IEICE TRANS. INF. & SYST., VOL.E105–D, NO.6 JUNE 2022

PAPER

An Attention Nested U-Structure Suitable for Salient Ship Detection in Complex Maritime Environment

Weina ZHOU[†a)], **Ying ZHOU**[†], ***Nonmembers***, ***and*** **Xiaoyang ZENG**[††], ***Member***

SUMMARY Salient ship detection plays an important role in ensuring the safety of maritime transportation and navigation. However, due to the influence of waves, special weather, and illumination on the sea, existing saliency methods are still unable to achieve effective ship detection in a complex marine environment. To solve the problem, this paper proposed a novel saliency method based on an attention nested U-Structure (AU^2Net). First, to make up for the shortcomings of the U-shaped structure, the pyramid pooling module (PPM) and global guidance paths (GGPs) are designed to guide the restoration of feature information. Then, the attention modules are added to the nested U-shaped structure to further refine the target characteristics. Ultimately, multi-level features and global context features are integrated through the feature aggregation module (FAM) to improve the ability to locate targets. Experiment results demonstrate that the proposed method could have at most 36.75% improvement in F-measure (F_{avg}) compared to the other state-of-the-art methods.

key words: *salient ship detection, attention nested U-structure, attention modules, feature aggregation*

1. Introduction

Ship detection is essential for the management and monitoring of maritime transportation and navigation, however, it is still a challenging problem of salient object detection (SOD), which aims to extract the obvious region from the background. Until now, although many SOD methods have been proposed, few of them could effectively be used in salient ship targets detection. That's because the ship targets are always affected by noises like waves, ripples, light reflection, and so on. What's more, missed detection will usually occur when different-scaled ship targets appear in the same scene, due to that the existing model will pay more attention to the large ship target and ignore the smaller ones.

The traditional saliency detection task mainly relied on hand-crafted features, such as color contrast [1], texture difference [2], and center prior [3]. With the development of deep CNNs, the fully convolutional network (FCN)-based network [4]–[6] has become the mainstream framework for SOD. Li et al. [7] concatenate different scale features and generate pyramid features. Pang et al. [8] adopt an aggregate interaction module to integrate features from adjacent levels and use a self-interaction module to gain more efficient multi-scale features from the integrated features. Wei et al. [9] choose a selective fusion strategy to aggregate multi-level important features by element-wise multiplication and refine low-level and high-level features. Wu et al. [10] present to use the random forest to extract spatial and intensity features. Gao et al. [11] utilize a multi-scale context fusion scheme to fuse the attentive features adaptively. Zhao et al. [12] propose an image-scale-symmetric cooperative network (IS2CNet) with hierarchical feature integration and bi-directional, which can effectively combine hierarchical features to gradually optimize the homogeneous region detection. However, the FCN-based methods cannot extract effective feature information from different layers, and cannot make low-level features and high-level features complement each other. In addition to FCN-based methods, U-shape based structures [13] gradually receive more attention due to their ability to utilize the multi-level information obtained from the top-down path to construct feature maps. Ueda et al. [14] improve the detection performance by changing the U-shaped structure of the encoder and applying the loss function considering the balance between classes. Wu et al. [15] incorporate stacked cross refinement units (CRUs) with the typical U-Net structures to refine multi-level features of salient object detection and edge detection simultaneously. Zhao et al. [16] present an edge guidance network (EGNet) to focus on the complementarity between salient edge information and salient object information. Zhou et al. [17] propose feature aggregation network FANet with region enhanced module (REM) to differentiate the salient regions and backgrounds. Although these saliency detection methods based on U-Net have achieved better performance in some aspects, they still could not perform well in the complex maritime scenario.

In that case, Cruz et al. [18] use two strategies to separate the foreground from the complex background. Cane et al. [19] present a maritime object detection and tracking algorithm to reduce false detections from wake and reflections. Wang et al. [20] adopt antijitter spatiotemporal saliency generation with parallel binarization (ASSGPB) methods to detect maritime targets in strong sun glitters. Liu et al. [21] propose an enhanced CNN-enabled learning method to improve ship detection in maritime video surveillance. However, these models always pay more attention to the large-scale ships in the foreground rather than the small ones. Recently, some detection methods for small ships based on the CNN have emerged. Liu et al. [22] present Rotated Region based CNN for ship detection in remote

Manuscript received August 29, 2021.
Manuscript revised January 15, 2022.
Manuscript publicized March 23, 2022.

[†]The authors are with the Information Engineering College, Shanghai Maritime University, China.
[††]The author is with the State Key Laboratory of ASIC and System, Fudan University, China.
a) E-mail: wnzhou@shmtu.edu.cn

DOI: 10.1587/transinf.2021EDP7181

Copyright © 2022 The Institute of Electronics, Information and Communication Engineers

sensing images. Zhang et al. [23] propose a new ship detection algorithm based on range-Doppler (RD) images. Zhang et al. [24] propose an S-CNN-Based ship detector that combines specially designed proposals extracted from the ship model with an improved saliency detection method. Wang et al. [25] present a single-channel SAR image unsupervised ship detection method based on multi-scale saliency and complex signal kurtosis (MSS-CSK). However, these researches are mainly based on remote sensing images and radar images, and could not be applied in saliency target detection.

Based on the above analysis, this paper proposes an improved attention nested U-Structure (AU^2Net) for salient targets detection in a complex environment. The entire network structure includes a nested U-shaped structure, channel attention module (CA), spatial attention module (SA), pyramid pooling module (PPM), global guidance paths (GGPs), and feature aggregation module (FAM). First, U^2Net is used as the backbone of the network, it can make the network deeper and achieve high-resolution target detection without significantly increasing memory and computational costs. Second, the attention modules are added to the nested U-shaped structure to capture detailed information about the target. Third, FAM integrates high-level semantic features, low-level detailed features, and global context features. Finally, the global context features are generated by PPM and transmitted by GGPs to accurately locate salient targets at each stage. The experimental results show that, compared with state-of-the-art methods based on saliency, the method proposed in this paper could achieve the best performance on salient ship target detection in the maritime scenario.

2. Proposed Method

The AU^2Net proposed in this paper can make ship positioning more accurate and the detected salient areas more complete. The overall pipeline is shown in Fig. 1.

We build a framework based on the [26], the right side of the network structure is the saliency probability map generated by the decoders at all levels. The network framework mainly includes codecs (6 encoders (En-1, ..., En-6) and 5 decoders (De-1, ..., De-5)) and PPM modules, and each level of codecs are embedded with a new U-shaped structure (shown in Fig. 2). The nested U-shaped structure not only could retain features similar to U-net [10] structure but also could extract multi-scale features from all levels of codecs without reducing the resolution of the feature map. The attention modules CA and SA are added after every convolution in all levels of codecs. In addition, by adding CA and SA after En-1, ..., En-6 encoders, the attention mechanism proposed in this paper can enable the network to quickly locate ship targets and extract the distinctive features of ship targets. After the En-1, ..., En-6 encoder structure, this paper also introduces PPM [27] to capture the context information, and pass it to the decoders of all levels through up-sampling GGPs, so as to further refine the feature map of each level to locate the salient target accurately. This paper also designs a FAM that can be used to integrate high-level semantic features, low-level detailed information features, and global context features. The module is set between the decoders of two adjacent stages and is used to aggregate the output characteristics of the encoder, the output characteristics of the decoder, and the context information, so as to recover the complete salient ship information. Finally, the six saliency probability maps generated by En-6, De-5, De-4, De-3, De-2, and De-1 are cascaded to fuse the saliency maps.

2.1 Attention Mechanism

Although the U-shaped network has the advantage of a lightweight structure, its computing power is limited. To allocate computing resources to relatively important tasks, that is, to detect all ship targets under limited resources, this paper accepts attention mechanism and pays attention to ship targets by adding spatial attention module SA and channel attention module CA. The attention mechanism has a powerful ability to select and refine features, so it is very suitable for saliency detection. Researchers have made many successful attempts in related areas. Wu et al. [28] adopt a holistic attention module to enlarge the coverage

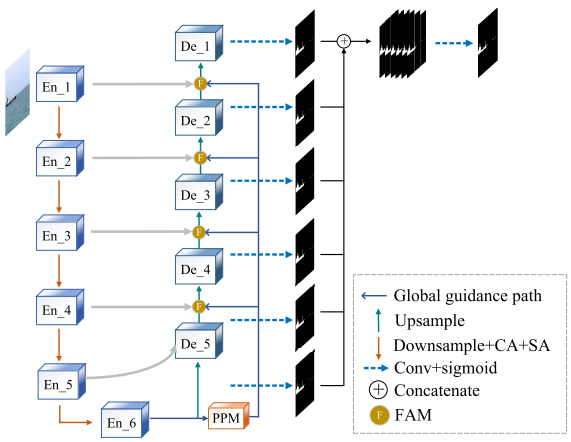


Fig. 1 The overall pipeline of our proposed method.

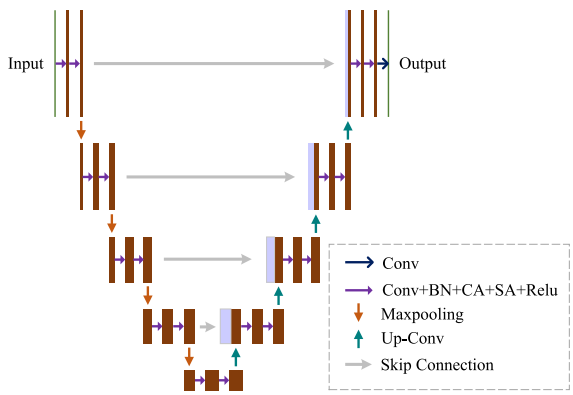


Fig. 2 Details of the nested U-shaped structure.

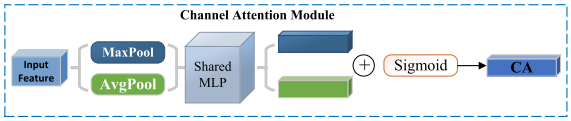


Fig. 3 Diagram of the channel attention (CA) module.

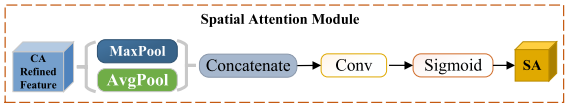


Fig. 4 Diagram of the spatial attention (SA) module.

area of the initial saliency map. Zhang et al. [29] utilize the spatial and channel attention mechanisms in the progressive attention guided network to generate layer-wise attentive features. Zhao et al. [30] combine spatial attention with edge information, focusing on effective low-level features, and selects the appropriate scale and receptive field on the channel-wise attention to generate saliency regions. Therefore, this paper uses SA and CA modules to focus on the detailed features and position information of the ship in the image.

To effectively calculate the CA, the spatial size of the input feature map needs to be squeezed. So far, average pooling has been widely used to aggregate spatial information, while max-pooling can collect another important clue about unique object features to infer finer channel-wise attention [31]. Therefore, both the average-pooled and max-pooled features are used in the proposed network structure in this paper. The details of the CA module are as follows:

It can be seen from Fig. 3 that, in CA module, the average pooling and max pooling operations are used to aggregate the spatial information of the feature mapping, and then the two generated spatial context descriptors will be forwarded to the shared multilayer perceptron (MLP) respectively. Next, after merging the output feature using an element-wise summation, the final CA feature map is generated through a sigmoid activation operation. The CA is computed as:

$$F_c(f) = \sigma(MLP(AvgPool(f)) + MLP(Maxpool(f))) \quad (1)$$

Where σ denotes the sigmoid function, f represents an intermediate feature map, and F_c is the CA map. $AvgPool$ and $MaxPool$ denote average pooling and max pooling operations respectively.

To calculate SA, we apply global max pooling and global average pooling operations along the channel axis and concatenate them to generate effective feature descriptors. Applying pooling operations along the channel axis could highlight the information area effectively. Then the average-pooled features and max-pooled features are concatenated and convolved by the standard convolution layer to generate an SA map. The detailed diagram of the SA module is shown in Fig. 4. The SA is computed as:

$$F_s(f) = \sigma(f^{7\times7}([Avgpool(f); Maxpool(f)])) \quad (2)$$

Where $f^{7\times7}$ represents a convolution operation with the filter size of 7×7 and F_s is the SA map.

2.2 Global Guidance Paths

In the classic FCN structure, there is insufficient contextual information of the scene, which leads to poor processing in the segmentation of objects of different scales. Thus, we introduce the PPM to capture contextual information and then pass it to each level through GGPs to help the feature map of each level determine the location of salient objects.

Different from the conventional information guidance using interpolation operations, GGPs in this paper mainly transfer the acquired context information to the decoders of each level by using up-sampling operation. It could help the feature maps of each level determine the location of salient objects.

Considering that the conventional interpolation operation uses adjacent pixels to restore the target pixel value, it will be more suitable for images with less information missing. However, since the feature map generated during the sampling process of the U-shaped network structure could only restore part of the original information, the conventional interpolation operation will cause image distortion. Therefore, the up-sampling operation is selected in this paper to restore the original resolution mask to obtain a better information guidance effect, which can make full use of most of the information in the feature map.

2.3 Feature Aggregation Module

Generally speaking, low-level features have rich detailed information but contain more background noises. While high-level features contain more semantic information but could only locate ship targets roughly. Recently, papers [5], [32], [33] build some simple framework structures, combining shallow and deep CNN features to capture low-level spatial details and high-level semantic information respectively. This fusion method can compensate for the defects between different layers, and ultimately improve the detection rate. Such a multi-level feature fusion mechanism is widely used in edge detection [34] and semantic segmentation [13]. Zhang et al. [35] utilize Amulet to integrate multi-level feature maps into multiple resolutions. Chen et al. [36] believe that the existing archaic fusion is incompetent for saliency detection in complex scenes, especially when overcoming multi-scale salient objects. The existing methods just simply concatenate multi-level features or element-wise addition of different layers for feature fusion, ignoring the gap between different features. Based on this, this paper proposes a FAM that integrates high-level semantic features, low-level detail features, and global context features at the same time. Through the cascading fusion of more features,

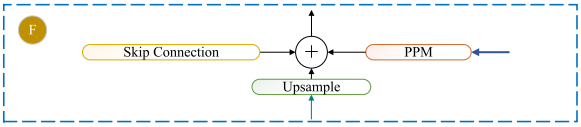


Fig. 5 Diagram of the feature aggregation module (FAM).

the network can suppress noise more effectively and detect a complete saliency ship silhouette. The detailed diagram of the FAM is shown in Fig. 5. The FAM is computed as:

$$F_{fam} = concat(f_{En-i}, f_{ppm}, f_{De-i+1}) \tag{3}$$

Where f_{ppm} denotes the global context feature generated by PPM, f_{En-i} and f_{De-i+1} represent the different levels of features generated by the encoder and decoder, where $i = 1, \ldots, 4$. $concat()$ represents the concatenation operation of the output feature of the codec and the global context feature.

2.4 Loss Function

In the training process, this paper adopts a deep supervision method similar to [26]. Our entire training loss is defined as:

$$\mathcal{L} = \omega_{fuse}\ell_{fuse} + \sum_{i=1}^{6} \omega_i \ell_i \tag{4}$$

Where ℓ_i is the loss of the i^{th} output saliency maps and ℓ_{fuse} is the loss of the fusion output saliency map. ω_{fuse} represents the weight of the fusion item and ω_i denotes the weight of the loss item of six different outputs. Each item in ℓ_i uses standard binary cross-entropy to calculate the loss:

$$\begin{aligned}\ell = -\sum_{(r,c)}^{(H,W)} [&P_{G(r,c)} \log P_{S(r,c)} \\ &+ (1 - P_{G(r,c)}) \log (1 - P_{S(r,c)})]\end{aligned} \tag{5}$$

Where (r, c) denotes the pixel coordinates and (H, W) is the height and width of the image. $P_{G(r,c)}$ and $P_{S(r,c)}$ represent the pixel values of the ground truth (GT) and predicted saliency probability maps respectively. During the test process, we choose the fusion map that can minimize the loss $\mathcal{L}$ as the final prediction map.

3. Experiment Results and Discussions

3.1 Datasets

At present, few existing datasets could be directly used for salient ship detection in marine. The dataset used in this paper is constructed from the Singapore Maritime Dataset (SMD) [37], which is open and commonly used in the maritime field. The constructed dataset contains 700 images with various marine environments, which are also manually labeled by our team. Figure 6 shows some ship images and their GT in our ship dataset. It can be seen from Fig. 6 that all the images in the dataset include ship targets and complex backgrounds. The ships are always disturbed by waves, ripples, weather, lighting, buoys, etc. In the experiment, data enhancement method is used to generate a total of 3,500 pictures such as the horizontal/vertical flip and rotation. For performance evaluation, 2800 pictures are randomly selected for training, and the left 700 pictures are used for test.

Fig. 6 Some ship images and their GT in our ship dataset.

3.2 Evaluation Criteria

To evaluate the pros and cons of each model comprehensively, the evaluation indicators used in this paper include F-measure (F_β), Weighted F-measure (F_β^ω), E-measure (E_m), Mean Absolute Error (MAE), and S-measure (S_m). F_β is defined as the weighted harmonic mean of precision and recall:

$$F_\beta = \frac{(1+\beta^2)(Precision \times Recall)}{\beta^2 Precision + Recall} \tag{6}$$

Where β^2 is set to 0.3 as suggested in [38] to emphasize precision rather than recall. The formula for precision and recall are as follows:

$$Precision = \frac{sumA(S,G)}{sumB(S)} \tag{7}$$

$$Recall = \frac{sumA(S,G)}{sumB(G)} \tag{8}$$

Where $sumA(S, G)$ is the result of multiplying and adding the values of the corresponding pixels of the saliency map and the GT map, $sumB(S)$ represents the sum of all pixel values on the saliency map, and $sumB(G)$ represents the sum of all pixel values on the GT map. Similar to previous work [6], [9], [26], the maximum F-measure (F_{max}) from the PR curve is also used as one of the evaluation indicators. To be fair, we also compute the average F-measure (F_{avg}) by using an adaptive threshold twice the predicted average value.

To solve the problems of interpolation flaw, dependency flaw, and the equal-importance flaw in existing metrics, F_β^ω defines the weighted Precision, and weighted Recall to improve the existing metric F-measure [39]. It is defined as:

$$F_\beta^\omega = \frac{(1+\beta^2)(Precision^\omega \cdot Recall^\omega)}{\beta^2 Precision^\omega + Recall^\omega} \tag{9}$$

E_m combines local pixel values with the image-level mean value to capture the two properties (pixel-level matching and image-level statistics) of a binary map evaluating the

foreground map and noise [40]. MAE is calculated based on the average per-pixel difference between the normalized saliency map S and the ground truth G:

$$MAE = \frac{1}{W \times H} \sum_{x=1}^{W} \sum_{y=1}^{H} |S(x,y) - G(x,y)| \tag{10}$$

Where W and H are the width and height of the saliency map, respectively.

The structural similarity between the S_m prediction and the ground truth is closer to the human visual system than F-measure. Therefore, we incorporate S_m into a more comprehensive evaluation. S_m is defined as:

$$S_m = \alpha \times S_0 + (1 - \alpha) \times S_r \tag{11}$$

Where α is set to 0.5 [41], S_0 and S_r denote the object-aware and region-aware structural similarity.

Table 1 Comparative experiment between networks with different modules

Model	F_{max}	F_{avg}	F_β^ω	E_m	S_m	MAE
U^2Net	0.899	0.894	0.894	0.977	0.911	0.0052
+PGM	0.902	0.890	0.922	0.986	0.929	0.0040
+FAM	0.920	0.916	0.920	0.986	0.929	0.0041
+CA-SA	0.927	0.925	0.928	0.992	0.993	0.0041
+PGM+FAM	0.923	0.920	0.924	0.988	0.932	0.0039
Ours	**0.935**	**0.934**	**0.936**	**0.993**	**0.936**	**0.0035**

3.3 Implementation Details

The proposed framework is implemented based on PyTorch. All training and test images are uniformly resized to 512 × 512. All the experiments are performed using the Adam optimizer [42] and its hyperparameters are set to default (initial learning rate lr=1e-3, betas=(0.9, 0.999), eps=1e-8, weight decay=0). The network was trained for a total of 180 epochs on the NVIDIA TeslaT4 GPU. And the test speed is 6 FPS on NVIDIA GTX1070 GPU.

3.3.1 Ablation Analysis

To demonstrate the effectiveness of each proposed module, we conduct a comparison between methods with different modules. As shown in Table 1, when PGM (PPM-GGPs),

Table 2 Comparison results with existing methods

Model	F_{max}	F_{avg}	F_β^ω	E_m	S_m	MAE
CPD	0.918	0.804	0.881	0.944	0.930	0.0057
EGNet	0.933	0.799	0.891	0.939	**0.939**	0.0052
F^3Net	0.917	0.842	0.893	0.966	0.926	0.0055
MINet	0.921	0.860	0.903	0.974	0.928	0.0052
R^2Net	0.904	0.683	0.822	0.865	0.914	0.0077
PoolNet	0.920	0.778	0.871	0.929	0.931	0.0060
Ours	**0.935**	**0.934**	**0.936**	**0.993**	0.936	**0.0035**

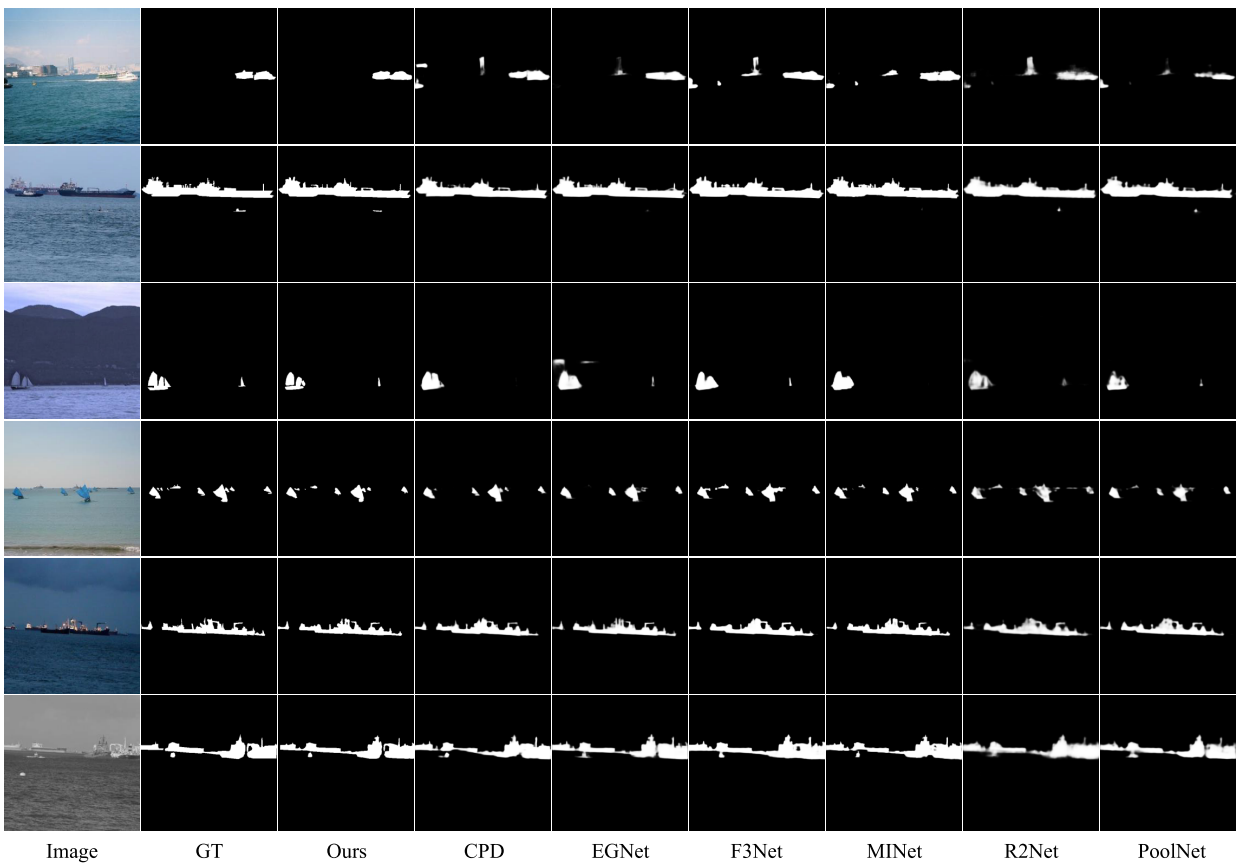


Fig. 7 Visual comparisons of different methods.

FAM, and attention modules CA and SA are added to the backbone network, all the indicators in the experimental results will be improved to varying degrees. However, compared to the method of introducing PGM and FAM separately, introducing PGM and FAM into the backbone network at the same time can better improve network performance. In the case when both PGM and FAM have been added, introducing the attention module can further improve the network in refining and extracting ship features to obtain the best results. Based on these ablation experiments, it is not difficult to conclude that the PGM, FAM, and attention modules proposed in this paper are beneficial to the detection of salient ship targets in complex sea conditions.

3.3.2 Comparison with State-of-the-Arts

We compare the proposed framework with six state-of-the-art saliency detection methods, including the CPD [28], EGNet [16], F^3Net [9], MINet [8], R^2Net [43], PoolNet [44]. To ensure the consistency of the comparison, all the recurring methods are trained by the constructed ship dataset, and the comparison results are shown in Table 2. From this table, we can see that the F_{max}, F_{avg}, F_{β}^{ω} and E_m values of our method are all above 0.9, which are all higher than the other existing methods respectively. And our method also obtains the smallest MAE, which means our predicted value is the closest to the real situation. In addition, although the S_m index value obtained by the method proposed in this paper is slightly lower than that of EGNet, the method proposed in this paper performs well on most evaluation indexes based on all the experimental results. Obviously, the method proposed in this paper has an excellent performance in salient ship detection tasks. In Fig. 7, we show the qualitative comparison of some representative challenging cases in complex ocean scenes. These scenes are interferenced by buoys (1^{st} and 6^{th} rows), buildings or mountains (1^{st} and 3^{rd} rows), small ships (2^{nd}, 3^{rd} and 4^{th}), wake waves (1^{st} row), severe weather (6^{th}), light changes (5^{th}), dense targets (2^{nd}, 5^{th} and 6^{th}), and other complex marine environments. Moreover, these images usually contain small-/middle- and large-scale objects simultaneously. By comparing between the results, it is not difficult to find out that the proposed algorithm is closer to the ground truth map than other methods, and it can consistently produce more accurate and complete saliency maps with sharp boundaries and coherent details. With the interference of buildings, buoys, and waves in the test scene, the method proposed in this paper not only highlights the important ship area but also suppresses the background noise well. In addition, in multi-target detection, all the smaller ship targets also have been detected with the interference of noise such as wave tailing.

4. Conclusion

In the paper, a novel attention nested U-Structure (AU^2Net) is proposed for salient target detection in complex environments. To improve the detection rate of the network and overcome the problem of missed detection of small targets during multi-target detection, this paper introduces an attention mechanism into the nested U-shaped structure and adds SA and CA modules to capture the target's information quickly. A PPM module is added after the encoder of the outer U-shaped structure to generate global context features, which are transmitted to the outer decoder by GGPs. In addition, the FAM module is used on the decoder side to integrate multi-level features and global context features to accurately locate salient targets at each stage. The experimental results on the ship dataset show that the proposed framework outperforms other existing saliency methods in six evaluation criteria in detecting ship targets in complex maritime scenarios. And in the future, we will try to further improve the speed of the method while keeping its accuracy.

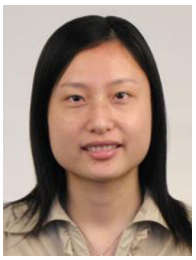
Acknowledgments

This study was supported by the National Natural Science Foundation of China (Grant No. 52071200, 61404083) and State Key Laboratory of ASIC and System (2021KF010). And the authors would like to thank Dr. Kun Wang for his advice to this Article, and thank Wanyu Song, Xiu Wang, and Zhongkun Zhang for manually labeled images in the experiments.

References

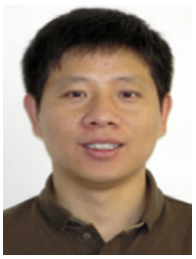
[1] M.-M. Cheng, N.J. Mitra, X. Huang, P.H. Torr, and S.-M. Hu, "Global contrast based salient region detection," IEEE Trans. Pattern Anal. Mach. Intell., vol.37, no.3, pp.569–582, 2014.
[2] Q. Yan, L. Xu, J. Shi, and J. Jia, "Hierarchical saliency detection," Proc. IEEE conference on computer vision and pattern recognition, pp.1155–1162, 2013.
[3] Z. Jiang and L.S. Davis, "Submodular salient region detection," Proc. IEEE Conference on Computer Vision and Pattern Recognition, pp.2043–2050, 2013.
[4] N. Liu and J. Han, "Dhsnet: Deep hierarchical saliency network for salient object detection," Proc. IEEE conference on computer vision and pattern recognition, pp.678–686, 2016.
[5] Q. Hou, M.-M. Cheng, X. Hu, A. Borji, Z. Tu, and P.H. Torr, "Deeply supervised salient object detection with short connections," Proc. IEEE conference on computer vision and pattern recognition, pp.3203–3212, 2017.
[6] N. Liu, J. Han, and M.-H. Yang, "Picanet: Pixel-wise contextual attention learning for accurate saliency detection," IEEE Trans. Image Process., vol.29, pp.6438–6451, 2020.
[7] Z. Li and F. Zhou, "Fssd: feature fusion single shot multibox detector," arXiv preprint arXiv:1712.00960, 2017.
[8] Y. Pang, X. Zhao, L. Zhang, and H. Lu, "Multi-scale interactive network for salient object detection," Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp.9413–9422, 2020.
[9] J. Wei, S. Wang, and Q. Huang, "F^3net: Fusion, feedback and focus for salient object detection," Proc. AAAI Conference on Artificial Intelligence, pp.12321–12328, 2020.

[10] J. Wu, G. Li, H. Lu, and T. Kamiya, "A supervoxel classification based method for multi-organ segmentation from abdominal ct images," Journal of Image and Graphics, vol.9, no.1, 2021.
[11] G. Gao, G. Xu, Y. Yu, J. Xie, J. Yang, and D. Yue, "Mscfnet: A lightweight network with multi-scale context fusion for real-time semantic segmentation," arXiv preprint arXiv:2103.13044, 2021.
[12] F. Zhao, H. Lu, W. Zhao, and L. Yao, "Image-scale-symmetric cooperative network for defocus blur detection," IEEE Trans. Circuits Syst. Video Technol., 2021.
[13] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," International Conference on Medical image computing and computer-assisted intervention, pp.234–241, Springer, 2015.
[14] A. Ueda, H. Lu, and T. Kamiya, "Deep-learning based segmentation algorithm for defect detection in magnetic particle testing images," Proc. International Conference on Artificial Life and Robotics, pp.235–238, 2021.
[15] Z. Wu, L. Su, and Q. Huang, "Stacked cross refinement network for edge-aware salient object detection," Proc. IEEE/CVF International Conference on Computer Vision, pp.7264–7273, 2019.
[16] J.X. Zhao, J.-J. Liu, D.-P. Fan, Y. Cao, J. Yang, and M.-M. Cheng, "Egnet: Edge guidance network for salient object detection," Proc. IEEE/CVF International Conference on Computer Vision, pp.8779–8788, 2019.
[17] X. Zhou, H. Wen, R. Shi, H. Yin, J. Zhang, and C. Yan, "Fanet: Feature aggregation network for rgbd saliency detection," Signal Processing: Image Communication, vol.102, p.116591, 2021.
[18] G. Cruz and A. Bernardino, "Image saliency applied to infrared images for unmanned maritime monitoring," International Conference on Computer Vision Systems, pp.511–522, Springer, 2015.
[19] T. Cane and J. Ferryman, "Saliency-based detection for maritime object tracking," Proc. IEEE conference on computer vision and pattern recognition workshops, pp.18–25, 2016.
[20] B. Wang, Y. Motai, L. Dong, and W. Xu, "Detecting infrared maritime targets overwhelmed in sun glitters by antijitter spatiotemporal saliency," IEEE Trans. Geosci. Remote Sens., vol.57, no.7, pp.5159–5173, 2019.
[21] R.W. Liu, W. Yuan, X. Chen, and Y. Lu, "An enhanced cnn-enabled learning method for promoting ship detection in maritime surveillance system," Ocean Engineering, vol.235, p.109435, 2021.
[22] Z. Liu, J. Hu, L. Weng, and Y. Yang, "Rotated region based cnn for ship detection," 2017 IEEE International Conference on Image Processing (ICIP), pp.900–904, IEEE, 2017.
[23] W. Zhang, Q.M.J. Wu, Y. Yang, T. Akilan, W.G.W. Zhao, Q. Li, and J. Niu, "Fast ship detection with spatial-frequency analysis and anova-based feature fusion," IEEE Geosci. Remote Sens. Lett., vol.19, pp.1–5, 2022.
[24] R. Zhang, J. Yao, K. Zhang, C. Feng, and J. Zhang, "S-cnn-based ship detection from high-resolution remote sensing images.," International Archives of the Photogrammetry, Remote Sensing & Spatial Information Sciences, vol.41, 2016.
[25] Z. Wang, R. Wang, X. Fu, and K. Xia, "Unsupervised ship detection for single-channel sar images based on multiscale saliency and complex signal kurtosis," IEEE Geosci. Remote Sens. Lett., vol.19, pp.1–5, 2022.
[26] X. Qin, Z. Zhang, C. Huang, M. Dehghan, O.R. Zaiane, and M. Jagersand, "U2-net: Going deeper with nested u-structure for salient object detection," Pattern Recognition, vol.106, p.107404, 2020.
[27] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," Proc. IEEE conference on computer vision and pattern recognition, pp.2881–2890, 2017.
[28] Z. Wu, L. Su, and Q. Huang, "Cascaded partial decoder for fast and accurate salient object detection," Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp.3907–3916, 2019.
[29] X. Zhang, T. Wang, J. Qi, H. Lu, and G. Wang, "Progressive attention guided recurrent network for salient object detection," Proc. IEEE Conference on Computer Vision and Pattern Recognition, pp.714–722, 2018.
[30] T. Zhao and X. Wu, "Pyramid feature attention network for saliency detection," Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp.3085–3094, 2019.
[31] S. Woo, J. Park, J.-Y. Lee, and I.S. Kweon, "Cbam: Convolutional block attention module," Proc. European conference on computer vision (ECCV), pp.3–19, 2018.
[32] G. Li, Y. Xie, L. Lin, and Y. Yu, "Instance-level salient object segmentation," Proc. IEEE Conference on Computer Vision and Pattern Recognition, pp.2386–2395, 2017.
[33] H. Xiao, J. Feng, Y. Wei, M. Zhang, and S. Yan, "Deep salient object detection with dense connections and distraction diagnosis," IEEE Trans. Multimedia, vol.20, no.12, pp.3239–3251, 2018.
[34] J. He, S. Zhang, M. Yang, Y. Shan, and T. Huang, "Bi-directional cascade network for perceptual edge detection," Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp.3828–3837, 2019.
[35] P. Zhang, D. Wang, H. Lu, H. Wang, and X. Ruan, "Amulet: Aggregating multi-level convolutional features for salient object detection," Proc. IEEE International Conference on Computer Vision, pp.202–211, 2017.
[36] S. Chen, X. Tan, B. Wang, and X. Hu, "Reverse attention for salient object detection," Proc. European Conference on Computer Vision (ECCV), pp.234–250, 2018.
[37] D.K. Prasad, D. Rajan, L. Rachmawati, E. Rajabally, and C. Quek, "Video processing from electro-optical sensors for object detection and tracking in a maritime environment: A survey," IEEE Trans. Intell. Transp. Syst., vol.18, no.8, pp.1993–2016, 2017.
[38] R. Achanta, S. Hemami, F. Estrada, and S. Susstrunk, "Frequency-tuned salient region detection," 2009 IEEE conference on computer vision and pattern recognition, pp.1597–1604, IEEE, 2009.
[39] R. Margolin, L. Zelnik-Manor, and A. Tal, "How to evaluate foreground maps?," Proc. IEEE conference on computer vision and pattern recognition, pp.248–255, 2014.
[40] D.P. Fan, C. Gong, Y. Cao, B. Ren, M.M. Cheng, and A. Borji, "Enhanced-alignment measure for binary foreground map evaluation," arXiv preprint arXiv:1805.10421, 2018.
[41] D.-P. Fan, M.-M. Cheng, Y. Liu, T. Li, and A. Borji, "Structure-measure: A new way to evaluate foreground maps," Proc. IEEE international conference on computer vision, pp.4548–4557, 2017.
[42] D.P. Kingma and J. Ba, "Adam: A method for stochastic optimization," arXiv preprint arXiv:1412.6980, 2014.
[43] M. Feng, H. Lu, and Y. Yu, "Residual learning for salient object detection," IEEE Trans. Image Process., vol.29, pp.4696–4708, 2020.
[44] J.-J. Liu, Q. Hou, M.-M. Cheng, J. Feng, and J. Jiang, "A simple pooling-based design for real-time salient object detection," Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp.3917–3926, 2019.



Weina Zhou received the M.S. degree in electronic information engineering from Shanghai Maritime University, China, in 2006, and the Ph.D. degree from the State Key Laboratory of ASIC and System, Microelectronics Department, Fudan University, China, in 2012. From 2014 to 2017, she worked as a post-doctor researcher at the Computer Science Department, Fudan University, P.R. China. And since 2006, She has been a lecture/associate professor in Information Engineering College, Shanghai Maritime University. Her research interests include object detection algorithms and ASIC design.

Ying Zhou received the B.S. degree in Internet of Things Engineering from the Jiangsu Second Normal University in 2015-2019. She is currently working toward the M.S. degree in Information and Communication Engineering with the Shanghai Maritime University. Her main research interests include image processing and deep learning.



Xiaoyang Zeng received the B.S. degree from Xiangtan University, Xiangtan, China, in 1992, and the Ph.D. degree from the Changchun Institute of Optics and Fine Mechanics, Chinese Academy of Sciences, Beijing, China, in 2001. Since 2007, he has been a Full Professor and the Director with the State Key Laboratory of ASIC and System, Fudan University, Shanghai, China, where he was a Postdoctoral Researcher from 2001 to 2003, and later was an Associate Professor. His research interests include information security chip, VLSI signal processing, communication systems design, image and video signal processing. He is the Steering Committee Member of the Asia and South Pacific Design Automation Conference (ASP-DAC), and the TPC member of the IEEE Asian Solid-State Circuits Conference (A-SSCC).