IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received October 10, 2020, accepted October 19, 2020, date of publication October 23, 2020, date of current version November 5, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.3033469

Analysis on Saliency Estimation Methods in High-Resolution Optical Remote Sensing Imagery for Multi-Scale Ship Detection

ZEZHONG LI, (Student Member, IEEE), YANAN YOU, (Member, IEEE), AND FANG LIU

Beijing Laboratory of Advanced Information Networks, School of Artificial Intelligence, Beijing University of Posts and Telecommunications, Beijing 100876, China

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

Corresponding author: Yanan You (youyanan@bupt.edu.cn)

This work was supported in part by the Ministry of Education-China Mobile Communication Corp (MoE-CMCC) Artificial Intelligence Project under Grant MCM20190701.

ABSTRACT Ship detection is of considerable significance in both military and civilian application domains. Deep Convolutional Neural Network (DCNN) with region proposal mechanism, e.g., Faster R-CNN, performs outstandingly in ship detection with high-resolution images. However, the accuracy limitation is induced by the region proposal restricted by the training set for multi-scale target detection. Therefore, the method of multi-scale ship object detection is proposed based on saliency estimation in our work. Saliency Estimation Algorithms (SEAs) are often used to extract saliency features in images. In ship detection of high-resolution remote sensing images, these algorithms can extract information such as the scale and position of the targets, and then help the DCNN-based ship detection method to obtain better performance. To verify the effectiveness of the saliency estimation algorithm in multi-scale ship detection of high-resolution remote sensing images, and analyze the advantages of different SEAs. This paper introduces 13 classic saliency estimation algorithms and 2 DCNN-based ones to compare them by using evaluation indicators. At last, in order to demonstrate the performance of different SEAs, the extracted saliency feature maps are used to assist the DCNN-based target detection under multi-scale ships condition. In general, this framework can improve the detection accuracy of large-scale ships under scenarios of training only with small scale ships or without enough datasets.

INDEX TERMS Saliency estimation, remote sensing image, multi-scale ship detection, deep learning.

I. INTRODUCTION

With the development of remote sensing technology, it has been increasingly mature for people to observe ship activities by using remote sensing images. And now, ship detection in remote sensing images are critical in both the military and civilian marine fields [1]. As far now, the research on marine areas has attracted much attention because of its unique strategic position, rich resource distribution, and extensive maritime trade. Notably, the ship target, as an indispensable correlative condition between the human and the marine environment, has made a significant distribution to marine research. And their position information and behavioral knowledge of ships have played an important role in maritime safety, marine transportation, marine pollution, traffic management, etc. [2]

In recent researches, the ship detection task is collectively referred to as the object detection task, which purpose is to extract the targets of interests in images or videos. With the development of deep learning (DL), especially convolution neural networks (CNN), some intelligent object detection methods have become mainstream because of its powerful ability in extracting the features of objects. Girishick *et al.* [3] proposed R-CNN and integrated the three steps of traditional object detection into the DL algorithms. It took advantage of the neural network to automatically extract the features of each window obtained by selective search, and used SVM to realize the classification. Then, He *et al.* [4] proposed the SPP method, which enabled the model to extract the feature with same size from pictures of different sizes. Inspired by the

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

above researches, Girshick *et al.* proposed two methods based on R-CNN. The one was Fast R-CNN [5] that can conduct convolution on the entire image to obtain the larger feature maps and then choose the candidate regions on the feature maps. The other was Faster R-CNN [6], which introduced the mechanism of Region Proposal Network (RPN) that used fixed-size anchors to select candidate regions on the feature map quickly.

At the same time, researchers began to dig deeper into the application of DL of object detection tasks. Broke through the way of Faster R-CNN, SSD proposed by Liu *et al.* [7] creatively fused the candidate region proposal and object classification into a single network, which further improved the processing rate of object detection. Afterwards some researchers proposed a series of improvements based on a practical problem, e.g., DSSD [8], which used the top-down structure to fuse multi-layer features and improved the detection ability of small objects. Another famous model is the YOLO series [9]–[11], which all adopted a more concise network to perform object feature extraction and classification simultaneously. And during the development of the YOLO series, it also absorbed the advantage of Faster R-CNN to have a better detection accuracy. Also, there are some methods aimed at the actual problem. For example, the FPN [12] combined high and low features to detect small-scale objects; the SNIP [13] changed the size of objects and got better detection results on multi-scale objects; He *et al.* [14] proposed Mask R-CNN to extract the shapes and contours of objects.

As a branch of object detection, the ship detection task also experiences the development process of object detection. However, there are some huge differences between the ships in high-resolution remote sensing images and ships in natural images: Firstly, the ship objects in remote sensing images are the top view, which limits its angular diversity. Secondly, ship objects present the characteristics of multi-direction, multi-scale, and multi-shape in the remote sensing images, and the background of the ships is more complicated. Thirdly, the high-resolution remote sensing images are always affected by some factors such as sunlight, clouds, and mist, which reduce the imaging quality of ship objects. Therefore, many different methods are proposed to solve these problems. Both Jiang *et al.* [15] and Yang *et al.* [16] proposed some methods to integrate rotate information of ship objects so as to improve the accuracy of multi-direction ships. Zhang *et al.* [17] combined the deep convolutional neural network (DCNN) and manually ship features to enhance the accuracy of ships. Lavalle *et al.* [18] introduced GIS data into the DL algorithm to realize the division of ocean and land areas, which show significantly performance on false alarm suppression. And You *et al.* [19] used scene segmentation to achieve the same thing. In addition, our previous work [2] proposed a useful algorithm framework to deal with the problems that may occur in the ship detection task.

Furthermore, there are some interesting and excellent object detection algorithms that can be effectively applied to detection tasks on high-resolution remote sensing images. A method [20] introduced a learnable rotation-invariant layer on the basis of existing algorithms to alleviate the problems caused by angles of objects in remote sensing images. Reference [21] introduced a rotation-invariant regularizer and fisher discrimination regularizer to construct a new objective function to improve the performance of object detection. Reference [22] proposed ORCNN (Oriented R-CNN) that applying an additional classifier on the detected regions to get orientations from the six predefined angle classes. Reference [23] exploited RPN by adding multi-angle anchor boxes beside the conventional ones to address rotation variations and appearance ambiguity. Besides, many impressive methods, comparisons, and analyses can be found in [24].

Although many innovations have been implemented, the problem of multi-scale ship detection still hinders the development of ship detection in high-resolution remote sensing images. In this paper, we will focus on this critical issue of the excessive dependence of training datasets on the DL network. Among the existing datasets, small-scale ships, e.g., yacht, often occupy the vast majority. However, the large-scale ships, e.g., aircraft carriers, occupy only a small proportion, like in DOTA [25] and DIOR [26]. Therefore, it always leads to a phenomenon that the performance on large ships is weak, while the performance on small ships is quite satisfied when these datasets are used to train ship detection network.

Aiming at the data imbalance problem, a method that combines the classic saliency estimation algorithm (SEA) and the DL network is proposed in our work. The SEA is used to process the images which is tested in the target detection network and get the saliency feature maps. Then one discriminant formula proposed to determine whether the image contains large-scale ship targets by calculate it on saliency feature maps. At last, the images which contains large-scale ships are resized to reduce the size of ships and improve the detection accuracy. Compared with other methods, the focus of our paper is that the traditional SEA can achieve the saliency feature maps without training on the dataset and alleviate the detection network's dependence on the size of ship targets in the training dataset. Furthermore, this article is an extended experiment of our conference paper [27] and paper [2]. Some extended experiments are designed for the verification of applicability and effectiveness of different SEAs in remote sensing images which contain two DCNN-based SEA. And, these SEAs are combined with a variety of detection networks to verify the effectiveness of the target detection task such as YOLOv3. Furthermore, there are many existing SEAs now, it is hard to present a complete introduction due to length constraints. Some detailed overview can be found in [28].

The rest of this paper is presented as follows. The Section II shows the problem and architecture. The Section III performs the analysis of the performance of SEA on high-resolution remote sensing images. The high-resolution datasets are introduced and the performance of SEAs on this dataset are presented in this section. The Section IV gives the results of experiments, which

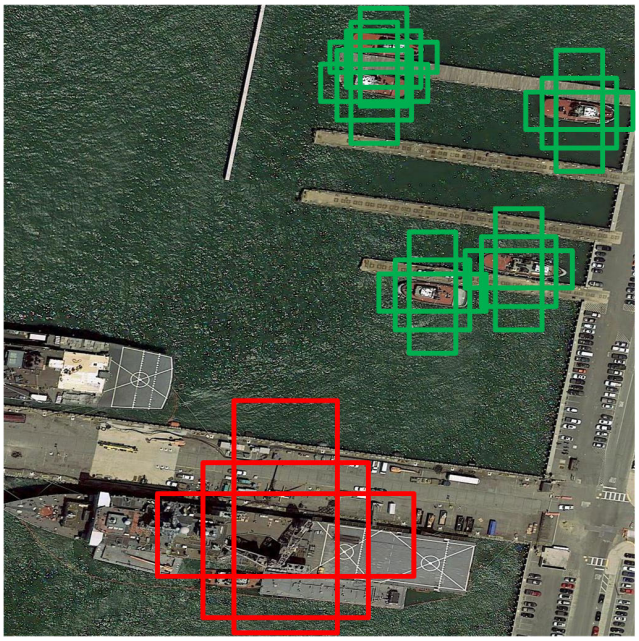


FIGURE 1. The mechanism of the region proposal network. The red boxes present large-scale ship that is not fully covered and the green boxes present the small-scale ships with the appropriate proposals.

include the performance of the method combining the SEAs and DL-based object detection algorithms on multi-scale ship detection tasks. And the Section V is the summary.

II. PROBLEMS AND ARCHITECTURE

A. PROBLEMS

Due to the multi-scale characteristic of ship objects in high-resolution remote sensing images, for example, the size of the aircraft carriers is about 300 meters. In contrast, the size of the yacht range in length from 10 to 100 meters and this result in a huge difference when they appear in images at the same time. However, the number of small-scale ships is far more than the large ship in the dataset. And this leads to a result that the performances of DCNN-based object detection algorithms on multi-scale ship detection are unsatisfied. Take Faster R-CNN as an example, the module of RPN, which contains many fixed-size anchors is used to search ships in the feature map. And during the training, these anchors gradually learn to classification and regression to cover the truth ship. However, in the case of imbalance samples, it is hard to get enough chances to learn the feature of large ships, and this result in poor performance on large ships. And the diagram of the mechanism of RPN is shown in Fig. 1.

B. ARCHITECTURE

In order to improve the detection accuracy of large-scale ship targets under the condition that the training set contains a large number of small-scale ship targets. SEAs are used to obtain the saliency feature maps of the image before ship detection, then these saliency feature maps are processed by one discriminant formula presented in (10), to determine whether the image contains large-scale ship targets. At last, the image with the large-scale ships is resized to reduce the size of ship targets. Finally, these images are served as the input of the object detection network and obtain the features of these ships. Therefore, a framework combined SEA and DL-based object detection method is proposed in this paper to improve the performance of multi-scale ship detection in high-resolution remote sensing images. In conclusion, the purpose of this architecture is to detect large-scale ship targets in the case that there are many small-scale ships in the training set. And the highlight is that the traditional SEAs which cannot be affected by the training set can differentiate the scale of ships. The diagram of architecture is shown in Fig. 2 and some implementation details will be analyzed in later sections.

Furthermore, the purpose of the SEA is to describe the distribution of saliency targets in the image. Some methods obtain the saliency value of each pixel by comparing each pixel with other pixels in the image or surrounding pixels. And some methods use the map after the domain changes (wavelet transform) to compare the pixels to obtain saliency value. Some detailed analysis of these SEAs can be found in [28]. Moreover, the SEA in our architecture is served as the data preparation and is used to find out the images with large-scale ships. Then it resizes these images to reduce the size of ships and send the results to the object detection network. Furthermore, this architecture can be applied to different object detection algorithms. Some results will be presented in the last experiment. And the Faster R-CNN is an example in this paper to introduce the training and process of the whole architecture.

During the training of this example, a multi-task loss function, which is calculated by forwarding propagation, is used to monitor the progress of training. Then, the parameters in the network are updated by gradient descent. In addition, it is unnecessary to train the SEA because it is based on digital image processing. Thus, the loss function is the sum of the regression loss function and the classification loss function of the object detection network. The calculation method is shown in (1).

$$Loss = L_{cls} + L_{reg} \tag{1}$$

where L_{cls} represents the classification loss function, and its calculation method is shown in (2), and L_{reg} represents the regression loss function, its calculation method is shown in (3) and (4).

$$L_{cls} = -\frac{1}{N_{cls}} \sum_{p} log[O^{(p)}C^{(p)} + (1 - O^{(p)})(1 - C^{(p)})] \tag{2}$$

where O is the probability of that the candidate box p is recognized as target, C is the ground truth. And when the p candidate box is ground truth, this value is 1, otherwise, it is 0. N_{reg} presents the number of candidate boxes.

$$L_{reg} = \frac{\beta}{N_{reg}} \sum_{p} C^{(p)} smooth_{L1}(B^{(p)} - reg^{(p)}) \tag{3}$$

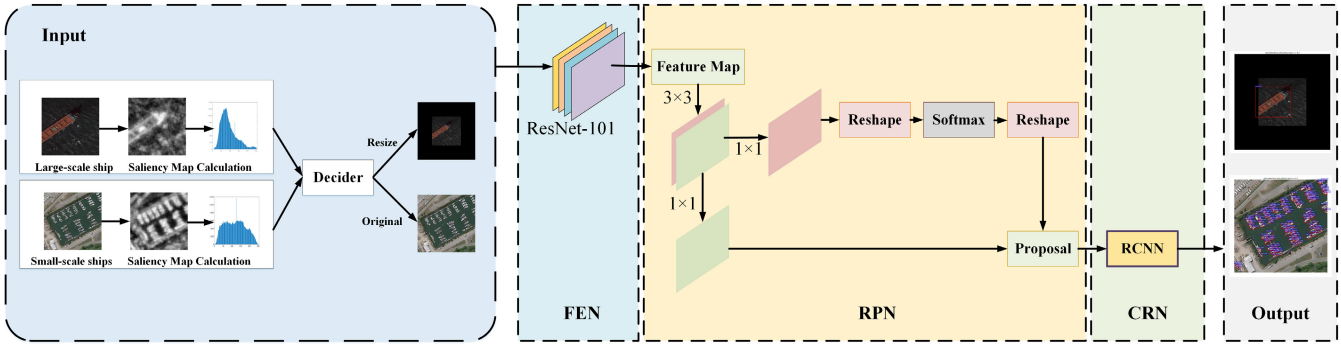


FIGURE 2. The architecture of SEA-based object detection method. The Decider is presented in (10), the FEN is feature extraction network (FEN) and CRN is classification and regression network.

$$smooth_{L1}(x) = \begin{cases} 0.5x^2, & IF\ |x| < 1 \\ |x| - 0.5, & OTHERWISE \end{cases} \tag{4}$$

where N_{reg} represents the number of candidate boxes, β is used to balance the ratio of classification loss and regression loss. $reg^{(p)} = reg_x^{(p)}, reg_y^{(p)}, reg_w^{(p)}, reg_h^{(p)}$ is a vector that contains the coordinates of each box, and $B^{(p)} = B_x^{(p)}, B_y^{(p)}, B_w^{(p)}, B_h^{(p)}$ represents the coordinates of ground truth.

During the inference process, the SEA is used to deal with input images and obtain the saliency feature maps. Then the saliency feature maps are processed by a discriminant formula explained in equation (10) to judge the ship size in images. At last, the image pyramid is used to resize the image with large-scale ships so that the ship can be detected by ship detection network and get better detection results.

III. ANALYSIS

In this section, some analyses about the performance of SEAs on high-resolution are presented. And one SEA, which is most suitable for multi-scale object detection tasks, are chosen to apply in our architecture.

A. SEAs

Many saliency estimation algorithms have been proposed. However, the performance of different SEAs on remote sensing images is still unproved. Therefore, some SEAs are picked to conduct a comparative analysis to verify that SEAs can effectively extract the saliency of ship objects on remote sensing images. In experiments, 15 SEAs in total are tested on high-resolution remote sensing images, containing 13 classic SEAs and 2 DL-based SEAs. And some information about these SEAs is shown in TABLE 1.

The calculation method of each SEA is different. The SIM algorithm extracts the saliency feature map by wavelet transform and calculates the distance of different pixels; The CA algorithm obtains local and global information by comparing the differences between different regions in the image. While the DCNN-based SEAs use scene masks to construct multi-layer neural networks to implement saliency estimation, e.g., PoolNet extracts the saliency information in different convolutional layers and finally merges to achieve saliency feature map.

TABLE 1. Outline information for SEAs involved in our study.

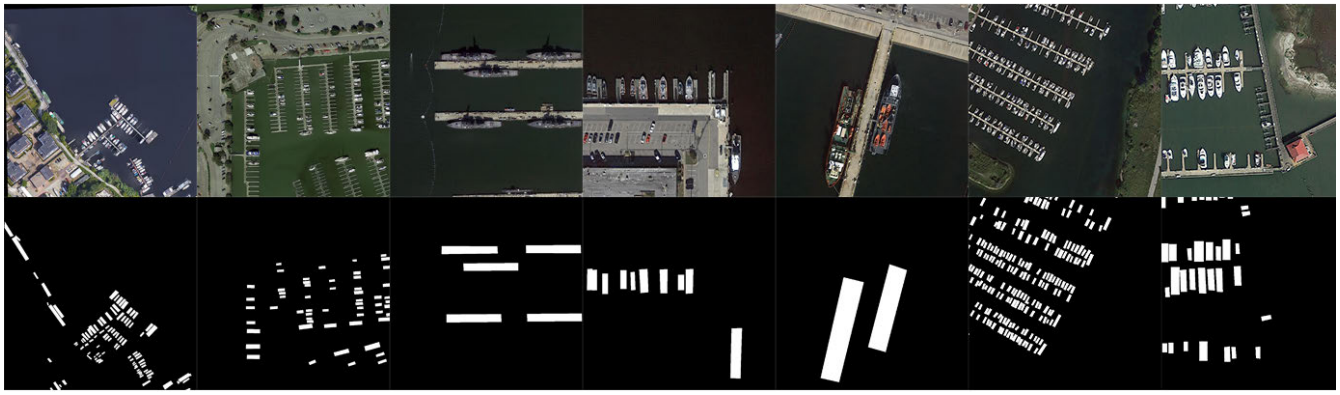
	Model	Publication	Year	categories
1	GR [33]	SPL	2013	Salient Object Detection
2	DSR [34]	ICCV	2013	
3	PCA [35]	CVPR	2013	
4	MC [36]	ICCV	2013	
5	COV [41]	JOV	2013	
6	SWD [31]	CVPR	2011	
7	FES [32]	Img.Anal	2011	
8	SIM [40]	CVPR	2011	
9	SEG [29]	ECCV	2010	
10	CA [30]	CVPR	2010	
11	SeR [39]	JOV	2009	
12	SUN [38]	JOV	2008	
13	SR [37]	CVPR	2007	
14	PoolNet [42]	CVPR	2019	DL-based Model
15	DHSNet [43]	CVPR	2016	

B. DATASETS

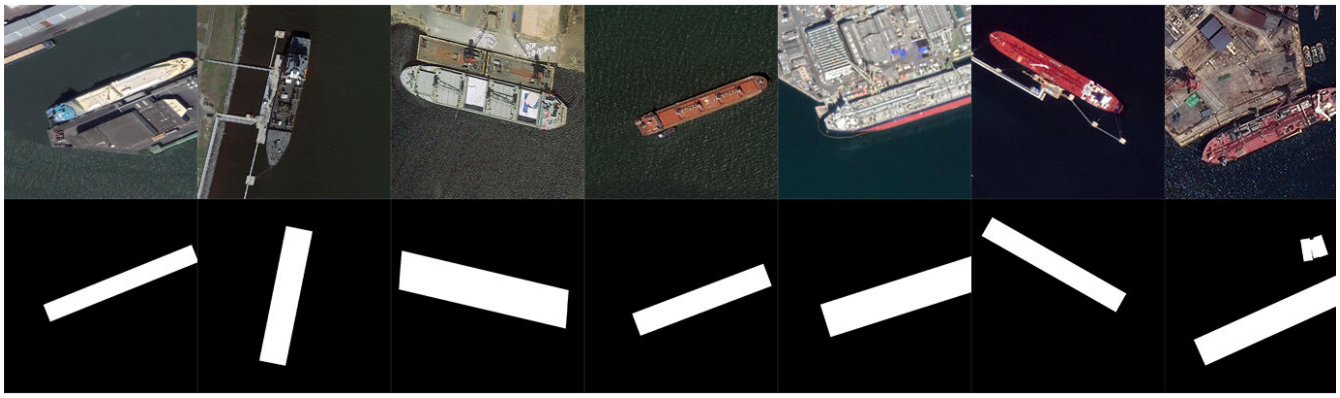
The DOTA dataset with around 400 images is used as the training dataset and the ratio of training/testing is 3:1. And the resolution of each image in DOTA is better than 1m. In addition, the DOTA dataset contains much more small-scale ships than large-scale ships. Some examples are shown in Fig. 3(a).

And the testing dataset consists of 310 images randomly picked from the DOTA and 152 images obtained manually from the Earth. The resolution of Google Earth is 0.3m, and they are used as supplementary data to present the performance of large-scale ships. Some examples of Google Earth dataset are shown in Fig. 3(b).

In order to ensure that the input image can be adequately trained on GPU, all images in both datasets are cut into a fixed size of 1024 × 1024, and thus more than 2000 images are prepared for training and testing.



(a) Samples collected from DOTA dataset



(b) Samples collected from Google Earth dataset

FIGURE 3. Samples in our datasets.

C. EVALUATION

Then, some details of evaluation indexes are introduced to verify the performance of different SEA algorithms on remote sensing images. It should be noted that the indexes mentioned below are used to measure the performance of SEA in remote sensing images containing ship objects. The saliency map obtained from SEAs is converted to the binarization mask before the assessment. The method of fixed threshold conversion is used in this experiment. It means that the pixel is labeled as the positive pixel if its value is higher than the threshold. In contrast, the pixel with the value smaller than the threshold will be labeled as the negative pixel. The calculation method is shown in (5).

$$M_b(i,j) = \begin{cases} 255, & S(i,j) > T_{mask} \\ 0, & S(i,j) < T_{mask} \end{cases} \tag{5}$$

where $M_b(i,j)$ represents the value of the pixel (i,j) in binary saliency map M_b, $S(i,j)$ represents the value of the pixel (i,j) in saliency map S, and T_{mask} is used to generate the binary saliency map from the saliency map and the range of this value is 0 to 255.

In terms of the indicators for evaluating the performance of each SEA, the PR (Precision-Recall) curve and the ROC (Receiver operating characteristic) curve are jointly used as evaluating indexes. In addition, the value of F-measure and AUC (Area Under Curve), which are calculated according to the PR curve and ROC curve, are used as the comprehensive indicators.

The PR curve consists of the value of precision and recall, which are calculated by binary saliency map and ground truth. The calculation method is shown in (6).

$$Precision = \frac{|GT \bigcap M_b|}{M_b}, Recall = \frac{|GT \bigcap M_b|}{GT} \tag{6}$$

where GT represents the ground truth of the saliency map, and M_b represents the binary saliency map, which can be changed by changing T_{mask}.

The value of F-measure is commonly calculated from the value of precision and recall, and its calculation method is shown in (7).

$$F - measure = \frac{(\beta^2 + 1) \times Precision \times Recall}{\beta^2 \times Precision + Recall} \tag{7}$$

And the value of ROC is calculated from TPR (True Positive Rate) and FPR (False Positive Rate), and these values are also calculated from the binary saliency map and ground truth. The calculation method is shown in (8).

$$\begin{aligned} TPR &= \frac{TP}{TP + FN} = \frac{M_b \bigcap GT}{GT}, \\ FPR &= \frac{FP}{FP + TN} = \frac{M_b \bigcap \overline{GT}}{\overline{GT}} \end{aligned} \tag{8}$$

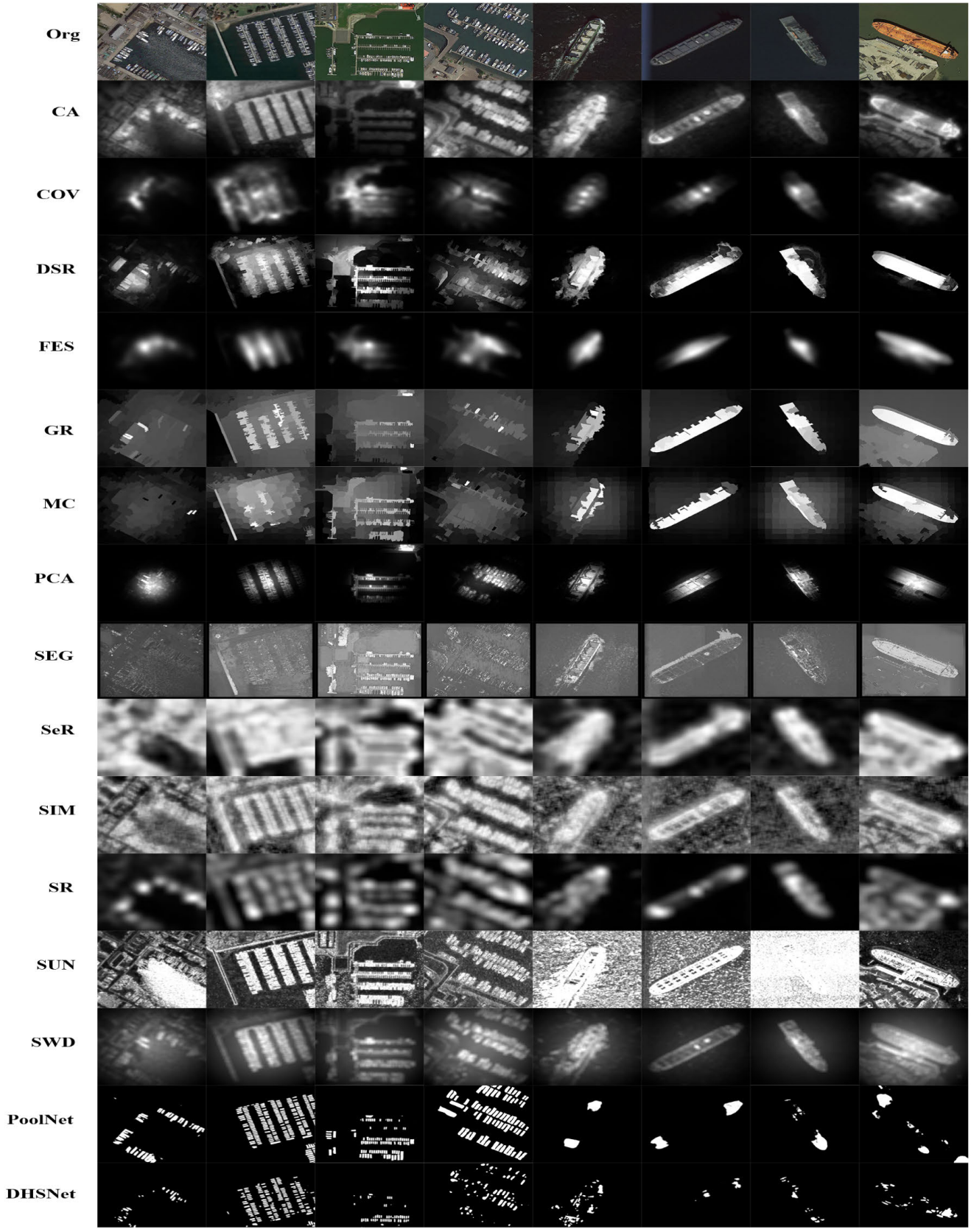


FIGURE 4. Some results of all SEAs on high-resolution remote sensing images. The first two columns are samples from the DOTA dataset and the last two columns are samples from the Google Earth dataset.

where $\overline{GT}$ is the inverse set of the ground truth of saliency map. And the value of *TPR* and *FPR* can be changed by changing T_{mask} so that the ROC curve can be obtained.

The AUC is a probability value, which represents the area of the ROC curve, and this index can intuitively show the advantage of different SEAs. The calculation method is shown in (9).

$$AUC = \sum_{T_{mask}=0}^{254} (FPR_{T_{mask}+1} - FPR_{T_{mask}}) \times (TPR_{T_{mask}+1} + TPR_{T_{mask}}) \quad (9)$$

where $FPR_{T_{mask}}$ and $TPR_{T_{mask}}$ presents the value of *FPR* and *TPR* when the threshold is T_{mask}, and the area of the ROC can be obtained by the way of approximate superposition.

D. COMPARATIVE EXPERIMENT

In the section, a comparative experiment is elaborated by using the SEAs, datasets, and evaluating indexes that we have introduced above. And the goal of this experiment is to analyze the performance of classic SEAs and DCNN-based SEAs and pick one SEA that is most suitable for multi-scale ship detection in high-resolution remote sensing images.

Firstly, all the classic SEAs are directly tested on our test dataset, and the DCNN-based SEAs are trained on the DOTA training dataset and then tested on the test dataset. The result of saliency maps is shown in Fig. 4.

It can be seen from Fig. 4 that DCNN-based SEAs obtain accurate results on the DOTA testing dataset, however, have a weak performance on the Google Earth dataset, which contains many images quite different from the training dataset (DOTA), some broken feature and disappearing feature are observed in Google Earth dataset. Therefore, it is hard to use these saliency feature map to determine the large-scale ships in this image. On the contrary, the results of some classic SEAs still have a good performance on both datasets, such as SIM, SUN. Meanwhile, the saliency feature map of these SEAs is reliable for the ship scale indication.

Then, these saliency feature maps are used to calculate the evaluating indexes, the PR curve, and the ROC curve, as shown in Fig. 5. The values of F-measure and AUC are shown in TABLE 2. Furthermore, the bold highlighted values in TABLE 2 present the value of SIM on each indicator. And the diagram of F-measure and AUC are shown in Fig. 6.

Firstly, It should be noted that the value of classic SEAs is commonly lower than DCNN-based SEAs. The salient feature map obtain through traditional SEAs is a continuous grayscale image. In contrast, the results of DCNN-based SEAs are binary feature map, which are determined by their calculation mechanism and the ground truth.

In Fig. 5(a), it is clear to see that the PR curves of PoolNet and DHSNet commonly are higher than the classic SEAs on the DOTA dataset. It means that the power of DCNN-based SEAs to extract the saliency feature on DOTA dataset are

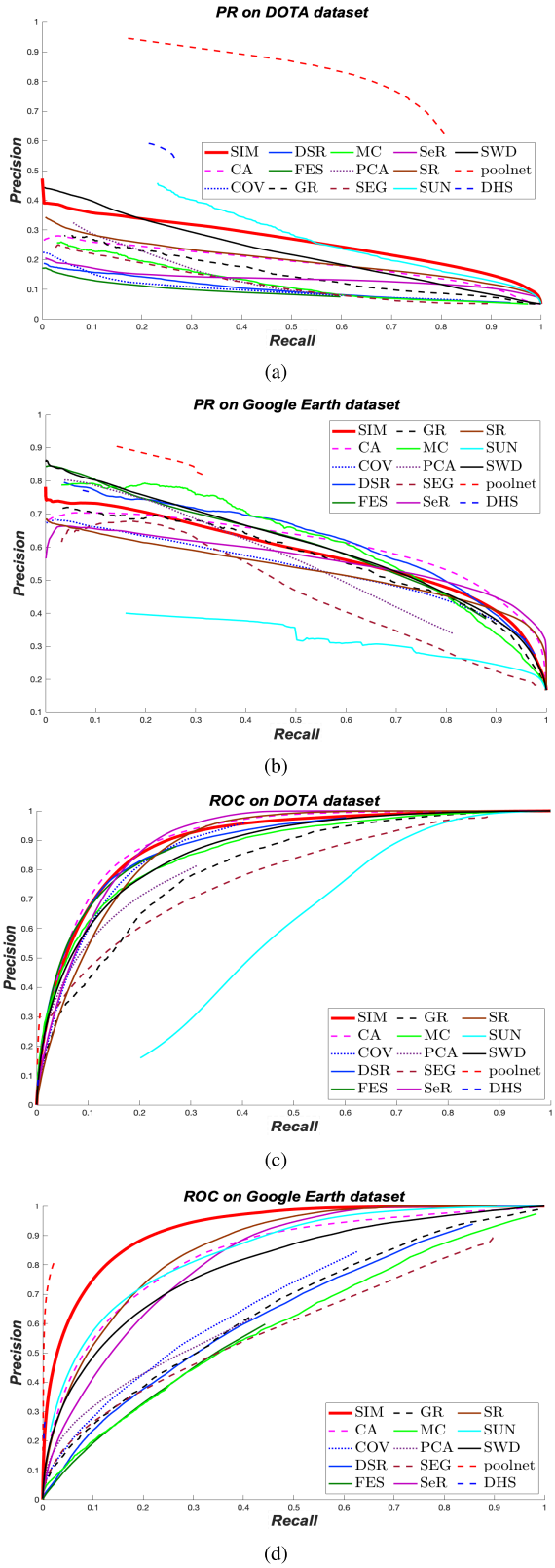
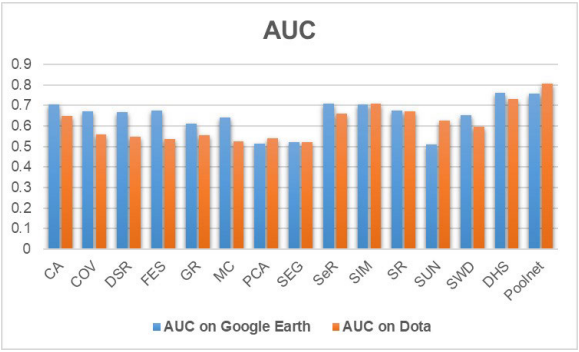


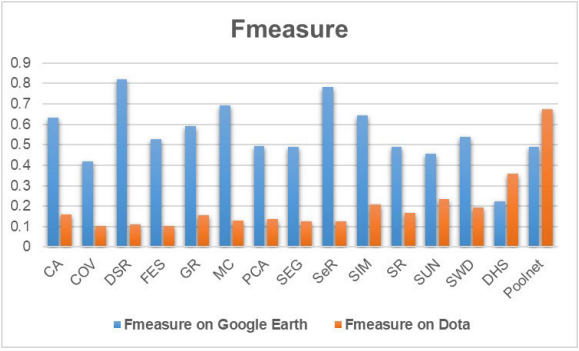
FIGURE 5. The performance of all SEAs. (a) is the PR curves of all SEAs on DOTA dataset, (b) is the PR curves of all SEAs on the Google Earth dataset, (c) is the ROC curve of all SEAs on the DOTA dataset and (d) is the ROC curve of all SEAs on the Google Earth dataset.

TABLE 2. The value of F-measure and AUC of all SEAs on Google Earth dataset and DOTA dataset.

	F-measure		AUC	
	Google Earth	DOTA	Google Earth	DOTA
GR	0.592	0.156	0.611	0.554
DSR	0.821	0.110	0.667	0.546
PCA	0.494	0.138	0.512	0.541
MC	0.693	0.129	0.642	0.525
COV	0.418	0.102	0.673	0.557
SWD	0.541	0.192	0.652	0.595
FES	0.527	0.105	0.675	0.537
SIM	**0.645**	**0.207**	**0.706**	**0.709**
SEG	0.489	0.126	0.520	0.520
CA	0.631	0.160	0.707	0.648
SeR	0.782	0.124	0.708	0.659
SUN	0.457	0.233	0.509	0.625
SR	0.490	0.166	0.674	0.672
PoolNet	0.489	0.674	0.759	0.808
DHSNet	0.222	0.360	0.760	0.730



(a)



(b)

FIGURE 6. The diagram of the F-measure and the AUC of all SEAs on the Google Earth dataset and the DOTA dataset.

higher than classic SEAs, while in Fig. 5(b), the power of classic have a massive improvement leading to better performance of some SEAs, like MC and DSR. This phenomenon also proves that the DCNN-based SEAs only have a perfect performance on the dataset, which is similar to their training dataset. And according to the statistics of F-measure, which are shown in TABLE 2 and Fig. 6, the value of F-measure of DCNN-based SEA has a dropping trend on the Google Earth dataset. The value of PoolNet has changed from 0.674 to 0.489, and the value of DHSNet drops from 0.360 to 0.222. On the contrary, almost each classic SEAs both have a huge improvement on the Google Earth dataset, such as the increased values of SIM (from 0.207 to 0.645), and DSR (from 0.11 to 0.821). By the analysis of this phenomenon, the broken features and the disappearing features appeared in the results of DCNN-based SEAs play a fatal role, because of the imbalanced training dataset (there are many small-scale ships and little large-scale ships). In addition, the samples in the Google Earth are large-scale ships whose energy is gathered in the images. Therefore, there is an improvement on the performance of SEAs.

Furthermore, the ROC curves shown in Fig. 5(c) and Fig. 5(d) are used to be another criterion. It should be noted that the value of ROC must be higher than 0.5, and its curve is higher than the straight line with a slope of 1, which lead to small difference between DCNN-based SEAs and classic SEAs. Besides, it can be seen in Fig. 5(c) and Fig. 5(d), the trend of the ROC curve of each SEA is similar with the PR curve: the performance of DCNN-based SEAs is lower on the Google Earth dataset than the DOTA dataset, while the trend of the classic is reversed, and the values of AUC shown in TABLE 2 have the same trend as the F-measure. Although the AUC of DCNN-based SEAs are the highest on both two datasets, the phenomenon of broken feature results in that these algorithms cannot be used to determine whether there are large-scale ships in images. Then, among the classic SEAs, the SIM is better than other SEAs because of its good performance on both datasets and all indexes, like 0.207 of the F-measure on the DOTA dataset and 0.706 of the AUC on the Google Earth dataset.

It is reasonable to believe that the SIM has the best performance on high-resolution remote sensing images based on its valid on two datasets. Therefore, the SIM is used as our data preparation to a multi-scale ship detection network. According to the saliency feature maps extracted by SIM, the feature distribution of the input images with large-scale ships commonly represents the energy accumulation. In contrast, the images which contain small-scale ships present the energy dispersion. Thus, a statistical histogram method is adopted to judge feature distribution. Firstly, the gray-scale histogram statistics on saliency feature maps are calculated. Then, a judgment model is established based on this gray-scale histogram. Finally, each saliency feature map is classified by the judgment model shown in (10). Some examples of statistics histogram are shown in Fig. 7.

$$f_s = \begin{cases} 1, & IF \ \dfrac{1}{L_u - L_l} \sum_{i=L_i}^{L_u} M_i \geq \dfrac{1}{R_u - R_l} \sum_{i=R_l}^{R_u} M_i \geq T \\ 0, & OTHERWISE \end{cases} \tag{10}$$

where M represents the information of saliency feature map, L_u and L_l represent the high and low boundaries of the left peak in histogram, R_u and R_l represent the high and low

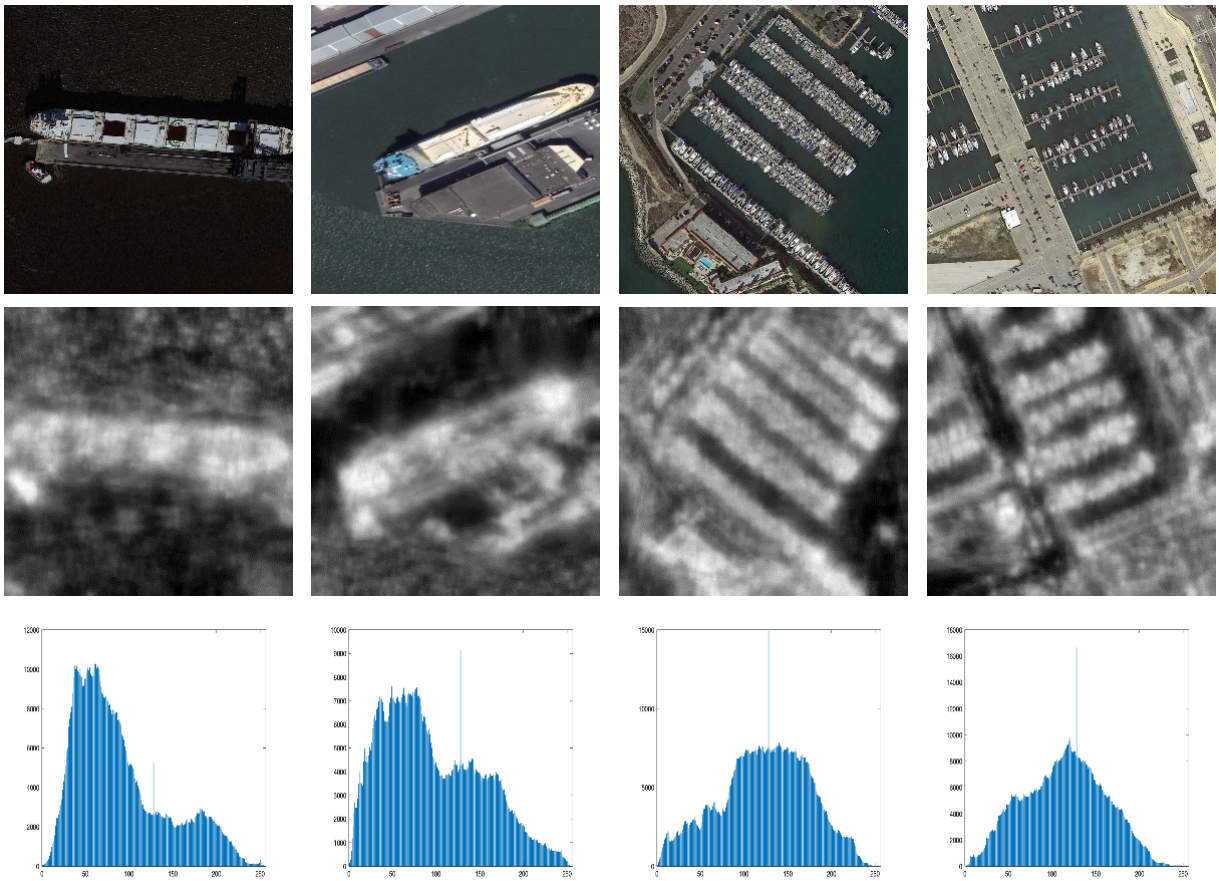


FIGURE 7. The statistic histogram of saliency feature maps of SIM on the Google Earth and the DOTA dataset. the first row represents the images in dataset, second row presents the saliency feature maps, and the third row is statistic histograms.

boundaries of the right peak in histogram. T is the threshold to differentiate the ship scale. The image is judged as a scene containing large ships when $f_s = 1$. And in the experiments, we adopt the parameters of $L_l = 30$, $L_u = 100$, $R_l = 130$, $R_u = 170$, $T = 2000$. In addition, these parameters are determined in DOTA and Google Earth datasets, and this parameter configuration should be changed with different datasets.

Based on the above discussion, the SIM, which extracts the saliency feature maps and obtains the scale information of image, is used to deal with the input data in ship detection progress, and its structure has been shown in Fig. 2.

IV. EXPERIMENTAL RESULTS

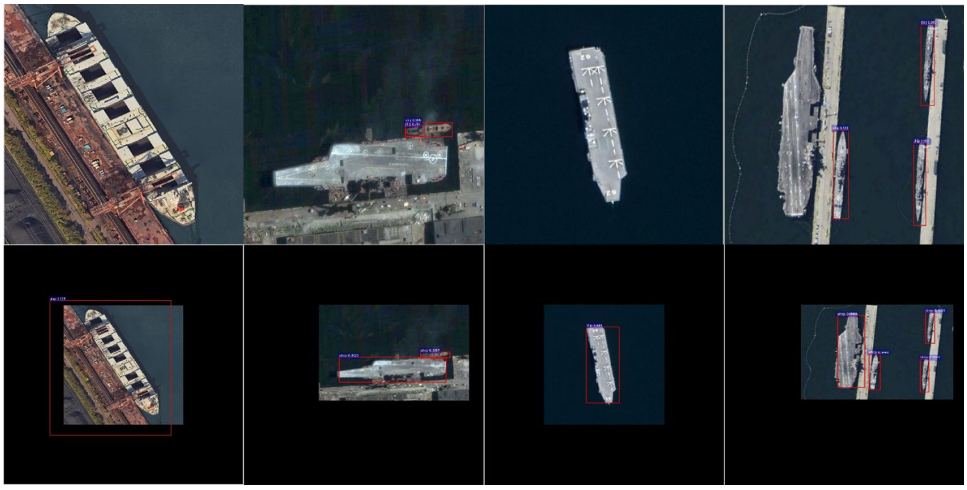
In this section, the performance of SEA on multi-scale ship object detection is introduced. And the model used in experimental is constructed on Tensorflow and trained on GeForce GTX 1080 Ti GPU.

Then, the Faster R-CNN, FPN and YoloV3 are used to verify the effectiveness because these methods are famous and relatively effective in tow-stage and one-stage methods. And in order to ensure the consistency of training, we set the hyper-parameters of different networks to the same: the input size is 1024, the learning rate is 0.001, the number of iterations is 80000, the anchor scale is [32,64,128,256,512] and the anchor ratio is [0.5,1,2] except for YoloV3. Furthermore, AP^L is adopted to present the performance of object detection algorithms on the Google Earth dataset. The results are shown in TABLE 3.

TABLE 3. The *APL* results of SEA-based object detection algorithms and Faster R-CNN.

-	Faster R-CNN	FPN	YoloV3
None	**0.664**	**0.642**	**0.61**
GR	0.677	0.649	0.623
DSR	0.669	0.645	0.621
PCA	0.671	0.647	0.617
MC	0.667	0.642	0.615
COV	0.684	0.652	0.63
SWD	0.69	0.65	0.638
FES	0.668	0.645	0.614
SIM	**0.727**	**0.685**	**0.673**
SEG	0.664	0.642	0.61
CA	0.697	0.651	0.643
SeR	0.708	0.682	0.652
SUN	0.702	0.676	0.648
SR	0.715	0.68	0.657
DHSNet	0.664	0.642	0.61
PoolNet	0.664	0.642	0.61

Furthermore, it should be noted that the AP^L of Faster R-CNN on our previous paper [27] is only 0.386. It's because



(a)

(b)

(c)

FIGURE 8. The results of different baselines and comparison results combined with our method. The baseline of (a) is Faster R-CNN, the baseline of (b) is FPN and the baseline of (c) is Yolov3.

some more detailed data preparation has been done during the experiments of this paper. And only the performance of the Google Earth dataset is shown in this table because there are many small-scale ships in the DOTA dataset, which

cannot be affected by our method. We observed that, in TABLE 3, the index value of the SIM-based Faster R-CNN, SIM-based YoloV3 and SIM-based FPN is 0.727, 0.673 and 0.685, respectively. These are the highest values among all SEA-based detection algorithms. At the same time, the index value of SEG-based Faster R-CNN, YoloV3 and FPN is only 0.664, 0.61 and 0.642 respectively, this is a strong proof for the effectiveness of our framework. The results of all the classic SEA-based algorithms are higher than the results of the original network. This proves that the object detection algorithms are limited by the imbalance training dataset, which reduce their detection ability on large-scale ships. In other words, it is effective to rescale the input image containing large-scale ships to improve their detection accuracy. In contrast, the DCNN-based SEAs cannot contribute to the detection performance, and it is limited by the incomplete features caused by the saliency extraction DL mechanism. These results effectively prove that some classic SEAs can apply in object detection of high-resolution remote sensing images. Some detection results of different DCNNs are shown in Fig. 8.

Furthermore, the efficiency of SEAs is still a problem with this detection framework. As we all know, the traditional SEA is a time-consuming operation, because it has to execute some pixel-level operations on the image, and the time consumption increases with the image scale In our detection framework, the process time of each input image in the SEA step is around one second, and this is a heavy price compared to the 0.2s of the detection step. Thus, this detection framework should consider the balance of time-consuming and accuracy when it is used in actual scenes. In conclusion, it effectively alleviates the problem proposed above, and it confirms the effectiveness of the saliency estimation algorithm on high-resolution remote sensing images.

V. CONCLUSION

To alleviate the problem of multi-scale ship detection in high-resolution remote sensing images. We propose a new method, combining the SEAs and the DCNN object detection, to ensure the extraction of large-scale ships. The SEA is used to represent the scale information of ships. Then the image pyramid is established to rescale those input images with large-scale ships so that the size of all ships are matching the training dataset, so as to get a better detection performance. Furthermore, the effectiveness of different SEAs in extracting the saliency feature map has been verified and analyzed in this paper. 15 SEAs, containing 13 classic SEAs and 2 DL-based SEAs, are used to conduct comparative experiments on both DOTA and Google Earth datasets. It can be seen from the saliency feature maps obtained by 15 SEAs, the DL-based SEAs can get the accurate feature of ship object, however, it always appears the incomplete features on the Google Earth dataset, caused by imbalance training samples. Therefore, the existing DL-based SEAs are not suitable for our detection framework. On the contrary, the classic SEAs can break the limitation of the supervised learning because these methods are realized by manual feature. One of the best is the SIM algorithm, it has an excellent performance on both datasets. The experiment result proves that our detection method, combining SIM algorithm and DCNN-based detection method, can effectively improve the performance of multi-scale ship detection.

ACKNOWLEDGMENT

The authors would like to thank the Data Science Center of Beijing University of Posts and Telecommunications for providing experimental equipment.

REFERENCES

[1] U. Kanjir, H. Greidanus, and K. Oštir, "Vessel detection and classification from spaceborne optical images: A literature survey," *Remote Sens. Environ.*, vol. 207, pp. 1–26, Mar. 2018.
[2] Y. You, Z. Li, B. Ran, J. Cao, S. Lv, and F. Liu, "Broad area target search system for ship detection via deep convolutional neural network," *Remote Sens.*, vol. 11, no. 17, p. 1965, Aug. 2019.
[3] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 580–587.
[4] K. He, X. Zhang, S. Ren, and J. Sun, "Spatial pyramid pooling in deep convolutional networks for visual recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 9, pp. 1904–1916, Sep. 2015.
[5] R. Girshick, "Fast R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2015, pp. 1440–1448.
[6] 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.
[7] 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.*, 2016, pp. 21–37.
[8] C.-Y. Fu, W. Liu, A. Ranga, A. Tyagi, and A. C. Berg, "DSSD: Deconvolutional single shot detector," 2017, *arXiv1701.06659*. [Online]. Available: https://arxiv.org/abs/1701.06659
[9] 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.*, Jun. 2016, pp. 779–788.
[10] J. Redmon and A. Farhadi, "YOLO9000: Better, faster, stronger," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2017, pp. 7263–7271.
[11] J. Redmon and A. Farhadi, "YOLOv3: An incremental improvement," 2018, *arXiv:1804.02767*. [Online]. Available: http://arxiv.org/abs/1804.02767
[12] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2017, pp. 2117–2125.
[13] B. Singh and L. S. Davis, "An analysis of scale invariance in object detection–SNIP," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 3578–3587.
[14] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis.*, Oct. 2017, pp. 2961–2969.
[15] Y. Jiang, X. Zhu, X. Wang, S. Yang, W. Li, H. Wang, P. Fu, and Z. Luo, "R2CNN: Rotational region CNN for orientation robust scene text detection," 2017, *arXiv:1706.09579*. [Online]. Available: http://arxiv.org/abs/1706.09579
[16] X. Yang, H. Sun, X. Sun, M. Yan, Z. Guo, and K. Fu, "Position detection and direction prediction for arbitrary-oriented ships via multitask rotation region convolutional neural network," *IEEE Access*, vol. 6, pp. 50839–50849, 2018.
[17] R. Zhang, J. Yao, K. Zhang, C. Feng, and J. Zhang, "S-CNN-based ship detection from high-resolution remote sensing images," *Int. Arch. Photogramm. Remote Sens. Spat. Inf. Sci.*, vol. 41, pp. 423–430, Jul. 2016.
[18] C. Lavalle, C. R. Gomes, C. Baranzelli, and F. E. B. Silva, "Coastal zones policy alternatives impacts on European coastal zones 2000–2050," JRC, Hong Kong, Tech. Note 64456, 2011.
[19] Y. You, J. Cao, Y. Zhang, F. Liu, and W. Zhou, "Nearshore ship detection on high-resolution remote sensing image via scene-mask R-CNN," *IEEE Access*, vol. 7, pp. 128431–128444, 2019.

[20] 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, doi: 10.1109/TGRS.2016.2601622.
[21] G. Cheng, J. Han, P. Zhou, and D. Xu, "Learning rotation-invariant and Fisher discriminative convolutional neural networks for object detection," *IEEE Trans. Image Process.*, vol. 28, no. 1, pp. 265–278, Jan. 2019, doi: 10.1109/TIP.2018.2867198.
[22] C. Chen, W. Gong, Y. Hu, Y. Chen, and Y. Ding, "Learning oriented region-based convolutional neural networks for building detection in satellite remote sensing images," *ISPRS-Int. Arch. Photogramm., Remote Sens. Spatial Inf. Sci.*, vols. 42-1/W1, pp. 461–464, May 2017, doi: 10.5194/isprs-archives-XLII-1-W1-461-2017.
[23] K. Li, G. Cheng, S. Bu, and X. You, "Rotation-insensitive and context-augmented object detection in remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 4, pp. 2337–2348, Apr. 2018.
[24] 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.
[25] G.-S. Xia, X. Bai, J. Ding, Z. Zhu, S. Belongie, J. Luo, M. Datcu, M. Pelillo, and L. Zhang, "DOTA: A large-scale dataset for object detection in aerial images," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 3974–3983.
[26] K. Li, G. Wan, G. Cheng, L. Meng, and J. Han, "Object detection in optical remote sensing images: A survey and a new benchmark," *ISPRS J. Photogramm. Remote Sens.*, vol. 159, pp. 296–307, Jan. 2020.
[27] Z. Li, Y. You, and F. Liu, "Multi-scale ships detection in high-resolution remote sensing image via saliency-based region convolutional neural network," in *Proc. IEEE Int. Geosci. Remote Sens. Symp. (IGARSS)*, Jul. 2019, pp. 246–249.
[28] A. Borji, M.-M. Cheng, Q. Hou, H. Jiang, and J. Li, "Salient object detection: A survey," *Comput. Vis. Media*, pp. 1–34, Jun. 2019.
[29] E. Rahtu, J. Kannala, M. Salo, and J. Heikkilä, "Segmenting salient objects from images and videos," in *Proc. Eur. Conf. Comput. Vis.*, 2010, pp. 366–379.
[30] S. Goferman, L. Zelnik-Manor, and A. Tal, "Context-aware saliency detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 10, pp. 1915–1926, Oct. 2012.
[31] L. Duan, C. Wu, J. Miao, L. Qing, and Y. Fu, "Visual saliency detection by spatially weighted dissimilarity," in *Proc. CVPR*, Jun. 2011, pp. 473–480.
[32] H. R. Tavakoli, E. Rahtu, and J. Heikkilä, "Fast and efficient saliency detection using sparse sampling and kernel density estimation," in *Proc. Scandin. Conf. Image Anal.*, 2011, pp. 666–675.
[33] C. Yang, L. Zhang, and H. Lu, "Graph-regularized saliency detection with convex-hull-based center prior," *IEEE Signal Process. Lett.*, vol. 20, no. 7, pp. 637–640, Apr. 2013.
[34] X. Li, H. Lu, L. Zhang, X. Ruan, and M.-H. Yang, "Saliency detection via dense and sparse reconstruction," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2013, pp. 2976–2983.
[35] R. Margolin, A. Tal, and L. Zelnik-Manor, "What makes a patch distinct?" in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2013, pp. 1139–1146.
[36] B. Jiang, L. Zhang, H. Lu, C. Yang, and M.-H. Yang, "Saliency detection via absorbing Markov chain," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2013, pp. 1665–1672.
[37] X. Hou and L. Zhang, "Saliency detection: A spectral residual approach," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2007, pp. 1–8.
[38] L. Zhang, M. H. Tong, T. K. Marks, H. Shan, and G. W. Cottrell, "SUN: A Bayesian framework for saliency using natural statistics," *J. Vis.*, vol. 8, no. 7, p. 32, Dec. 2008.
[39] H. J. Seo and P. Milanfar, "Static and space-time visual saliency detection by self-resemblance," *J. Vis.*, vol. 9, no. 12, p. 15, 2009.
[40] N. Murray, M. Vanrell, X. Otazu, and C. A. Parraga, "Saliency estimation using a non-parametric low-level vision model," in *Proc. CVPR*, Jun. 2011, pp. 433–440.
[41] E. Erdem and A. Erdem, "Visual saliency estimation by nonlinearly integrating features using region covariances," *J. Vis.*, vol. 13, no. 4, p. 11, 2013.
[42] J.-J. Liu, Q. Hou, M.-M. Cheng, J. Feng, and J. Jiang, "A simple pooling-based design for real-time salient object detection," 2019, *arXiv:1904.09569*. [Online]. Available: http://arxiv.org/abs/1904.09569
[43] N. Liu and J. Han, "DHSNet: Deep hierarchical saliency network for salient object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 678–686.

ZEZHONG LI (Student Member, IEEE) is currently pursuing the master's degree with the Beijing University of Posts and Telecommunications, China. His research interests include deep learning and ship detection on remote sensing optical images.

YANAN YOU (Member, IEEE) received the Ph.D. degree from the School of Electronic and Information Engineering, Beihang University, China, in 2015. From 2015 to 2017, he held a postdoctoral position with Beihang University. Since September 2017, he has been a Lecturer with the School of Information and Communication Engineering, Beijing University of Posts and Telecommunications. His research interests include remote sensing image processing, SAR interferometry processing, deep learning, big data technology, and so on.

FANG LIU received the Ph.D. degree from Nankai University, Tianjin, China, in 1997. She is currently an Associate Professor with the School of Information and Communication Engineering, Beijing University of Posts and Telecommunications, Beijing, China. Her research interests include broadband IP networks, network traffic monitoring, machine learning, and data mining.