A deep attention mechanism method for maritime salient ship detection in complex sea background*

ZHOU Weina (周薇娜)** **and CHEN Peiqiu** (陈培秋)

College of Information Engineering, Shanghai Maritime University, Shanghai 201306, China

(Received 1 September 2020; Revised 11 November 2020)

Saliency ship detection has received increasing attention due to its important applications in maritime field in recent years. Up to now, numerous studies on saliency detection have been done based on traditional methods and deep learning methods. But these previous research works are still not competent enough in detecting ship targets with complex backgrounds and noises. In this letter, we propose a deep attention mechanism method for more accurate and faster maritime salient ship detection. We optimize the initial ship saliency map by using a feature attention module to focus on salient objects. We reduce and improve the convolution kernel in refinement residual module to enhance the detection efficiency. In addition, Leaky ReLU is selected as the activation function to increase the non-linear capability of the method. Experiment results show that, the proposed method could obtain outstanding performance in salient ship detection in complex sea background.

Document code: A **Article ID:** 1673-1905(2021)07-0438-6

DOI https://doi.org/10.1007/s11801-021-0137-z

Salient ship detection aims to highlight the most prominent ship target in an image. It has been widely used to monitor and ensure the safety of ocean and inland rivers in maritime field. However, it is still a challenging branch of saliency detection[1,2], especially in complex sea environment.

Methods used before could been divided into two kinds: the traditional methods[3-5] and deep learning methods[6-9]. The traditional methods of extracting ships' shape, contour and texture features are usually not robust enough. For instance, Cane et al[3] presented a method for ship detection and tracking based on visual saliency to suppress the wave and reflected light in maritime environments. But it could not detect distant and dim objects well. Xu et al[4] also proposed an unsupervised ship detection method based on visual saliency. Ship targets are characterized by histograms of oriented gradient (HOG) descriptor. But the result of the method still could not meet the demand in reality. Bao et al[5] proposed a cabin feature detector. It created a detector by training ship features and combined context and motion saliency analysis. But this detector is completely dependent on the training set size. It is not a universally applicable method. Compared to traditional methods, deep learning methods have greatly promoted the effect of saliency detection, but it still has some unsolved problems when the detection is affected by distance, sea clutter, light intensity, weather changes and so on. For example, Bi et al[6] extracted salient candidate regions from the entire detection scene by using a bottom-up visual attention mechanism. Its appearance and neighborhood similarity features are combined to further discriminate the salient regions. But its real-time performance is not satisfying. Lin et al[7] implemented a partitioning task model with the deep path for attention/saliency maps and the shallow path for detection. But this method is unable to provide predictions on the direction of the proposed candidates. Mumtaz et al[8] used the graph-based visual saliency algorithm to calculate the saliency map, and then the saliency map is processed by multi-level threshold to obtain the ordered saliency area of input image. However, multiple experiments should be done to obtain an appropriate cluster selection threshold. Shao[9] proposed a saliency-aware convolutional neural network (CNN) framework for ship detection. It extracted coastline features and incorporate them into CNN to improve the robustness and efficiency of the ship detection. However, It is most effective only in inshore ship detection.

In this letter, we propose a deep attention mechanism (DAM) framework. Different from existing salient methods for ship detection, our framework adds a feature attention module (FAM) after extracting high-level semantic features, which not only could focus the operation on the specific ship area, but also could enhance the regional features of this part. At the same time, a refinement residual module (RRM) with Leaky ReLU[10] activation function is also used in this framework to refine the initial saliency map. The architecture of DAM is

* This work has been supported by the National Natural Science Foundation of China (Nos.61404083 and 52071200).

** E-mail: wnzhou@shmtu.edu.cn

shown in Fig.1.

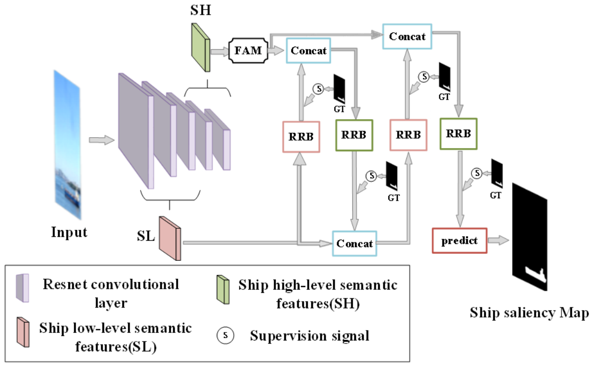

SH
FAM
Concat
Concat
S
GT
RRB
RRB
RRB
RRB
S
GT
S
GT
S
GT
Input
SL
Concat
predict
Ship saliency Map
Resnet convolutional layer
Ship high-level semantic features(SH)
Ship low-level semantic features(SL)
S
Supervision signal

Fig.1 The proposed DAM framework

As shown in Fig.1, the framework consists of three parts. The first part is the feature extraction network, which collects both low-level semantic features (SL) and high-level semantic features (SH) maps. Resnet-101 is used as the backbone network and is divided into five stages. The first three stages are used to extract SL, and latter two stages are used to extract SH. The characteristics of different layers of network are complementary [11,12]. In ship detection, SH usually contains global context-aware information, which are suitable for locating ship area correctly, while SL contains spatial structure details, which are more suitable for locating ship boundary. In the second part, an FAM is added after extracting SH to generate an initial ship saliency map. After that, in the third part, RRM is then designed to refine the initial saliency map by integrating with SL information. It is a gradual optimization process of the initial saliency map. The "Concat" model shown in Fig.1 is used to do concatenation operation. A supervision signal[13] is also applied at each step to improve the final ship saliency prediction map gradually. This could help to compute the loss between the predicted saliency map and Ground Truth (GT) mask during the training process. In addition, the Leaky ReLU activation function is selected in RRM to improve the nonlinear capability of the framework and reduce the training loss. The details of the framework will be explained as follows.

As we mentioned above, the initial saliency map should be refined for an accurate result. However, the refinement will be severely affected by different kinds of noises at the beginning of the detection. That's one of the major difficulties in detecting marine targets in complex backgrounds. Considering that there are few semantic differences between SL compared to SH, FAM is designed and used just after extracting the SH information to improve the effectiveness of the initial saliency map. FAM simulates the human visual attention mechanism, which ignores the global information temporarily, and focuses on the ship target we are concerned with. The architecture of the FAM is shown in Fig.2. Once FAM receives SH information, it will be convoluted with Gaussian kernel[14]. Then the results will be normalized to obtain the maximum value of the input, which makes the system focus the attention on local salient ship area, and extract the most prominent ship target from the image.

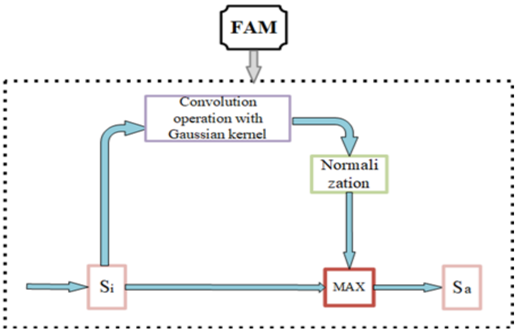


Fig.2 Architecture of FAM

FAM could also be expressed by

$$S_a = MAX(f_{\min-\max}(Con(g,\ S_i)),\ S_i), \tag{1}$$

$$f_{\min-\max}(x) = \frac{x - x_{\min}}{x_{\max} - x_{\min}}, \tag{2}$$

where S_a is the initial salient signal we obtained by FAM, $Con(g,\ S_i)$ is a convolution operation with a Gaussian kernel g, which is set to 32 initially. Gaussian's bias is set to 0, and the standard bias is set to 4 in our framework. S_i represents the SH. $f_{\min-\max}$ is a normalization function, which is illustrated in Eq.(2), and its result will be mapped in [0, 1]. *MAX* is a function that takes the maximum value as the salient region of the input dataset. Eq.(2) is the definition for $f_{\min-\max}(x)$ function, $x_{\min}$ and $x_{\max}$ respectively represent the minimum and maximum values of the sample data.

RRM could optimize the initial salient map output by FAM. This model learns from the experiences of edge and feature optimizing, and processes the feature map alternatively from SL and SH with "*concat*" operation. To reduce the overfitting, GT is used as a supervised signal during training. The function of RRM module could also be defined as

$$S_r = Conv_j(concat(S_{j-1},\ F)), \tag{3}$$

$$S_j = S_{j-1} + S_r, \tag{4}$$

where S_j represents the saliency signal of jth RRM, F is the feature maps, *concat* represents concatenation operation. $Conv_j$ indicates the concatenation of the predicted saliency map S_{j-1} and the feature maps F, the value range of j is [1, 3]. S_r is the residual. Eq.(4) indicates that the residual is added with S_{j-1} to compute the output S_j in RRM.

The architecture of RRM is shown in Fig.3. It is composed of three 3×3, one 1×1 convolution kernels and Leaky ReLU activation function. The advantage of using three 3×3 convolutions is that the nonlinear ability of the network could be increased without changing the receptive field of the convolutional layer. And following of the 1×1 convolution kernel can lower the dimension and reduce the computational cost greatly. In addition, due to the fact that an appropriate activation function could help to improve the expressive ability of our network, Leaky ReLU is selected to ensure the ship detection efficiency in complex backgrounds. In RRM, Leaky ReLU activation function is used between every two convolution kernels. It is based on ReLU and has solved the hard saturation problem of ReLU. Tab.1 shows the training loss of Leaky ReLU and other three activation functions, which are all typically used in deep learning methods. The values are obtained in the same configuration environment and using the same dataset, the number of training step is 10 000. After comparison, we can find out that the training loss of Leaky ReLU is the smallest, and it is beneficial for reducing the loss of ship features during the detection process.

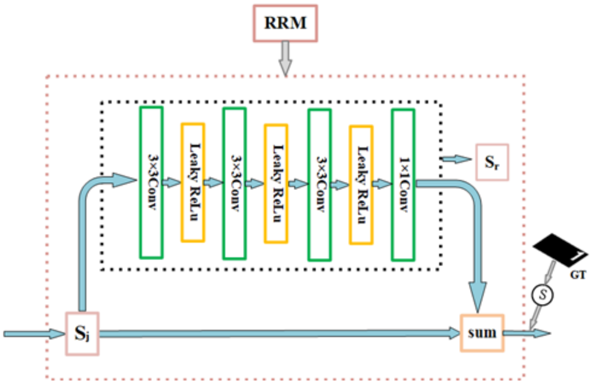


Fig.3 Architecture of RRM

Tab.1 Comparison of training loss of four activation functions

Activation	Train loss
ReLU	0.101 85
ELU	0.102 62
PReLU	0.101 63
Leaky ReLU	0.100 84

To test the performance of DAM, we constructed a new challenging dataset by riching different ship images. Images of the dataset are mainly captured by our team or collected from Singapore maritime dataset (SMD)[15]. It contains 400 ship images with more than 1000 ship targets in different kinds of ocean environments. Fig.4 shows some ship images and their GT in our dataset. It's worth mentioning that ships usually occupy less area compared with background in the images. And the sea and sky background of the image are always disturbed by noises of waves, islands, ripples and so on. For our performance evaluation, 320 images are randomly selected as training images and 80 as test images. So the appearance of ships in the test images are randomly different with ships in training set, what could further test the generalization of our network.

Besides our ship dataset, MSRA10K dataset[16] is also used to pre-train the DAM. MSRA10K dataset contains 10 000 images of different scenes. Its large scale could help improve the detection performance and remedy the deficiency of our ship dataset. Our evaluation experiments are also done on other five saliency detection benchmark datasets for comparison, including ECSSD[17], HKU-IS[11], PASCAL-S[18], DUT-OMRON[19], SOD[19].

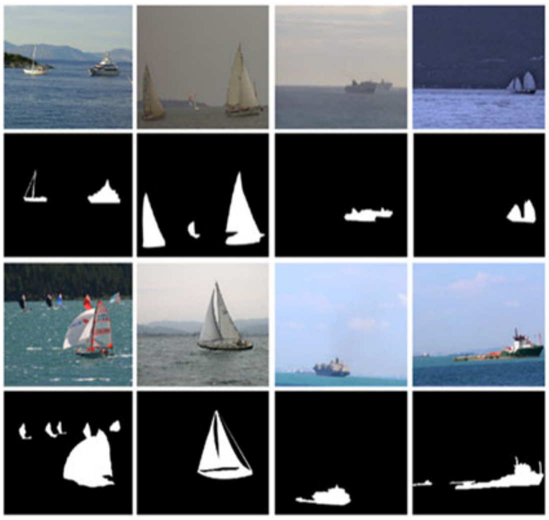


Fig.4 Ship images and their GT in our ship dataset

In the letter, two widely used indicators are selected as our evaluation indexes: mean absolute error (*MAE*) and *F*-measure (F_β). F_β is the weighted harmonic mean of the recall and precision under non-negative weight β. β^2 is taken as 0.3 here as suggested by previous works[20,21]. As we know, a good saliency detection model should have a large F_β and a small *MAE*.

The definitions of *MAE* and F_β are

$$MAE=\frac{1}{W\times H}\sum\nolimits_{x=1}^{W}\sum\nolimits_{y=1}^{H}\left|S(x,y)-G(x,y)\right|, \tag{5}$$

$$F_\beta=\frac{\left(1+\beta^2\right)precision\times recall}{\beta^2 precision+recall}, \tag{6}$$

where $S(x, y)$ is the saliency map output by the network, $G(x, y)$ represents GT, W and H are the width and height of saliency map. And precision and recall are defined as

$$precision=\frac{|S\cap G|}{|S|}, \tag{7}$$

$$recall=\frac{|S\cap G|}{|G|}. \tag{8}$$

From Eqs.(7) and (8), we can find out that "*precision*" is the ratio of the number of successfully detected targets to the number of all detected targets, "*recall*" is the ratio of the number of successfully detected targets to the number of ground truth targets.

Our training and test are operated based on Pytorch1.0, Ubuntu 16.04 system, and GTX 1080Ti GPU. ResNet-101 network is used to initialize parameters of feature extraction network, and the default setting of Pytorch is used to initialize other convolutional layers to speed up the training process and avoid over-fitting problems. The proposed model was trained by Adam optimizer[22] with a Momentum of 0.9, a decay of 0.000 5, a batch size of 14. The basic learning rate is set to 10^{-3}, we use the "poly" learning rate policy and stop the training procedure after 10k iterations.

Fig.5 is the loss curve obtained during training. From it, we can see that, the loss decreased rapidly in the first 2k iterations, and then it flattens out. Finally, its value stabilizes at 0.1 after 8k—10k iterations. Thus, we determine to stop the train after 10k iterations.

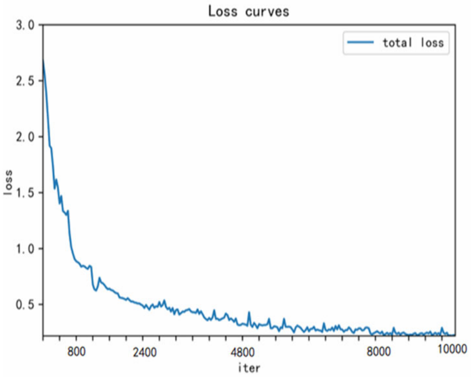
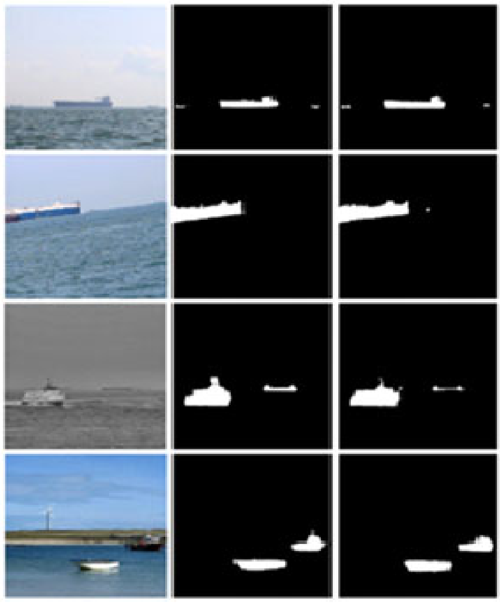


Fig.5 Training loss curve

The test results of our method are shown in Tab.2. *MAE* and F_β in Tab.2 reflect the detection capability of DAM. Fig.6 shows a part of detection results with our ship dataset. From it, we can easily find out that, in most real marine environment, the ship edge could be well detected using our algorithm, and the difference is small between our results and GT.

Tab.2 The indicators result of DAM with our ship dataset

Method	*MAE*	F_β
DAM	0.002 1	0.938



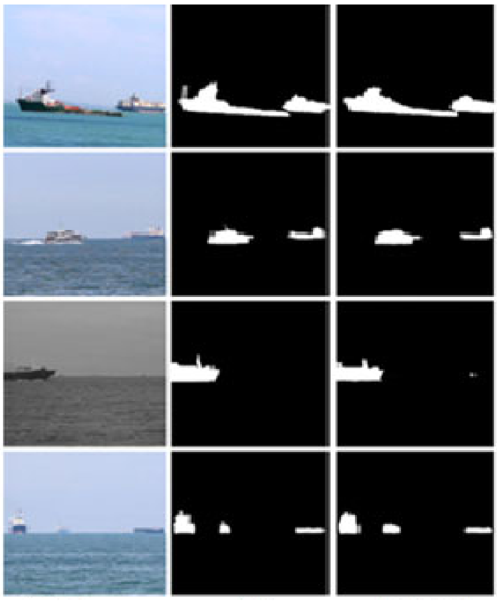


Fig.6 Some test results of our DAM framework

In order to fully verify the validity of the proposed method and prove the effect of our method in ship detection, we compared the processing result of DAM with the result of other ship detection frameworks using the same evaluation indicators –*MAE*, F_β and detection time. The comparison results are shown in Tab.3, Figs.7 and 8. Figs.7 and 8 are respectively the comparison result of *MAE* and *F* in graph and histogram style. They could provide more intuitive result. From them three, we can see that DAM obtained a relative small *MAE* and the largest F_β. It also has a relatively little time consumption in ship detection. Therefore, it can be inferred that our proposed framework has a strong detection capability and a good performance on the practical application of ship detection.

Tab.3 Comparison of detection result

Method	MAE	F_β	Time (s)
ALB[23]	0.075 1	/	/
BMS[24]	0.047 0	/	/
LIU[25]	0.241 7	/	/
MAK[26]	0.009 1	/	/
SDMOT[3]	0.005 7	/	/
IR[27]	/	0.8	/
STSD[28]	/	0.815	/
SDE[4]	/	0.73	/
ISR[29]	/	0.924	0.453
CMBSD[29]	/	0.785	0.185
CFD[5]	/	0.717	0.365
Cabin-detector[30]	/	0.777	/
GBVS[27]	/	0.58	/
AMISD[31]	/	0.899	/
TPISD[32]	/	0.92	/
SBAS[8]	/	0.90	/
DAM (ours)	**0.002 1**	**0.938**	**0.162**

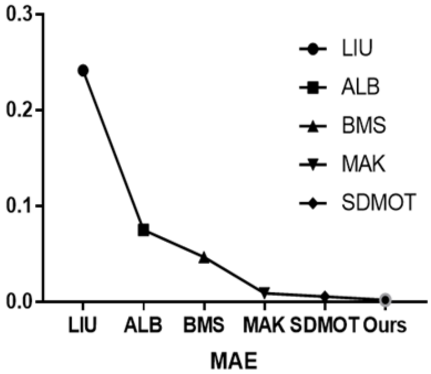


Fig.7 *MAE* comparison between DAM and other ship detection algorithms

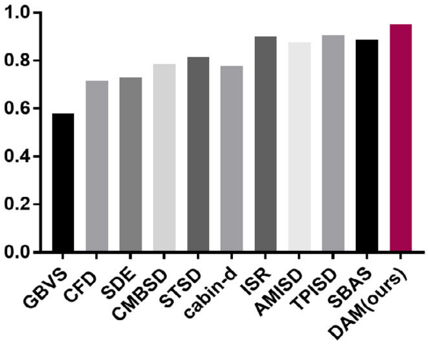


Fig.8 F_β comparison between DAM and other ship detection algorithms

In this paper, we propose a DAM framework for accurate and fast maritime salient ship detection. Considering the particularity of ship targets, DAM adds FAM after feature extraction network to extract the initial saliency map, and improves the convolution layer and activation function of RRM to optimize the saliency map step-by-step. The experimental results show that our method can obtain a good performance in maritime saliency ship detection. Although our method executes better compared with other ship detection algorithms, it still would lose some ship edge information. In the future, we will further improve RRM to gain a better result.

References

[1] Cheng M. M., Zhang G. X., Niloy J. M., Huang X. L. and Hu S. M., Global Contrast Based Salient Region Detection, Conference on Computer Vision and Pattern Recognition, 409 (2011).

[2] Sun M. J., Zhou Z. Q., Hu Q. H., Wang Z. and Jiang J. M., IEEE Transactions on Cybernetics **49**, 2900 (2018).

[3] Tom Cane and James Ferryman, Saliency-Based Detection for Maritime Object Tracking, Conference on Computer Vision and Pattern Recognition, 18 (2016).

[4] Xu Fang and Liu Jing-hong, Optoelectronics Letters **12**, 473 (2016).

[5] X. Bao, S. Zinger, R. Wijnhoven and Peter H. N. de With, Ship Detection in Port Surveillance Based on Context and Motion Saliency Analysis, SPIE, 86630D (2013).

[6] Bi F., Zhu B., Gao L. and Bian M., IEEE Geoscience and Remote Sensing Letters **9**, 749 (2012).

[7] Lin Hao-ning, Shi Zhen-wei and Zou Zheng-xia, IEEE Geoscience and Remote Sensing Letters **14**, 1665 (2017).

[8] A. Mumtaz, A. Jabber, Z. Mahmood, R. Nawaz and Q. Ahsan, Saliency Based Algorithm for Ship Detection in Infrared Images, 13th International Bhurban Conference on Applied Sciences and Technology, 167 (2016).

[9] Shao Zhen-feng, Wang Ling-gang, Wang Zhong-yuan, Du Wan and Wu Wen-jing, IEEE Transactions on Circuits and Systems for Video Technology **30**, 781 (2020).

[10] P. Isola, J. Y. Zhu, Zhou T. H. and A. Efros Alexei, Image-to-Image Translation with Conditional Adversarial Networks, Conference on Computer Vision and Pattern Recognition, 1125 (2017).

[11] Li Guan-bin and Yu Yi-zhou, Visual Saliency Based on Multiscale Deep Features, Conference on Computer Vision and Pattern Recognition, 5455 (2015).

[12] Zhang P., Wang D., Lu H., Wang H. and Ruan X., Amulet: Aggregating Multi-Level Convolutional Features for Salient Object Detection, IEEE International Conference on Computer Vision, 202 (2017).

[13] Xie Sai-ning and Tu Zhou-wen, Holistically-Nested Edge Detection, IEEE International Conference on Computer Vision, 1395 (2015).

[14] V. Mnih, N. Heess and A. Graves, Recurrent Models of Visual Attention, Advances in Neural Information Processing Systems, 2204 (2014).

[15] Dilip K. Prasad, Singapore Maritime Dataset, https://sites.google.com/site/dilipprasad/home/singapore-maritime-dataset, 2016.

[16] Cheng Ming-ming, Niloy J. Mitra, Huang Xiao-lei, Philip H. S. Torr and Hu Shi-Min, IEEE Transactions on Pattern Analysis and Machine Intelligence **37**, 569 (2014).

[17] Yan Qiong, Xu Li, Shi Jian-ping and Jia Jia-ya, Hierarchical Saliency Detection, Conference on Computer Vision and Pattern Recognition, 1155 (2013).

[18] Li Yin, Hou Xiao-di, Christof. Koch, James M. Rehg and Alan L. Yuille, The Secretsof Salient Object Segmentation Conference on Computer Vision and Pattern Recognition, 280 (2014).

[19] Yang Chuan, Zhang Li-he, Lu Hu-chuan, Ruan Xiang and Yang Ming-Hsuan, Saliency Detection via Graph-Based Manifold Ranking, Conference on Computer Vision and Pattern Recognition, 3166 (2013).

[20] Hou Qi-bin, Cheng Ming-ming, Hu Xiao-wei, Ali Borji, Tu Zhuo-wen and Philip H.S., IEEE Transactions on Pattern Analysis and Machine Intelligence **41**, 4 (2019).

[21] Zhao Ting and Wu Xiang-qian, Pyramid Feature Attention Network for Saliency Detection, Conference on Computer Vision and Pattern Recognition, 1 (2019).

[22] D. P. Kingma and J. Ba. Adam, A Method for Stochastic Optimization, International Conference on Learning

Representations, 1 (2015).

[23] T. Albrecht, G. A. W. West, T. Tan and T. Ly, Visual Maritime Attention Using Multiple Low-Level Features and Naive Bayes Classification, IEEE International Conference on Digital Image Computing: Techniques & Applications, 243 (2011).

[24] Zhang Jian-ming, Stan Sclaroff, Lin Zhe, Shen Xiaohui, Price Brian and Mech Radomír, Minimum Barrier Salient Object Detection at 80 FPS, IEEE International Conference on Computer Vision, 2015.

[25] Liu Hai-ying, Omar Javed, Geoff Taylor, Cao Xiao-chun and Niels Haering, Omni-Directional Surveillance for Unmanned Water Vehicles, Proc. Eighth Int. Workshop Visual Surveillance, 1 (2008).

[26] K. Makantasis, A. Doulamis and N. Doulamis, Vision-Based Maritime Surveillance System Using Fused Visual Attention Maps and Online Adaptable Tracker, IEEE 14th International Workshop on Image Analysis for Multimedia Interactive Services, 1 (2013).

[27] Hou Xiao-di, Jonathan Harel and Christof Koch, IEEE Transactions on Pattern Analysis and Machine Intelligence **34**, 194 (2011).

[28] Guo Shao-jun, Lou Shu-li and Liu Feng, Chinese Journal of Liquid Crystals and Displays **31**, 1006 (2016). (in Chinese)

[29] Lv Fu-qiang, Wang Hui and Liu Hong, Journal of Shanghai Jiaotong University **46**, 1920 (2012). (in Chinese)

[30] Rob Wijnhoven, Kris van Rens, Egbert G.T. Jaspers and Peter H.N. de With, Online Learning for Ship Detection in Maritime Surveillance, Symposium on Information Theory in the Benelux, 73 (2010).

[31] Liu Ge, Zhang Ya-sen, Zheng Xin-wei, Sun Xian, Fu Kun and Wang Hong-qi, IEEE Geoscience and Remote Sensing Letters **11**, 617 (2014).

[32] Lin Hao-ning, Shi Zhen-wei and Zou Zheng-xia, IEEE Geoscience and Remote Sensing Letters **14**, 1665 (2017).