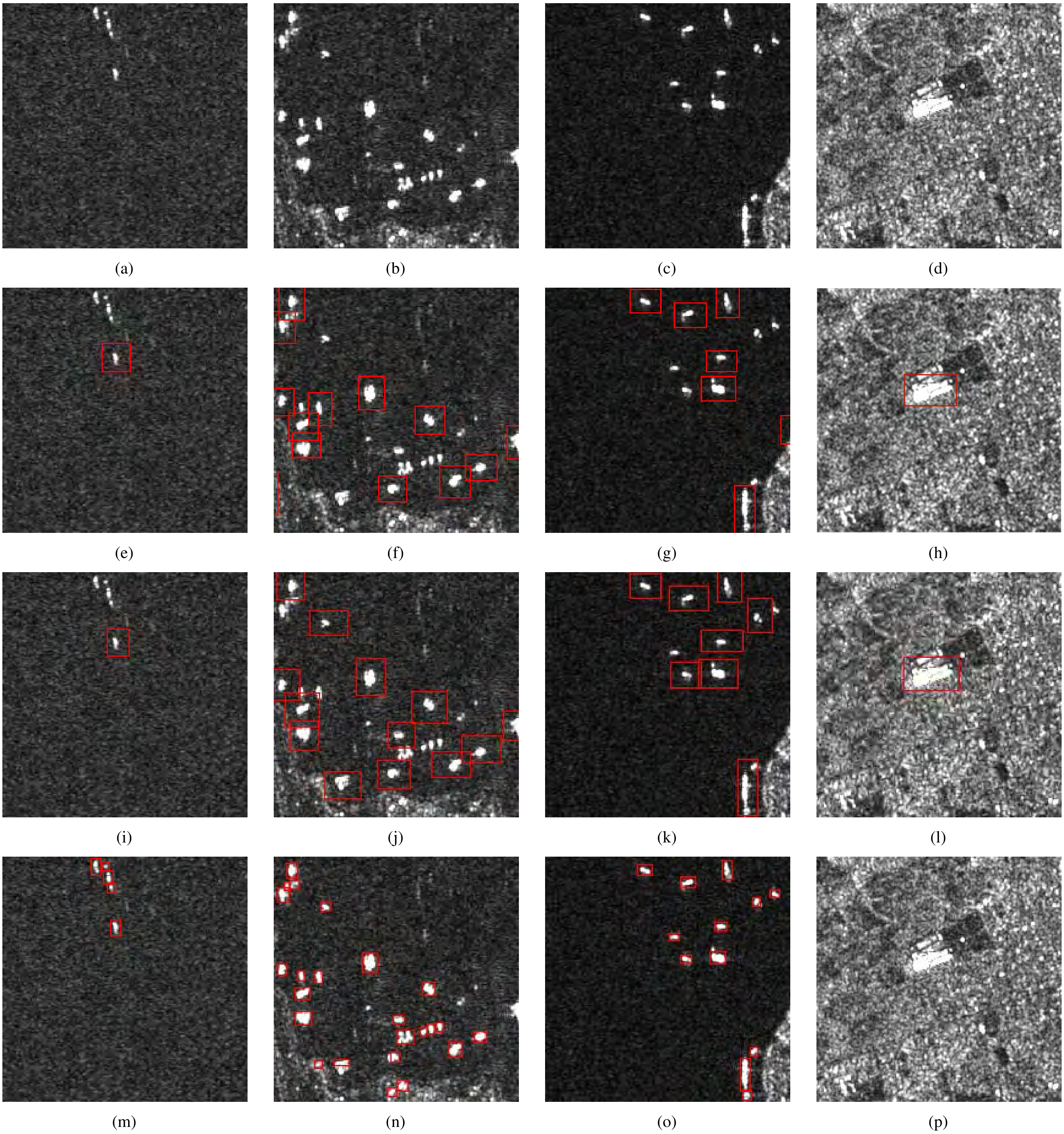
1) THE COMPARATIVE OVERALL DETECTION PERFORMANCE

In this evaluation system, twenty-five 200×200 image blocks with 324 ground-truths tiled from Sentinel-1 images are selected to test the performance of the proposed method and the baseline methods. Image blocks with the complex background conditions (e.g., high clutter ocean environments, and areas, where the ships are mostly small in size and densely clustered), are chosen in this experiment (see in Fig. 7(a) to Fig. 7(d) as representative samples). Table 2 shows the quantitative evaluation results of the proposed method and the baselines. As demonstrated in this table, we can conclude that: 1) the missing detections and the false alarms of our proposed method are significantly decreased, compared with the CCNN method and the 3C2N method; 2) the detection performance of the 3C2N method has been improved by using the cascade structure. As a comprehensive evaluation metric, the value of F_1 score obtained by the 3C2N-guided visual attention method, equal to 0.9605, is higher than the value of the 3C2N method the CCNN method, i.e., 0.8984 and 0.8524, respectively, see Table 2.

Fig. 7 exhibits the detection results of the 3C2N-guided visual attention method and other baselines of four classical original image patches of size 200 × 200. The detection results of each method is visualized with red rectangles

TABLE 2. The quantitative detection results.

Methods	Ground Truth	True Positive	False Positive	Precision	Recall	F_1 score
CCNN	324	260	26	0.9090	0.8025	0.8524
3C2N	324	283	23	0.9248	0.8735	0.8984
3C2N-guided visual attention method	324	316	18	0.9461	0.9753	**0.9605**



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

(e) (f) (g) (h)

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

(m) (n) (o) (p)

FIGURE 7. Detection results on four image patches, shown by the red rectangles. (a), (e), (i) and (m) denote the original image patch tiled from Sentinel-1 imagery, the CCNN detection result, the 3C2N detection result, and the bounding boxes detected by the 3C2N-guided visual attention method, respectively. (b), (f), (j) and (n) display the second image patch example and its corresponding results, respectively. (c), (g), (k) and (o) exhibit the third image patch example and its corresponding results, respectively. (d), (h), (l) and (p) show the fourth example image patch and its corresponding results, respectively. (a) Sample1: image patch. (b) Sample2: image patch. (c) Sample3: image patch. (d) Sample4: image patch. (e) Sample1: CCNN method. (f) Sample2: CCNN method. (g) Sample3: CCNN method. (h) Sample4: CCNN method. (i) Sample1: 3C2N method. (j) Sample2: 3C2N method. (k) Sample3: 3C2N method. (l) Sample4: 3C2N method. (m) Sample1: Ours. (n) Sample2: Ours. (o) Sample3: Ours. (p) Sample4: Ours.

which indicate the detected bounding boxes. In this figure, each column illustrates one example, including the original image patch and the detection results of each method, respectively. The first row (Fig. 7(a) to Fig. 7(d)) in this figure shows the original image patches. The second row (Fig. 7(e) to Fig. 7(h)) and the third row (Fig. 7(i) to Fig. 7(l)) show the detection results of the CCNN method and the 3C2N method, respectively. The detection performance of

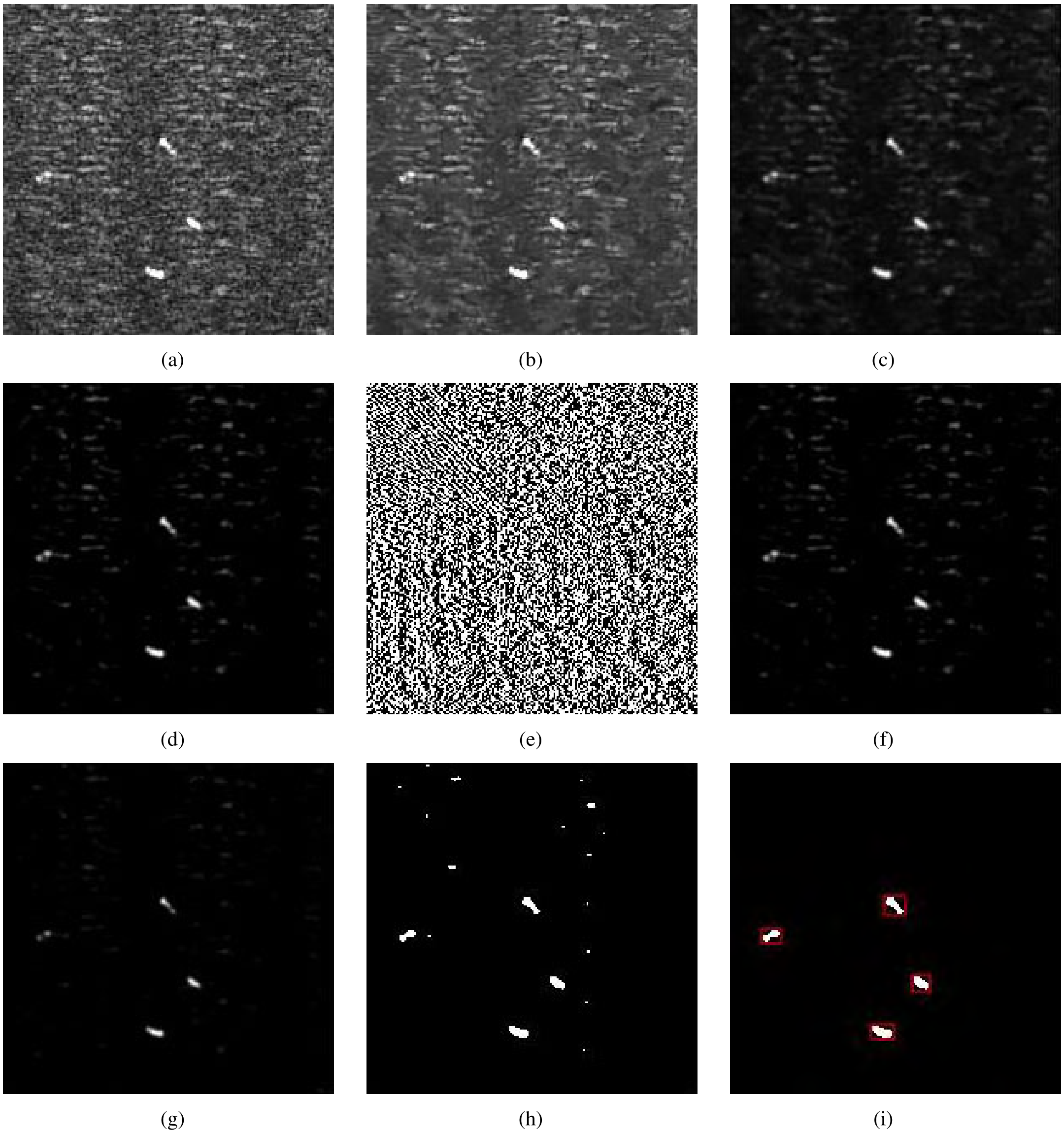


FIGURE 8. The intermediate results of PCT model. (a) One image patch tiled from the original Sentinel-1 imagery. (b) The image patch after image despeckling by using the NLM algorithm. (c) The image patch after preprocessing by (7). (d) The flooding map. (e) The image patch in frequency domain. (f) The transformed image patch reversed to the space domain. (g) The image patch after gaussian filtering. (h) The image patch after morphology. (i) The detection results by some constraints.

the 3C2N-guided visual attention method is provided as the fourth row (Fig. 7(m) to Fig. 7(p)). The background conditions in the first and the second examples are both facing extreme ocean clutters. In the third image patch example, several ships are located at inshore area, besides many offshore situated ships. The fourth example shows a non-ship object situated on land, which has been mistakenly regarded as a ship by the CNN-based method, including the CCNN method and the 3C2N method, because it is similar to a ship in appearance. It is of high interest to note that the majority of ships in the four image patches are small in size and sometimes are packed densely and disorderedly, which makes it very difficult to distinguish.

In comparison with the detection results of the CCNN method, the 3C2N method, and the 3C2N-guided visual attention method, one can easily obtain the results as follows:

- Even though the background is no more than complex, the 3C2N-guided visual attention model could detect the ship targets with accurate rectangles.
- The missing detections by the 3C2N-guided visual attention method are declined dramatically, when compared to the results via the CCNN method and the 3C2N method.
- The false alarms on land have been eliminated very well by using the 3C2N-guided visual attention method.

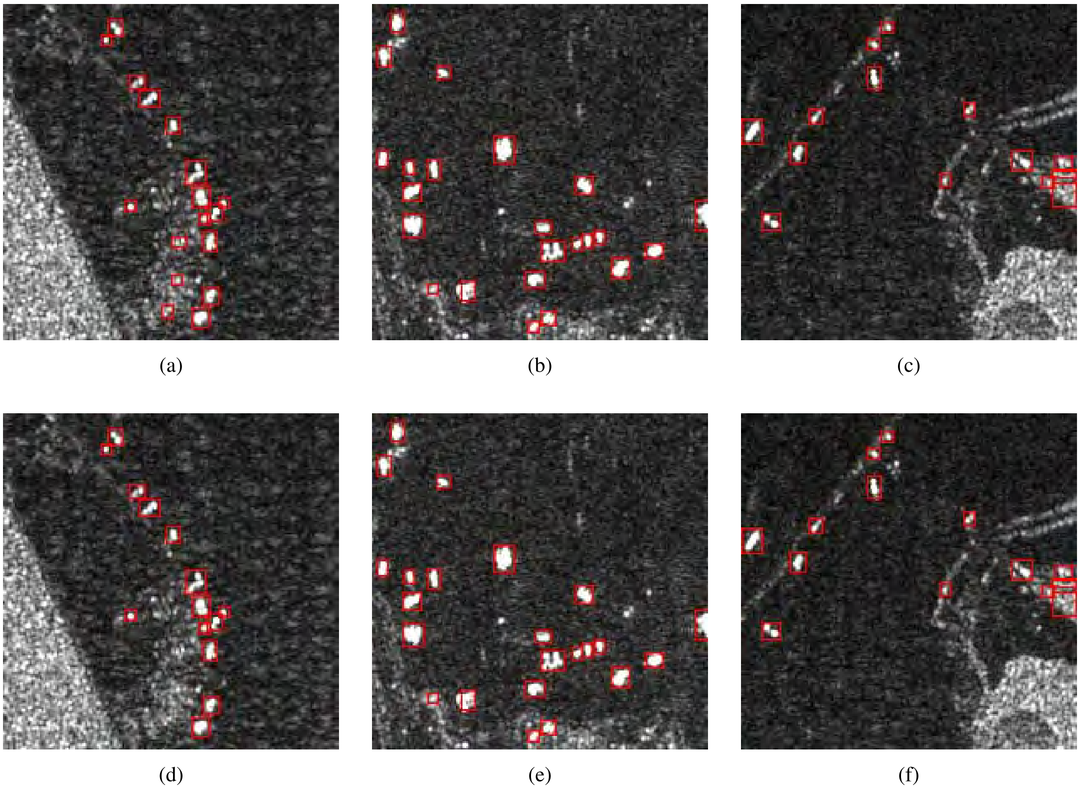


FIGURE 9. Detection results of the proposed method with and without the NLM algorithm for despeckling on three image patches are shown by the red rectangles. The first and the second rows display the detection results of the 3C2N-guided visual attention method without and with NLM for image despeckling, respectively. (a) Sample1: Ours without the NLM for despeckling. (b) Sample2: Ours without the NLM for despeckling. (c) Sample3: Ours without the NLM for despeckling. (d) Sample1: Ours with the NLM for despeckling. (e) Sample2: Ours with the NLM for despeckling. (f) Sample3: Ours with the NLM for despeckling.

- When compared with the CCNN method, the 3C2N method could improve the detection result to a certain extent.

2) ANALYSIS OF THE PCT MODEL

One of the core algorithm in this method is the PCT-based visual attention model. In order to analyze the mechanism and identify its influence, some intermediate results hidden in this model are visually displayed in Fig. 8. This figure displays the visualized intermediate results of the PCT model step by step. Fig. 8(a) shows an original image patch of size 200 × 200 tiled from the wide-swath Sentinel-1 imagery. The image patch shows an extremely complex sea clutter, where the real ship targets with small-size are emerged. The high clutter has brought great challenge to locate and recognize ships. Fig. 8(b) visualizes the despeckled image by using the NLM algorithm, where the sea clutter has been depressed and the ship targets are more clear. This makes the following detections easier. The filtered image after preprocessing via (7) is provided in Fig. 8(c). Fig. 8(d) shows the flooding image operated by (10) and (11), where the prominent scatters are highlighted and others are weakened. Then, the image patches after the two-dimensional DCT transformation and the quantification by the sign function is presented in Fig. 8(e), which visualizes the results in frequency domain. In the next step, the image patch is transformed back to the space domain, as shown in Fig. 8(f). Then the gaussian filter is applied to the squared patch obtained from the last step. The result can be easily seen in Fig. 8(g). Finally, Fig. 8(h). and Fig.8(i) exhibit the detection results after morphology and the constraints consecutively. It is clear that each step is helpful for ship detection from SAR images filled with the extreme background conditions. As a result, the final detected ship targets' bounding boxes are desirable.

3) THE FUNCTION OF THE NON-LOCAL-MEAN ALGORITHM

This section aims to emphasize the importance of despeckling in SAR ship detection. Particularly, the despeckling algorithm in this paper has adopted the NLM algorithm, which provides an efficient way for reducing speckle noise in SAR images. To illustrate this problem, three pairs of comparative detection results are represented in Fig. 9. Fig. 9(a) and Fig. 9(d) provide the detection results of the first image patch example, including the 3C2N-guided visual attention method without and with using the NLM algorithm for despeckling, respectively. The corresponding results of the second example are shown in Fig. 9(b) and Fig. 9(e), respectively. The results for the third example are in the similar way as the former

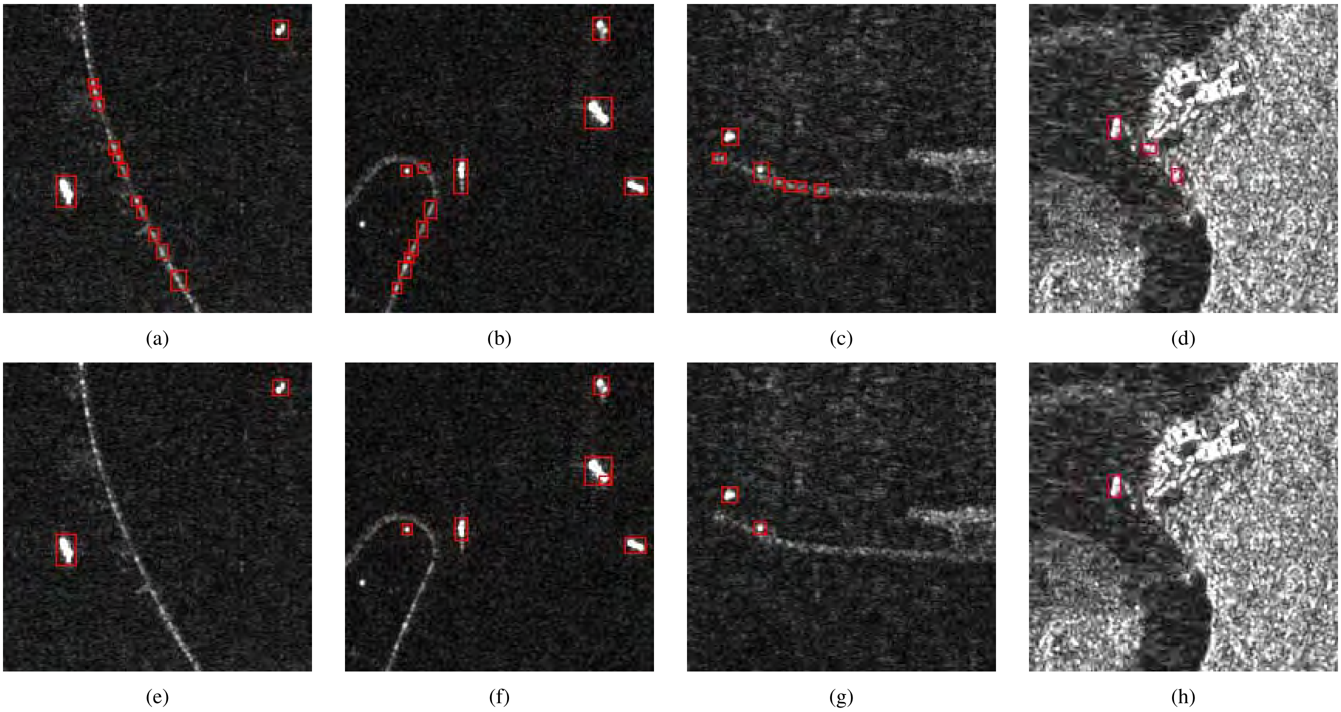


FIGURE 10. Detection results on four image patches, shown by the red rectangles. Results visualized in the first row denote the detected bounding boxes on four different image patches by the 3C2N-guided visual attention method without preprocessing, respectively. Results listed in the second row display the corresponding detections by the 3C2N-guided visual attention method with preprocessing. (a) Sample1: Ours-without-PP. (b) Sample2: Ours-without-PP. (c) Sample3: Ours-without-PP. (d) Sample4: Ours-without-PP. (e) Sample1: Ours-with-PP. (f) Sample2: Ours-with-PP. (g) Sample3: Ours-with-PP. (h) Sample4: Ours-with-PP.

two examples. For all the three examples, by comparing with the detection results in the first row and the second row, one can easily obtain that the false alarms are ruled out significantly by using the NLM algorithm so as to reduce the speckle noise.

4) THE FUNCTION OF IMAGE PREPROCESSING

Image enhancement, considering as a preprocessing step, plays an essential role in the 3C2N-guided visual attention method. It is no exaggeration to say that the effects caused by the image enhancement are very severe. This section aims at discussing the effects caused by image preprocessing. Fig. 10 presents the detection results of the 3C2N-based method with and without preprocessing via (7). In this figure, the first row (Fig. 10(a) to Fig. 10(d)) shows the detection results on the four image patches, and the second line (Fig. 10(e) to Fig. 10(h)) displays the corresponding results of the 3C2N-guided visual attention method with preprocessing. One observes from the comparative results that the false alarms are dramatically decreased in the condition that image patches are preprocessed.

IV. CONCLUSION

The CCNN method has achieved the state-of-the-art detection performance in SAR ship detection. However, there are still many challenges to be addressed. Firstly, the detected bounding boxes are not very compact, which makes the detection results not very accurate, and hence leads to big overlaps between two densely packed ships. Secondly, there are quite a few missing detections in areas, where ships are small in size and they are densely clustered. Thirdly, there still exist some false alarms on land. These can be attributed to the following reasons: 1) The spatial information contained in the state-of-the-art CNN-based method, namely CCNN, is not utilized sufficiently; 2) The CNN-based ship detection method only extract the image information in space domain, while neglecting the information in frequency domain completely; 3) Information attached in the meta-data file, such as the geographical locations, is overlooked. In this paper, in order to improve the CNN-based SAR ship detection performance, we firstly present a 3C2N deep learning method. Based on this, a 3C2N-guided visual attention method is proposed for accurate ship detection from SAR images. The main contribution of the proposed approach is that the pre-trained 3C2N model, which could quickly regress the coarse locations while only characterizes the SAR image in space domain, is employed as a ship proposal generator. The 3C2N method combines the CCNN method and the recently published cascade RCNN method. Thus, it could spatially leverage the SAR image information more sufficiently. Complementally, a PCT-based visual attention method is added to perform ship discrimination from the perspective of frequency domain. In addition, the DEM data, which could be indexed by the geographical locations in the meta-data file of Sentinel-1 imagery, is employed to exclude ship-like targets on land. It is notable that in order to improve ship discrimination and without bringing too much burden on computation, the NLM algorithm is utilized on the adaptive ship regions to reduce the speckle noise in SAR images. The experiments are evaluated on twenty-five image blocks of size 200 × 200 in pixel tiled from Sentinel-1 imagery. The F1 score of our approach could reach 0.9605, over 0.0621 when compared with the 3C3N method. This leads to the significant decrease of both the missing detections and the false alarms. Meanwhile, the visualized qualitative results reveal that the detected bounding boxes by our approach are much more accurate than other baselines.

ACKNOWLEDGMENT

The authors would like to thank the European Space Agency for providing the Sentinel-1 data. The authors appreciate the anonymous reviewers for their valuable comments and suggestions for further improving the paper.

REFERENCES

[1] W. Ao, F. Xu, Y. Li, and H. Wang, "Detection and discrimination of ship targets in complex background from spaceborne ALOS-2 SAR images," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 11, no. 2, pp. 536–550, Feb. 2018.

[2] C. H. Gierull and I. Sikaneta, "A compound-plus-noise model for improved vessel detection in non-Gaussian SAR imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 3, pp. 1444–1453, Mar. 2018.

[3] D. Gleich and M. Datcu, "Despeckling and information extraction from SLC SAR images," *IEEE Trans. Geosci. Remote Sens.*, vol. 52, no. 8, pp. 4633–4649, Aug. 2014.

[4] 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.

[5] D. L. McCann and P. S. Bell, "A simple offset 'calibration' method for the accurate geographic registration of ship-borne X-band radar intensity imagery," *IEEE Access*, vol. 6, pp. 13939–13948, 2018.

[6] 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.

[7] D. Xiang, T. Tang, Y. Ban, Y. Su, and G. Kuang, "Unsupervised polarimetric SAR urban area classification based on model-based decomposition with cross scattering," *ISPRS J. Photogramm. Remote Sens.*, vol. 116, pp. 86–100, Jun. 2016.

[8] 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.

[9] 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.

[10] X. Leng, K. Ji, X. Xing, S. Zhou, and H. Zou, "Area ratio invariant feature group for ship detection in SAR imagery," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 11, no. 7, pp. 2376–2388, Jul. 2018.

[11] D. J. Crisp, "The state-of-the-art in ship detection in synthetic aperture radar imagery," Defence Sci. Technol. Org. Salisbury (Australia) Info Sci. Lab, Canberra, ACT, Australia, Tech. Rep. DSTO-RR-0272, 2004.

[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] A. Gambardella, F. Nunziata, and M. Migliaccio, "A physical full-resolution SAR ship detection filter," *IEEE Geosci. Remote Sens. Lett.*, vol. 5, no. 4, pp. 760–763, Oct. 2008.

[14] M. Migliaccio, F. Nunziata, A. Montuori, and R. L. Paes, "Single-look complex COSMO-SkyMed SAR data to observe metallic targets at sea," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 5, no. 3, pp. 893–901, Jun. 2012.

[15] C. S. Yang, J. H. Park, and R. A. Harun-Al, "An improved method of land masking for synthetic aperture radar-based ship detection," *J. Navigat.*, vol. 71, no. 4, pp. 788–804, 2018.

[16] D. E. Molina, D. Gleich, and M. Datcu, "Gibbs random field models for model-based despeckling of SAR images," *IEEE Geosci. Remote Sens. Lett.*, vol. 7, no. 1, pp. 73–77, Jan. 2010.

[17] X. Qin, S. Zhou, H. Zou, and G. Gao, "A CFAR detection algorithm for generalized gamma distributed background in high-resolution SAR images," *IEEE Geosci. Remote Sens. Lett.*, vol. 10, no. 4, pp. 806–810, Jul. 2013.

[18] P. Iervolino, R. Guida, and P. Whittaker, "A novel ship-detection technique for sentinel-1 SAR data," in *Proc. IEEE 5th Asia–Pacific Conf. Synth. Aperture Radar (APSAR)*, Sep. 2015, pp. 797–801.

[19] C. Hu, L. Ferro-Famil, and G. Kuang, "Ship discrimination using polarimetric SAR data and coherent time-frequency analysis," *Remote Sens.*, vol. 5, no. 12, pp. 6899–6920, 2013.

[20] Z. Wang *et al.*, "Visual attention-based target detection and discrimination for high-resolution SAR images in complex scenes," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 4, pp. 1855–1872, Apr. 2018.

[21] Y. Yu, B. Wang, and L. Zhang, "Hebbian-based neural networks for bottom-up visual attention and its applications to ship detection in SAR images," *Neurocomputing*, vol. 74, no. 11, pp. 2008–2017, 2011.

[22] M. Amoon, A. Bozorgi, and G. Rezai-Rad, "New method for ship detection in synthetic aperture radar imagery based on the human visual attention system," *J. Appl. Remote Sens.*, vol. 7, no. 1, p. 071599, 2013.

[23] C. Wang, Z. Wang, H. Zhang, B. Zhang, and F. Wu, "A PolSAR ship detector based on a multi-polarimetric-feature combination using visual attention," *Int. J. Remote Sens.*, vol. 35, no. 22, pp. 7763–7774, 2014.

[24] X. Huang, P. Huang, L. Dong, H. Song, and W. Yang, "Saliency detection based on distance between patches in polarimetric SAR images," in *Proc. IEEE Int. Geosci. Remote Sens. Symp. (IGARSS)*, Jul. 2014, pp. 4572–4575.

[25] Q. Zheng, M. Yang, J. Yang, Q. Zhang, and X. Zhang, "Improvement of generalization ability of deep CNN via implicit regularization in two-stage training process," *IEEE Access*, vol. 6, pp. 15844–15869, 2018.

[26] A. Ullah, J. Ahmad, K. Muhammad, M. Sajjad, and S. W. Baik, "Action recognition in video sequences using deep Bi-directional LSTM with CNN features," *IEEE Access*, vol. 6, pp. 1155–1166, 2018.

[27] Z. Lin, K. Ji, M. Kang, X. Leng, and H. Zou, "Deep convolutional highway unit network for sar target classification with limited labeled training data," *IEEE Geosci. Remote Sens. Lett.*, vol. 14, no. 7, pp. 1091–1095, Jul. 2017.

[28] Z. Deng, H. Sun, S. Zhou, J. Zhao, and H. Zou, "Toward fast and accurate vehicle detection in aerial images using coupled region-based convolutional neural networks," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 10, no. 8, pp. 3652–3664, Aug. 2017.

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

[30] Y. Long, Y. Gong, Z. Xiao, and Q. Liu, "Accurate object localization in remote sensing images based on convolutional neural networks," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 5, pp. 2486–2498, May 2017.

[31] G. Cheng and J. Han, "A survey on object detection in optical remote sensing images," *ISPRS J. Photogramm. Remote Sens.*, vol. 117, pp. 11–28, Jul. 2016.

[32] M. Kang, K. Ji, X. Leng, and Z. Lin, "Contextual region-based convolutional neural network with multilayer fusion for SAR ship detection," *Remote Sens.*, vol. 9, no. 8, p. 860, 2017.

[33] J. Jiao *et al.*, "A densely connected end-to-end neural network for multiscale and multiscene SAR ship detection," *IEEE Access*, vol. 6, pp. 20881–20892, 2018.

[34] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2015, pp. 91–99.

[35] G. Huang, Z. Liu, L. van der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, vol. 1, Jul. 2017, no. 2, pp. 2261–2269.

[36] Z. Juanping, G. Weiwei, Z. Zenghui, and Y. WenXian, "A coupled convolutional neural network for small and densely clustered ship detection in SAR images," *Sci. China Inf. Sci.*, 2018. [Online]. Available: http://engine.scichina.com/publisher/scp/journal/SCIS/doi/10.1007/s11432-017-9405-6?slug=abstract

[37] T. Y. Lin, P. Dollár, R. Girshick, S. Belongie, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proc. CVPR*, vol. 1, 2017, no. 2, p. 4.

[38] T. Chen, S. Lu, and J. Fan, "S-CNN: Subcategory-aware convolutional networks for object detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 10, pp. 2522–2528, Oct. 2017.

[39] 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.

[40] P. Sermanet and Y. LeCun, "Traffic sign recognition with multi-scale convolutional networks," in *Proc. Int. Joint Conf. IEEE Neural Netw. (IJCNN)*, Jul./Aug. 2011, pp. 2809–2813.

[41] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2009, pp. 248–255.

[42] Z. Cai and N. Vasconcelos. (2017). "Cascade R-CNN: Delving into high quality object detection." [Online]. Available: https://arxiv.org/abs/1712.00726

[43] D. Cozzolino, S. Parrilli, G. Scarpa, G. Poggi, and L. Verdoliva, "Fast adaptive nonlocal SAR despeckling," *IEEE Geosci. Remote Sens. Lett.*, vol. 11, no. 2, pp. 524–528, Feb. 2014.

[44] Y. Yu, Z. Ding, B. Wang, and L. Zhang, "Visual attention-based ship detection in SAR images," in *Advances in Neural Network Research and Applications*. Berlin, Germany: Springer, 2010, pp. 283–292.

[45] A. Neubeck and L. Van Gool, "Efficient non-maximum suppression," in *Proc. 18th Int. Conf. IEEE Pattern Recognit. (ICPR)*, vol. 3, Aug. 2006, pp. 850–855.

[46] K. Simonyan and A. Zisserman. (2014). "Very deep convolutional networks for large-scale image recognition." [Online]. Available: https://arxiv.org/abs/1409.1556

[47] Y. Jia *et al.*, "Caffe: Convolutional architecture for fast feature embedding," in *Proc. 22nd ACM Int. Conf. Multimedia*, 2014, pp. 675–678.

[48] M. Zuhlke *et al.*, "SNAP (sentinel application platform) and the ESA sentinel 3 toolbox," in *Proc. Sentinel-3 Sci. Workshop*, vol. 734, 2015, p. 21.

[49] Y. Yu, B. Wang, and L. Zhang, "Pulse discrete cosine transform for saliency-based visual attention," in *Proc. IEEE 8th Int. Conf. Develop. Learn. (ICDL)*, Jun. 2009, pp. 1–6.

[50] P. A. M. Berry, J. D. Garlick, and R. G. Smith, "Near-global validation of the SRTM DEM using satellite radar altimetry," *Remote Sens. Environ.*, vol. 106, no. 1, pp. 17–27, 2007.

[51] A. Neubeck and L. Van Gool, "Efficient non-maximum suppression," in *Proc. 18th Int. Conf. Pattern Recognit. (ICPR)*, vol. 3, Aug. 2006, pp. 850–855.

[52] P. W. Vachon, J. Wolfe, and H. Greidanus, "Analysis of sentinel-1 marine applications potential," in *Proc. IEEE Int. Geosci. Remote Sens. Symp. (IGARSS)*, Jul. 2012, pp. 1734–1737.

JUANPING ZHAO received the B.Sc. degree in electronic information engineering from Xidian University, Xi'an, China, in 2014. She is currently pursuing the Ph.D. degree with the School of Electronic Information and Electrical Engineering, Shanghai Jiao Tong University, Shanghai, China.

Her research interests include synthetic aperture radar image interpretation, pattern recognition, and machine learning.

ZENGHUI ZHANG (M'13) received the B.Sc. degree in applied mathematics, the M.Sc. degree in computational mathematics, and the Ph.D. degree in information and communication engineering from the National University of Defense Technology (NUDT), Changsha, China, in 2001, 2003, and 2008, respectively.

From 2008 to 2012, he was a Lecturer with the Department of Mathematics and System Science, NUDT. He is currently an Associate Professor with the School of Electronic Information and Electrical Engineering, Shanghai Jiao Tong University, Shanghai, China. His main research interests include radar image interpretation, radar signal processing, and compressed sensing theory and applications.

WENXIAN YU was born in Shanghai, China, in 1964. He received the B.Sc. degree in radio measurement and control and data transmission, the M.Sc. degree in communication and electronic system, and the Ph.D. degree in communication and information system from the National University of Defense Technology (NUDT), Changsha, China, in 1995, 1988, and 1993, respectively.

From 1996 to 2008, he was a Professor with the College of Electronic Science and Engineering, NUDT, where he served as the Deputy Head of the College and the Assistant Director of the National Key Laboratory of Automatic Target Recognition. He was the Executive Dean of Shanghai Jiao Tong University, Shanghai, from 2009 to 2011. He is currently with the School of Electronic Information and Electrical Engineering, Shanghai Jiao Tong University, where he is a Yangtze River Scholar Distinguished Professor and the Head of Research. His current research interests include radar target recognition, remote sensing information processing, multisensor data fusion, and integrated navigation system. In these areas, he has published over 200 research papers.

TRIEU-KIEN TRUONG (M'82–SM'83–F'99–LF'13) received the B.Sc. degree from National Cheng Kung University, Taiwan, in 1967, the M.Sc. degree from Washington University, St. Louis, MO, USA, in 1971, and the Ph.D. degree from the University of Southern California, Los Angeles, CA, USA, in 1976, all in electrical engineering.

From 1975 to 1992, he was a Senior Member of Technical Staff (E6) with the Communication System Research Section of the JPL, Pasadena, CA, USA. From 1976 to 1995, he was an Adjunct Professor with the Department of Electrical Engineering System, Communications Science Institute, USC, and was a Consultant with the Department of Radiology, Memorial Hospital of Long Beach. From 1995 to 2010, he was a Chair Professor and the Dean of the College of Electrical and Information Engineering, I-Shou University, Taiwan, where he is currently a Visiting Distinguished Chair Professor. He is also a Visiting Chair Professor with the School of Electronic Information and Electrical Engineering, Shanghai Jiao Tong University.

He holds many patents in different fields. He has authored over 200 journal papers, among them about 26 papers are published in the IEEE Transactions on Information Theory. His main research interests include error correcting code, VLSI architecture design, communication systems, signal and image processing, synthetic aperture radar digital professor, aerial images, and CT-aided robotic stereotaxis system. He was a recipient of many honors, including 23 NASA awards for outstanding technical contributions.