This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCSVT.2019.2897980, IEEE Transactions on Circuits and Systems for Video Technology

Saliency-Aware Convolution Neural Network for Ship Detection in Surveillance Video

Zhenfeng Shao, Linggang Wang*, Zhongyuan Wang*, Wan Du, and Wenjing Wu

***Abstract*— Real-time detection of inshore ships plays an essential role in the efficient monitoring and management of maritime traffic and transportation for port management. Current ship detection methods which are mainly based on remote sensing images or radar images hardly meet real-time requirement due to timeliness of image acquisition. In this paper, we propose to use visual images captured by an on-land surveillance camera network to achieve real-time detection. However, due to the complex background of visual images and the diversity of ship categories, the existing convolution neural network (CNN) based methods are either inaccurate or slow. To achieve high detection accuracy and real-time performance simultaneously, we propose a saliency-aware CNN framework for ship detection, comprising comprehensive ship discriminative features, such as deep feature, saliency map and coastline prior. This model uses CNN to predict the category and the position of ships, and uses the global contrast based salient region detection to correct the location. We also extract coastline information and respectively incorporate it into CNN and saliency detection to obtain more accurate ship locations. We implement our model on Darknet under CUDA 8.0 and CUDNN V5 and use a real-world visual image dataset for training and evaluation. The experimental results show that our model outperforms representative counterparts (Faster R-CNN, SSD, and YOLOv2) in terms of accuracy and speed.**

***Index Terms*—Ship Detection, Saliency Detection, Coastline Extraction, Object Location, CNN.**

I. Introduction

Ship detection is of great value in many application fields, such as ocean surveillance, port management, and navigation safety. In the field of port management, ship detection can monitor and assist in the management of maritime traffic and transportation. As for marine surveillance, ship detection plays a strong supervisory role in fisheries dumping of pollutants and illegal smuggling. In the navigation safety, ship detection can judge whether there are abnormal sailing behaviors such as landing or stagnation, so as to ensure the safety on the coast and at sea. Real-time detection of ships is also very important and has the ability to proactively alert. The real-time detection system can be connected with other systems, especially the emergency dispatch system, which is helpful for responding to abnormal behaviors and emergencies in time to avoid possible adverse consequences. It can also be integrated with space-time systems to process and analyze previous surveillance videos in real time and make timely decisions. According to the image generation source, images based ship target detection methods are roughly classified into three categories: radar images [1], remote sensing images, and visual images. The acquisition and preprocessing of radar images and remote sensing images always takes time and cannot be detected in real time. Compared with other categories, visual images are generally obtained more easily from continuous monitoring video, and so they can be used as real-time detection. However, because the background of visual images is more complicated and less clear, there exists severe interference for foreground detection. Therefore, accurate ship object detection from surveillance video faces huge challenges.

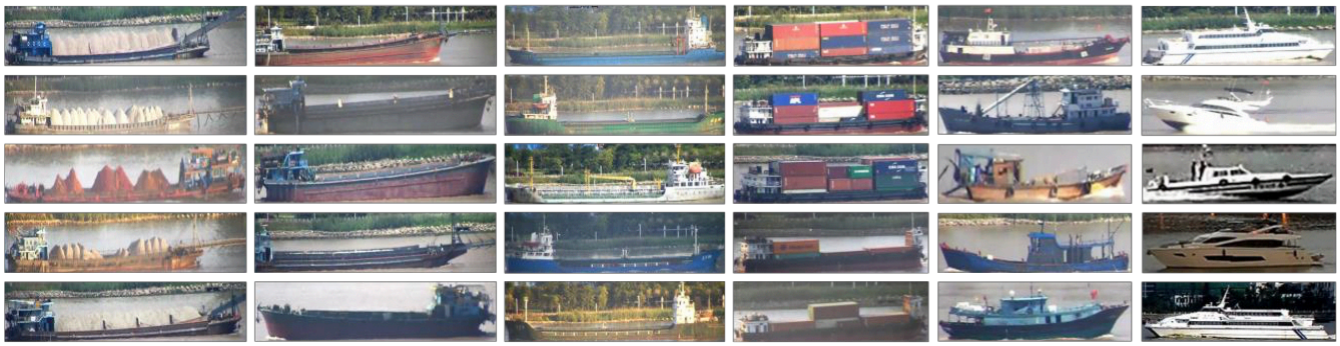
There appear some ship detection methods based on visual images [2]-[5]. They usually use ship features, such as the contextual information of the image, the temporal-spatial information of the ship, and the geographical environment prior (e.g., coastline). In recent few years, convolutional neural network (CNN) has achieved great success in natural image classification [6]-[8] and object detection [9]-[15]. In contrast to traditional methods using manual pre-defined features, CNN based methods are able to automatically represent and extract discriminative and robust features for object detection. However, there are special difficulties for ship detection task in the marine environment. First, due to waves and floating objects, the background is very complicated so that ships are easily mixed with them and even visually overlap the nearshore buildings. Second, the ships vary in categories and sizes, which range from dozens to hundreds pixels in size and may cross or occlude with each other. Finally, because marine climate and lighting conditions are variable, low visibility weather such as clouds and fog often degrade the acquired video quality. Therefore, CNN

This work was supported in part by the National key R & D plan on strategic international scientific and technological innovation cooperation special project under Grants 2016YFE0202300, the National Natural Science Foundation of China under Grants 61671332, 41771452, and 41771454, Guangzhou Science and Technology Project under Grant 201604020070, and the Key Research and Development Program of Hubei Province of China under Grant 2016AAA018.

Z. Shao, L. Wang, and W. Wu are with the State Key Laboratory for Information Engineering in Surveying, Mapping and Remote Sensing, Wuhan University, Wuhan 430079, China. (e-mail: shaozhenfeng@whu.edu.cn, wanglinggang95@163.com, wuwenjing94@163.com).

Z. Wang is with the National Engineering Research Center for Multimedia Software, Wuhan University, Wuhan 430072, China. (e-mail: wzy_hope@163.com).

W. Du is with Computer Science and Engineering, University of California, Merced. (e-mail: wdu3@ucm.edu).



(a) ore carrier (b) bulk cargo carrier (c) general cargo ship (d) container ship (e) fishing boat (f) passenger ship

Fig. 1. The ship category in our used ship dataset.

approaches to regular object detection often fail to provide satisfactory ship detection performance.

Recently, some researchers have used CNN for ship detection in remote sensing images. Li *et al*. [16] pioneers to introduce CNN into ship detection of remote sensing images. They propose a novel parallelogram image cropping (PIC) method to generate parallelogram samples, each of which only contains single ship or dock. Lin *et al*. [17] propose to divide the detection among the network layers at different depths to combine the advantages of deep network used for location and shallow network used for detecting. However, different from the top-down perspective of the remote sensing images, most of the visual images are in frontal perspectives. In addition, for detection task from the continuous video, they must sacrifice accuracy to guarantee real-time performance.

To address the ship detection under seashore surveillance video conditions, we develop a novel saliency-aware CNN framework, which is built on the YOLOv2 pipeline. Essentially, our proposed ship detection model follows a classification and localization driven coarse-to-fine idea. It first uses CNN to predict the ship class and the rough position. However, because the onshore buildings and ships are similar in color and appearance, YOLOv2 often judges ashore buildings either as the target ships or as part of the nearshore ships, which results in detection confusion, including false detection (false positive) or inaccurate localization and size of the actual ship region. In a further examination, inaccurate location and size will reduce the confidence score of bounding box, which in turn leads to missed ships (i.e., low recall rate) because YOLOv2 tends to discard low scoring candidates. To this end, considering that the coast surveillance image contains sea areas and land areas, but ships only appear in the sea area, we extract coastline features and incorporate them into CNN to improve the robustness and efficiency of the ship detection. More specifically, only cells (by YOLOv2) at sea are produced and the classification is further examined, but excluding onshore cells. Furthermore, since ships differ much from water in terms of visual saliency, we further incorporate saliency detection technique to refine a more accurate location. Owning to the improved localization, the missed ship suffering from low confidence score is accordingly recalled. At last, as for ship detection in continuous video, because the position of the ship in the video frame is spatially coherent, we use temporal continuity to set the initial observation position of each frame instead of re-traversing the entire video frame. Extensive validations on real-world coast surveillance video datasets (as shown in Fig. 1) from Hengqin Island in Zhuhai in China show the proposed model's capability in terms of detection accuracy and speed [18].

The main contributions of this paper are highlighted as follows:

1) To the best of our knowledge, we are the first to introduce CNN into ship detection in surveillance video.

2) Based on the YOLOv2 pipeline, we propose a saliency-aware CNN framework to improve the accuracy and robustness of ship detection under complex seashore surveillance conditions, where the ship's category and location are first predicted by CNN and then are refined with saliency detection.

3) We propose coastline segmentation to reduce the inspection range and further improve the detection efficiency.

The rest of the paper is organized as follows. In Section II, we introduce the related work of ship detection. In Section III, we give detailed explanations of our proposed model. Section IV illustrates experimental results and comparisons against other state-of-the-art methods. Section V draws a conclusion.

II. RELATED WORK

A. Ship Detection

Some methods using hand-crafted features are widely studied for ship detection. W. Krüger *et al*. [2] first use color segmentation and edge detection to detect the sea level feature, and then use the image registration and subtraction to separate the ship from the water. Bao *et al.* [3] detect ship with the contextual information and the ship space-time information. They manually determine the mean and variance threshold for each category, and then based on contextual information analyze the regional-level movement of the ship and its corresponding local context for detection. Chen *et al*. [4] proposed a new method based on mean shift and the peak of grayscale for ship automatic detection and tracking, but it needs to be optimized in real time. Zhang *et al.* [5] detect the horizon line by exploiting the characteristics of discrete cosine transform (DCT) blocks and extract the sea-surface background regions below the horizon. They simply remove the background to obtain ship targets and the results are unreliable. These features leverage human attention mechanisms to obtain the saliency of the ship in the entire image with respective features designed for different ship conditions, but they are not suitable for detecting diverse

category of ships in our scenario.

B. Convolutional Neural Network

Convolutional neural network (CNN) has been successfully applied to object detection [9]-[15]. In recent years, the development of deep learning has been driven by the regional proposal method and the regional proposal-based CNN (R-CNN) [9]. R-CNN is the first network to use the CNN feature for classification. In order to improve efficiency, R. Girshick further proposed Fast R-CNN based on R-CNN [10]. Fast R-CNN maps the proposal region to the feature map of the last convolutional layer of CNN. In this way, an image needs to be extracted only once which greatly increases speed. Based on Fast R-CNN, R. Girshick also proposed the Faster R-CNN [11], which is composed of Region Proposal Network (RPN) and Fast R-CNN. Two models share the features and the RPN module tells the Fast R-CNN module where to look.

The accuracy of the R-CNN framework is getting higher and higher, especially the Faster R-CNN. The bottleneck of the R-CNN framework is that it cannot fully utilize the context information of the local object in the entire image after transforming the decomposition problem into the classification problem of the image local area. Therefore, J. Redmon and R. Girshick [15] proposed the YOLO (You Only Look Once) network together. The idea is handling object detection problem as regression problem, separating object locations and categories from space. The detection speed of the network is very fast and can achieve real-time video processing, but the accuracy is not high enough. J. Redmon [19] used a series of methods to improve YOLO and proposed YOLOv2, which improves the accuracy and maintains the speed. However, the accuracy still cannot meet the requirements. Qi [20] proposed a novel paradigm of deep network to explore various scales of spatial contexts adjusted to pixels at different locations. This model constructs multiple layers of memory cells, whose outputs are hierarchically gated on different scales before recursively feeding to higher layers. Then the pixel labels at different locations are decided based on the spatial contexts of the customized scales. Compared with the general object detection algorithm, it can make full use of the context information to obtain a good pixel-level detection effect. Nevertheless, the required dataset must be fully labeled on individual pixel level, which does not meet the situation of our ship dataset. In the view of the above discussion, we follow the YOLOv2 framework to construct our ship detection pipeline.

C. Saliency Detection

Salient object detection can help people quickly locate target region of interest in an image. It has been widely used as a preprocess step in many computer vision tasks such as super-pixel segmentation [21]-[24], object recognition [25]-[27], image retrieval [28]-[30], etc.

Inspired by these works, many researchers began to harness saliency information in ship detection. Bi [31] extracted salient candidate regions across the entire detection scene using a bottom-up visual attention mechanism. Then appearance and neighborhood similarity features are combined to discriminate the selected salient regions. Jiang [32] used the salient corner features at ship bow to precisely detect in-shore ships and separate multiple docked targets. Lin [33] implemented a task partitioning model similar to the attention model in FCN [34] network. With deep path for attention/saliency maps and the shallow path for detection, the integrated FCN can detect ships robustly and simply.

The above works have all proved that adding salient information to ship detection problem can effectively improve the detection performance in optical remote sensing images. However, few studies have introduced this idea into ship detection in natural images. Walther [35] proposed a biologically plausible model for forming and attending proto-objects in natural scenes. But this method is hardly generalized to other computer vision tasks, such as image segmentation and object detection. Achanta [36] adopted a frequency-tuned approach to compute full resolution saliency maps with well-defined boundaries, which uses an image-dependent adaptive threshold to binarize the generated saliency map. Rahtu [37] firstly generated saliency maps using a statistical framework and local feature contrast in illumination, color, and motion information, and then segmented the salient object with a conditional random field. Goferman [38] detected context-aware saliency maps based on four principles observed in the psychological literatures. The approach was evaluated in two applications where the context of the dominant objects is just as essential as the objects themselves. Cheng [39] proposed a method that considers both appearance similarity and spatial distribution of image pixels, which produces perceptually accurate salient region detection. For more detailed literature discussion of some state-of-the-art saliency detection algorithms, we refer readers to [40] and [41].

In this paper, we adopted a regional contrast-based saliency extraction algorithm [42] which simultaneously evaluates global contrast differences and spatial coherence. We compared the improvements of this algorithm with previously mentioned methods [36]-[39] on the final ship detection results. And experimental results showed that regional contrast-based method can achieve better recall and precision rates and more accurate location.

III. Proposed Method

Our proposed model is shown in Fig. 2. Our model mainly consists of CNN and saliency detection, both combined with coastline features. The former is used to predict the category and preliminary position of the ship, and the latter is used to determine the exact position. The model first resizes the input image to a fixed size and passes it to CNN to extract feature maps, where CNN learns from extracted coastline features to exclude onshore feature maps. On the final feature map, we examine the spatial relationship between each cell and the coastline, and only generate a bounding box for the sea part. Then, the corresponding grids of the remaining feature maps

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCSVT.2019.2897980, IEEE Transactions on Circuits and Systems for Video Technology

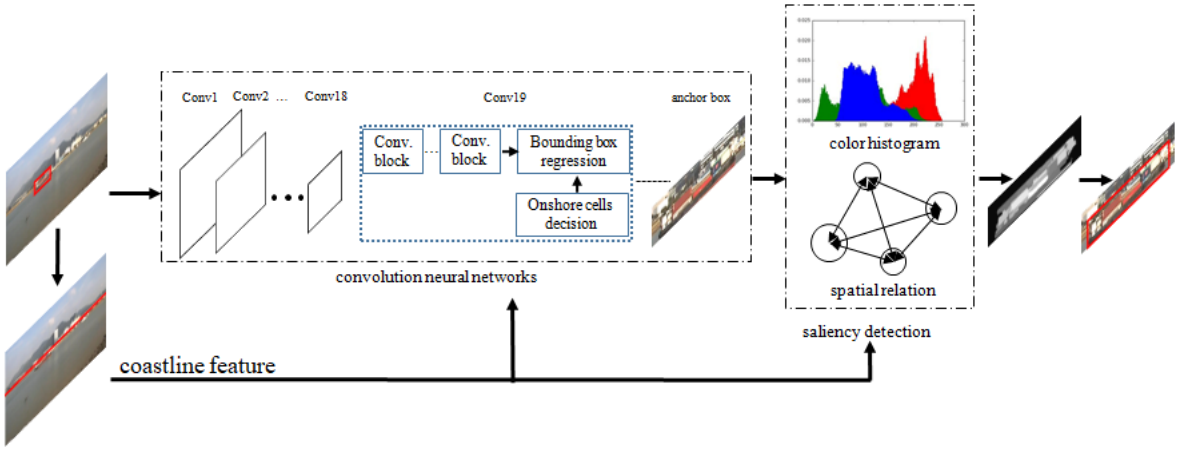
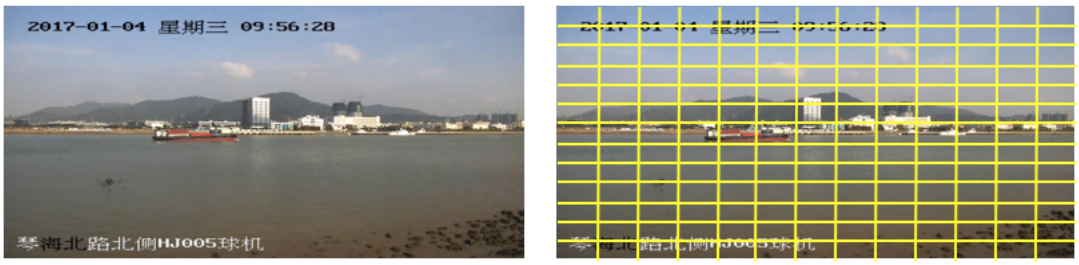


Fig. 2. The proposed ship detection pipeline. We first input the image into a convolutional neural network and generate anchor boxes combined with coastline feature. Then we use saliency detection which uses the spatial relationship and color space to produce more accurate ship location in combination with coastline feature.



(a) (b)

Fig. 3. Feature map of whole image. (a) The original image. (b) The 13x13 feature map.

generate several anchor boxes of different sizes and output the most likely box category and the position to be corrected. Finally, we use the salient region detection of the proposals to rectify the ship localization generated by CNN. Because the color information of the house is also very salient, we attempt to combine coastline with saliency to reduce the impact of onshore houses on saliency detection. We intersect the coastline with the detection box and only perform saliency detection on the offshore part. The detection result takes the outer rectangle as the position of the ship.

In the practical application of ship detection in video, we can further leverage temporal continuity to set the initial inspection position for individual frames. We estimate the ship's moving displacement between adjacent frames based on the speed, heading, and interframe interval and then determine the approximate location of the ship in the next frame.

A. YOLOv2 Based Classification

YOLOv2 integrates bounding box generation, feature extraction, target classification and target position into convolutional neural network. It directly extracts bounding boxes from the image, and predicts the position and probability of ship through the entire image feature. It converts ship detection problems into regression problems, which is truly end-to-end detection. So, we use YOLOv2 to predict classification of ship in images.

YOLOv2 resizes the input image to 416x416 and the downsampling rate is 32, so finally the feature map covers 13 × 13 cells. As shown in Fig. 3 (b), many cells are on the shore, which actually have nothing to do with the ship detection task. The coastline is a very useful feature that distinguishes between sea and land. Therefore, we can take advantages of coastline prior to exclude unnecessary generation of onshore cells, for reducing both computational burden and the interference of onshore buildings to ship detection.

1) Coastline Extraction

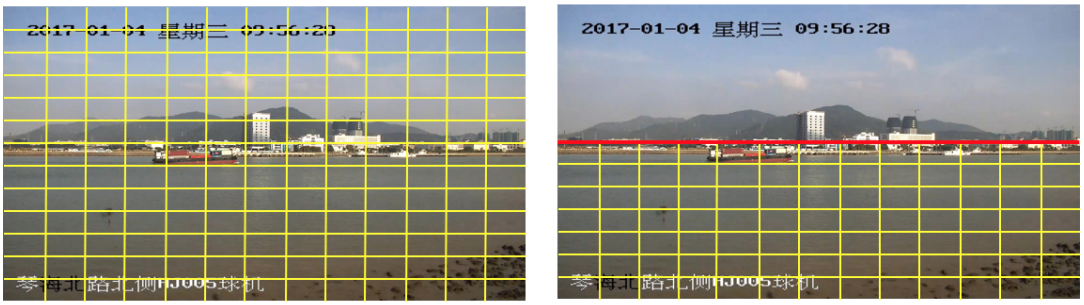
We first use the Canny operator to detect edges in the image. Then we use Hough transform to obtain all line segments based on the edges. Suffering from the complex coast background, Hough transform used for extracting line segments usually produces multiple segments rather than unique one. Considering that the generated line segments are mostly concentrated near the coastline and are relatively random, we need to figure out the accurate coastline. We find that the slope k in the coastline does not exceed 0.15 and the intercept b does not exceed 800 from the origin in the upper left corner of the image (1920x1080). Therefore, we calculate k and b of all generated line segments, and then exclude some outliers that impossibly form a reasonable coastline based on this observation.

Through the above steps, we have excluded the outliers, and then we are to fit a line segment (from the remaining segments)

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCSVT.2019.2897980, IEEE Transactions on Circuits and Systems for Video Technology

(a) (b)

Fig. 4. Extraction results of the coastline. (a) All detected segments. (b) The final coastline which is up 30 pixels.



(a) (b)

Fig. 5. Coastline feature in generating feature map. (a) The original feature map. (b) The feature map combined with coastline.

that is closest to the true coastline. Due to the difference in position and slope of the remaining segments, they unevenly contribute to the resulting coastline. To do this, we need to establish a judgment criterion to find the most valuable line segment to fit the final coastline. The position of the line is determined by the slope and intercept, which also make differences on the position, and so we consider the following discriminant function:

$$L_i = \lambda \times k_i^2 + (1-\lambda) \times b_i^2 \tag{1}$$

where i denotes the i-th segment, and L_i denotes the cost function. k_i and b_i represent the slope and intercept of the i-th segment after normalization, and λ indicates the weight parameter.

Then we arrange the line segments from small to large according to the cost function. Because too small values in cost function have a negative impact on the resulted coastline, we only use the first fraction of ρ as effective line segments for generating the coastline. We set the average k and b of these line segments as the slope and intercept of the extracted coastline. We will move the coastline up 30 pixels, considering that the ship may intersect the coastline. As a concrete example, the coastline extraction results are shown in Fig. 4.

2) Bounding Box Regression with Coastline

After the coastline is extracted, we input the coastline feature into CNN together with the image, which are jointly used to conduct bounding box regression and classification by YOLOv2.During the classification, both in training and testing, coastline features can assist YOLOv2's decision-making to reduce detection time and improve accuracy. Based on coastline feature, we use Eq. (2) to determine whether the cell is on the shore or not. If the cell satisfies Eq. (2), we mark it as ashore cell for which we do not generate bounding boxes to reduce the wrong ship classification, like the shown example in Fig. 5.

$$\begin{cases} \dot{k} \times i - \dot{b} - j \geq 0 \\ \dot{k} \times (i+1) - \dot{b} - j \geq 0 \end{cases} \tag{2}$$

where $\dot{k}$ and $\dot{b}$ denote the coastline parameters (slope and intercept) transformed into the feature map. i and j are the numbers of the cells, from 1 to 13. We need to resize the coastline along with the image and get the coastline parameters on the feature map. Since the final feature map size is 13x13, the transformation of k and b obeys

$$\begin{cases} \dot{k} = \frac{w}{h} * k \\ \dot{b} = \frac{13}{h} * b \end{cases} \tag{3}$$

k and b refer to the slope and intercept of coastline in original image and w and h denote the width and height of original image, respectively.

During training stage, each of cells on the sea predicts 5 detected bounding boxes with their confidence score for containing a specific category of ship. In order to get better and more representative prior boxes, YOLOv2 uses the IoU-based K-means clustering method to train bounding boxes, which can automatically find the width and height of 5 boxes more properly.

The IoU represents the overlap ratio of the resulted bounding box to the ground truth.

$$IoU = \frac{area(BB_{dt} \cap BB_{gt})}{area(BB_{dt} \cup BB_{gt})} \tag{4}$$

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCSVT.2019.2897980, IEEE Transactions on Circuits and Systems for Video Technology

(a) (b) (c)

Fig. 6. Examples on bounding boxes of the missed ships. (a) The bounding box is larger than the ship (the IoU is 0.466). (b)The bounding box contains a portion of the ship (the IoU is 0.397). (c) The bounding box is smaller than the ship (the IoU is 0.462).

(a) (b)

Fig. 7. Salient detection result of an image. (a) The original image. (b) The salient result of the entire image.

BB_{gt} denotes the ground truth of box, BB_{dt} denotes the detected bounding box and $area(\cdot)$ represents the area operator.

During testing stage, each of cells directly generates 5 bounding boxes according to the model. YOLOv2 treats the bounding box with a confidence probability above a certain threshold as the valid detection. The boxes which do not satisfy the threshold are directly discarded. Because several regressed bounding boxes may correspond to the same target, NMS (Non-Maximum Suppression) [43] is further used to find the most suitable bounding box.

The labeled bounding boxes are used for ground truth in training. But unlike bounding boxes, the coastlines are not used as ground truth during the training of YOLOv2. They are thus only extracted online by our proposed method, both in training and testing phases.

B. Salient Region Detection based Location

Suffering from complex onshore backgrounds (e.g., similar color and appearance between ashore buildings and nearshore ships), YOLOv2 often judges the ashore building either as the target ship or as part of the nearshore ship. The former leads to false detection (false positive) while the latter results in inaccurate localization and size of the actual ship region. In other words, for the latter, the resulting bounding box does not match the actual area of the ship, which further leads to missing detection due to the scoring mechanism of YOLOv2.

The rule for YOLOv2 to determine whether the bounding box is available according to its confidence score. When the confidence score of the bounding box is larger than the threshold (typically 0.24), YOLOv2 then calculates its IoU value. Only those with IoU greater than the IoU threshold (usually 0.5) are considered usable. In other words, when the bounding box is considered unavailable, its corresponding object will be missed. Experimentally, we do observe that the missing detection occurs due to the low IoU, with some examples shown in Fig. 6. In Fig. 6, some of the bounding boxes are much larger than the ship (with the IoU value being 0.466), others are smaller (with the IoU value being 0.462), and some contain only a portion of the ship (with the IoU value being 0.397). In all of these cases, their IoU values are less than 0.5, so they are abandoned. The reason for these problems is that the localization of YOLOv2 is not very accurate due to the interference of the complex ashore backgrounds. If we can improve the location, the missed ships will be detected correctly or the recall rate will be increased.

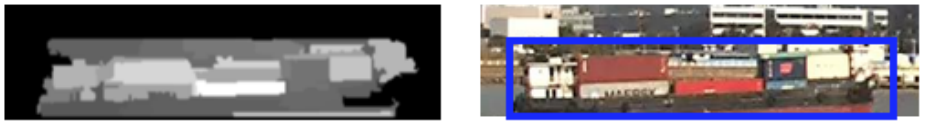
Salient object detection can help people quickly locate target region of interest in an image. So we thus combine the saliency features to rectify the preliminary location generated by YOLOv2. Due to the complexity of the image, we still get poor results when we input the entire image in saliency extraction, as shown in Fig. 7. Instead, we use the bounding boxes as the target detection range. Additionally, we do not intend to perform saliency detection on all boxes, but instead choose those that are expected to be corrected. If the IoU is too low, which means that the box intersects with the boat very little, we think this is a completely wrong box, without further correction value. Therefore, we set a low threshold for the IoU and choose those boxes whose IoU values fall between lower limit and 0.5 for further re-examination using saliency detection. With the help of preliminary location given by YOLOv2, we will expand the box slightly and adaptively to contain the complete ship as much as possible for the saliency detection in all conditions.

We perform salient region detection based on global contrast. The idea of using region based contrast (RC) [42] to produce saliency maps comes from a common sense that saliency of a region mainly depends on its contrast to its nearby regions. RC first segments the input image into regions

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCSVT.2019.2897980, IEEE Transactions on Circuits and Systems for Video Technology

(a) (b)

Fig. 8. Coastline feature in salient region. (a) The salient region combined with coastline. (b) The original salient region.



(a) (b)

Fig. 9. Saliency detection result of a bounding box. (a) The salient region. (b) The location of the ship.

(a) (b) (c)

Fig. 10 mproved bounding boxes of missed ships by saliency. (a) The IoU of the box is increased to 0.610. (b) The IoU of the box is increased to 0.533. (c) The IoU of the box is increased to 0.698.

through a graph-based super-pixel segmentation algorithm [44]. Then, the saliency value of each region r_k is computed as follows:

$$S(r_k) = w_s(r_k)\sum_{r_k \neq r_i} e^{\frac{D_S(r_k,r_i)}{-\sigma_s^2}} \omega(r_i) D_r(r_k,r_i) \quad (5)$$

where $w_s(r_k)$ is a spatial prior weighting term; $D_s(r_k,r_i)$ is the spatial Euclidean distance between centroids of two regions r_k and r_i, and σ_s adjusts the influence of spatial distance weights; $\omega(r_i)$ is the weight of region r_i which is measured by the number of pixels in r_i.

The color distance between two regions $D_r(r_k,r_i)$ is defined as follows:

$$D_r(r_1,r_2) = \sum_{i=1}^{n_1}\sum_{j=1}^{n_2} f(c_{1,i}) f(c_{2,j}) D(c_{1,i},c_{2,j}) \quad (6)$$

where $f(c_{k,i})$ is the probability of the i-th color $c_{k,i}$ among all n_k colors in the k-th region r_k, $k = \{1, 2\}$. It plays as a weighting role in the distance computation to emphasize the color differences between dominant colors.

In saliency detection, because the information of the ashore buildings is also salient, we manage to make use of the coastline features to reduce the impact of the houses, making the saliency area consistent with the real ship area. We process the bounding boxes which intersect the coastline and if the pixels satisfy Eq. (7), we mark them as ashore pixels and set them to 0.

$$k \times p_i - b - p_j \geq 0 \quad (7)$$

where k and b denote the coastline parameters (slope and intercept). p_i and p_j are the horizontal and vertical coordinates of the pixel, respectively. We detect the processed boxes and produce the result which is not affected by the buildings and more robust, as shown in Fig. 8. In Fig. 9, we then take the outer rectangle of the salient region as the ultimate location of the ship.

Finally, to especially confirm the role of saliency refinement, we further show the corresponding improved counterparts on bounding boxes in Fig. 6, as shown in Fig. 10. Accordingly, their IoU values are promoted to 0.610, 0.533 and 0.698 from original 0.466, 0.397 and 0.462, respectively. Since the improved IoU values are all above the IoU threshold 0.5, the bounding boxes will be considered valid and the associated ships will be recalled.

IV. EXPERIMENTAL RESULTS

To prove the effectiveness of our proposed method, we designed experiments and evaluated our method quantitatively on our own ship data set. Subjective and objective results are reported in this section.

A. Dataset

We use our own new large ship dataset called SeaShips. The dataset currently consists of 11126 images and covers 6 common ship categories (ore carrier, bulk cargo carrier, general cargo ship, container ship, fishing boat, and passenger ship). All the images are from about 5400 real-world video segments, which are acquired by 156 monitoring cameras in the coastline video surveillance system deployed on Hengqin

Island in Zhuhai in China. They are carefully selected to mostly cover all possible imaging variations, e.g., different scales, hull parts, illumination, viewpoints, backgrounds, and occlusions. All images are annotated with ship category labels and high-precision bounding boxes.

B. Test Environment

We conduct experiments based on learning platform Darknet in Windows 10. All our experiments are performed on a workstation with TitanX GPU cards under CUDA 8.0 and CUDNN V5. Our testing uses only one card.

Our network structure is modified from darknet19. To train the hyper-parameters, the mini-batch size is set to 16. The base learning rate is 0.0001 when iteration number is low than 20k and steps to 0.00001 when iteration number is low than 26k. The poly learning rate policy is adopted with power 0.9 together with the maximum iteration number 26k. Momentum is 0.9 and weight decay is 0.0001. Data augmentation contains random mirror and rand resizing is between 0.5 and 2.

During the coastline extraction, we use the grid search method to determine the final λ and ρ. In all candidate parameters, the best performing parameters are the final result by looping through each possibility of the parameters. λ is set to 0.1, 0.3, 0.5, 0.7, 0.9 and ρ is set to 1, 1/3, 1/5, 1/7, 1/9. The final results show that only when λ is equal to 0.3 and ρ is equal to 1/3, the coastline is completely correct. Because the average width of all bounding boxes of ships in training set is in 30 pixels, we move the coastline up 30 pixels.

C. Evaluation Indicators

There are some typical quantitative indicators for evaluating an object detection model, which are briefly described below.

1) Average Precision

Given an IoU threshold, there are two indicators called recall and precision. We manually marked the ground truth of ships, whose total number is defined as *NP*. If the bounding box has an IoU overlapping with the ground truth over 0.5, we mark these as true positive (*TP*). Each bounding box can only match one ground truth. Therefore, false detections of the same ground truth are defined as false positives (*FP*). So, recall and precision follow:

$$Recall = \frac{TP}{NP} \tag{8}$$

$$Precision = \frac{TP}{TP+FP} \tag{9}$$

For each category, we can draw a precision-recall curve according to recall and precision values. *AP* is the area surrounded by the curve.

$$\mathrm{AP} = \int_0^1 P(R)dR \tag{10}$$

2) Mean Average Precision

mAP denotes the average values of AP_i of each class i.

$$\mathrm{mAP} = \frac{\sum_{i=1}^{n} AP_i}{n} \tag{11}$$

where n is the number of classes that need to be detected.

3) Frame Per Second(FPS)

In addition to evaluate the accuracy, we also consider the model speed as one of the evaluation criteria. FPS indicates the number of image frames that models detect in one second. We use this indicator to measure the model speed.

D. Results and analysis

1) Comparison with Other Detection Methods

We compare our method with other detection methods, such as Fast R-CNN [10], Faster R-CNN [11], SSD [14], and YOLOv2 [19]. For the Fast R-CNN algorithm, we choose the VGG training detection model. For Faster R-CNN, we set a convolutional neural network that has been pre-trained on ImageNet as pre-trained model, then use ZF net [45] (5 convolutional layers and 3 fully connected layers) and VGG16 net [46] (13 convolutional layers and 3 fully connected layers) to retrain the detection model. For SSD, we use the MobileNet [47] and VGG16 net [46] to retrain the detection model. For YOLO v2, we use our pre-trained weights to retrain the detection model while using some common data enhancement methods to increase the amount of data and improve model robustness such as hue, saturation, and exposure shifts. Our own method uses the parameters described above for training. All experiments were performed on four Titan Xp. We recorded the results of each model according to the previous evaluation indicators, as shown in Table I. Fig. 11 shows AP performance for each ship with the IoU threshold set to 0.5.

As can be seen from Table I, Fast R-CNN is much worse in mAP performance than others by a large margin. The performance of the Faster R-CNN series is significantly better than YOLOv2 and SSD. On average, Faster R-CNN's mAP is 14.52% higher than YOLOv2 and 11.12% higher than SSD. Our approach significantly improves the performance of YOLOv2, narrows the gap with Faster R-CNN, and performs better than the Faster R-CNN on general cargo ships.

Our proposed model is based on YOLOv2, and the mAP of each category in our model has a good improvement. Among the six categories of ship, ore and container ships can achieve better results. Because these two categories of ships are mainly used to transport cargo such as ore and containers, and these goods have very distinct features that are distinguished from other ships. In contrast, the performance of fishing ships is worse than other categories. The main reason is that fishing boats are generally small, occupying only 70x130 pixels in a 1920x1080 image. Detectors usually have poor detection results of small targets. After many forward convolutions layers, the feature of the small targets becomes blurred, even worse in the YOLOv2 series.

We use saliency detection to increase 5% on the basis of YOLOv2. For fishing boats, we have increased from 73.3% to 78.3%, which almost achieves the mAP of the SSD method. Although there are still gaps compared with other categories, the results are still good. For the passenger ships, our method has increased by up to 10%. We hold that the color feature of the passenger ships is generally very salient, so the

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCSVT.2019.2897980, IEEE Transactions on Circuits and Systems for Video Technology

TABLE I
DETECTION RESULTS OF DIFFERENT DETECTORS ON THE SEASHIP DATASET

Model	mAP	ore carrier	bulk cargo carrier	general cargo ship	container ship	fishing boat	passenger ship	FPS (Titan Xp)
Fast[a](VGG)	0.710	0.771	0.713	0.771	0.868	0.617	0.522	0.5
Faster[b](ZF)	0.892	**0.905**	0.900	0.908	**0.909**	0.857	0.871	15
Faster[b](VGG)	**0.901**	0.894	**0.903**	0.907	**0.909**	**0.888**	**0.906**	6
SSD	0.794	0.750	0.767	0.877	0.907	0.718	0.744	7
YOLOv2	0.830	0.849	0.850	0.881	0.888	0.733	0.781	**83**
Ours	0.874	0.881	0.876	**0.917**	0.903	0.783	0.886	49

Fast[a] means Fast R-CNN. Faster[b] means Faster R-CNN.

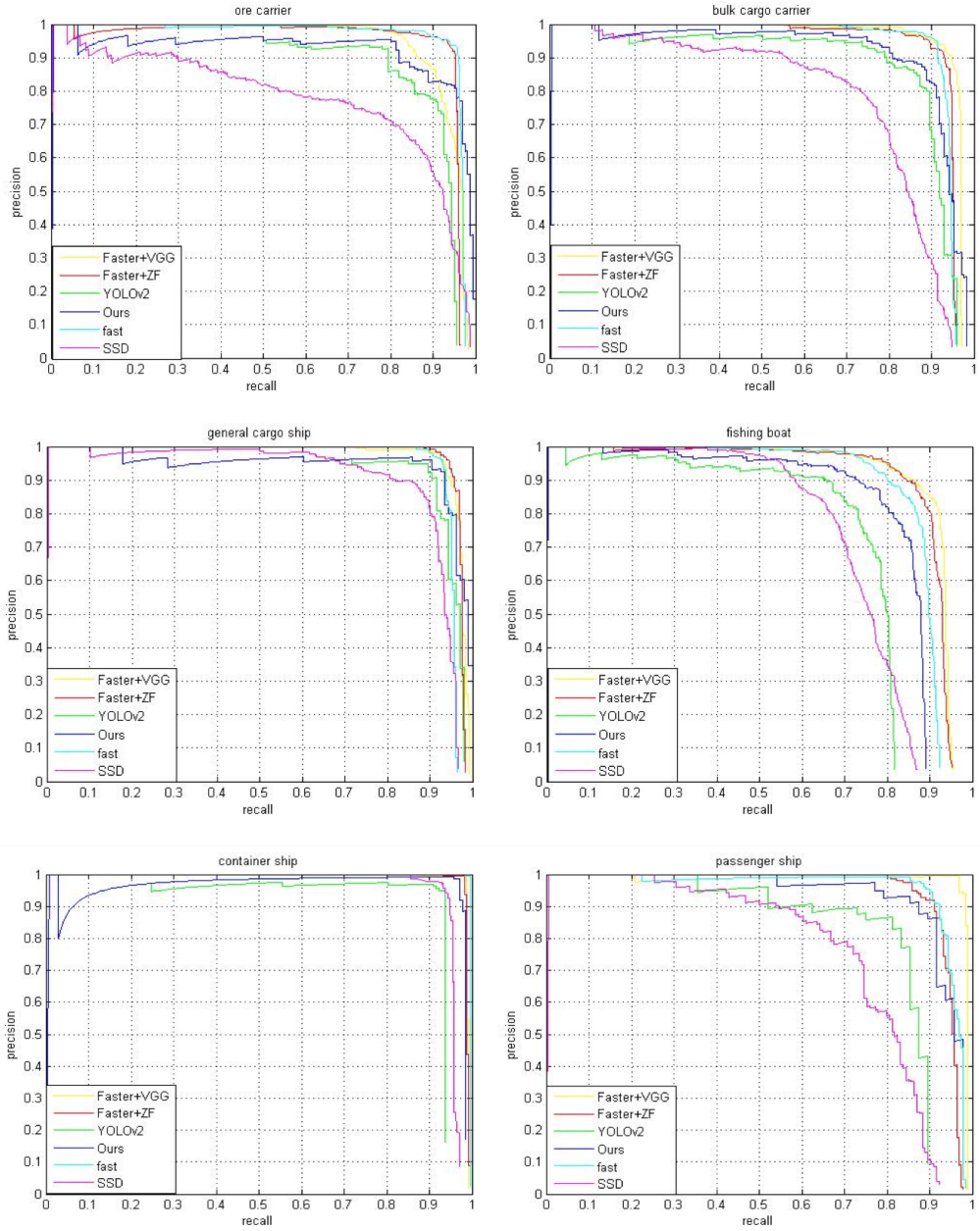


Fig. 11. Precision-recall curves for each detector on six ship categories.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCSVT.2019.2897980, IEEE Transactions on Circuits and Systems for Video Technology

(a) (b)

Fig. 12. Ship detection results. (a) The original YOLOv2. (b) Our proposed method.

performance of the saliency detection is particularly good and the accuracy is higher. For general cargo ships which perform well on the original method, our method has also improved notably over the Faster R-CNN.

In terms of speed, Detector FPS of 24 is considered to be real-time detector in object detection. In terms of real-time performance, the detection speed of the YOLOv2 is much better than other methods, and the FPS reaches 83, but the detection effect is not good. Since the SSD and the Faster R-CNN use the end-to-end training method, the detection effect is better, but the FPS is respectively 7 and 15, lower than the requirement. Our method gives FPS of 49, which not only guarantees the real-time performance, but also increases the accuracy by 5% against YOLOv2.

As shown examples in Fig. 12, we can see the visual improvement of our proposed method against YOLOv2. Specifically, when ships intersect, the bounding box of YOLOv2 is much larger than the ship, but our method can predict a more accurate box. When the ship is small, YOLOv2 is prone to misdetection, but our method remains good. When the ship is similar to the background, YOLOv2 can easily detect the background as a ship by mistake, but our method can eliminate false detection.

2) Comparison with Other Saliency Detection Methods

We combined other saliency detection methods with YOLOv2 and compared them with our method. These methods include FT [36], SEG [37], CA [38], and GC [39]. All experiments use the same dataset and are performed on four Titan Xp. We recorded the results of each model based on previous assessment indicators, shown in Table II. Fig. 13 shows AP performance where the IoU threshold for each ship is set to 0.5.

As we see in Table II, our method is basically the best for each category of ships. The IoU values have been improved by saliency detection, indicating that the saliency detection has refined the location, with the refinement from 70.69% to 74.53%. For FT [36], SEG [37] and CA [38], in addition to fishing boats and passenger ships with salient colors and features, the mAP values of other categories of ships are basically the same as those of the YOLOv2, indicating that these three methods make little improvement of detection and cannot meet the demand at all. FT [36] is mainly based on local features for detection, without considering global features. It tends to produce small salient objects other than the main ship body, so that the outer rectangle has a large error. The advantage of SEG [37] is that the feature between the video sequences can be used for saliency detection, but mainly based on local features, so the effect is also not good. CA [38] combines global features and local features, but it not only extracts the salient region, but also extracts the background information. In fact, it is not suitable for our dataset because our salient region generally occupies the main part of bounding box. GC [39] considers both global uniformity and

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCSVT.2019.2897980, IEEE Transactions on Circuits and Systems for Video Technology

TABLE II
DETECTION RESULTS OF DIFFERENT SALIENCY METHODS ON THE SEASHIP DATASET

Model	IoU	mAP	ore carrier	bulk cargo carrier	general cargo ship	container ship	fishing boat	passenger ship	FPS (Titan Xp)
YOLOv2+FT[36]	0.7133	0.835	0.840	0.850	0.880	0.888	0.768	0.781	1.5
YOLOv2+SEG[37]	0.7217	0.841	0.840	0.850	0.881	0.888	0.735	0.853	3
YOLOv2+CA[38]	0.7162	0.839	0.841	0.857	0.880	0.888	0.742	0.828	0.3
YOLOv2+GC[39]	0.7309	0.862	0.872	0.870	0.902	**0.907**	0.781	**0.888**	40
YOLOv2	0.7069	0.830	0.849	0.850	0.881	0.888	0.733	0.781	**83**
Ours	**0.7453**	**0.874**	**0.881**	**0.876**	**0.917**	0.903	**0.783**	0.886	49

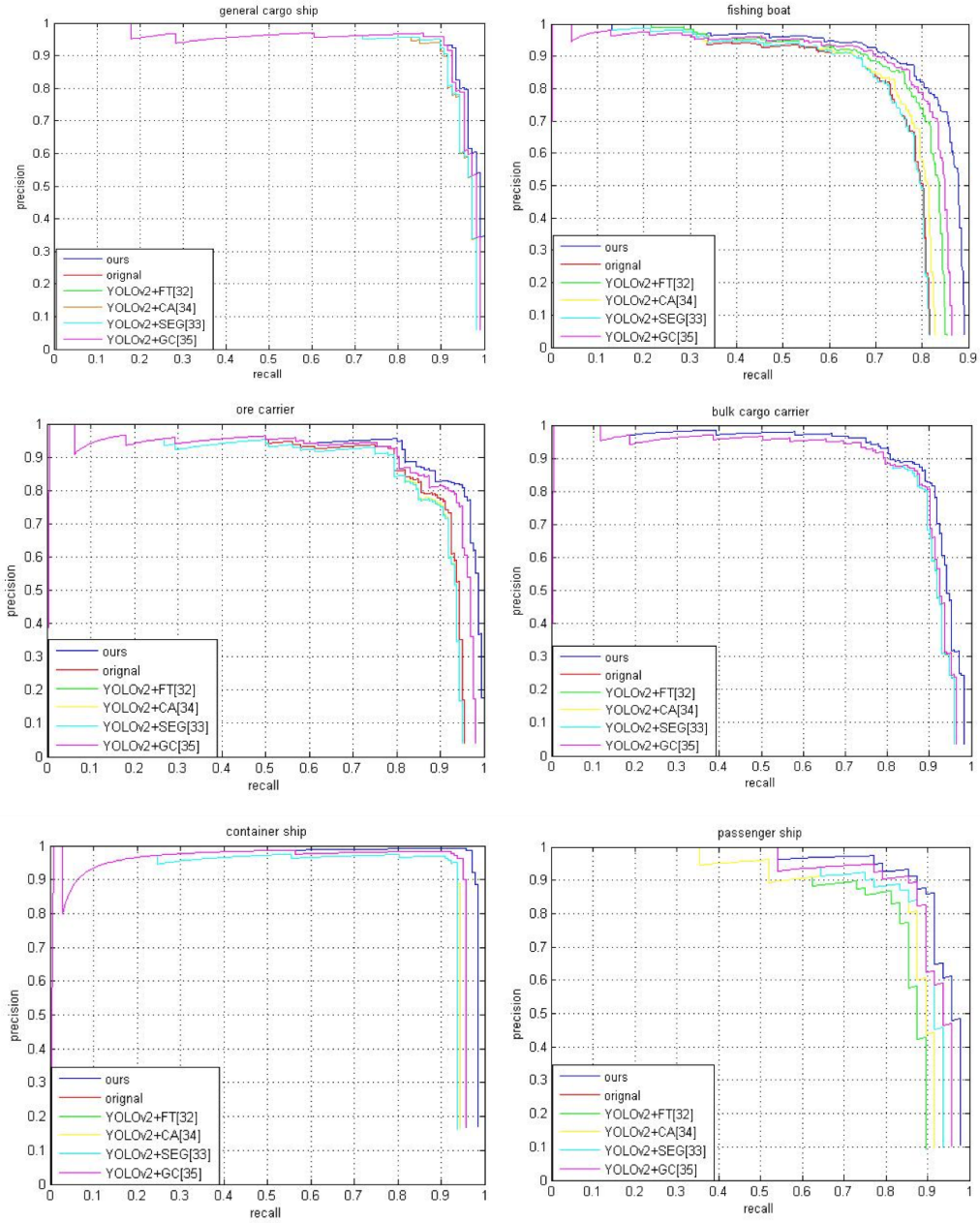


Fig. 13. Precision-recall curves for each salient method with YOLOv2 on six ship categories.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCSVT.2019.2897980, IEEE Transactions on Circuits and Systems for Video Technology

TABLE III
DETECTION RESULTS WITH OR WITHOUT COASTLINE ON THE SEASHIP DATASET

Model	mAP	ore carrier	bulk cargo carrier	general cargo ship	container ship	fishing boat	passenger ship	FPS (Titan Xp)
without coastline	0.862	0.872	0.857	0.870	0.902	0.781	**0.888**	**54**
with coastline	**0.874**	**0.881**	**0.876**	**0.917**	**0.903**	**0.783**	0.886	49

TABLE IV
DETECTION RESULTS OF DIFFERENT SHIP DETECTION METHODS

Model	mAP	ore carrier	bulk cargo carrier	general cargo ship	container ship	fishing boat	passenger ship	FPS (Titan Xp)
Ours	**0.874**	**0.881**	**0.876**	**0.917**	**0.904**	**0.783**	**0.886**	**49**
Zhang's[5]	0.487	0.414	0.432	0.387	0.462	0.583	0.502	2

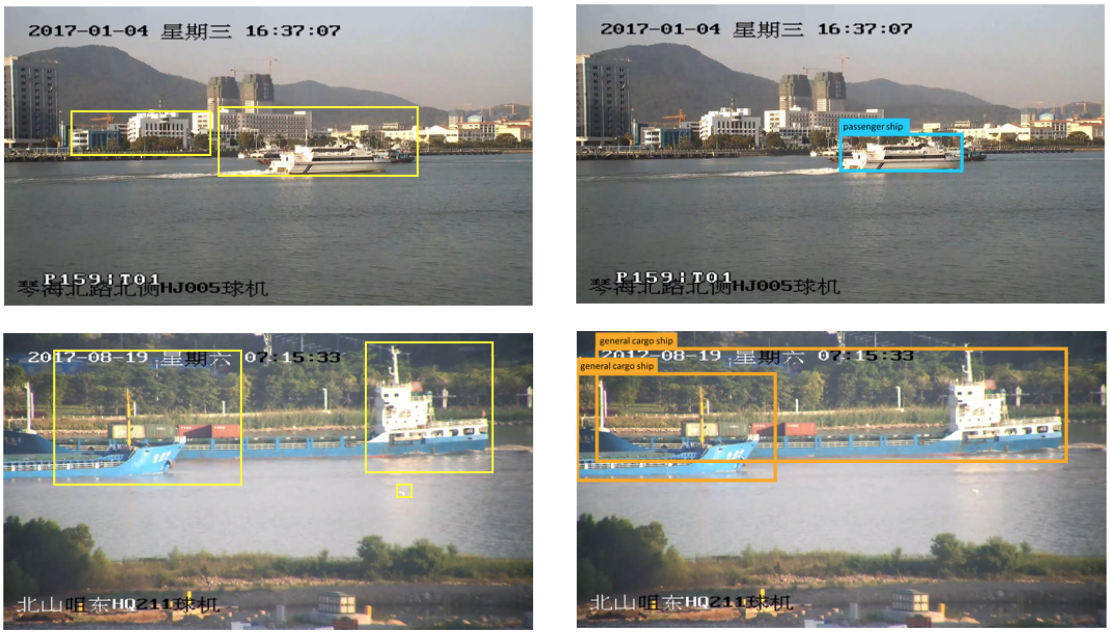


Fig. 14. The typical results of our proposed method (right) and Zhang's method [5] (left).

color distribution, so the improvement of detection is almost the same as ours. Nevertheless, its FPS is 40, which is lower than ours.

3) Comparison with or without Coastline

In the process of generating the anchor boxes, the number of boxes can be reduced by using the coastline feature, which can save the detection time. Coastline feature can also be used to remove disturbances from shore houses and improve accuracy during saliency detection. We conducted experiments to compare the results of the models with or without coastline feature. The results are shown in Table III. After adding coastline feature, mAP increases by 1%, and the biggest increase is from the general cargo ship, which is 5%. We think that the color of the general cargo ship's bow is usually white and there are several rows of windows on it, which are very similar to the shore house. Thus, it is easy to falsely detect the house as part of ship. The FPS of model without coastline is 5 higher than the FPS of our method, which is due to the fact that the time required to generate the coastline exceeds the time saved by reducing the amount of anchor generation.

4) Comparison with other ship detection methods

Since there are no public source codes or executables for ship detection, we compare the proposed algorithm with Zhang et al. [5]. Because the method cannot detect the category of ship, we separately detect the images of different categories. The results are shown in Table IV. As can be seen from the table, our method is far better than comparison method for each category. Among six categories of ship, the result for fishing boat is relatively best for the comparison method, which is possibly due to the fact that the fishing boat does not intersect with the coastline, thus less affected by ashore background. Some typical detection results are shown

in the Fig.14, where the buoy is mistaken as ship by the comparison method. At the same time, its FPS is much lower than ours, which confirms that our approach is comparably effective and fast under complex environments.

V. Conclusion

In this paper, in order to realize the real-time detection of ships in many application fields, we propose method based on convolution neural network and saliency detection. Our method generates the bounding boxes based on YOLOv2 and proposes saliency detection to predict the location of the ships in the bounding boxes. When the probability of bounding boxes is low, we use salient features to predict more accurate location in combination with the coastline feature. We train the model on the real-world ship dataset built by our own and compare it with other methods. Experimental results prove that our method is able to simultaneously result in high accuracy and fast speed against typical CNN based methods.

In the future, we will apply the detection model to achieve multi-target tracking of ships.

References

[1] Z. Huang, "An Ship Detection And Analysis System Based On Synthetic Aperture Radar Image", *Microcomputer Information*, vol. 25, no. 6, pp. 298-300, 2009.

[2] W. Krüger and Z. Orlov, "Robust layer-based boat detection and multi-target-tracking in maritime environments," *International WaterSide Security Conference*, 2010, pp. 1-7.

[3] X. Bao *et al.*, "Context modeling combined with motion analysis for moving ship detection in port surveillance," *Journal of Electronic Imaging*, 2013, 22(4):041114.

[4] Z. Chen, B. Li, L. Tian and D. Chao, "Automatic detection and tracking of ship based on mean shift in corrected video sequences," 2*017 2nd International Conference on Image, Vision and Computing*, 2017, pp. 449-453.

[5] Y. Zhang, Q. Li, F,Zhang, "Ship detection for visual maritime surveillance from non-stationary platforms." *Ocean Engineering* ,141(2017):53-63.

[6] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," *International Conference on Neural Information Processing Systems*. Curran Associates Inc. 2012:1097-1105.

[7] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *Computer Science*, 2014.

[8] C. Szegedy, W. Liu, Y. Jia, et al., "Going deeper with convolutions," *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2015, pp. 1-9.

[9] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," *IEEE Conference on Computer Vision and Pattern Recognition IEEE Computer Society*, 2014:580-587.

[10] R. Girshick, "Fast R-CNN," *Computer Science*, 2015.

[11] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," *International Conference on Neural Information Processing Systems* MIT Press, 2015:91-99.

[12] P. Sermanet et al., "OverFeat: Integrated recognition, localization and detection using convolutional networks," Proc. ICLR, 2014, pp. 1-16.

[13] 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, Sept. 2015.

[14] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, et al., "SSD: Single Shot MultiBox Detector," *Proc. Eur. Conf. Comput. Vis.*, 2015, pp. 21-37.

[15] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You Only Look Once: Unified, Real-Time Object Detection," *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2016, pp. 779-788.

[16] X. Li, S. Wang, B. Jiang and X. Chan, "Inshore ship detection in remote sensing images based on deep features," *2017 IEEE International Conference on Signal Processing, Communications and Computing*, 2017, pp. 1-5.

[17] H. Lin, Z. Shi and Z. Zou, "Fully Convolutional Network With Task Partitioning for Inshore Ship Detection in Optical Remote Sensing Images," *IEEE Geoscience and Remote Sensing Letters*, 2017, vol. 14, no. 10, pp. 1665-1669.

[18] Shao Z, Wu W, Wang Z, et al. SeaShips: A Large-Scale Precisely-Annotated Dataset for Ship Detection[J]. IEEE Transactions on Multimedia, 2018, 20(10): 2593-2604.

[19] J. Redmon, A. Farhadi, "YOLO9000: Better, Faster, Stronger," *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2017, pp. 6517-6525.

[20] Guo-Jun Qi, "Hierarchically Gated Deep Networks for Semantic Segmentation," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 2267-2275.

[21] M. Donoser, M. Urschler, M. Hirzer, and H Bischof, "Saliency driven total variation segmentation," *Proc. Int. Conf. on Computer Vision.*, 2009, pp. 817-824.

[22] Q. Li, Y. Zhou, and J. Yang, "Saliency based image segmentation," *Proc. Int. Conf. on Multimedia Technology*., 2011, pp. 5068-5071.

[23] C. Qin, G. Zhang, Y. Zhou, W. Tao, and Z. Cao, "Integration of the saliency-based seed extraction and random walks for image segmentation," *Neurocomputing*, vol. 129, no. 4, pp. 378-391, 2013.

[24] M. Johnson-Roberson, J. Bohg, M. Bjorkman, and D. Kragic, "Attention-based active 3d point cloud segmentation," *Proc. IEEE Conf. Intelligent Robots & Systems*., 2010, pp. 1165–1170.

[25] U. Rutishauser, D. Walther, C. Koch, and P. Perona, "Is bottom-up attention useful for object recognition?" *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2004.

[26] C. Kanan and G. Cottrell, "Robust classification of objects, faces, and flowers using natural image statistics," *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2010, pp. 2472–2479.

[27] Z. Ren, S. Gao, L.-T. Chia, and I. Tsang, "Region-based saliency detection and its application in object recognition," *IEEE transactions on Circuits & Systems for Video Technology*, 2014, vol. 24, no. 5, pp. 769-779.

[28] T. Chen, M. M. Cheng, P. Tan, A. Shamir, and S.-M. Hu, "Sketch2photo: Internet image montage," *Acm Transactions on Graphics*, 2009, vol. 28, no. 5, pp. 1-10.

[29] S. M. Hu, T. Chen, K. Xu, M. M. Cheng, and R. R. Martin, "Internet visual media processing: a survey with graphics and vision applications," *Visual Computer International Journal of Computer Graphics*, 2013, vol. 29, no. 5, pp. 393-405.

[30] Y. Gao, M. Wang, Z.-J. Zha, J. Shen, X. Li, and X. Wu, "Visual-Textual Joint Relevance Learning for Tag-Based Social Image Search," *IEEE Transactions on Image Processing A Publication of the IEEE Signal Processing Society*, vol. 22, no. 1, pp. 363–376, 2013.

[31] F. Bi, B. Zhu, L. Gao, & M. Bian, "A Visual Search Inspired Computational Model for Ship Detection in Optical Satellite Images," *IEEE Geoscience & Remote Sensing Letters*, vol. 9, no 4, pp.749-753, 2012.

[32] L. Jiang, "In-shore ship extraction from HR optical remote sensing image via salience structure and GIS information," *International Symposium on Multispectral Image Processing & Pattern Recognition, MIPPR 2015: Remote Sensing Image Processing, Geographic Information Systems, and Other Applications.*, 2015, vol. 9815, pp. 98150U.

[33] H. Lin, Z. Shi, and Z. Zou, "Fully Convolutional Network With Task Partitioning for Inshore Ship Detection in Optical Remote Sensing Images," *IEEE Geoscience & Remote Sensing Letters*, vol. 99, pp. 1-5, 2017.

[34] E. Shelhamer, J. Long, and T. Darrell, "Fully Convolutional Networks for Semantic Segmentation," *IEEE Trans. Pattern Anal. Mach Intell.*, vol. 39, no. 4, pp. 640-651, 2014.

[35] D. Walther, and C. Koch, "Modeling attention to salient proto-objects," *Neural Networks the Official Journal of the International Neural Network Society*, vol. 19, no. 9, pp. 1395, 2006.

[36] R. Achanta, S. Hemami, F. Estrada , and S. Susstrunk, "Frequency-tuned salient region detection," *Computer vision and pattern recognition,* vol. 22, no. 9-10, pp. 1597-1604, 2009.

[37] E. Rahtu, J. Kannala, M. Salo, and J. Heikkilä, "Segmenting Salient Objects from Images and Videos," *Proc. Eur. Conf. Comput. Vis.*, 2010, pp. 366-379.

[38] S. Goferman, L. Zelnikmanor, and A. Tal, "Context-Aware Saliency Detection," *IEEE Trans Pattern Anal Mach Intell*, vol. 34, no. 10, pp. 1915-1926, 2012.

[39] M.-M. Cheng, J. Warrell, W.-Y. Lin, S. Zheng, V. Vineet, and N. Crook, "Efficient salient region detection with soft image abstraction," in *IEEE ICCV*, 2013, pp. 1529–1536.

[40] A. Borji, M.-M. Cheng, H. Jiang, and J. Li. (Nov. 2014). "Salient object detection: A survey." [Online]. Available: http://arxiv.org/abs/1411.5878.

[41] A. Borji and L. Itti, "State-of-the-art in visual attention modeling," *IEEE TPAMI*, 2012.

[42] M. M. Cheng, N. J. Mitra, X. Huang, et al., "Global Contrast based Salient Region Detection," *IEEE Trans. Pattern Anal. Mach Intell.*, vol. 37, no. 3, pp. 409-416, 2015.

[43] A. Neubeck and L. Van Gool, "Efficient Non-Maximum Suppression," *18th International Conference on Pattern Recognition* (ICPR 2006), 2006, pp. 850–855.

[44] P. Felzenszwalb, and D. Huttenlocher, "Efficient graph-based image segmentation," *International Journal of Computer Vision*, vol. 59, no. 2, pp. 167–181, 2004.

[45] M. D. Zeiler, R Fergus, "Visualizing and Understanding Convolutional Networks," *Proc. Eur. Conf. Comput. Vis.*, 2014, pp. 818-833.

[46] K. Simonyan, A. Zisserman, "Very Deep Convolutional Networks for Large-Scale Image Recognition," *Computer Science*, 2014.

[47] A. G. Howard, M Zhu, B Chen, D Kalenichenko, W Wang, et al., "MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications," *arXiv:1704.04861*, 2017.

Zhenfeng Shao received the PhD degree from Wuhan University, China, in 2004. He is now a professor of State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing, Wuhan University, China. His research interest is computer vision.

Linggang Wang received the BEng degree in school of geodesy and geomatics from Wuhan University, Wuhan, China, in 2017. He is currently working toward the MEng degree in photogrammetry and remote sensing from the State Key Lab oratory of Information Engineering in Surveying, Mapping and Remote Sensing, Wuhan University. his research interests include image processing and object detection.

Zhongyuan Wang received the PhD degree in communication and information system from Wuhan University and is currently working as a professor with Computer School at Wuhan University. His research interests include video compression, image processing and multimedia big data analytics as well.

Wan Du is currently an Assistant Professor in Computer Science and Engineering at the University of California, Merced. He had worked as a Research Fellow in the School of Computer Science and Engineering, Nanyang Technological University, Singapore, from 2011 to 2017. He received the Ph.D. degree in Electronics from the University of Lyon (Ecole centrale de Lyon), France, in 2011. His research interests include the Internet of Things, cyber-physical system, wireless networking systems and mobile computing systems.

Wenjing Wu received the BEng degree in remote sensing science and technology from Wuhan University, Wuhan, China, in 2016. She is currently working toward the MEng degree in photogrammetry and remote sensing from the State Key Lab oratory of Information Engineering in Surveying, Mapping and Remote Sensing, Wuhan University. Her research interests include image processing and object detection.