IGARSS 2024 - 2024 IEEE International Geoscience and Remote Sensing Symposium | 979-8-3503-6032-5/24/$31.00 ©2024 IEEE | DOI: 10.1109/IGARSS53475.2024.10642328

EDGE COMPLEMENTARY MULTI-SCALE AGGREGATION NETWORK FOR SALIENT OBJECT DETECTION IN OPTICAL REMOTE SENSING IMAGES

Bei Cheng[1], *Zao Liu*[1], *Chengbiao Fu*[1], *Tao Shen*[1,*]
[1] School of Information Engineering and Automation, Kunming University of Science and Technology of China
chengbei@kust.edu.cn, liuzao@stu.kust.edu.cn, fcb@kust.edu.cn, *shentao@kust.edu.cn

ABSTRACT

In recent years, salient object detection (SOD) has attracted more and more attention. However, the SOD in remote sensing images (RSI-SOD) faces various issues, including large scene span, cluttered background and changeable object scale. To address these challenges, an edge complementary multi-scale aggregation network (ECMANet) is proposed in this paper. Specifically, a multi-scale feature aggregation module (MFAM) is designed to extract hierarchical multi-scale information and reduce the noise interference of different scale information. In addition, foreground edge guidance module (FEGM) is designed to cross-refine foreground information and edge information. Finally, the foreground, edge, and background are generated by background-foreground fusion module (BFFM) to complement the overall network information. Extensive experiments are conducted on two popular datasets demonstrate that the proposed method outperforms other state-of-the-art methods.

Index Terms— Salient object detection, Edge Complementary, Multi-scale, remote sensing

1. INTRODUCTION

Salient object detection (SOD) aims at locating and segmenting the most visually attractive areas in an image, and it has shown successful applications in various computer vision tasks, such as semantic segmentation [1], image quality assessment [2], change detection [3] and so on. Concurrently, SOD method has made vigorous development in natural images. However, due to the variable object scale and large scene span in remote sensing images (RSI), the salient object detection in remote sensing images (RSI-SOD) is facing more challenges than that in natural images.

In recent years, researchers have begun to focus on RSI-SOD. Zhang *et al.* [4] proposed a global context-aware attention module, which was used to capture the remote semantic context adaptively. Cong *et al.* [5] proposed a parallel multi-scale attention scheme to recover detailed information Wang *et al.* [6] proposed a method of multi-scale feature integration with the explicit and implicit assistance for salient target detection. This method explicitly expresses rich multi-scale depth features by integrating significant edge clues. However, most of the above methods ignore multi-scale information of objects of different sizes, and it is easy to cause feature redundancy in the integration process.

On the other hand, edge information can produce a clear and accurate salience map. Zhou *et al.* [7] introduced an additional decoder to detect edge features, which was then used to guide the decoder to accurately locate salient objects. Further, Gong *et al.* [8] performs multi-scale high-level feature integration with the help of the significant edge extraction module and skeleton extraction module to further improve the accuracy of the significance map. Luo *et al*. [9] proposed a semantics-edge interaction model, which enable close interaction between semantic and edges. However, most of these methods adopt a single edge detection method and the foreground information closely related to the edge is ignored which either cannot make full use of the edge information, or the edge information generated is too rough to fully extract the boundary information of significant objects.

To solve the above problems, an edge complementary multi-scale aggregation network (ECMANet) is proposed in this paper. Firstly, a multi-scale feature aggregation module (MFAM) is designed to extend the depth in-formation of the encoder to different scales, and it uses up-and-down information propagation to integrate and branch across parallel volumes, to effectively keep the local details of salient objects. This module makes each branch learn different features from other branches and reduces redundant features when it effectively adapts to targets of different sizes. At the same time, the foreground edge guidance module (FEGM) is proposed to cross-refine foreground information and edge information to achieve complementary effect, realize accurate extraction of edge information. Finally, Background-Foreground Fusion Module (BFFM) is designed to merge the whole scene information. ECMANet achieves balance by integrating hierarchical multi-scale features and foreground edge perception, enhances salient regional features, and retains edge details.

*Corresponding author. The research work was supported by the Open Foundation of Yunnan Key Laboratory of Computer Technology Application and National Natural Science Foundation of China, No. 42067029.

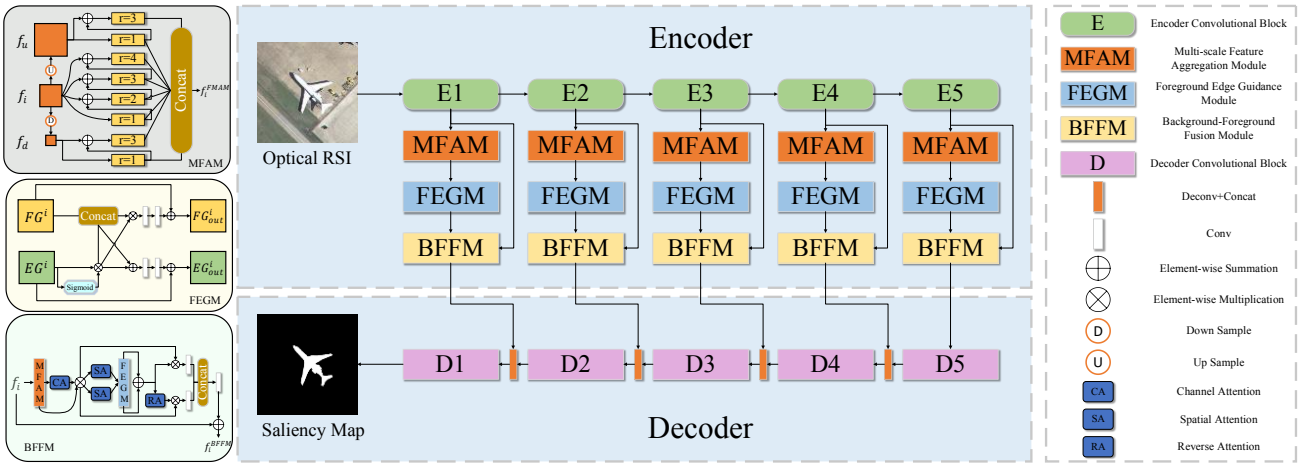

Fig. 1: Overall framework of the proposed ECMANet.

2. PROPOSED METHOD

2.1. Network Overview

The framework of ECMANet is illustrated in Fig. 1. Firstly, the encoder is composed of modified VGG-16, that is, the last four layers of convolutional blocks are removed. Then, the enhancer will extract multi-scale information and generate foreground, edge, and background features. Finally, our decoder network infers salient objects through a progressive resolution restoration process.

2.2. Multi-Scale Feature Aggregation Module

Inspired by [10], we propose MFAM to extract features of different scales and reduce the interference of noise. Initially, it consists of eight parallel dilated convolutions, where f_i is the input image. Up sampling and down sampling are designed to obtain a larger feature map f_u and a smaller feature map f_d The generated two feature maps pass through dilated convolutions with dilation rates $r = 1$ and $r = 3$, respectively, while the original feature map f_i passes through dilated convolutions with dilation rates $r = 1,2,3,4$. Furthermore, for dilated convolutions with dilation rates other than $r = 1$, we add their feature maps to the result of dilated convolution of the previous layer to form a residual cascade structure. The process can be summarized as follows:

$$\begin{aligned} f_u^r &= \mathrm{DiConv}_r(f_u), r \in \{1\} \\ f_u^r &= \mathrm{DiConv}_r(f_u + f_u^1), r \in \{3\} \\ f_i^r &= \mathrm{DiConv}_r(f_i), r \in \{1\} \\ f_i^r &= \mathrm{DiConv}_r(f_i + f_i^{r-1}), r \in \{2,3,4\} \\ f_d^r &= \mathrm{DiConv}_r(f_d), r \in \{1\} \\ f_d^r &= \mathrm{DiConv}_r(f_d + f_d^1), r \in \{3\} \\ f_{out}^i &= \mathrm{Concat}(f_u^1, f_u^3, f_i^1, f_i^2, f_i^3, f_i^4, f_d^1, f_d^3) \end{aligned} \quad (1)$$

where DiConv_r is a dilated convolution with dilation rate r.

2.3. Foreground Edge Guidance Module

Given the certain correlation between foreground features and edge features, we propose FEGM to guide the edge information and foreground information for mutually refine each other. The refining process can be calculated as follows:

$$EG_{\text{gate}}^i = \mathrm{Sigmoid}(EG^i) \otimes EG^i \quad (2)$$

$$FG_{out}^i = FG^i \oplus \mathrm{Conv}_{3\times3}\left(\mathrm{Conv}_{3\times3}(\mathrm{Concat}(FG^i, EG_{\text{gate}}^i) \otimes EG^i)\right) \quad (3)$$

$$EG_{out}^i = EG^i \oplus \mathrm{Conv}_{3\times3}\left(\mathrm{Conv}_{3\times3}(\mathrm{Concat}(FG^i, EG_{\text{gate}}^i) \oplus EG^i)\right) \quad (4)$$

where EG^i represents the edge map, FG^i represents the foreground map, $\oplus$ is the elementwise summation, and $\otimes$ is the elementwise multiplication.

2.4. Background-Foreground Fusion Module

To address the complexity of scenes in RSI, inspired by [11], we propose BFFM to establish the relationship among foreground, background and edges, which can jointly guide the detection of RSI-SOD.

The original feature map is output as f_i^{MFAM} after passing through MFAM, and then generates foreground map a_f^{MFAM} and the edge map a_e^{MFAM} through channel attention and spatial attention. The edge map E_i is supervised by the true edge map G_e, where G_e is obtained based on paper [12]. Finally, a_f^{MFAM} and a_e^{MFAM} are cross-refined and fused to realize the mutual complement a_f^{FEGM} and a_e^{FEGM}. Different from [11], BFFM realizes the complementarity of foreground and edge through cross thinning, which can make full use of the correlation of objects in the scene, as follows:

$$\begin{aligned} f_{\text{ca}}^{MFAM} &= \mathrm{CA}(f_i^{MFAM}) \odot f_i^{MFAM} \\ a_f^{MFAM} &= \mathrm{SA}(f_{\text{ca}}^{MFAM}) \\ a_e^{MFAM} &= \mathrm{SA}(f_{\text{ca}}^{MFAM}) \\ a_{fe}^{FEGM} &= a_f^{FEGM} \oplus a_e^{FEGM} \end{aligned} \quad (5)$$

Table 1: Quantitative comparison of our method with seven other methods. ↑ and ↓ denote larger and smaller, respectively. The top three results are marked in red, blue, and green, respectively.

	ORSSD						EORSSD					
	$S_\alpha \uparrow$	$F_\beta^{\max} \uparrow$	$F_\beta^{\text{adp}} \uparrow$	$E_\xi^{\max} \uparrow$	$E_\xi^{\text{adp}} \uparrow$	$\mathcal{M} \downarrow$	$S_\alpha \uparrow$	$F_\beta^{\max} \uparrow$	$F_\beta^{\text{adp}} \uparrow$	$E_\xi^{\max} \uparrow$	$E_\xi^{\text{adp}} \uparrow$	$\mathcal{M} \downarrow$
DAFNet_{21}[4]	0.9191	0.8928	0.7876	0.9771	0.9360	0.0113	0.9166	0.8614	0.6427	0.9861	0.8446	0.0060
SARNet_{21}[14]	0.9134	0.8850	0.8512	0.9557	0.9464	0.0187	0.9240	0.8719	0.8304	0.9620	0.9536	0.0099
RRNet_{21}[5]	0.9311	0.9011	0.8252	0.9722	0.9479	0.0104	0.9266	0.8781	0.7251	0.9716	0.9034	0.0075
CorrNet_{22}[15]	0.9380	0.9129	0.8875	0.9790	0.9721	0.0098	0.9289	0.8778	0.8311	0.9696	0.9593	0.0083
ERPNet_{22}[7]	0.9254	0.8974	0.8356	0.9710	0.9520	0.0135	0.9210	0.8632	0.7554	0.9603	0.9228	0.0089
CRNet_{23}[16]	0.9389	0.9107	0.8695	0.9793	0.9711	0.0091	0.9370	0.8873	0.8140	0.9743	0.9563	0.0063
SEINet_{23}[9]	0.9382	0.9090	0.8816	0.9771	0.9682	0.0097	0.9283	0.8788	0.8587	0.9723	0.9678	0.0076
SeaNet_{23}[17]	0.9260	0.8942	0.8625	0.9767	0.9670	0.0098	0.9208	0.8649	0.8304	0.9710	0.9651	0.0073
Ours	0.9461	0.9161	0.8879	0.9814	0.9728	0.0096	0.9391	0.8935	0.8474	0.9758	0.9679	0.0064

where CA denotes the channel attention, $\odot$ is the channel-wise multiplication, and SA is the spatial attention.

In BFFM, the background attention opposite to the foreground edge attention is obtained by inversion. This and the foreground edge attention are multiplied with the channel-attention processed feature map f_{ca}^{MFAM}, enhancing foreground and background aspects. The process can be calculated as follows:

$$\begin{aligned} f_{fe} &= a_{fe}^{FEGM} \otimes f_{ca}^{MFAM} \\ a_b^{FEGM} &= 1 \ominus a_{fe}^{FEGM} \\ f_b &= a_b^{FEGM} \otimes f_{ca}^{MFAM} \\ f_i^{BFFM} &= \text{Conv}_{3\times3}\left(\text{Concat}\left(f_{fe}, f_b\right) \oplus f_i\right) \end{aligned} \tag{6}$$

Where 1 refers to a matrix of the same size as a_{fe}^{FEGM}, and all elements are 1, $\ominus$ is the element-wise subtraction.

3. EXPERIMENTS

3.1. Dataset and implementation Details

The datasets of ORSSD [13] and EORSSD [4] are used in this experiment. ORSSD contains 800 images and EORSSD contains 2000 images. We trained our method on the PyTorch platform using an NVIDIA RTX 4060Ti GPU (16GB memory). The batch size is 8 and initial learning rate is 1e−4, which was divided by 10 after 30 epochs. For the dataset, we trained for 38 epochs.

For comprehensive and fair evaluation, the widely used metrics are applied in this paper, including S-measure (S_α), max F-measure (F_β^{max}), adaptive F-measure (F_β^{adp}), max E-measure (E_β^{max}), adaptive E-measure (E_β^{adp}) and mean absolute error (MAE, $\mathcal{M}$)

3.2. Comparison with State-of-the-arts

The proposed method is compared with seven RSI-SOD methods, including DAFNet [4], SARNet [14], RRNet [5], Cornnet [15], ERPNet [7], CRNet [16], SEINet [9] and SeaNet [17]. The results are shown in Table 1. It can be seen that the proposed method achieves better scores on two datasets. Fig. 2 shows a visual comparison of the saliency maps predicted by different methods. The proposed method can accurately locate the salient target and delineate the boundary. Additionally, we compared the Precision-Recall (PR) curves of various methods across two datasets. As illustrated in Fig. 3 and Fig. 4, our method consistently shows excellent performance. In summary, the proposed method can extract more complete and clear significance objects in complex background.

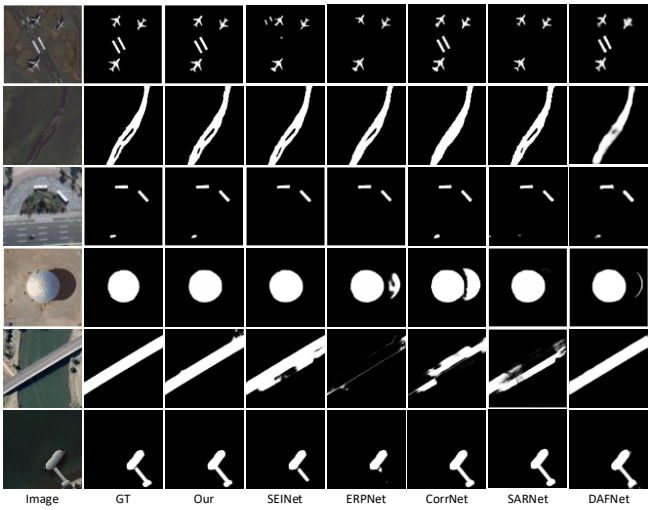

Fig. 2: Visual comparison between our results and other methods.

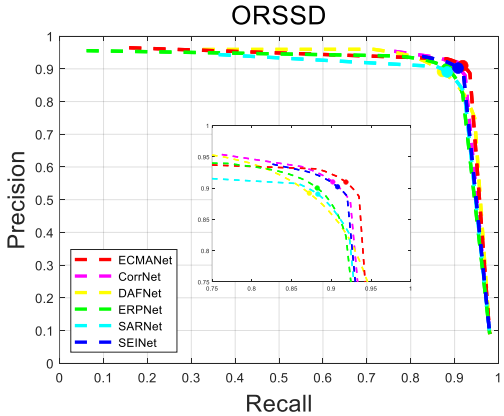

Fig. 3: Comparison of PR curves of ours with other SOTA methods on ORSSD.

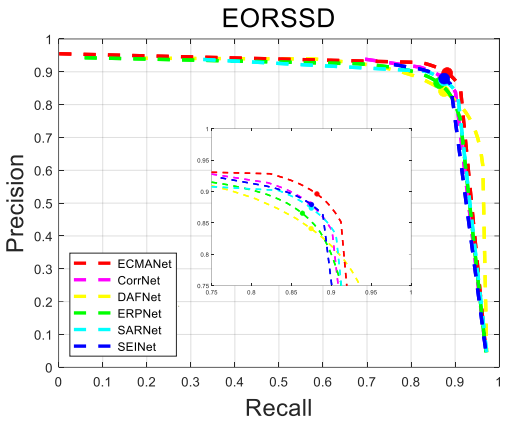

Fig. 4: Comparison of PR curves of ours with other SOTA methods on EORSSD.

4. CONCLUSION

This paper presents a novel ECMANet to detect salient objects in RSI. The key is to solve the feature redundancy problem of multi-scale fusion and integrate multiple contents in the scene. MFAM can reduce the noise interference of different scale information while extracting hierarchical multi-scale information. FEGM guides the edge information and foreground information to cross-refine each other for obtain clearer boundary features. BFFM integrates multiple contents including background to make full use of the information in the scene. These components adaptively complement each other and can accurately detect salient objects in optical RSI. Experimental results on two public datasets confirm the superiority of ECMANet.

5. REFERENCES

[1] Zeng Y, Zhuge Y, Lu H, et al., "Joint learning of saliency detection and weakly supervised semantic segmentation," in Proc. *IEEE Int. Conf. Comput. Vis*., 2019, pp. 7222–7232.

[2] Yang S, Jiang Q, Lin W, et al., "SGDNet: An end-to-end saliency-guided deep neural network for no-reference image quality assessment," in Proc. *27th ACM Int. Conf. Multimedia*., Oct. 2019, pp. 1383–1391.

[3] Feng Y, Xu H, Jiang J, et al., "ICIF-Net: Intra-scale cross-interaction and inter-scale feature fusion network for bitemporal remote sensing images change detection," *IEEE Trans. Geosci. Remote Sens*., vol. 60, 2022, Art. no. 4410213.

[4] Zhang Q, Cong R, Li C, et al., "Dense attention fluid network for salient object detection in optical remote sensing images," *IEEE Trans. Image Process*., vol. 30, pp. 1305–1317, 2021.

[5] Cong R, Zhang Y, Fang L, et al., "RRNet: Relational reasoning network with parallel multi-scale attention for salient object detection in optical remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5613311.

[6] Wang Z, Guo J, Zhang C, et al., "Multiscale feature enhancement network for salient object detection in optical remote sensing images,". *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5634819.

[7] Zhou X, Shen K, Weng L, et al., "Edge-guided recurrent positioning network for salient object detection in optical remote sensing images," *IEEE Trans. Cybern.*, vol. 53, no. 1, pp. 539–552, Jan. 2022.

[8] Gong A, Nie J, Niu C, et al., "Edge and Skeleton Guidance Network for Salient Object Detection in Optical Remote Sensing Images," *IEEE Transactions on Circuits and Systems for Video Technology.*, vol. 33, 2023.

[9] Luo, Huilan, and Bocheng Liang, "Semantic-Edge Interactive Network for Salient Object Detection in Optical Remote Sensing Images," *IEEE J. Sel. Top. Appl. Earth Obs. Remote. Sens*., vol. 16, pp. 6980-6994, 2023.

[10] Zeng X, Xu M, Hu Y, et al., "Adaptive Edge-Aware Semantic Interaction Network for Salient Object Detection in Optical Remote Sensing Images," *IEEE Trans. Geosci. Remote. Sens*., vol. 61, 2023, Art. no. 5617416.

[11] Li G, Liu Z, Lin W, et al., "Multi-content complementation network for salient object detection in optical remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5614513.

[12] Gu Y, Xu H, Quan Y, et al., "ORSI Salient Object Detection via Bidimensional Attention and Full-Stage Semantic Guidance," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5603213.

[13] Li C, Cong R, Hou J, Zhang S, Qian Y, et al., "Nested Network with Two-Stream Pyramid for Salient Object Detection in Optical Remote Sensing Images," *IEEE Trans. Geosci. Remote. Sens*., vol. 57, no. 11, 2019.

[14] Huang Z, Chen H, Liu B, et al., "Semantic-guided attention refinement network for salient object detection in optical remote sensing images," *Remote Sens.*, vol. 13, no. 11, pp. 2072–4292, 2021.

[15] GongyangLi Z, Bai Z, Lin W, et al., "Lightweight salient object detection in optical remote sensing images via feature correlation," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5601111.

[16] Sun L, Wang Q, Chen Y, et al., "CRNet: Channel-Enhanced Remodeling-Based Network for Salient Object Detection in Optical Remote Sensing Images," *IEEE Trans. Geosci. Remote. Sens*., vol. 61, 2023, Art. no. 5618314.

[17] Li G, Liu Z, Zhang X, et al., "Lightweight salient object detection in optical remote-sensing images via semantic matching and edge alignment," *IEEE Trans. Geosci. Remote Sens*., vol. 61, 2023, Art. no. 5601111.