Uncertainty-Aware Graph Reasoning With Global Collaborative Learning for Remote Sensing Salient Object Detection

Yanfeng Liu, *Student Member, IEEE*, Yuan Yuan, *Senior Member, IEEE*, and Qi Wang, *Senior Member, IEEE*

***Abstract*— Recently, fully convolutional networks (FCNs) have contributed significantly to salient object detection (SOD) in optical remote sensing images (RSIs). However, owing to the limited receptive fields of FCNs, accurate and integral detection of salient objects in RSIs with complex edges and irregular topology is still challenging. Moreover, suffering from the low contrast and complicated background of RSIs, existing models often occur ambiguous or uncertain recognition. To remedy the above problems, we propose a novel hybrid modeling approach, i.e., uncertainty-aware graph reasoning with global collaborative learning (UG2L) framework. Specifically, we propose a graph reasoning pipeline to model the intricate relations among RSI patches instead of pixels and introduce an efficient graph reasoning block (GRB) to build graph representations. On top of it, a global context block (GCB) with a linear attention mechanism is proposed to explore the multiscale and global context collaboratively. Finally, we design a simple yet effective uncertainty-aware loss (UAL) to enhance the model's reliability for better prediction of saliency or nonsaliency. Experimental and visual results on three datasets show the superiority of the proposed UG2L. Code is available at https://github.com/lyf0801/UG2L.**

***Index Terms*— Global collaborative learning, graph reasoning, optical remote sensing image (RSI), salient object detection (SOD), uncertainty-aware loss (UAL).**

I. Introduction

SALIENT object detection of optical remote sensing images (RSI-SOD) aims to locate visually attractive regions with irregular topology and produce accurate saliency maps from RSIs with complicated background [1], [2], [3], [4], serving as an essential and challenging preprocessing technique to foster other vision topics, such as object detection [5], [6] and semantic segmentation [7] in the remote sensing community.

In recent years, benefiting from the rapid development of fully convolutional networks (FCNs), SOD in natural scene images (NSIs) has experienced considerable progress and extensive research efforts. For instance, Luo et al. [8] propose a nonlocal deep feature modeling approach for SOD, which greatly facilitates the application of deep learning-based methods in this field. Hou et al. [9] introduce residual short connections with deep supervision paradigms to further explore the effectiveness of convolutional networks for SOD tasks. A reverse attention mechanism is designed for feature extraction of nonsalient regions in NSIs [10], while a pooling-based approach [11] is proposed to replace some convolutional processing steps for lightweight SOD. Recently, Ma et al. [12] propose a novel pyramidal shrinking network for multiscale SOD in a sophisticated feature interaction manner. The above-mentioned methods have greatly improved the detection performance of NSI-SOD, providing some valuable insights for the research on RSI-SOD models.

RSI-SOD needs to discover various man-made objects with accurate region boundaries which attract human attention, e.g., airplanes, storage tanks, roads, rivers, bridges, vehicles, lakes, vessels, and buildings. Owing to the unique imaging mechanism and variations of sizes, types, and arrangements of man-made objects in optical RSIs, RSI-SOD witnesses more challenging difficulties than NSI-SOD. To address these issues, many researchers have proposed a range of approaches based on FCNs [13], [14], [15]. For instance, Huang et al. [16] introduce an attention-based refinement mechanism in both spatial and channel aspects to guide semantic feature aggregation. MCCNet [17] proposes a multicontent complementation strategy to fuse salient features in different scales. RRNet [18] introduces graph convolution to build the relational reasoning module at the channel and spatial level to boost the performance, while bringing huge computational costs and parameter complexity. Recently, Wang et al. [13] propose a hybrid feature encoder by using the advantages of both FCNs and Transformers, and designing an adjacent feature alignment approach to combine the above components effectively.

The above algorithms have significantly contributed to this research topic, yet still remain some limitations. First, most of these methods are based on FCN paradigms, which are limited by the spatially fixed receptive fields and lack effective recognition of regions with complex edges and irregular topology. Second, they cannot capture the long-range dependence among various spatial regions and are also unable to interpret the correlations between salient foreground and background as well as the relations among different saliency objects effectively and efficiently. Although RRNet [18] and HFANet [13] introduce graph convolutional networks (GCNs) and Transformers for feature mapping at the pixel level, respectively, their complicated structure and numerous parameters are not conducive to practical applications. Third, existing models often suffer from ambiguous and uncertain predictions for saliency and nonsaliency, which greatly restricts the reliability of models. Thus, more focused research efforts need to be further explored to meet the above open and particular issues.

Manuscript received 3 April 2023; revised 13 June 2023; accepted 24 July 2023. Date of publication 27 July 2023; date of current version 15 August 2023. This work was supported by the National Natural Science Foundation of China under Grant U21B2041, Grant U1864204, and Grant 61825603. *(Corresponding author: Qi Wang.)*

Yanfeng Liu is with the School of Computer Science and School of Artificial Intelligence, Optics and Electronics (iOPEN), Northwestern Polytechnical University, Xi'an 710072, China (e-mail: liuyanfeng99@gmail.com).

Yuan Yuan and Qi Wang are with the School of Artificial Intelligence, Optics and Electronics (iOPEN), Northwestern Polytechnical University, Xi'an 710072, China (e-mail: y.yuan1.ieee@gmail.com; crabwq@gmail.com).

Digital Object Identifier 10.1109/LGRS.2023.3299245

To this end, we propose the **U**ncertainty-aware **G**raph reasoning with **G**lobal collaborative **L**earning (UG2L) for RSI-SOD in this letter. First of all, we do not adopt the paradigm of FCNs to construct the saliency encoder. Instead, we introduce a novel hybrid modeling approach. Specifically, a feature-patches-oriented graph reasoning pipeline is proposed to model the intricate relations among RSI patches instead of pixels, which can better represent the semantic independence of irregular topological objects. Unlike pixel-oriented GCNs that cope with large input items (i.e., pixels) and complicated parameters, our graph reasoning blocks (GRBs) have two advantages as follows: 1) we use convolutional layers to extract shallow features of RSIs and explore more flexible structural correlations at deep layers oriented to image patches using the proposed GRBs, which are more generalized and alleviate numerous parameters and computational costs; and 2) we insert fully connected layers and feed-forward block into GRB for feature transformation and enhance its nonlinear capability and feature diversity, making it adaptive to RSI-SOD tasks. On top of it, to explore long-distance relationships, a global context block (GCB) with atrous pyramid pooling layers and a novel linear global attention mechanism is proposed for multiscale and global context learning. Finally, based on ambiguity theory, we design a simple yet effective uncertainty-aware loss (UAL) as a solid auxiliary constraint of binary cross-entropy loss to enhance the confidence and reliability of the model for better saliency or nonsaliency prediction.

The contributions of this letter are presented as follows.

1) We propose a hybrid modeling UG2L, which utilizes convolution as local encoder in shallow stages and GRBs in deep stages oriented to image patches to better learn the salient objects of scale variety and irregular topology.
2) We propose GCB, an efficient module to obtain multiscale contexts while modeling long-distance relations with linearly global attention for complex backgrounds.
3) A simple yet effective UAL is designed to guide the model's training phase and foster the accurate prediction of blurred edges or sophisticated background pixels.

II. Methodology of the Proposed Framework

A. Overview of the Proposed UG2L

As shown in Fig. 1(a), the proposed UG2L employs an encoder-decoder framework and performs multilevel supervision with a pyramidal decoder. Inspired by the inductive bias of convolutional networks, we equip the first three stages of ResNet50 as shallow parts of the encoder to produce the local contexts named $S_1 - S_3$. On top of it, we introduce visual graph modules [19] to construct GRB and stack it as the deep stages to produce the graph-based context named S_4 and S_5. Besides, to explore the long-distance independence, GCB is connected at the end of the encoder, feeding S_3 as input, and outputs the global context as S_6. Thus, our encoder can capture the local context, the graph-based context, and the global context in a hybrid and collaborative manner. We adopt the classical pyramid decoder, i.e., upsampling and channel concatenation, to generate multiscale saliency maps as $P_1 - P_5$, where P_1 is the final prediction. In the training phase, UAL is utilized to boost original binary cross-entropy loss and facilitate uncertain or ambiguous pixel prediction.

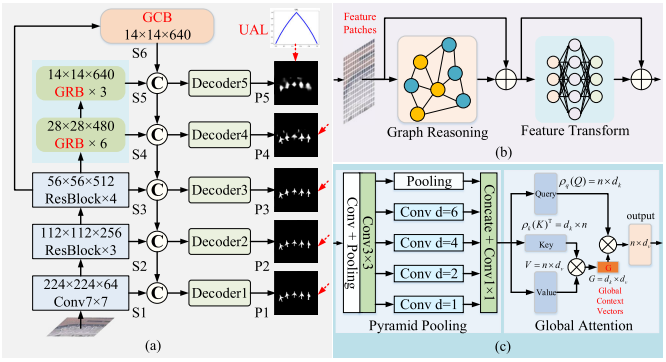

Fig. 1. (a) Framework of the proposed UG2L. (b) Framework of GRB. (c) Framework of GCB.

B. GRB With RSI Patches as Nodes

The fixed and local receptive fields of convolution cannot explore the dependency among different parts and edges of saliency regions, especially for objects with complex edges and irregular topology. GCNs usually repeat graph convolutions with huge parameters and thus cannot efficiently model the graph relations among each pixel in RSIs, but can reduce computational budgets and work well in modeling image patches. To more flexibly represent salient objects in optical RSIs, we first introduce the graph reasoning mechanism as the deep part of the encoder and deploy image patches as nodes to generate graph relations as revealed in [19], further improving the representation of various semantic categories.

As illustrated in Fig. 1(b), dissimilar to the original GCNs, the introduced GRB adds nonlinear activation functions as feature transformation after max-relative graph convolution [20] to increase the feature diversity. Assuming that the input is X, the processing function of GRB can be expressed as follows:

$$Y = \sigma(\text{Graph}(X \cdot W_1))W_2 \oplus X \tag{1}$$

where W_1 and W_2 indicate the weights of fully connected layers, Graph$(\cdot)$ denotes the function of max-relative graph convolution, σ represents GELU activation function, and $\oplus$ is element-wise summation. Additionally, a two fully connected layer-based multilayer perceptron, is employed as a feed-forward module on nodes, to boost the feature transformation capacity, as well as alleviate over-smoothing, as follows:

$$Z = \sigma(Y \cdot W_3)W_4 \oplus Y \tag{2}$$

where W_3 and W_4 are the weights of fully connected layers, and Z is the final output of a GRB. As revealed in Fig. 1(a), a stack of GRBs constitutes the block which serves as the basic stage for constructing the deep part of the encoder. Compared with the basic GCNs, the introduced GRB illustrates feature patches as nodes for graph representation and can maintain the diversity of features to learn discriminative description.

C. GCB With Efficient Linear Attention

Multiscale context, as well as global context, is critical for recognizing various objects and extracting semantic features in optical RSIs with complex background [5], as Chen et al. [7] propose dilated convolutions to facilitate semantic segmentation. However, the direct integration of multiscale contexts may weaken the expressiveness of salient regions due to the interference of nonsalient backgrounds. For this reason, we propose to impose global spatial attention on multiscale

context to focus more on visually attractive regions as well as to capture long-distance independence. Specifically, we introduce a linearly lightweight attention mechanism to project global context vectors [21] and enhance multiscale features, building a GCB as shown in Fig. 1(c).

GCB, feeding S_3 as input, consists of convolutions with a pooling layer for downsampling, atrous pyramid pooling layers for multiscale context modeling, and efficient linear attention for global context learning, collaboratively. First, S_3 performs nonlinear mapping after several convolutions, is fed into a 4×4 max-pooling layer for downsampling, and then passes through a 3×3 convolution to obtain the intermediate feature as $F \in \mathbb{R}^{256\times14\times14}$. After that, it enters the multiscale context learning phase with five parallel branches consisting of a global pooling layer and four convolutions with dilation rates [1, 2, 4, 6], and the processing function is defined as follows:

$$F_m = \mathcal{C}_1\left[\mathcal{C}_3^1(F), \mathcal{C}_3^2(F), \mathcal{C}_3^4(F), \mathcal{C}_3^6(F), \mathcal{P}_g(F)\right] \tag{3}$$

where F_m is the multiscale context, $\mathcal{C}_1(\cdot)$ denotes the function of 1×1 convolution, $\mathcal{C}_3^i(\cdot)$ indicates the process of 3×3 convolution with dilation of i, $\mathcal{P}_g(\cdot)$ represents global average pooling, and $[\cdot, \cdot]$ is the channel concatenation, respectively.

To efficiently capture global spatial attention and focus on visually salient areas, we impose the efficient attention into GCB to model global context vector [21]. As shown in Fig. 1(c), we first reshape F_m into $\mathbb{R}^{n\times d}$, and utilize three linear layers to produce the query $Q \in \mathbb{R}^{n\times d_k}$, key $K \in \mathbb{R}^{n\times d_k}$, value $V \in \mathbb{R}^{n\times d_v}$, the same as self-attention. Differently, GCB does not deploy these features for self-attention, but as weights of all spatial locations to generate a global context vector as $G = \rho_k(K)^T V$, a global description of the input. The global attention mechanism is expressed as follows:

$$E(Q, K, V) = \rho_q(Q)(\rho_k(K)^T V) \tag{4}$$

where $\rho_q(\cdot)$ and $\rho_k(\cdot)$ indicate the normalization functions which combine scaling and softmax as below

$$\rho_q(Q) = \text{softmax}(Q/\sqrt{n}), \quad \rho_k(K) = \text{softmax}(K/\sqrt{n}). \tag{5}$$

Compared with self-attention, GCB changes the computational complexity from $O(n^2)$ terms to linear complexity $O(d_k d_v)$, which dramatically reduces the number of operations and is more efficient. For a better adaptive aggregation of multiscale contexts and global contexts, we introduce residual learning to derive the final S_6, which is defined as follows:

$$S_6 = F_m \oplus \gamma \cdot E(Q, K, V) \tag{6}$$

where γ is a differentiable factor during the training phase.

D. Multisupervision With UAL

In general, due to the challenges and complexity of optical RSIs, the model cannot accurately perceive the complete boundary of salient objects and generates severe ambiguity when trained only under cross-entropy loss, and thus decrease the reliability of SOD. Inspired by the prediction uncertainty [22], to enhance the ability of RSI-SOD models for probability distribution and confidence of saliency and nonsaliency, as well as penalizing inaccurate pixel predictions, we design a simple yet effective auxiliary supervised loss, i.e., UAL, as a strong constraint to train the model.

The SOD model defines saliency probabilities ranging from 0 to 1, where predictions closer to 0 belong to the background and those closer to 1 belong to salient objects. Hence, if the predicted probability is closer to 0.5, then a pixel is more likely to be ambiguous and uncertain. Based on this observation, we define the Euclidean distance between probability and 0.5 as the uncertainty measure of the pixel, and gradually constrain these hard samples in the training phase. Obviously, the uncertainty measure function must have a minimum value of 0 at 0 and 1, a maximum value of 1 at 0.5, while taking 0.5 as its symmetry axis. In addition, it must be smooth and continuous with few nondifferentiable points. Here, we introduce the logarithmic function in base 2 to construct the proposed UAL, and define the uncertainty measure as $2 - |2p_{i,j} - 1|$ for a prediction $p_{i,j}$. Thus, UAL can be expressed as follows:

$$\mathcal{L}_{\text{UAL}} = \log_2(2 - |2p_{i,j} - 1|). \tag{7}$$

The curve of UAL is illustrated in Fig. 1(a), and the total loss function of ith prediction of decoder is denoted as follows:

$$\mathcal{L}_i = \mathcal{L}_{\text{BCE}} + \mathcal{L}_{\text{IoU}} + \lambda_{\text{iter}} \cdot \mathcal{L}_{\text{UAL}} \tag{8}$$

where $\mathcal{L}_{\text{BCE}}$ and $\mathcal{L}_{\text{IoU}}$ indicate the binary cross-entropy loss and IoU loss, respectively, and λ_{iter} is a dynamic factor to balance the learning process of the model, which we define the cosine increase strategy to optimize UAL as follows:

$$\lambda_{\text{iter}} = 0.5 \times (1 - \cos(\text{iter}/T \times \pi)) \tag{9}$$

where T is the total number of iterations and iter indicates the current iterative step during training. For the total loss function about five stages of the decoder, we apply a weighted summation for better balance and deep supervision, i.e.,

$$\mathcal{L}_{\text{total}} = \mathcal{L}_1 + \sum_{i=2}^{5}(\mathcal{L}_i/2^{i-2}). \tag{10}$$

III. Experiments and Analysis

A. Datasets, Metrics, and Implementation Details

We conduct experiments on three public datasets, namely, ORSSD [1], EORSSD [2], and ORSI-4199 [3], which contain 800, 2000, and 4199 RSIs, respectively. Three quantitative indicators are used to evaluate these models, i.e., MAE, F-measure (F_β), and S-measure (S_m). Furthermore, we show PR curves to reveal performance differences among all competitors. As for implementation, we follow the data augmentation strategy as previous work [13], [23] to reproduce all compared methods for a fair comparison, and train and test on three datasets separately. All competitors are deployed on an NVIDIA RTX 3090 GPU with Pytorch 1.8 toolbox, and all RSIs are uniform as 448×448. We employ the Adamw optimizer with a cosine learning rate scheduler to train our model, and the batch size, epochs, initial learning rate, momentum, and weight decay are 8, 300, 5e-4, 0.9, and 0.05, respectively.

B. Comparison With State-of-the-Art Methods

As illustrated in Fig. 2, it can be found that PR curves of UG2L marked in red are closer to the upper right corner on both ORSSD and EORSSD datasets. The above results indicate that the presented UG2L has better quantitative superiority among these competitors, and due to the space limitation, we do not show the curve of the ORSI-4199 dataset.

TABLE I

QUANTITATIVE RESULTS ON THREE DATASETS. TOP THREE RESULTS ARE HIGHLIGHTED IN RED, GREEN, AND BLUE, RESPECTIVELY

Methods	ORSSD Dataset [1]			EORSSD Dataset [2]			ORSI-4199 Dataset [3]		
	$F_\beta \uparrow$	MAE $\downarrow$	$S_m \uparrow$	$F_\beta \uparrow$	MAE $\downarrow$	$S_m \uparrow$	$F_\beta \uparrow$	MAE $\downarrow$	$S_m \uparrow$
NLDF [8]	0.8352	0.0267	0.8702	0.8060	0.0154	0.8706	0.7639	0.0584	0.8053
DSS [9]	0.8469	0.0268	0.8688	0.7921	0.0167	0.8371	0.7672	0.0561	0.8115
RAS [10]	0.8841	0.0185	0.8896	0.8636	0.0114	0.8800	0.7930	0.0595	0.8142
PoolNet [11]	0.8291	0.0268	0.8610	0.8121	0.0207	0.8279	0.7777	0.0573	0.8184
PFSNet [12]	0.9153	0.0101	0.9303	0.8979	0.0077	0.9287	0.8496	0.0374	0.8686
FSMINet [24]	0.8623	0.0178	0.8949	0.8527	0.0100	0.8989	0.8046	0.0451	0.8344
SARNet [16]	0.8963	0.0185	0.8976	0.8865	0.0102	0.9097	0.8309	0.0448	0.8536
DAFNet [2]	0.8717	0.0161	0.8982	0.8378	0.0106	0.8824	0.8169	0.0473	0.8477
RRNet [18]	0.8857	0.0142	0.9110	0.8511	0.0101	0.8964	0.8122	0.0448	0.8449
MCCNet [17]	0.9005	0.0212	0.9040	0.8976	0.0083	0.9306	0.8284	0.0439	0.8506
MJRBM-R [3]	0.9058	0.0129	0.9128	0.8685	0.0092	0.8980	0.8406	0.0379	0.8685
EMFINet-R [14]	0.9132	0.0107	0.9350	0.8972	0.0075	0.9286	0.8469	0.0352	0.8678
HFANet [13]	0.9224	0.0113	0.9324	0.9007	0.0082	0.9292	0.8419	0.0379	0.8659
UG2L (Ours)	0.9313	0.0095	0.9385	0.9102	0.0068	0.9357	0.8508	0.0333	0.8724

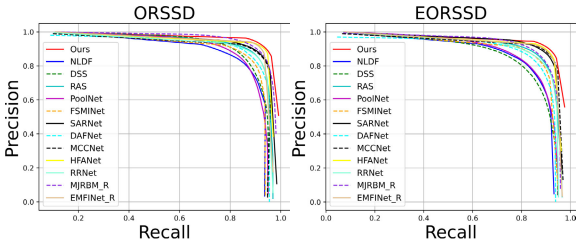

Fig. 2. PR curves of 13 models on the ORSSD and EORSSD datasets.

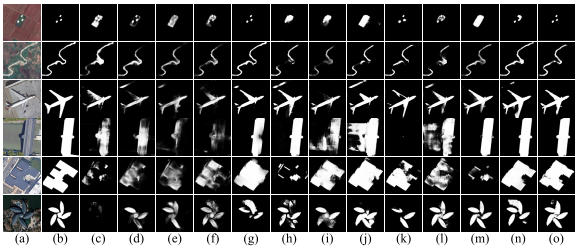

Fig. 3. Typical predicted results of our proposed UG2L with 12 models on the most challenging ORSI-4199 dataset. (a) RSIs. (b) GTs. (c) NLDF. (d) DSS. (e) RAS. (f) PoolNet. (g) FSMINet. (h) SARNet. (i) DAFNet. (j) MCCNet. (k) HFANet. (l) RRNet. (m) MJRBM. (n) EMFINet. (o) UG2L.

Besides, we report five NSI-SOD methods (NLDF [8], DSS [9], RAS [10], PoolNet [11], PFSNet [12]), and eight RSI-SOD methods (FSMINet [24], SARNet [16], DAFNet [2], MCCNet [17], HFANet [13], RRNet [18], MJRBM [3], and EMFINet [14]) for comparison in Table I. It can be observed that the most competitive methods are PFSNet, HFANet, and EMFINet, and UG2L reaches the best performance in terms of all metrics. We attribute this excellent performance to the novel design of UG2L, i.e., the hybrid modeling framework with graph representation on image patches, efficient global context learning, and the practical uncertainty-aware function.

Finally, some typical visual predicted results on the most challenging ORSI-4199 dataset are shown in Fig. 3. By comparison, it can be demonstrated that the proposed model can predict the most complete object regions with few ambiguous or blurred regions; however, other methods suffer from it.

C. Ablation Study

In this section, we conduct ablation studies to verify the effects of the proposed modules as presented in Fig. 4 and Table II. We find that the model's performance declines considerably by removing any of the introduced modules, including GCB, GRB, and UAL. Among them, GCB makes the most prominent performance gain, which also illustrates the significance of multiscale and global contexts. As shown in

TABLE II

PERFORMANCE ON ORSSD AND EORSSD FOR ABLATION STUDY

Ablation Methods	ORSSD Dataset [1]			EORSSD Dataset [2]		
	$F_\beta \uparrow$	MAE $\downarrow$	$S_m \uparrow$	$F_\beta \uparrow$	MAE $\downarrow$	$S_m \uparrow$
w/o global context branch	0.8991	0.0149	0.9164	0.8968	0.0082	0.9252
w/o graph relation branch	0.9200	0.0104	0.9290	0.8998	0.0075	0.9278
w/o uncertainty-aware loss	0.9217	0.0100	0.9308	0.9036	0.0071	0.9319
fully proposed UG2L	**0.9313**	**0.0095**	**0.9385**	**0.9102**	**0.0068**	**0.9357**

TABLE III

PERFORMANCE ON ORSSD AND EORSSD FOR MODEL ANALYSIS

Models	ORSSD Dataset [1]			EORSSD Dataset [2]		
	$F_\beta \uparrow$	MAE $\downarrow$	$S_m \uparrow$	$F_\beta \uparrow$	MAE $\downarrow$	$S_m \uparrow$
ResNet50	0.9081	0.0119	0.9197	0.8883	0.0080	0.9199
ResNet50 w/o UAL	0.9021	0.0122	0.9163	0.8863	0.0089	0.9184
ResNet50+GCB	0.9258	0.0104	0.9307	0.9010	0.0072	0.9275
ResNet50+GCB w/o UAL	0.9162	0.0113	0.9269	0.8945	0.0078	0.9263
fully proposed UG2L	**0.9313**	**0.0095**	**0.9385**	**0.9102**	**0.0068**	**0.9357**

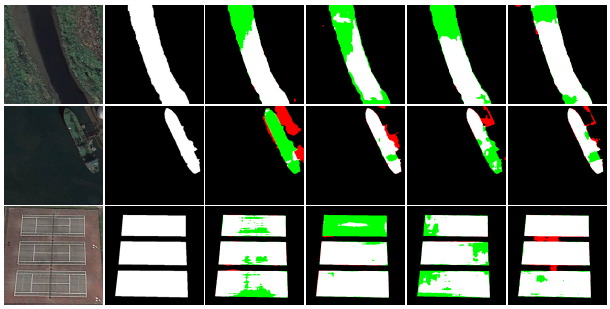

Fig. 4. Typical results of ablation models on the EORSSD dataset, where false positives and missing parts are marked in red and green, respectively. (a) RSIs. (b) GTs. (c) w/o GCB. (d) w/o GRB. (e) w/o UAL. (f) UG2L.

Fig. 4, we find that either the lack of global context modeling or graph representation modeling, the presented UG2L suffers from severely missed detections and thus loses accurate and complete detection of salient objects. Also, with the assistance of UAL, the model's ability to discriminate between saliency and nonsaliency is enhanced, i.e., the number of false positives and missed samples is greatly reduced by observation.

To further analyze the proposed encoder, we organize experiments to compare it with the generic encoder ResNet50 as reported in Table III. Among them, ResNet50 + GCM still does not perform as well as the proposed UG2L, which particularly illustrates the superiority of our introduced graph representation modeling with image patches as nodes.

D. Visual Analysis

In this subsection, we present more extensive visual results to verify the effects of the proposed modules. First, we show the model's results regarding saliency pixel values with and without UAL for training, as shown in Fig. 5. Clearly, with the help of UAL, the model predicts fewer sample values in [15, 240], i.e., it embodies closer to the binarized histogram of 0 and 255, indicating more confidence and stability of the model. Then, we draw the visual maps at different encoder stages as shown in Fig. 6. In contrast, S_3 derived from the convolutional layer focuses on the edges of salient objects and some locally detailed information, while S_6 derived from GCB can focus on the whole salient objects, which is the difference between the local receptive field and global context. With respect to Fig. 6(d), GRB can discard limited receptive

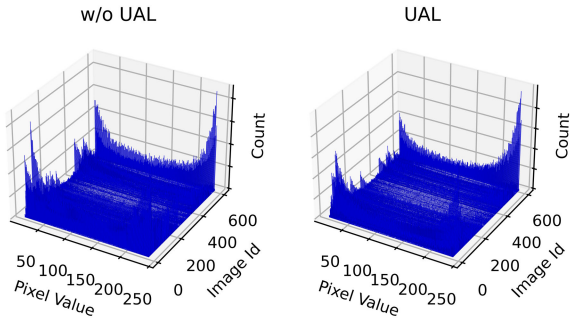

Fig. 5. Visual comparison of stacked histograms on EORSSD test dataset. Good results show as few samples as possible in the middle. For a clearer presentation, only intervals with pixel values in [15, 240] are presented here.

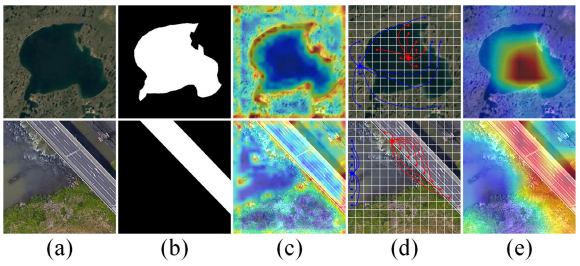

Fig. 6. Typical visual samples in different encoder stages. (a) RSIs. (b) GTs. (c) Feature maps in S_3. (d) Graph structures in S_5. (e) Feature maps in S_6.

TABLE IV
MODEL COMPLEXITY COMPARISON WITH RRNET

Methods	FLOPs (G)	Params (M)
RRNet [18]	692.15	86.27
UG2L (Ours)	**102.53**	**57.18**

fields and establish edge relations among image patches of the same semantic categories with more generalized ability.

Moreover, we compare another model that introduces graph representation learning, i.e., RRNet [18], in terms of FLOPs and Params, as revealed in Table IV. RRNet builds graph structures with each pixel as a graph node, which suffers from a huge number of parameters and computational costs, i.e., 692.15 G and 86.27 M, respectively. By contrast, we use 16×16 and 32×32 image patches as nodes, which greatly reduces the complexity of the model and achieves superior performance.

IV. CONCLUSION

In this letter, we propose UG2L for RSI-SOD, which introduces graph representation to model the intricate correlation among RSI patches and uncertainty theory to alleviate the ambiguous prediction of saliency, bringing several research insights for the remote sensing community. Experimental and visual analysis shows its effectiveness compared with state-of-the-art methods.

REFERENCES

[1] C. Li, R. Cong, J. Hou, S. Zhang, Y. Qian, and S. Kwong, "Nested network with two-stream pyramid for salient object detection in optical remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 11, pp. 9156–9166, Nov. 2019.

[2] Q. Zhang 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.

[3] Z. Tu, C. Wang, C. Li, M. Fan, H. Zhao, and B. Luo, "ORSI salient object detection via multiscale joint region and boundary model," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5607913.

[4] Z. Xiong, Y. Liu, Q. Wang, and X. X. Zhu, "RSSOD-Bench: A large-scale benchmark dataset for salient object detection in optical remote sensing imagery," in *Proc. IEEE Int. Geosci. Remote Sens. Symp. (IGARSS)*, Jul. 2023, pp. 1–4.

[5] Y. Liu, Q. Li, Y. Yuan, Q. Du, and Q. Wang, "ABNet: Adaptive balanced network for multiscale object detection in remote sensing imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5614914.

[6] Y. Liu, Q. Li, Y. Yuan, and Q. Wang, "Single-shot balanced detector for geospatial object detection," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2022, pp. 2529–2533.

[7] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 4, pp. 834–848, Apr. 2018.

[8] Z. Luo, A. Mishra, A. Achkar, J. Eichel, S. Li, and P.-M. Jodoin, "Non-local deep features for salient object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 6593–6601.

[9] Q. Hou, M.-M. Cheng, X. Hu, A. Borji, Z. Tu, and P. H. S. Torr, "Deeply supervised salient object detection with short connections," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 4, pp. 815–828, Apr. 2019.

[10] S. Chen, X. Tan, B. Wang, and X. Hu, "Reverse attention for salient object detection," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Sep. 2018, pp. 234–250.

[11] J.-J. Liu, Q. Hou, M.-M. Cheng, J. Feng, and J. Jiang, "A simple pooling-based design for real-time salient object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 3912–3921.

[12] M. Ma, C. Xia, and J. Li, "Pyramidal feature shrinking for salient object detection," in *Proc. AAAI Conf. Artif. Intell. (AAAI)*, May 2021, pp. 2311–2318.

[13] Q. Wang, Y. Liu, Z. Xiong, and Y. Yuan, "Hybrid feature aligned network for salient object detection in optical remote sensing imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5624915.

[14] X. Zhou, K. Shen, Z. Liu, C. Gong, J. Zhang, and C. Yan, "Edge-aware multiscale feature integration network for salient object detection in optical remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5605315.

[15] Y. Liu, Z. Xiong, Y. Yuan, and Q. Wang, "Distilling knowledge from super-resolution for efficient remote sensing salient object detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5609116.

[16] Z. Huang, H. Chen, B. Liu, and Z. Wang, "Semantic-guided attention refinement network for salient object detection in optical remote sensing images," *Remote Sens.*, vol. 13, no. 11, p. 2163, May 2021.

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

[18] R. Cong, Y. Zhang, L. Fang, J. Li, Y. Zhao, and S. Kwong, "RRNet: Relational reasoning network with parallel multiscale attention for salient object detection in optical remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5613311.

[19] K. Han, Y. Wang, J. Guo, Y. Tang, and E. Wu, "Vision GNN: An image is worth graph of nodes," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, Dec. 2022, pp. 8291–8303.

[20] G. Li, M. Müller, A. Thabet, and B. Ghanem, "DeepGCNs: Can GCNs go as deep as CNNs?" in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 9267–9276.

[21] Z. Shen, M. Zhang, H. Zhao, S. Yi, and H. Li, "Efficient attention: Attention with linear complexities," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Jan. 2021, pp. 3531–3539.

[22] Y. Pang, X. Zhao, T.-Z. Xiang, L. Zhang, and H. Lu, "Zoom in and out: A mixed-scale triplet network for camouflaged object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 2160–2170.

[23] Y. Liu, Z. Xiong, Y. Yuan, and Q. Wang, "Transcending pixels: Boosting saliency detection via scene understanding from aerial imagery," *IEEE Trans. Geosci. Remote Sens.*, early access, Jul. 25, 2023, doi: 10.1109/TGRS.2023.3298661.

[24] K. Shen, X. Zhou, B. Wan, R. Shi, and J. Zhang, "Fully squeezed multiscale inference network for fast and accurate saliency detection in optical remote-sensing images," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, 2022, Art. no. 6507705.