An Important Pick-and-Pass Gated Refinement Network for Salient Object Detection in Optical Remote Sensing Images

Mo Yang, Ziyan Liu, Wen Dong, and Ying Wu

***Abstract*—Salient object detection in optical remote sensing images (ORSI-SOD) is a very challenging task due to the complex scale, shape, details, uncertainty of the predicted location of the object, etc. In this article, we propose a novel important pick-and-pass gated refinement network model for SOD in ORSIs, named IP2GRNet, including a detail refinement stage and a feature pick-and-pass refinement stage. Specifically, the detail refinement stage, named refinement synchronization, uses the self-modality attention refinement module and the dynamic weight refinement module to accurately describe and capture approximate coordinate positions and feature information of salient objects. The feature pick-and-pass and refinement stage progressively picks and refines the prediction results in a coarse to a fine manner by combining high-level semantic information and low-level semantic information under the guidance of attention and counter-attention. In addition, the aggregation operation is used to fuse detail refinement information as well as mixture loss for supervised network training, which effectively improves the model performance from three perspectives: pixel, region, and statistics. Extensive experiments on two benchmarks datasets demonstrate that our proposed IP2GRNet, both qualitatively and quantitatively, outperforms the state-of-the-art saliency detectors.**

***Index Terms*—Detail refinement stage, important pick-and-pass gate, optical remote sensing images (ORSIs), salient object detection (SOD).**

I. Introduction

SALIENT object detection (SOD) is similar to people being attracted by some objects/regions in whole field of view [1], which imitates the visual attention system, and is used for handling visual subtasks, i.e., segmentation [2], recognition [3], tracking [4], and retrieval [5], etc. SOD as an essential image preprocessing process has been applied to natural scene images (NSIs) and optical remote sensing images (ORSIs), and encouraging results have been achieved [6]. Therefore, in this article, we focus on the salient object detection in optical remote sensing images (ORSI-SOD) method to solve the task of SOD in optical remote sensing, which will provide useful guidance and reference for many downstream remote sensing tasks.

Fig. 1. Visual illustration of SOD results in RSIs. (a) ORSIs. (b) GT. (c) Proposed IP2GRNet. (d) DAFNet. (e) CorrNet.

In the past decades, many traditional manual feature extraction SOD methods had made significant progress in NSIs, but the research of SOD in ORSIs is limited. Since ORSIs are taken from a high-altitude area, there are three major challenges in SOD tasks. First, the object scale changes greatly, and the RSIs-SOD model needs to be very sensitive to the scale change of salient objects, just as the four airplanes in the first row of Fig. 1. Second, the complex background brings more difficulties to detection. Third, because RSI is captured from the aerial view of the ground, such as the lake in the second row of Fig. 1, long and thin salient objects, such as rivers cross the image boundary, and the edges of aircraft parking details overlap, such as the river in the third row of Fig. 1, which brings greater challenges to the complete the detection of these objects. There are significant differences between NSI and ORSI in terms of captured devices, scene, and view orientation, which result in differences in resolution, object type, and object proportion. At the same time, this leads to unsatisfactory direct migration from the NSI-SOD model to the RSI-SOD model.

To solve these issues, some meaningful methods have been proposed to specifically solve RSI-SOD tasks. CNN-based methods are dedicated to exploring the feature of the effective interaction strategies, in order to overcome the ORSI complex topology structure and special scenario. The nested network [7]

Manuscript received 19 March 2023; revised 15 May 2023; accepted 18 June 2023. Date of publication 29 June 2023; date of current version 21 July 2023. This work was supported in part by the Natural Science Foundation of Guizhou Province under Grant 20161054, in part by the Joint Natural Science Foundation of Guizhou Province under Grant LH20177226, and in part by the 2017 Special Project of New Academic Talent Training and Innovation Exploration of Guizhou University under Grant 20175788. *(Corresponding author: Ziyan Liu.)*

The authors are with the College of Big Data and Information Engineering, Guizhou University, Guiyang 550025, China (e-mail: yangmo97@126.com; gzucomm@gmail.com; dw251211119@163.com; wuying8477@163.com).

Digital Object Identifier 10.1109/JSTARS.2023.3290675

fuses multiresolution feature information to improve performance. The parallel bottom-up fusion network [8] mainly focuses on the cross-path interaction between adjacent features from the bottom path to the top path. The dense attention fluid network [9] transmits the shallow attention clues of the underlying features, and captures the edge and texture deep information, and hence high-level feature information capture of semantic and object location is achieved. However, the influence of higher level features on lower level features is easily overlooked and there is insufficient coverage in terms of feature interaction, which may also result in ORSI exploration of context information to be incomplete. In addition, the scale of the object in RSI diverse, complex shape, the fuzzy boundaries are the unique characteristics of the object, which hindered the object details accurate prediction.

Inspired by the results of the abovementioned observations, we propose a novel RSI-SOD specific solution, an important pick-and-pass gated refinement network (IP2GRNet), which focuses on refining detail features and capturing context information to accommodate different scales and types in ORSIs. In this article, our core idea is to accurately describe and capture object location and feature information, further explore the information features contained in the background, expand the coverage of feature interaction, and improve the details, and also, use aggregation operations to concentrate detail refinement information to improve pick-and-pass refinement stage context capture. Specifically, we process the object location and feature information in the detail refinement stage. In this way, we determine the object location and then use previous features to provide comprehensive auxiliary information for the current features. In addition, we use aggregation operations and introduce mixture loss for supervised network training to effectively improve model feature information processing.

In summary, our proposed SOD approach for ORSI is unique because the modal feature layer information refinement is tightly coupled in a comprehensive and deep manner. In the detail refinement stage, to determine the importance of object proximity location and contextual information cues for SOD, we construct an embeddable refinement synchronization structure to refine features of ORSI through a self-modality aggregation perspective, and in the feature pick-and-pass fusion stage, to progressively learn semantic information from the detail refinement stage, we propose an important pick-and-pass gated refinement (IP2GR) model guided by attention and reverse attention high-level semantic information and low-level semantic pick-and-pass refinement predict saliency maps. The main contributions are summarized as follows.

1) An end-to-end IP2GRNet is proposed to achieve ORSI-SOD that sufficiently captures and exploits contextual information by fusing it in a refined and progressive pick-and-pass manner.
2) The refinement synchronization architecture is equipped with a self-modality attention refinement (SAR) module and a dynamic weighted refinement (DWR) module designed to determine object locations and refine encoded features by encoding a 3-D attention tensor of self-modality and contextual dependencies of feature layers, in addition to capturing information from different feature layers centrally using aggregation operations to improve detail refinement information.
3) An IP2GR module is proposed to progressively learn the contextual information of each feature layer sampled on the pick-and-pass refinement stage, and based on the previous feature guidance, the saliency detection performance is further improved.
4) We evaluate the proposed IP2GRNet against two typical ORSI-SOD datasets on the existing state-of-the-art methods. Experiments demonstrate that our proposed IP2GRNet compared with existing CNN and specialized RSI-based methods can achieve better or competitive performance.

II. Related Works

A. Traditional and CNN-Based for SOD in NSIs

1) Traditional Methods: In NSIs, SOD explores artificial features in many traditional methods for NSI-SOD. Traditional methods are mainly composed of three categories: unsupervised, semisupervised, and supervised. Unsupervised methods extend many principles and techniques, such as the saliency tree [10], directional information [11], the sparse graph [12], and the hybrid sparse learning [13]. And it is different from unsupervised learning, semisupervised and supervised are less common. Zhou et al. [14] used the boundary uniformity model to generate pseudolabels to establish the relationship between the control state and the salience map on a linear feedback-based control system model. Liang and Hu [15] supervised learning to select features by training a support vector machine to remove redundant.

2) CNN-Based Methods: Unlike traditional methods, most CNN-based methods utilize supervised learning to greatly improve detection accuracy. To explore the internal relationship multiscale feature interaction, feature suppression/balance, sparse/dense label aggregation, and a large number of feature processing measurement methods are proposed [16]. Li et al. [17] proposed a deep nonlocal network that used a single CPU thread to achieve competitive performance.

B. Traditional and CNN-Based Methods for ORSI-SOD

1) Traditional Methods: In terms of feature extraction, the traditional ORSI-SOD method relies as heavily on handcrafting as the NSI-SOD method. Faur et al. [18] proposed rate distortion and mean shift algorithm estimation method to segment the RSIs. Zhang et al. [19] proposed a content analysis method for color information, which obtained the result map by calculating the saliency score of the color channel and fusing the color content. Zhang et al. [20] introduced a self-adaptive fusion method to integrate color, intensity, texture, and global contrast features in ORSI.

2) CNN-Based Methods: Many ORSI-SOD methods have shown good performance by exploiting the powerful feature representation capability of CNN. Because the method based on CNN has urgent data requirements, the authors in [7] and [9]

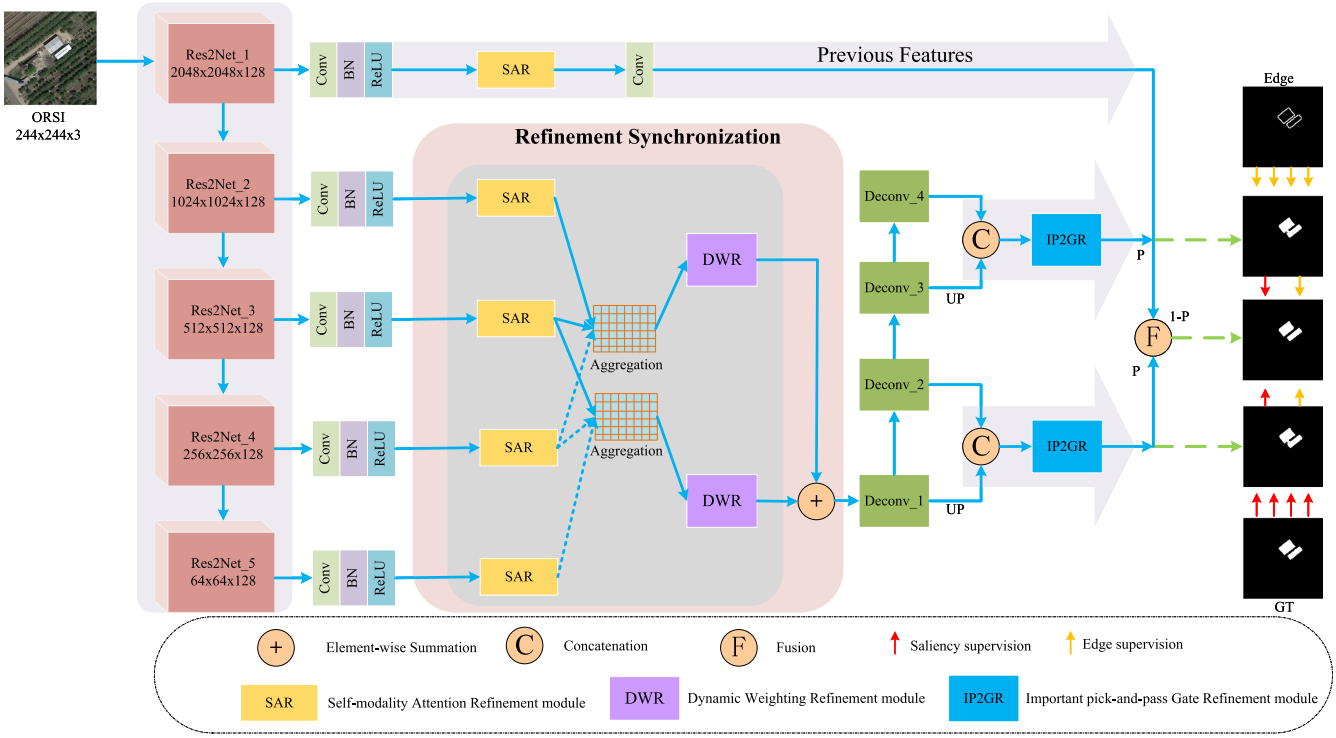

Fig. 2. Proposed pipeline for IP2GFNet consists of three key components: SAR module, DWR module, and IP2GR module. First, SAR extracts low-level features on five different scales. Then, this feature information is activated by aggregation to coordinate functions. Then, the aggregated features are input to the DWR module for dynamic weight refinement. Finally, DWR output context features are transmitted to IP2GR to further capture context messages and infer significant objects. In the training stage, we adopted a GT supervision strategy and attached pixel-level supervision to the final salient object prediction operation.

constructed ORSI datasets, namely ORSSD and EORSSD; there are more challenging scenarios. Li et al. [7] extracted the features in five different resolutions ORSIs and proposed a two-stream pyramid model and a detail refinement stage—pick-and-pass refinement stage with connections module. As edge information plays a guiding and improving role in the SOD task, Wang et al. [21] introduced edge features into the feature layer to predict salient regions. In addition, Huang et al. [22] located multiscale objects guided by high-level features and refined objects by combining cross-level features and semantic information.

III. Proposed Method

First, in this section, we overview the proposed IP2GRNet, as shown in Fig. 2. Then, refinement synchronization, and the SAR module and DWR module that make it up, and IP2GR module are discussed in detail. Finally, a clear explanation of the loss function is given.

A. Overview

As shown in Fig. 2, we present the overall pipeline of our proposed IP2GRNet, a multilevel information aware architecture, similar to the detail refinement stage and the pick-and-pass refinement stage, equipped with a refinement synchronization in the detail refinement stage. We input ORSI images into the backbone network (e.g., Res2Net-50, VGG16) to extract the five different resolutions multilevel features, denoted as f_1, f_2, f_3, f_4, and f_5, which indexes the feature levels, respectively. Since computational complexity and parameter number directly influence the inference performance of the whole network, to reduce their negative effects, we downscale the channel by performing 1×1 convolutional layer computations. After that, in order to better match the feature detail refinement stage, the extracted feature maps of different channel numbers were uniformly converted into 128 channel numbers. Considering the redundancy of multilevel information extracted from the backbone network and the complementary content in single-level information, before the network feature extraction and decoding, the detailed synchronization structure is introduced to further highlight the effective information. Specifically, the refinement-accumulation-refinement mechanism, consisting of SAR, aggregation, and DWR modules, is designed to progressively refine the optimal feature encoding of multilevel information in a self-modality refinement and accumulation manner.

In the pick-and-pass refinement stage, we design a more detailed refinement structure in which the features' output from the refinement synchronization are decoded, and four sets of corresponding decoded features are obtained by upsampling, with two sets of shallow corresponding feature streams and two sets of higher corresponding feature streams input to achieve IP2GR. In the feature pick-and-pass refinement process, an IP2GR module is proposed to integrate shallow and high-level features and a previous decoded feature outputs in a dynamically weighted way. Finally, the prediction maps of the respective IP2GR modules are output and the fused prediction maps of the two IP2GR modules are supervised to infer the final saliency maps.

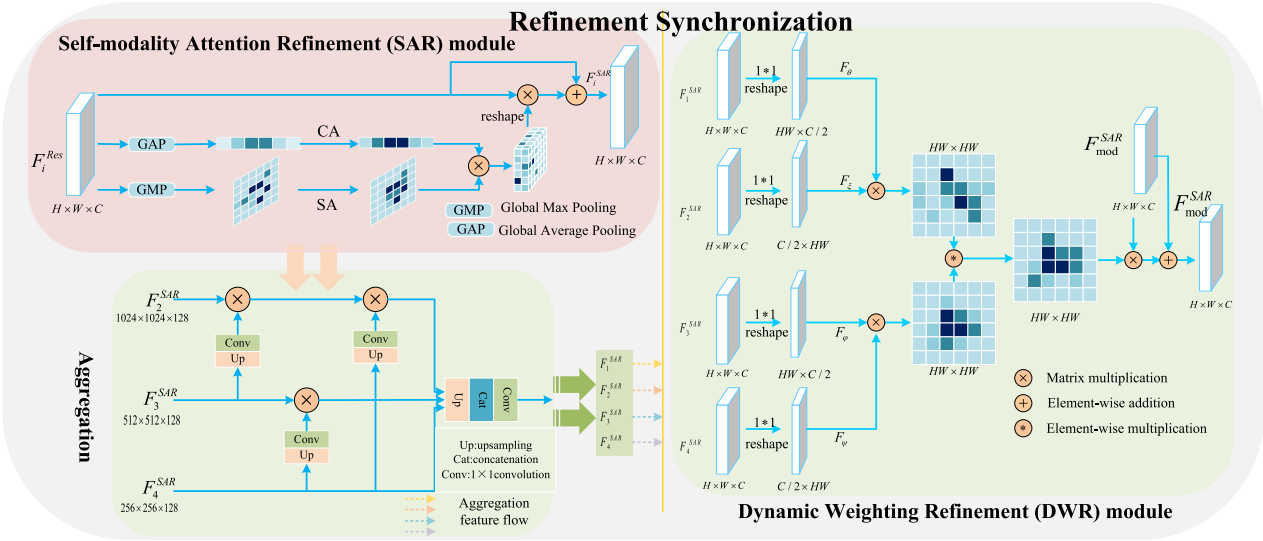

Fig. 3. Illustration of the refinement synchronization. SAR module and aggregation module are shown on the left, and a DWR module is shown on the right.

B. Refinement Synchronization

In order to input detail refinement stage features to the pick-and-pass refinement stage more efficiently, as shown in Fig. 3, we embed the refinement synchronization structure as a connecting bridge between the detail refinement stage and the pick-and-pass refinement stage to refine the detail refinement stage features for self-modality attention and dynamic weights. We consider the design of refinement synchronization from two aspects: 1) the detail refinement stage features of each different scale contain rich spatial and channel modality information, and learning effective feature representation is increased by the effects of undifferentiated transmission of information. Therefore, we rethink from a single-modality perspective and design the SAR module to suppress background noise and highlight important cues; 2) In view of the strong correlation and complementarity between different modality information, the multilevel features containing graphical background, color contrast, and internal consistency, we use aggregation operations to centralize modal features and facilitate feature representation, and from the perspective of global feature refinement, we design a DWR module to capture long-term correlation between features and refine modal information inter features.

C. SAR Module

ORSIs are coded by a backbone network, the given multilevel coded features of ORSIs contain rich spatial and channel channels that represent salient objects. However, it is prone to redundancy in the single modal information. In addition, the undifferentiated information transmission may again increase difficulty for feature learning and it is more likely to cause the subsequent decoding and inferencing process to suffer from serious deviations. To solve this problem, we rethink from a single-modality perspective and design the SAR module to suppress background noise in a take-away spatial channel 3-D attention manner and highlight important cues.

Attentional mechanisms have been widely applied in ORSI-SOD tasks, the most common of which are spatial attention (SA) and channel attention (CA). There are the following three common forms.

1) *Separate utilization:* SA is adopted to extract low-level features, whereas CA is adopted to extract high-level features.
2) *Serial utilization:* CA is used first for feature enhancement, and then, the final enhanced features are captured by SA.
3) *Parallel utilization:* CA and SA are used separately for input channel feature enhancement, and after fusing the enhanced features to generate the final features.

In all vision tasks, the use of CA and SA alone for different levels of functionality leads to not necessarily good results. However, the serial approach utilizes against a certain requirement of combining order, whereas the parallel utilization can only enhance the 1-D features of space or channel, which not only increases the computational complexity, but also generates certain information redundancy. To solve this problem, in the spatial channel 3-D attention tensor, we embed SA and CA into the spatial channel and consider the following three aspects.

1) Improving robustness and reducing the computational complexity of the 3-D attention approach through parallel utilization.
2) Refinement of single modal features in spatial and channel dimensions effectively improves prediction performance of the network.
3) Suppress background noise for 1-D feature information to solve the problem of diluted high-level semantics.

In Fig. 3, the features of the multilevel coded branches of the backbone network are embedded in the SAR module. First, in a parallel structure, CA and SA of input features are calculated, and corresponding spatial attention maps and channel attention maps are obtained. Then, the spatial feature information and channel feature information are fused directly to the attention space map by matrix multiplication to generate a 3-D attention

tensor. The specific process is as follows:

$$\mathrm{SAR}_{3\mathrm{D}} = \mathrm{SA}\left(f^{i}_{\mathrm{mod}}\right) \otimes \mathrm{CA}\left(f^{i}_{\mathrm{mod}}\right) \tag{1}$$

where f^{i}_{mod} denotes the information of each modality in the detail refinement stage layer after multiscale feature extraction, and mod denotes modality information. In the 3-D attention tensor calculation, we apply residual connectivity to refine the modal features for multiscale decoding:

$$f^{\mathrm{SAR}}_{\mathrm{mod}} = \mathrm{conv}\left(\mathrm{SAR}_{3\mathrm{D}} \odot f^{i}_{\mathrm{mod}} + f^{i}_{\mathrm{mod}}\right). \tag{2}$$

In the ablation section, we provide an ablation study of different connection ways of attention to demonstrate the effectiveness of our proposed network.

D. DWR Module

Although the SAR module refines the features extracted by detail refinement stage, it is difficult to take full advantage of the strong correlation and complementary between the multilevel codes. OSRI background color contrast, object texture, and object edge information provide highlighting the internal consistency of the object and spatial relationships between objects. After the aggregation operation to concentrate the image feature information of the context, but the refinement synchronization first stage refinement information need as a guide to obtain richer feature information. After the aggregation operation, the feature information is condensed again for further processing of the detail information. Therefore, we design the DWR module to further capture the long-term dependence of multilevel features on the condensation feature, and optimize the extraction of feature information from rough to fine refinement.

The detailed structural composition of DWR is shown on the right side of Fig. 3. DWR module accept the feature input from the SAR module to the feature F after the aggregation operation. First, we use the bottleneck convolution layer to reduce the number of channels by half and obtain different layer forms, which are mapped to a uniform feature space to complete the dynamic weight allocation. The specific calculation process is as follows:

$$\begin{aligned} F_{\theta} &= W_{\theta} f^{\mathrm{SAR}}_{\mathrm{mod}} \\ F_{\xi} &= W_{\xi} f^{\mathrm{SAR}}_{\mathrm{mod}} \\ F_{\varphi} &= W_{\varphi} f^{\mathrm{SAR}}_{\mathrm{mod}} \\ F_{\psi} &= W_{\psi} f^{\mathrm{SAR}}_{\mathrm{mod}} \end{aligned} \tag{3}$$

where W_{θ}, W_{ξ}, W_{φ}, and W_{ψ} denote the embedding weight parameters that can be learned by the modal information of the four input channels through the bottleneck convolutional layer, respectively.

Then, according to the scaling dot product attention calculation method, the correlation between multilevel ORSI features and the autocorrelation of modal information are calculated pixelwise:

$$\begin{aligned} M_1 &= \mathrm{softmax}\left(F^{T}_{\theta} \otimes F_{\gamma}\right) \\ M_2 &= \mathrm{softmax}\left(F^{T}_{\varphi} \otimes F_{\psi}\right) \end{aligned} \tag{4}$$

where $\otimes$ denotes the matrix multiplication, and softmax denotes the activation function. $M_1 \in \mathbb{R}^{HW \times HW}$ highlights the common response between the different scales information, and $M_2 \in \mathbb{R}^{HW \times HW}$ establishes a dependency on the self-modality itself. We separate M_1 and M_2 is the essential purpose of hope common expression feature information and the interaction mode dependence of the final similarity in the feature space, that we will be in the ablation study for validation.

Finally, we map the information from the correlations unified feature space to the pick-and-pass refinement stage, which jointly generate global correlation weights to refine the original input features:

$$f^{\mathrm{DWR}}_{\mathrm{mod}} = R\left(f^{\mathrm{SAR}}_{\mathrm{mod}}\right) \otimes \mathrm{softmax}\left(M_1 \odot M_2\right) + f^{\mathrm{SAR}}_{\mathrm{mod}} \tag{5}$$

where $\odot$ denotes the elementwise multiplication, and reshapes the feature space dimension from $\mathbb{R}^{C \times H \times W}$ to $\mathbb{R}^{C \times HW}$. $M_1 \odot M_2$ is generated by the dependency weights in the global modal architecture. By refining the original features from a global perspective, we can improve the integrity of the detection results and obtain better detection accuracy. We proved the advantages of DWR module in Table III through various experiments.

E. IP2GR Module

As previously highlighted, multilevel feature information refinement is critical to the ORSI-SOD task. Many existing methods generally choose to perform refinement in a single stage or intrastage, which is not sufficient. In our proposed network, the functions and information transfer of each stage are considered. The detail refinement stage and the pick-and-pass refinement stage have a different function in feature learning selection, where the detail refinement stage focuses more on regular feature extraction, whereas the pick-and-pass refinement stage focuses more on saliency-related feature learning. Therefore, in addition to feature integration through detailed synchronization structures in the detail refinement stage, we also perform modal information refinement in the pick-and-pass refinement stage to obtain distinguished saliency prediction features. We construct the context convergence aggregation structure in the detail refinement stage, which can further provide valid information such as refinement edge, sharpness, and internal consistency for the pick-and-pass stage, which is beneficial to the integrity and more comprehensive learning of the SOD detection task. Since the direct equal combination of modal feature information is uncontrollable and complex, it is a challenging and urgent problem to effectively pick-and-pass the most valuable and saturated feature information from the convergent stream in the context of convergent aggregation structure.

To solve the problems encountered, we design an IP2GR module to learn the importance map P, which is used to control the effects of different modality information in a dynamically weighted manner, as shown in Fig. 4. The IP2GR module can determine the contribution of different modal complementary information during the refinement of modal information in the pick-and-pass stage. Moreover, our proposed network using learnable weight parameters can resist to some extent the failure of modal features due to low-quality maps, etc. The brief process

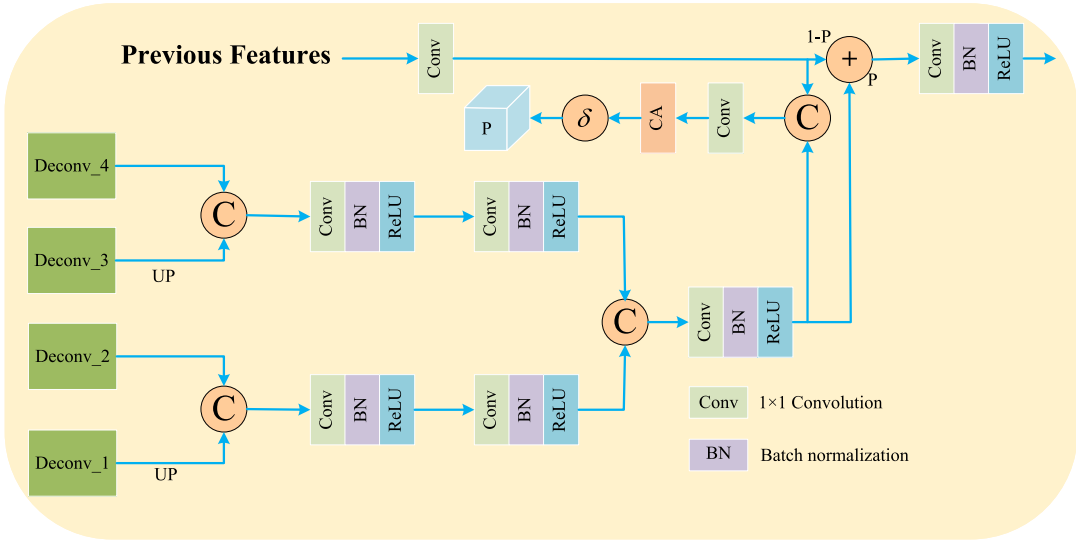

Fig. 4. Illustration of the IP2GR module.

is as follows: the decoded features are cascaded through a double layer of convolutional layers for contextual features, and the prior features are combined with learnable weights, and the pick-and-pass most valuable feature information is output using the residual network structure. The specific computational process is as follows:

$$f_{\mathrm{IP}}^{i} = \mathrm{conv}\left(P^{i} \odot H^{i} + (1 - P^{i}) \odot f_{\mathrm{IP}}^{i+1} \uparrow\right) \tag{6}$$

where f_{IP}^{i+1} denotes the IP output features at the $(i+1)$th pick-and-pass refinement stage level, $i \in \{5, 4, 3, 2, 1\}$, and P^{i} denotes the learnable weight parameters, which measures the importance of the pick-and-pass refinement stage features. In the weight parameter learning process, we apply 1×1 convolution for channel matching and connect H^{i} and f_{IP}^{i+1} features to obtain U^{i}. We input it to CA with a sigmoid activation function to obtain the pick-and-pass result map $P^{i} \in \mathbb{R}^{C \times H \times W}$ as follows:

$$P^{i} = \sigma\left(\mathrm{CA}(U^{i})\right) = \sigma\left(\mathrm{CA}\left(\mathrm{conv}\left(\left[H^{i}, f_{\mathrm{IP}}^{i+1} \uparrow\right]\right)\right)\right) \tag{7}$$

where σ denotes the sigmoid activation function. The pick-and-pass result map determines the contribution of different modalities to the supplementary information during the ith refinement stage.

F. Loss Function

In this article, we supervise the training of the network from pixel, region, and statistical perspectives using a hybrid loss function, in which cross-entropy loss, IoU loss, and recent F-measure value loss are extensively used to form the hybrid loss function. In addition, edge information and true value information are also added to the loss function

$$\mathrm{loss} = \mathrm{loss}_{\mathrm{bce}} + \lambda \cdot \mathrm{loss}_{\mathrm{iou}} + \mathrm{loss}_{F} \tag{8}$$

$$\mathrm{Loss} = \mathrm{loss}(P, G_P) + \mu \cdot \mathrm{loss}(E, G_E) \tag{9}$$

where P is the predicted saliency map, G_P is the saliency ground truth label, E is the edge saliency map, G_E is the edge saliency ground truth label, and λ and μ denote the hyperparameters that balance the contribution of the three losses and the contribution of the two supervised objects. According to the experience, we set $\lambda = 0.6$ and $\mu = 0.5$.

IV. Experiment and Discussion

A. Datasets and Evaluation Metrics

1) Datasets: We evaluate the detection task performance of the proposed model on the benchmark RSI-SOD dataset.

ORSSD: This dataset [6] is a collection of images from Google Earth and existing remote sensing, and is the first publicly available RSI dataset. It contains 800 remote sensing scenes and provides corresponding pixel-level annotations for each image, among which 600 images are used as training sets and 200 images are used as testing sets.

EORSSD: This dataset [9] extends 2000 images on the ORSSD dataset and has corresponding pixel-level GTs, which is also the largest public dataset for the RSI-SOD task. Among them, 1400 images were used as training set and 600 images as testing set.

2) Evaluation Metrics: The widely used F-measure mean (F_β, $\beta^2 = 0.3$) [23], S-measure value (S_α, $\alpha = 0.5$) [24], e-measure mean (S_m), and mean absolute error (MAE, $\mathcal{M}$) are adopted as evaluation indexes to comprehensively evaluate our IP2GRNet and compare the performance of other methods [22]. Specifically, F-measure is the weighted harmonic average of the precision and recall rates. This article focuses more on precision. S-measure measures the structural similarity of both area perception and object perception. E-measure comprehensively considers local pixel-level matching information and global image-level statistics information. MAE evaluates the average pixel-level error.

B. Implementation Details

We used 600 images from the ORSSD dataset for training and 200 images for testing, and 1400 images from the ORSSD dataset for training and 600 images for testing, respectively. The size of each sample is uniformly adjusted to 224×224 to

TABLE I
QUANTITATIVE COMPARISONS WERE MADE WITH 22 OF THE STATE-OF-THE-ART METHODS ON EORSSD AND ORSSD DATASETS, INCLUDING FOUR CONVENTIONAL NSI-SOD METHODS, THREE CONVENTIONAL ORSI-SOD METHODS, SEVEN CNNSI-SOD METHODS, AND EIGHT CNN-BASED ORSI-SOD METHODS

Methods (years)	Type	Input size	#Param (M)	EORSSD				ORSSD			
				$F_\beta^{\text{mean}}\uparrow$	$S_\alpha\uparrow$	$E_\xi^{\text{mean}}\uparrow$	$\mathcal{M}\downarrow$	$F_\beta^{\text{mean}}\uparrow$	$S_\alpha\uparrow$	$E_\xi^{\text{mean}}\uparrow$	$\mathcal{M}\downarrow$
RRWR15	T.N.	–	–	0.3686	0.5992	0.5943	0.1677	0.5125	0.6835	0.7017	0.1324
HDCT16	T.N.	–	–	0.4018	0.5971	0.6376	0.1088	0.4235	0.6197	0.6495	0.1041
SMD17	T.N.	–	–	0.5473	0.7101	0.7286	0.0771	0.6214	0.764	0.7745	0.0715
RCRR18	T.N.	–	–	0.3685	0.6007	0.5946	0.1644	0.5126	0.6849	0.7021	0.1277
VOS18	T.R.	–	–	0.2107	0.5082	0.4886	0.2096	0.2717	0.5366	0.5352	0.2151
SMFF19	T.R.	–	–	0.2992	0.5401	0.5197	0.1434	0.2684	0.5312	0.492	0.1854
CMC19	T.R.	–	–	0.2692	0.5798	0.5894	0.1057	0.3454	0.6033	0.6417	0.1267
R3Net18	C.N.	300×300	56.15	0.6302	0.8184	0.8294	0.0171	0.7383	0.8141	0.8681	0.0399
EGNet19	C.N.	380×320	108.57	0.6967	0.8601	0.8775	0.011	0.75	0.8721	0.9013	0.0216
PoolNet19	C.N.	400×300	53.63	0.6406	0.8207	0.8193	0.021	0.6999	0.8403	0.865	0.0358
MINet20	C.N.	320×320	47.56	0.8174	0.904	0.9346	0.0093	0.8574	0.904	0.9454	0.0144
GateNet20	C.N.	384×384	100.02	0.8228	0.9114	0.9385	0.0095	0.8679	0.9186	0.9538	0.0137
SUCA21	C.N.	256×256	117.71	0.7949	0.8988	0.9277	0.0097	0.8237	0.8989	0.94	0.0145
PA-KRN21	C.N.	600×600	141.06	0.8358	0.9192	0.9536	0.0104	0.8727	0.9239	0.962	0.0139
LVNet19	C.R.	128×128	–	0.7328	0.863	0.8801	0.0146	0.7995	0.8815	0.9259	0.0207
LDF20	C.R.	352×352	25.15	–	0.8859	0.94	0.0229	–	0.8799	0.9448	0.014
F3Net20	C.R.	352×352	25.54	–	0.8985	0.9498	0.0199	–	0.8887	0.9485	0.0118
DAFNet21	C.R.	128×128	29.35	0.7845	0.9166	0.9291	0.006	0.8511	0.9191	0.9539	0.0113
MSCNet22	C.R.	224×224	3.26	–	0.9226	0.9754	0.0132	–	0.9086	0.9684	0.009
AGNet22	C.R.	224×224	26.6	0.8736	0.9284	0.9614	**0.0069**	0.9109	0.9392	0.9707	0.0093
ACCoNet22	C.R.	256×256	–	0.8552	**0.929**	0.9653	0.0074	0.8971	**0.9437**	0.9754	**0.0088**
CorrNet23	C.R.	256×256	4.09	0.862	0.9289	0.9646	0.0083	0.9002	0.938	0.9746	0.0098
Ours	C.R.	224×224	25.6	**0.8962**	0.9248	**0.9756**	0.0071	**0.9121**	0.9406	**0.9778**	0.009

[1] '-' represents no experimental data. "↑" and "↓" indicate that the value changes better toward trend.
The bold value indicates the best results.

reduce computing resources. The proposed IP2GRNet is based on the PyTorch implementation and deployed on a workstation equipped with an NVIDIA GeForce RTX 3060 GPU. After the experimental environment was prepared, Res2Net-50 was used as the feature extraction network and the ADAM optimization strategy was used to conduct 60 epochs iterative training on the model parameters. The batch size was set to 8, the initial learning rate was set to 1e-4, and then decreased evenly to 5e-4. We use the method of initial weights in the Xavier strategy, initialized to constant deviation parameters.

C. Comparison With Sate-of-the-Arts

In the experiment, we have 22 kinds of NSI-SOD and RSI-SOD method and the proposed IP2GRNet makes a comprehensive comparison. These include seven traditional methods (RRWR [25], HDCT [26], SMD [27], RCRR [28], VOS [29], SMFF [20], and CMC [30]), fifteen CNN-based methods (R3Net [31], EGNet [32], PoolNet [33], MINet [34], GateNet [35], SUCA [36], PA-KRN [37], LVNet [7], LDF [38], F3Net [39], DAFNet [9], MSCNet [40], CorrNet [41], AGNet [42], and ACCoNet [43]). We employ the Res2Net backbone network on the testing sets of the ORSSD and EORSSD datasets to report the performance of the comparison methods, as well as using the source code or saliency maps provided by the authors. For a fair comparison, we retain the parameter settings to generate the results.

Quantitative comparison: In Table I, we report the F_β^{mean}, S_α, E_ξ^{mean}, and $\mathcal{M}$ of our proposed method drawn on 22 other methods on two RSI-SOD datasets. Our method shows impressive performance compared to other NSI-SOD and RSI-SOD methods. Specifically, IP2GRNet on the EORSSD dataset in F_β^{mean}, S_α [e.g., F_β^{mean}:0.8962 (IP2GRNet) versus 0.8552 (ACCoNet)] and S_α:0.9756 (IP2GRNet) versus 0.9653 (ACCoNet). Meanwhile, our method is only weaker than DAFNet and AGNet on $\mathcal{M}$. While on the ORSSD dataset, our method achieves

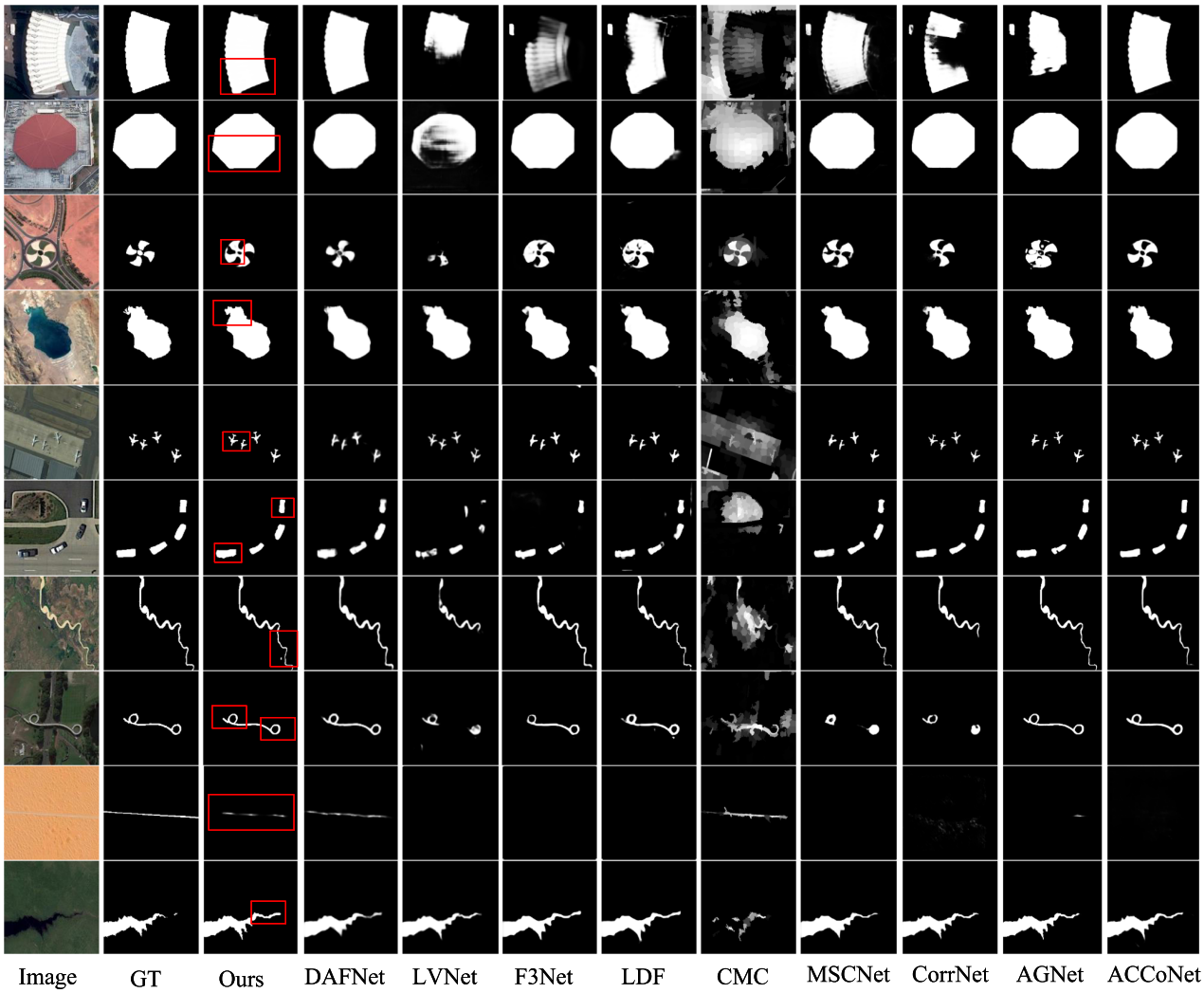

Fig. 5. Visual comparisons with nine representative state-of-the-art methods on EORSSD/ORSSD datasets. The red box can quickly identify the detection effect of our method.

optimal and suboptimal on F_β^{mean}, S_α, E_ξ^{mean}, and $\mathcal{M}$. Although our approach fails to achieve the best S_α score on the datasets, more impressive performance is achieved with a smaller parameter cost, e.g., parameters: 25.60M (IP2GRNet) versus 117.7M (SUCA). Our method also beats MINet with the same parameter size, about 25.6M for the former and 47.56M for the latter. Compared with the NSI-SOD method, the performance of our proposed model based on the RSI-SOD method is ahead of NSI, which further proves that the design of the special ORSI-SOD model is essential for the detection of remote sensing scenes.

Qualitative comparison: In Fig. 5, we present the visual comparisons with nine representative methods on some challenging scenes of ORSIs. The first and second rows of Fig. 5 show a large objects scene, in which most of the segmentation is incomplete and neglects the core part due to the large span of building coverage. The third and fourth rows of Fig. 5 show a cluttered background scene. The complex background confuses some other methods, causing the network to incorrectly include backgrounds or omit objects in the salient maps prediction process. The fifth and sixth rows of Fig. 5 show multiple and small salient object scenes, which is a highly challenging problem in the RSI-SOD task. Our method clearly highlights all salient objects, and other compared methods receive similar background and small-scale interference and fail in individual cases (i.e., CMC and LVNet). The seventh, eighth, ninth, and tenth rows of Fig. 5 show slender or complex structured object scenes, and some models occasionally lose objects (i.e., LVNet, F3Net, and LDF) and fail to outline objects (i.e., CorrNet and ACCoNet). Due to the presence of IP2GR, our method provides a more complete and clear saliency map than other detection results, which further confirms the importance of detail refinement for the RSI-SOD task.

D. Ablation Studies

1) Analysis of Different Modules: To evaluate the effectiveness and contribution of the modules in the proposed model, ablation studies were performed on the EORSSD dataset. Quantitative evaluation results and visualization examples are presented in Table II and Fig. 6, respectively. We constructed the

TABLE II
ABLATION RESULTS OF EVALUATING THE INDIVIDUAL

No.	Baseline	SAR	DWR	IP2GR	EORSSD			
					$F_\beta^{\text{mean}}\uparrow$	$S_\alpha\uparrow$	$E_\xi^{\text{mean}}\uparrow$	$\mathcal{M}\downarrow$
1	√				0.8747	0.9187	0.9567	0.0077
2	√	√			0.8794	0.9204	0.9643	0.0075
3	√	√	√		0.8813	0.9211	0.9663	0.0074
4	√	√		√	0.8817	0.9215	0.9708	0.0073
5	√	√	√	√	**0.8848**	**0.9248**	**0.9756**	**0.0071**

Note: Contribution of each module in IP2GRNet.
The bold value indicates the best results.

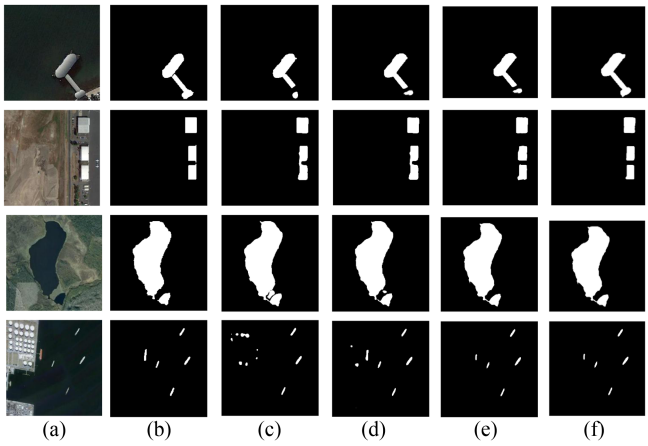

Fig. 6. Visual comparison of IP2GRNet variants equipped with different models. (a) ORSIs. (b) GT. (c) Baseline. (d) Baseline + SAR. (e) Baseline + SAR + DWR. (f) Baseline + SAR + DWR + IP2GR.

TABLE III
ABLATION RESULTS OF EVALUATING THE IMPORTANCE AND RATIONALITY OF DYNAMIC WEIGHTS OF THE CONNECTION MODES OF ATTENTION MECHANISMS IN THE SAR AND DWR MODULES OF THE REFINED SYNCHRONIZATION IN IP2GRNET

		EORSSD			
		$F_\beta^{\text{mean}}\uparrow$	$S_\alpha\uparrow$	$E_\xi^{\text{mean}}\uparrow$	$\mathcal{M}\downarrow$
	Full model	**0.8848**	**0.9248**	**0.9756**	**0.0071**
SAR	w/ CA, w/o SA	0.8624	0.8956	0.9521	0.0079
	w/o CA, w/SA	0.8663	0.9011	0.9543	0.0078
	SA-CA	0.8722	0.9068	0.9613	0.0078
DWR	w/ M1, w/o M2	0.8742	0.9026	0.9633	0.0075
	w/o M1, w/ M2	0.8763	0.9098	0.9642	0.0076

The bold value indicates the best results.

TABLE IV
ABLATION STUDIED ON AGGREGATION AND IP2GR MODULES

		EORSSD			
		$F_\beta^{\text{mean}}\uparrow$	$S_\alpha\uparrow$	$E_\xi^{\text{mean}}\uparrow$	$\mathcal{M}\downarrow$
	Full model	**0.8848**	**0.9248**	**0.9756**	**0.0071**
Aggregation	Single layer	0.8501	0.8852	0.9423	0.008
	Double layer	0.8632	0.8923	0.9632	0.0078
IP2GR	w/ add	0.8685	0.9022	0.9578	0.0076
	w/ cat	0.8698	0.9069	0.9602	0.0076

The bold value indicates the best results.

baseline model by simplifying the full model in the following two ways: 1) Remove the refinement synchronization structure including SAR and DWR modules; 2) Replace the IP2GR module with a simple deconvolution layer.

We conducted the ablation experiment by gradually increasing the design module. The SAR module is introduced into the baseline model (represented as "+SAR"), then the DWR and IP2GR modules are gradually added to the model. In other words, "+IP2GR" means "baseline +SAR+DWR+IP2GR." In addition, all the ablation experiment parameters were set with the same training configuration as IP2GRNet for training.

Fig. 6 shows that the salient object predicted by the baseline model is roughly located, resulting in the incomplete spatial structure, unclear boundary information, and difficult to suppress background areas. The introduction of the SAR module obtained more complete and consistent structural information (for example, the lake in the third image) compared with the baseline model, but still included many areas for error detection. From the quantitative results, in the optical remote sensing testing dataset, the F_β^{mean} metric value increased from 0.9187 to 0.9204. Then, we introduce the DWR module to refine the different modes, and some improvement in background suppression and object structure can be observed. In addition, the addition of IP2GR modules for modality feature integration in the pick-and-pass refinement stage results in clearer boundaries for salient objects (for example, bins in the second image) and significantly improved quantization performance. Specifically, on the EORSSD, the F_β^{mean} score was increased to 0.8848, and the percentage gain is 0.35% compared with the "+SAR+DWR" model. Moreover, the best performance was achieved by adding a SAR module to highlight the important hints of a self-modality perspective and adding a DWR module to refine the SAR module information, resulting in a 1.6% percentage gain in the F_β^{mean}. In summary, the ablation study further proves the effectiveness and contribution of our proposed module.

2) Analysis of Refinement Synchronization: To further verify the effectiveness of our proposed improved synchronization structure, we conducted a series of ablation experiments shown in Table III. In the SAR module, we use channel attention weight (represented by "w/CA, w/o SA"), spatial attention weight (represented by "w/o CA, w/SA"), and SA-CA serial combination (represented by "SA-CA") to ablate the replacement 3-D attention tensor. Compared with serial use of the SA-CA combination module on the EORSSD dataset, our full model of the SAR module achieved F_β^{mean} of 0.8848 and the percentage gain is 1.45%, the percentage gain is 2.0% for the S_α, 2.5% for E_ξ^{mean}, and 11.2% for $\mathcal{M}$.

To demonstrate the advantage of the DWR module, we use final weight (i.e., M1×M2) with the case of only M1 (represented by "w/ M1, w/o M1") and only M2 (represented by "w/o M2, w/ M2") to ablate the replacement final weight maps. In Table V, we can observe that the M1×M2 method is more efficient than

TABLE V
ABLATION STUDIES OF THE ADDITION OPERATION IN (2). (5), AND (6)

	EORSSD			
	$F_\beta^{\text{mean}}\uparrow$	$S_\alpha\uparrow$	$E_\xi^{\text{mean}}\uparrow$	$\mathcal{M}\downarrow$
w/o add in Eq.(2)	**0.8653**	0.9086	0.9662	0.0079
w/o add in Eq.(5)	0.8622	0.9025	0.9647	0.0079
w/o add in Eq.(6)	0.8617	0.9067	**0.9665**	**0.0076**

The bold value indicates the best results.

the individual weight map M1 or M2. The F_β^{mean} increased from 0.8742 to 0.8848, whose percentage gain is 4.3%, and the S_α increased from 0.9026 to 0.9248, whose percentage gain is 2.4%, 1.3% for E_ξ^{mean}, and 5.6% $\mathcal{M}$, compared with the M2 case alone.

3) Analysis of Feature Interaction Strategies on Aggregation and IP2GR: In order to verify the effectiveness of the design of aggregation and IP2GR modules, we conducted various ablation experiments, shown in Table IV. For the aggregation module, we added two ablation experiments. One is to verify the fusion of single-channel layers, and the other is to verify the combination and propagation of different channel layers. The F_β^{mean} score increased from 0.8501 to 0.8848 with a percentage gain of 4.1% compared with the propagation of the single-channel layer, whereas the S_α score increased from 0.8852 to 0.9248 with a percentage gain of 4.5% compared with the feature fusion strategy. The experimental results further prove that the double layer combination is more effective than the common feature fusion strategy.

IP2GR modules, we used addition (represented by "w/ add") or concatenation (represented by "w/ cat") instead of dynamic fusion strategies to demonstrate the effectiveness of IP2GR modules. In Table IV, we can see that the performance is improved with the help of the proposed IP2GR modules compared with the commonly used fusion strategies (add or cat). The F_β^{mean} score increased from 0.8685 to 0.8848 compared with the cascading operation (that is, w/ cat), with a percentage gain of 1.9%, and the S_α score increased from 0.9022 to 0.9248, with a percentage gain of 2.5%, for E_ξ^{mean}, the percentage gain is 1.9%, and for $\mathcal{M}$, the percentage gain is 7.1%.

4) Analysis of Residual Connection: To prove the validity of the residual structural connection, we remove the addition operation in (2), (5), and (6) for ablation experiments. The residual connection is shown in Table V. Compared with the simple addition, our residual connection achieved better quantization performance. In Table V, w/o (2), F_β^{mean} increases from 0.8653 to 0.8848, with a percentage gain of 2.3%, and for S_α, the percentage gain is 1.8%. Similarly, removing addition from (5) and (6) also degrades performance.

E. Failure Cases

Fig. 7 shows several typical failure cases. Locating salient objects is difficult in the following scenarios.

1) *Multiple and small salient objects:* As shown in the first row in Fig. 7, although there are features of multiple salient objects in the whole scene, the scale of the salient objects in the image is too small causing the corresponding depth map to fail to provide effective depth information, such that it is difficult to fully detect all the salient objects.
2) *High-contrast but not salient objects:* As shown in the second and third rows in Fig. 7, the roads and boats contrast with the background in the depth map. Therefore, the ambiguity caused by this conflict prevents our model from accurately detecting the salient objects.
3) *Complex background noise:* As shown in the fourth row of Fig. 7, the contrast between the salient object and the background is low, the depth information in the depth map is easily misleading, and our algorithm cannot effectively suppress the background. It is worth noting for the challenging scenarios described above.

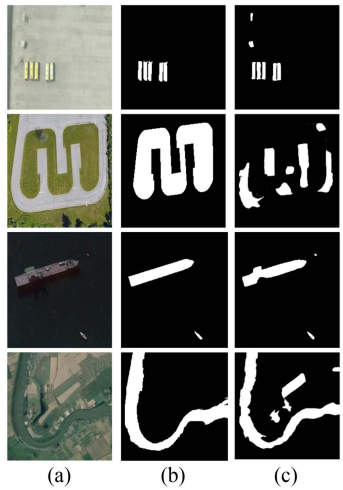

Fig. 7. Visual examples of failure cases. (a) ORSIs. (b) GT. (c) Saliency maps produced by IP2GRNet.

V. CONCLUSION

In this article, we propose a novel end-to-end method for SOD in ORSI that is capable of inferring semantic information and recovering refined details. The relationship between spatial and channel space is designed to detail the relationship between different salient objects or different parts of salient objects. Self-modality attention refinement and DWR modules are used to accurately describe and capture the approximate position and feature information of salient objects. In this process, aggregation operations are used to recover the details of objects of different scales. In addition, we propose an IP2GR model, which progressively refines the prediction results under the direction of attention and reverse attention. Experimental evaluation on two datasets demonstrate that our proposed method outperforms the existing state-of-the-art SODs. In addition, when this task is extended to RSIs, our model can optimize the semantic relationships of different objects or different parts of different objects from multiple scales, thus improving performance.

In the future, we will try to design high-precision and lightweight network structures to further promote the RSI-SOD model in real-life applications.

ACKNOWLEDGMENT

This work was carried out in part using computing resources at the State Key Laboratory of Public Big Data, Guizhou University.

REFERENCES

[1] A. Borji, M.-M. Cheng, Q. Hou, H. Jiang, and J. Li, "Salient object detection: A survey," *Comput. Vis. Media*, vol. 5, pp. 117–150, 2019.
[2] G. Yue, W. Han, B. Jiang, T. Zhou, R. Cong, and T. Wang, "Boundary constraint network with cross layer feature integration for polyp segmentation," *IEEE J. Biomed. Health Inform.*, vol. 26, no. 8, pp. 4090–4099, Aug. 2022.
[3] X. Sun et al., "FAIR1M: A benchmark dataset for fine-grained object recognition in high-resolution remote sensing imagery," *ISPRS J. Photogrammetry Remote Sens.*, vol. 184, pp. 116–130, 2022.
[4] P. Zhang, W. Liu, D. Wang, Y. Lei, H. Wang, and H. Lu, "Non-rigid object tracking via deep multi-scale spatial-temporal discriminative saliency maps," *Pattern Recognit.*, vol. 100, 2020, Art. no. 107130.
[5] Y. Fang, Z. Chen, W. Lin, and C.-W. Lin, "Saliency detection in the compressed domain for adaptive image retargeting," *IEEE Trans. Image Process.*, vol. 21, no. 9, pp. 3888–3901, Sep. 2012.
[6] L. Zhang and J. Ma, "Salient object detection based on progressively supervised learning for remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 11, pp. 9682–9696, Nov. 2021.
[7] 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.
[8] C. Li et al., "A parallel down-up fusion network for salient object detection in optical remote sensing images," *Neurocomputing*, vol. 415, pp. 411–420, 2020.
[9] 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.
[10] Z. Liu, W. Zou, and O. L. Meur, "Saliency tree: A novel saliency detection framework," *IEEE Trans. Image Process.*, vol. 23, no. 5, pp. 1937–1952, May 2014.
[11] M. Jian, K.-M. Lam, J. Dong, and L. Shen, "Visual-patch-attention-aware saliency detection," *IEEE Trans. Cybern.*, vol. 45, no. 8, pp. 1575–1586, Aug. 2015.
[12] L. Zhou, Z. Yang, Z. Zhou, and D. Hu, "Salient region detection using diffusion process on a two-layer sparse graph," *IEEE Trans. Image Process.*, vol. 26, no. 12, pp. 5882–5894, Dec. 2017.
[13] Q. Wang, Y. Yuan, P. Yan, and X. Li, "Saliency detection by multiple-instance learning," *IEEE Trans. Cybern.*, vol. 43, no. 2, pp. 660–672, Apr. 2013.
[14] Y. Zhou, S. Huo, W. Xiang, C. Hou, and S.-Y. Kung, "Semi-supervised salient object detection using a linear feedback control system model," *IEEE Trans. Cybern.*, vol. 49, no. 4, pp. 1173–1185, Apr. 2019.
[15] M. Liang and X. Hu, "Feature selection in supervised saliency prediction," *IEEE Trans. Cybern.*, vol. 45, no. 5, pp. 914–926, May 2015.
[16] K. Yan, X. Wang, J. Kim, and D. Feng, "A new aggregation of DNN sparse and dense labeling for saliency detection," *IEEE Trans. Cybern.*, vol. 51, no. 12, pp. 5907–5920, Dec. 2021.
[17] H. Li, G. Li, B. Yang, G. Chen, L. Lin, and Y. Yu, "Depthwise nonlocal module for fast salient object detection using a single thread," *IEEE Trans. Cybern.*, vol. 51, no. 12, pp. 6188–6199, Dec. 2021.
[18] D. Faur, I. Gavat, and M. Datcu, "Salient remote sensing image segmentation based on rate-distortion measure," *IEEE Geosci. Remote Sens. Lett.*, vol. 6, no. 4, pp. 855–859, Oct. 2009.
[19] L. Zhang, S. Wang, and X. Li, "Salient region detection in remote sensing images based on color information content," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, 2015, pp. 1877–1880.
[20] L. Zhang, Y. Liu, and J. Zhang, "Saliency detection based on self-adaptive multiple feature fusion for remote sensing images," *Int. J. Remote Sens.*, vol. 40, no. 22, pp. 8270–8297, 2019.
[21] Z. Wang, J. Guo, C. Zhang, and B. Wang, "Multiscale feature enhancement network for salient object detection in optical remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5634819.
[22] 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, 2021, Art. no. 2163.
[23] Y. Niu, Y. Geng, X. Li, and F. Liu, "Leveraging stereopsis for saliency analysis," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2012, pp. 454–461.
[24] D.-P. Fan, M.-M. Cheng, Y. Liu, T. Li, and A. Borji, "Structure-measure: A new way to evaluate foreground maps," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 4558-4567.
[25] C. Li, Y. Yuan, W. Cai, Y. Xia, and D. D. Feng, "Robust saliency detection via regularized random walks ranking," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 2710–2717.
[26] J. Kim, D. Han, Y.-W. Tai, and J. Kim, "Salient region detection via high-dimensional color transform and local spatial support," *IEEE Trans. Image Process.*, vol. 25, no. 1, pp. 9–23, Jan. 2016.
[27] H. Peng, B. Li, H. Ling, W. Hu, W. Xiong, and S. J. Maybank, "Salient object detection via structured matrix decomposition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 4, pp. 818–832, Apr. 2017.
[28] Y. Yuan, C. Li, J. Kim, W. Cai, and D. D. Feng, "Reversion correction and regularized random walk ranking for saliency detection," *IEEE Trans. Image Process.*, vol. 27, no. 3, pp. 1311–1322, Mar. 2018.
[29] Q. Zhang, L. Zhang, W. Shi, and Y. Liu, "Airport extraction via complementary saliency analysis and saliency-oriented active contour model," *IEEE Geosci. Remote Sens. Lett.*, vol. 15, no. 7, pp. 1085–1089, Jul. 2018.
[30] Z. Liu, D. Zhao, Z. Shi, and Z. Jiang, "Unsupervised saliency model with color Markov chain for oil tank detection," *Remote Sens.*, vol. 11, no. 9, 2019, Art. no. 1089.
[31] Z. Deng et al., "R3Net: Recurrent residual refinement network for saliency detection," in *Proc. 27th Int. Joint Conf. Artif. Intell.*, 2018, pp. 684–690.
[32] 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," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2019, pp. 8778–8787.
[33] 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/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 3912–3921.
[34] Y. Pang, X. Zhao, L. Zhang, and H. Lu, "Multi-scale interactive network for salient object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 9410–9419.
[35] X. Zhao, Y. Pang, L. Zhang, H. Lu, and L. Zhang, "Suppress and balance: A simple gated network for salient object detection," in *Proc. 16th Eur. Conf. Comput. Vis.*, 2020, pp. 35–51 .
[36] J. Li, Z. Pan, Q. Liu, and Z. Wang, "Stacked U-shape network with channel-wise attention for salient object detection," *IEEE Trans. Multimedia*, vol. 23, pp. 1397–1409, 2021.
[37] B. Xu, H. Liang, R. Liang, and P. Chen, "Locate globally, segment locally: A progressive architecture with knowledge review network for salient object detection," in *Proc. AAAI Conf. Artif. Intell.*, 2021, pp. 3004–3012.
[38] J. Wei, S. Wang, Z. Wu, C. Su, Q. Huang, and Q. Tian, "Label decoupling framework for salient object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 13022–13031.
[39] J. Wei, S. Wang, and Q. Huang, "F ^{3} net: Fusion, feedback and focus for salient object detection," in *Proc. AAAI Conf. Artif. Intell.*, 2020, pp. 12321–12328.
[40] Y. Lin, H. Sun, N. Liu, Y. Bian, J. Cen, and H. Zhou, "A lightweight multi-scale context network for salient object detection in optical remote sensing images," in *Proc. IEEE 26th Int. Conf. Pattern Recognit.*, 2022, pp. 238–244.
[41] G. Li, Z. Liu, X. Zhang, and W. Lin, "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.
[42] Y. Lin, H. Sun, N. Liu, Y. Bian, J. Cen, and H. Zhou, "Attention guided network for salient object detection in optical remote sensing images," in *Proc. 31st Int. Conf. Artif. Neural Netw. Mach. Learn.*, 2022, pp. 25–36.
[43] G. Li, Z. Liu, D. Zeng, W. Lin, and H. Ling, "Adjacent context coordination network for salient object detection in optical remote sensing images," *IEEE Trans. Cybern.*, vol. 53, no. 1, pp. 526–538, Jan. 2023.

Mo Yang received the B.S. degree in communication engineering in 2021 from Guizhou University, Guiyang, China, where he is currently working toward the M.S. degree in electronic information.

His research interests include deep learning and object detection.

Ziyan Liu received the B.S. degree in automation and the M.S. degree in control theory and control engineering from Guizhou University, Guiyang, China, in 1997 and 2000, respectively.

She is currently a Professor with Guizhou University. Her research interests include computer vision, image processing, and mobile robots.

Prof. Liu is a Senior Member of China Computer Federation and China Institute of Communications.

Wen Dong received the B.S. degree in communication engineering in 2020 from Guizhou University, Guiyang, China, where he is currently working toward the M.S. degree in electronic information.

His research interests include deep learning and instance segmentation.

Ying Wu received the B.S. degree in communication engineering from Southwest Minzu University, Chengdu, China, in 2021. She is working toward the M.S. degree in information and communication engineering from Guizhou University, Guiyang, China.

Her research interests include deep learning and channel estimation.