ORSI Salient Object Detection via Multiscale Joint Region and Boundary Model

Zhengzheng Tu, Chao Wang, Chenglong Li, Minghao Fan, Haifeng Zhao, and Bin Luo, *Senior Member, IEEE*

***Abstract*— Salient object detection (SOD) in optical remote sense images (ORSIs) is a valuable and challenging task. The factors in ORSI, such as background clutter, lighting shadows, imaging blur, and low resolution, significantly degrade the completeness and accuracy of salient objects. To handle this problem, we propose a novel model to learn robust multiscale region features of salient objects by simultaneously optimizing their boundaries. First, we extract multiscale region features of salient objects through a hierarchical attention module. Second, we generate the boundary features by combining the local cues and the global information generated by pyramid pooling. Finally, we embed the boundary features into region features at multiple scales. In particular, we design a joint learning scheme based on a bidirectional feature transformation to optimize boundary and region features simultaneously for accurate ORSI SOD. To provide a comprehensive evaluation platform, we construct a new dataset called ORSI-4199 for ORSI SOD. It contains 4199 finely annotated image pairs with diverse scenes, in which nine attributes (i.e., challenge types) are annotated to facilitate analyzing the strengths and weaknesses of SOD models from different perspectives. Extensive experiments on the public dataset ORSSD, EORRSD, and the newly created dataset ORSI-4199 show that the proposed approach achieves promising results against state-of-the-art methods. https://github.com/wchao1213/ORSI-SOD.**

***Index Terms*— Joint region and boundary learning, multiscale transformation, optical remote sense images (ORSIs), salient object detection (SOD).**

I. Introduction

SALIENT object detection (SOD) is to simulate the human visual system that selectively extracts the most visually noticeable and class-agnostic objects/regions from the whole field of view. SOD has been greatly developed in the natural image; the research for SOD in optical remote sense image (ORSI) also receives some attention. Different from natural images, ORSI has the following properties. First, it covers a wide scope and, thus, contains a lot of noise interference and complex backgrounds. Second, it is collected from a high-altitude overhead view through remote sensing satellites, resulting in a large variance of object scales and orientations. Finally, various weathers make it blurred. Therefore, ORSI SOD is a challenging problem in which many difficulties need to be solved. It is worth noting that SOD is different from practical object detection because the task of practical object detection is to identify class-specific objects in the image. Therefore, SOD cannot be directly extended to practical object detection, but it can be used as an auxiliary component to highlight important regions and suppress background information in practical object detection applications. In addition, SOD can also serve other visual tasks, such as object segmentation and recognition, visual tracking, and image retrieval.

Existing methods usually adopt full convolutional neural networks (FCNs) [1], [2] since combining the semantic information of deep features with the spatial information of shallow features is helpful for locating salient regions in complex scenes. LV-Net [1] adopts multiscale inputs to extract a set of complementary information hierarchically and then uses a nested connection to gradually fuse information of different scales. From Fig. 1, we observe that it is difficult for LV-Net [1] to capture the complete salient object in complex scenes, mainly because LV-Net [1] cannot control the transmission of multiscale information, resulting in the ambiguity of features of different scales. DANet [2] first enhances the high-level feature attention map with shallow cues by using dense attention fluid and then models the global context semantic relationship through the global feature aggregation module. Finally, the cascaded pyramid is used to handle the problem of multiple scales. From Fig. 1, we can find that, even though DANet [2] can locate salient objects, the predicted boundaries of salient objects are not accurate. Although DANet [2] adds a boundary loss to punish boundary errors, this simple way cannot refine boundaries very well.

To handle the above problems, we propose a novel model to learn robust multiscale region features of salient objects by simultaneously optimizing their boundaries. First, we design a hierarchical attention module (HAM) to extract effective multiscale features. The receptive field block (RFB) [3] can obtain different sizes of receptive fields on different branches. However, it concatenates all branches directly, which causes the ambiguity of the features with different scales. To learn

Manuscript received May 8, 2021; revised July 8, 2021 and July 20, 2021; accepted July 24, 2021. Date of publication August 11, 2021; date of current version January 17, 2022. This work was supported in part by the Natural Science Foundation of Anhui Higher Education Institution of China under Grant KJ2020A0033, in part by the Anhui Natural Science Foundation Anhui Energy Internet Joint Fund under Grant 2008085UD07, in part by the Joint Funds of the National Natural Science Foundation of China under Grant U20B2068, in part by the National Natural Science Foundation of China under Grant 61976003, in part by the NSFC Key Project of International (Regional) Cooperation and Exchanges under Grant 61860206004, and in part by the Natural Science Foundation of Anhui Higher Education Institution of China under Grant KJ2019A0005 and Grant KJ2019A0026. *(Corresponding author: Chenglong Li.)*

Zhengzheng Tu, Chao Wang, Chenglong Li, Haifeng Zhao, and Bin Luo are with the Information Materials and Intelligent Sensing Laboratory of Anhui Province, Anhui University, Hefei 230601, China, also with the Anhui Provincial Key Laboratory of Multimodal Cognitive Computation, Anhui University, Hefei 230601, China, and also with the School of Computer Science and Technology, Anhui University, Hefei 230601, China (e-mail: zhengzhengahu@163.com; wchao19971213@163.com; lcl1314@foxmail.com; senith@163.com; luobin@ahu.edu.cn).

Minghao Fan is with the Anhui Province Key Laboratory of Electric Fire and Safety Protection, State Grid Anhui Electric Power Research Institute, Hefei 230601, China (e-mail: mhfansfp@163.com).

Digital Object Identifier 10.1109/TGRS.2021.3101359

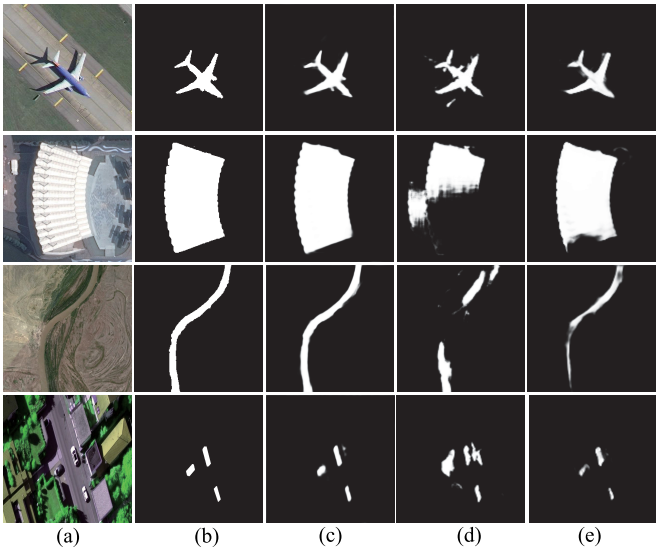


Fig. 1. Several visual examples of different SOD methods. (a) Input images. (b) GT. (c) Results of our MJRBM. (d) Results of LV-Net [1]. (e) Results of DAFNet [2].

more discriminative features and suppress background interference, we integrate the attention mechanism [4] into RFBs to improve the expression of multiscale information. Second, previous studies have shown that adopting the boundary loss alleviates the wrong boundary predicted to a certain extent; however, the problem of inaccurate boundary features still exists. Therefore, we pay more attention to the complementarity between salient boundary features and salient region features. Thus, we design a joint learning scheme based on two direction-specific feature transformations to learn the representations of boundary and region features simultaneously.

There are two ORSI SOD benchmark datasets based on ORSI SOD, including ORSSD [1] and EORSSD [2]. These datasets have several limitations. First, both datasets are not large enough. ORSSD has only 800 images, and EORSSD is an extension of ORSSD, including 2000 images. Second, most images are very simple and less diverse. For example, most images have clear contrast, objects are clearly distinguished from the background, objects are usually in the center of images, and scenes are simple. Finally, these two datasets have no challenge-based labels, and the analysis for ORSI SOD methods under different challenges cannot be performed.

To address these problems, we construct a large dataset for the purpose of the evaluation for ORSI SOD, which contains 4199 finely annotated images. Similar to [5] and [6], we mainly collect these images from Google Earth with the spatial resolution ranging from 0.5 to 2 m. There are large variances in spatial resolutions, object types, scales and directions, light shadows, blurred sensor imaging, and cluttered backgrounds in ORSI. Finally, we annotate nine different challenges (see Table I for the details) to facilitate analyzing the performance of ORSI SOD algorithms under different challenges.

The major contributions of this work can be highlighted as follows.

1) We propose a multiscale joint boundary and region model to generate high-quality boundary-aware ORSI SOD maps by simultaneously optimizing multiscale salient region and boundary features in a joint learning framework.
2) We design an HAM to extract effective multiscale features by integrating attention mechanisms into RFBs.
3) We construct a comprehensive benchmark dataset containing 4199 images with ground-truth (GT) annotations for performance evaluation of different ORSI SOD methods. This dataset will be released to the public for free academic usage.
4) We evaluate the proposed method against 12 state-of-the-art SOD methods on public dataset ORSSD, EORSSD, and the newly created dataset ORSI-4199. The results show that our method achieves the best performance under different evaluation metrics.

II. Related Work

Over the past two decades, some methods have been proposed to detect salient objects in nature images. Early methods usually employ some priors to highlight salient objects [10], such as contrast prior [7], boundary prior [8] and center prior [9]. However, these handcrafted methods have great limitations in capturing high-level semantic information, but semantic information is also very important for SOD tasks. Under these limitations, emerge of FCNs [11] demonstrates its excellent performance in visual tasks. Therefore, the methods based on handcrafted are gradually replaced by the methods based on FCNs [11]. Next, we will discuss the SOD methods based on deep learning in nature image and ORSI.

A. Saliency Object Detection in Nature Images

1) Multilayer Fusion Models: Scale variation is one of the challenges in SOD tasks. At present, many researchers propose different multilayer fusion strategies to simultaneously encode high-layer semantic information and low-layer spatial information. For example, Zhang *et al.* [12] propose a multilayer feature aggregation network, which combines features of different resolutions into multiscale features. Zhang *et al.* [13] propose a novel bidirectional message-passing model to integrate multilayer features, adopting a multiscale context-aware feature extraction module to capture rich context information. Finally, multilayer features are integrated to generate the final saliency map. Deng *et al.* [14] propose to build a sequence of residual refinement blocks to interactively use semantic information and spatial information to refine the prediction results. Wang *et al.* [15] detect the saliency from coarse to fine in a top-down way and also combine high-layer semantic information with shallow-layer detailed features in a bottom-up way to achieve collaborative optimization step by step. Wang *et al.* [16] propose an advanced network for polishing features, in which they modify the features of each level by directly aggregating all the context information in deeper layers. Wei *et al.* [17] design a cross-feature module that can selectively aggregate multilevel features, use a multistage feedback mechanism to enrich features of the previous layer, and eliminate the differences between features. Although these methods integrate the features of different layers to improve

SOD, they excessively use the information of different layers, which ultimately leads to information redundancy and spatial inconsistency in the predicted map. We find that the RFB [3] is helpful for solving the problem of scale difference, which is attributed to the fact that receptive fields with different sizes can obtain the features with different scales. Therefore, we also adopt a similar strategy to RFB [3]. However, as too many zeros are inserted into the convolution layer with a large dilated rate, RFB [3] leads to the decrease in correlation between pixels, called spatial inconsistency. Therefore, we utilize channel attention to alleviate the above problems.

2) Boundary-Aware Models: In order to improve the blurred boundary of salient objects, some methods take advantage of boundary features to get clear boundaries. Luo *et al.* [18] use the IOU-based boundary loss to penalize errors on the boundary and directly optimize the boundary of the predicted saliency map. Zhuge *et al.* [19] propose to use the deep convolutional network to extract and integrate multilevel features and then use boundary information to guide the generation of salient object features. Su *et al.* [20] propose to detect the boundary information and object information, respectively, and then use the transition compensation stream to correct the false boundary. Qin *et al.* [21] propose to generate coarse salient prediction maps with a simple encoder–decoder structure, then propose a new hybrid loss combining binary cross-entropy (BCE) loss, structural similarity (SSIM) index, and Intersection over Union (IoU), and use them to optimize the boundary in the fine-tuning process. Zhao *et al.* [22] use the boundary features to guide the learning for multilayer features that are then concatenated to generate the prediction map. Tu *et al.* [23] design the edge guide block to integrate the edge prior knowledge into the multilayer features and then get the final prediction map. These methods simply use the boundary information to improve the salient region features but pay less attention to optimize the boundary features. Once false boundary features are extracted, they will influence the features of the salient regions. The difference is that we use two specific branches to optimize the boundary features and object features and then combine the object features with the boundary features to get the final prediction map.

B. Salient Object Detection in Optical Remote Sensing Images

Studies on ORSI SOD have also started in recent years. For example, Zhao *et al.* [24] use the sparse representation to obtain global information and background clues and construct two dictionaries based on these two clues to describe global and background ORSI properties. Finally, the Bayesian formula is used to integrate predicted maps generated by the two clues. Li *et al.* [25] propose a parallel downward fusion network, which makes full use of low-level features and high-level features of parallel paths and multiresolution features of cross-paths to detect salient objects with different scales. Li *et al.* [1] propose a two-stream pyramid module to obtain multiscale features and then use the nested connections module to concatenate the multiscale features to get a predicted map. Zhang *et al.* [2] propose to adaptively capture the long-term semantic context and embed it into the dense attention flow model to guide the generation of advanced salient features.

The deep learning-based class-specific activation maps are also used to detect various objects in ORSI. Li *et al.* [26] learn discriminative convolution weights in the first stage of training and class-specific activation weights in the second stage, respectively. Finally, they propose a multiscale scene-sliding-voting strategy to calculate the class-specific activation maps based on the aforementioned weights. Li *et al.* [27] modify the trained deep networks in the testing phase by designing a local pooling pruning to generate high-quality cloud activation maps. Du *et al.* [28] propose a refinement method for weakly supervised object detection, which can obtain high-quality class-specific activation maps and generate nail boxes for suppressing the incorrect refinement direction. Zhang *et al.* [29] propose the pixel area index to assist in detecting the main water bodies and make all water pixels clustered as a guide map. Finally, the main water bodies and the guide map are merged to obtain the final water mask.

In addition, there are some other visual tasks based on saliency detection. Dong *et al.* [5] construct a saliency detection model based on the difference of statistical characteristics between RoIs and homogeneous backgrounds, which could guide the generation of candidate regions. Feng *et al.* [30] adopt saliency detection to help to search RoI in the initial difference image obtained by an improved robust change vector analysis algorithm. Li *et al.* [31] propose to generate three kinds of features based on the peak value, standard deviation, and comparison with the background, respectively, for generating the saliency map, finally segment the saliency map, and extract object regions with a global threshold. Hu *et al.* [32] use the saliency algorithm based on the background prior to obtaining RoI and its location information of the image, then map the location information onto the features, and, finally, classify and fine-tune the location of each RoI.

The SOD in ORSI can be taken as the preprocessing of many vision tasks, as the unique imaging condition of ORSI brings a big challenge to many vision tasks of ORSI. Therefore, the SOD in ORSI should be paid more attention. As far as we know, there are two public datasets, but these datasets lack sufficient complex scenarios and do not have a challenge label. Therefore, we select 4199 ORSIs to construct a larger and more challenging ORSI SOD dataset. In order to better evaluate the advantage of different models, we annotate challenges in each image.

III. Proposed Method

A. Overview of Network Architecture

The overall framework is shown in Fig. 2. As low-level features usually have high resolution and, thus, require a lot of computational costs. In our network, the decoders cascade the salient object features of the top three layers of the backbone, which can reduce the computational cost and make the network learn salient object features from deeper layers. However, if we only use one decoder, the network only gets the coarse salient objects because it is difficult to accurately segment objects in complex scenes of remote sensing images. Therefore, we add the second decoder to integrate multilevel features of the backbone effectively by holistic attention and, thus, reduce the interference of noise and get more accurate

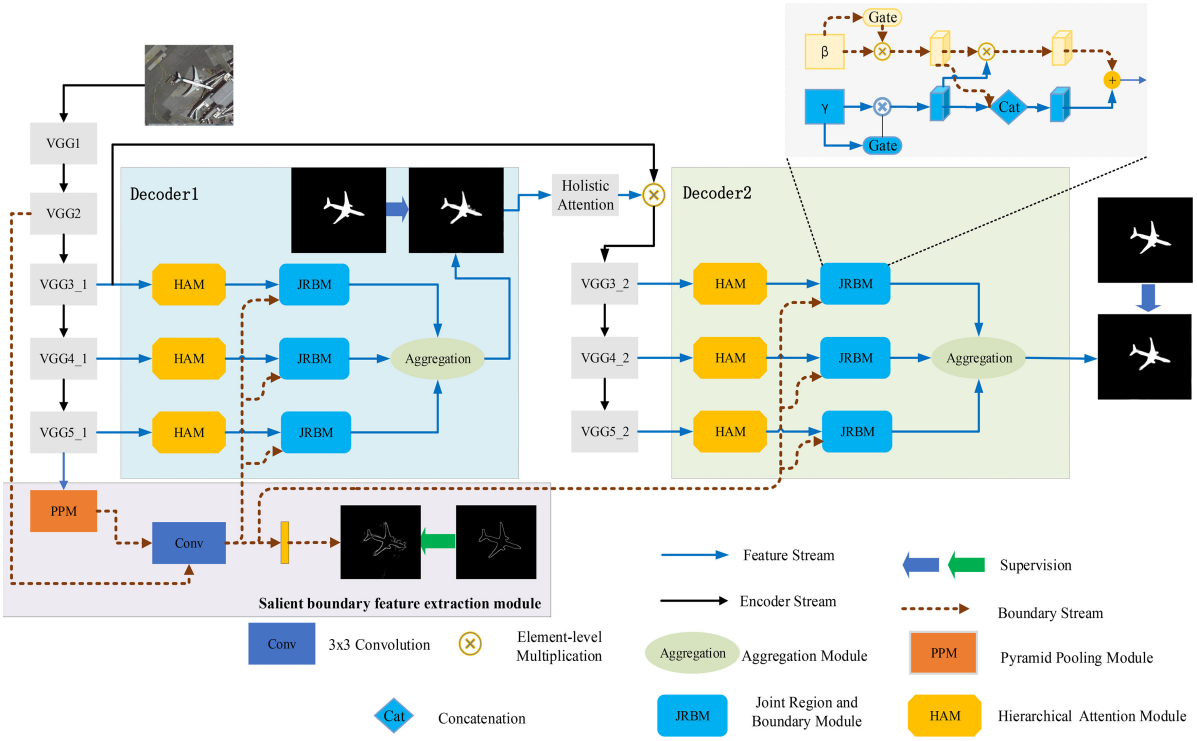


Fig. 2. Overall architecture of the MJRBM, where the encoder comes from CPD [33] (gray rectangle). We use VGG16 [34] as the backbone network. MJRBM consists of three parts: 1) the boundary feature is obtained by concatenating VGG2 and PPM [35]; 2) an HAM is proposed to extract fine multiscale salient region features; and 3) a JRBM is proposed for uniting boundary features and region features. Finally, the optimized features are aggregated to detect the salient map.

and unified highlighted saliency maps. In addition, the works in [33] and [36] also verify that making full use of high-level features and taking dual-decoder is beneficial to predicting more accurate saliency maps. Thus, our model includes two branches, each of which is constructed by encoder–decoder. We will start the description around the first decoder. First, we use PPM [35] to obtain good semantic information and then combine the semantic information with features from the second layer of VGG-16 [34] to extract boundary features. Then, we propose the HAM (see Fig. 3) to extract multiscale features and avoid semantic distinctions between different channels. Third, we design the joint region and boundary module (JRBM) (see Fig. 2) to jointly optimize the boundary and regional features. Finally, we use the aggregation module (see Fig. 4) to aggregate three parallel branches to get the accurate saliency map. In the following, we will elaborate on the boundary feature extraction, the HAM, the JRBM, and the aggregation module.

B. Salient Boundary Feature Extraction

In Fig. 2, the salient boundary feature extraction module aims to extract salient boundary features. The resolution size of VGG1 is the same as the input image size, which not only needs a large amount of calculation but also has a small receptive field. Thus, we discard VGG1 in the decoder. As VGG2 preserves robust boundary of objects [12], we extract the boundary feature from VGG2. However, local information cannot locate the object, while high-level semantic information can fulfill that. Thus, we use high-level semantic information to locate the object. As PPM [35] can capture the global information, which makes the position of the salient target more accurate, we introduce PPM [35] to guide the generation of boundary features of salient objects. As shown in Fig. 2 (brown dotted line), we use VGG2 and PPM to obtain boundary features. The fused features can be denoted as

$$F_e = \text{Conv}(\text{Concat}(F_2, \text{Upsample}(\text{PPM}(F_{5\_1})))) \quad (1)$$

where the Concat(,) means concatenating the feature maps according to the channel axis. Conv(∗) represents a 3×3 convolution layer, which aims to change the number of channels. Upsample(∗) is the bilinear interpolation, which aims to upsample ∗ to the same size as F_2. PPM(∗) is the pyramid pooling module for getting global information.

In order to get clearer boundary features, we add additional boundary supervision to guide the boundary features. The cross-entropy loss can be defined as

$$L_e = -\frac{1}{H \times W} \sum_{x=1}^{H} \sum_{y=1}^{W} \left[E_{xy} \log P_{xy}^{E} + (1 - E_{xy}) \log \left(1 - P_{xy}^{E}\right) \right] \quad (2)$$

where H and W represent the height and width of the GT of the boundary, respectively. $E_{xy} \in \{0, 1\}$ is the boundary label of the pixel (x, y), and P_{xy}^{E} is the predicted probability that a pixel (x, y) belongs to the boundary.

C. Hierarchical Attention Module

As we all know, large convolution kernels can capture large objects, and small convolution kernels are fit for small objects. However, there are various types and scales of salient objects

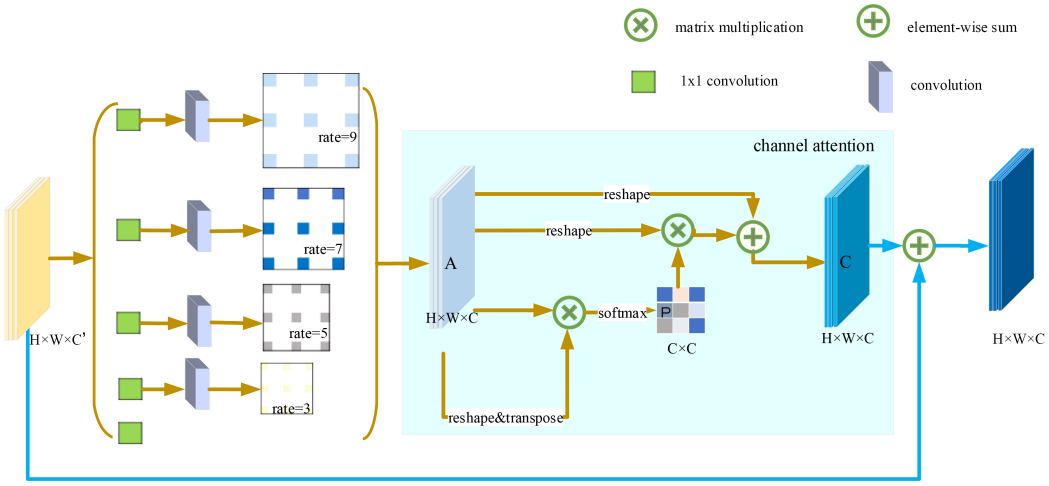


Fig. 3. Overview of the HAM. We first use two convolutional layers: 1×1 convolution layer and $(2b-1) \times (2b-1)$ convolutional layer. Then, we use a 3×3 convolutional layer with $(2b-1)$ dilated rate $\{b = 5, \ldots, 1\}$. We concatenate the outputs of these branches and then take 1×1 convolution layers to reduce the number of channels to 32. Finally, we use the CAM to capture the channel correlation between every two-channel maps. $\bigotimes$: matrix multiplication; $\bigoplus$: elementwise sum operation.

in ORSI; it is not ideal to use the kernel with a single scale. Therefore, we adopt a method similar to [3], which designs an RFB to get multiscale information. Compared with RFB, we use more branches to take advantage of the complementarity of receptive fields with different sizes. As shown in Fig. 3, we first use 1×1 convolution layer to reduce the number of channels to 32. Then, we use $((2b-1) \times (2b-1))$ convolutional layers and 3×3 convolutional layer with $(2b-1)$ dilation $\{b = 5, \ldots, 1\}$; we concatenate these branches and also reduce the number of feature channels to 32. The final concatenated features are ambiguous, as features obtained from different receptive fields cannot distinguish the subtle features, so the salient objects have spatial inconsistency. Then, we use the channel attention map (CAM) [4] to solve the above problems. We directly calculate the CAM with the original feature $\mathbf{A}$. Specifically, we reshape $\mathbf{A}$ to $\mathbb{Q}^{C \times N}$ and then perform the matrix multiplication $\mathbf{A}$ with the transposed $\mathbf{A}$. Finally, we use a softmax to obtain the channel attention map $P \in \mathbb{Q}^{C \times C}$

$$\mathrm{Im}_{ij} = \frac{\exp(A_i \bigotimes A_j)}{\sum_{j=1}^{C} \exp(A_i \bigotimes A_j)} \tag{3}$$

where Im_{ij} measures the impact of the jth channel on the ith channel. In addition, we multiply transpose of $\mathbf{P}$ and $\mathbf{A}$ and reshape their results to $\mathbb{Q}^{H \times W \times C}$. Then, we multiply the result by the parameter α and use the elementwise sum operation with $\mathbf{A}$ to get the final output $C \in \mathbb{Q}^{H \times W \times C}$

$$C_i = \alpha \sum_{j=1}^{C} (\mathrm{Im}_{ij} \times A_j) \bigoplus A_i \tag{4}$$

where α is a parameter with an initial value of 0, which can be learned later. It can be seen from (4) that the final feature of each channel can be obtained by weighting the features of all channels and the original features and then summing them up, and robust semantic correlation between features helps to generate more stable features. Finally, a short connection is added for training the network conveniently.

D. Joint Region and Boundary Module

Boundary information is very important for SOD and can help generate salient objects with fine boundaries. For example, Tu *et al.* [23] design the edge guidance module to use the edge information as prior knowledge to restore the salient object boundary. Li *et al.* [37] propose the gated featurewise transform (GFT), using the extracted boundary features guide accurate object segmentation. However, these methods simply use boundary information to improve the accuracy of prediction maps but do not pay enough attention to improving boundary features. Thus, we want to collaboratively optimize boundary features and regional features. For further verification, we show the comparison results in the experiment section (see Fig. 11 and Table IV).

Specifically, after obtaining multiscale salient region features and boundary features, we design a JRBM (see Fig. 2) to jointly optimize boundary features and salient region features. Since the boundary feature extraction module introduces some noises, we use a gate to make JRBM control the boundary information adaptively. We give the formula of this gate as follows:

$$\hat{\beta} = \sigma(\beta) \times \beta. \tag{5}$$

As extracted multiresolution features of salient regions have noises probably, we also use a gate to adaptively select features, and the specific formula is given as follows:

$$\hat{\gamma} = \sigma(\gamma) \times \gamma \tag{6}$$

where β represents the inputted boundary features and γ represents the inputted region features. $\hat{\beta}$ and $\hat{\gamma}$, respectively, represent the outputted boundary features and salient region features after the gate. $\sigma(*)$ represents the sigmoid function.

According to the logical interrelationship between region features and boundary features of objects [38], the boundary can be seen as a subset of the object region. The specific

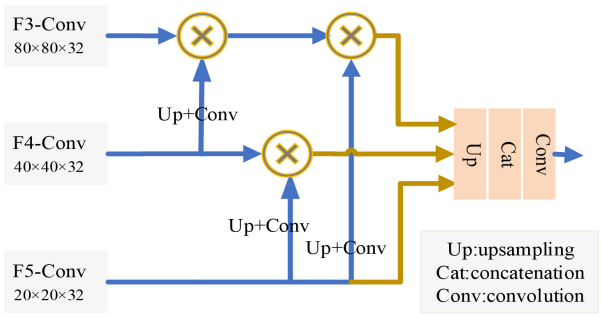


Fig. 4. Overview of aggregation module. Fi-Conv comes from JRBM output ($i = 3, 4, 5$). $\bigotimes$: multiply operation. Cat: concatenation operation.

formula can be represented by

$$\begin{cases} S_b \wedge S_f = S_b \\ S_b \vee S_f = S_f \end{cases} \tag{7}$$

where S_b represents the boundary feature and S_f represents the salient region feature. In addition, $\wedge$ and $\vee$, respectively, represent the AND operation and the OR operation. Since the resolution of boundary features is different from that of the salient region features, we use a convolution with stride to reduce the resolution of boundary features to the same size as the salient region features. When salient region features refine boundary features, we use the multiplication to approximate the Boolean AND operation. Unlike the above methods, we enhance the salient region feature by concatenating it with the boundary feature. The formulation can be written as follows:

$$\begin{cases} \text{feature} = \text{Conv}(\text{Concat}(\hat{y}, \hat{\beta})) \\ \text{boundary} = \text{Conv}\left(\hat{y} \bigotimes \hat{\beta}\right) \end{cases} \tag{8}$$

where Conv$(*)$ represents the 3×3 convolutional layer with 32 output channels. $\bigotimes$ is the elementwise multiplication. Concat$(,)$ is the concatenation operation among channel axis. Finally, the outputted features are generated by adding up fine-tuned salient region features and boundary features in the element level

$$\text{output} = \text{Conv}\left(\text{feature} \bigoplus \text{boundary}\right). \tag{9}$$

E. Aggregation Module

The inputs to the aggregation module (see Fig. 4) are derived from the outputs of JRBM. As MJRBM needs to generate the predicted map, the aggregation module aims to aggregate three parallel branches together. Specifically, it first uses high-level features to optimize low-level features and then concatenates multiple layers of features to make the final prediction. Taking the VGG-16 [34] backbone as an example, this operation is defined as follows:

$$\begin{cases} F4 - \text{Conv} = \text{Up}(\text{Conv}(F5 - \text{Conv})) \bigotimes F4 - \text{Conv} \\ F3 - \text{Conv} = \text{Up}(\text{Conv}(F5 - \text{Conv})) \\ \qquad \bigotimes \text{Up}(\text{Conv}(F4 - \text{Conv})) \bigotimes F3 - \text{Conv} \\ \text{Predict} = \text{Conv}(\text{Concat}(\text{Up}(F5 - \text{Conv}) \\ \qquad \text{Up}(F4 - \text{Conv}), F3 - \text{Conv})) \end{cases} \tag{10}$$

where Up$(*)$ increases the resolution of the high-level feature maps to the same size as the resolution of the low-level feature maps. Predict is the predicted map.

F. Loss Function

We jointly train two branches by taking the GT for supervision and do not share the parameters of the two branches. In addition, there is also boundary supervision, as shown in (2), so the total loss is calculated as

$$L_{\text{total}} = L_p + \omega L_c + \eta L_e. \tag{11}$$

L_c and L_p belong to cross-entropy loss. ω and η are the hyperparameters. c represents the coarse prediction branch from the first branch, and p represents the final prediction branch from the second branch

$$L_i = -\frac{1}{H \times W} \sum_{x=1}^{H} \sum_{y=1}^{W} \left[G_{xy} \log P_{xy}^{P} + (1 - G_{xy}) \log\left(1 - P_{xy}^{P}\right) \right] \tag{12}$$

where $i \in \{c, p\}$. H and W represent the height and width of the input, respectively. $G_{xy} \in \{0, 1\}$ is the GT label of the pixel (x, y), and P_{xy}^{P} is the probability that a pixel (x, y) belongs to salient regions.

IV. Proposed Dataset

To the best of our knowledge, the ORSSD dataset [1] and the EORSSD dataset [2] are publicly available for ORSI SOD. In these ORSI SOD images, most of the objects are in the center of the image and have clear boundaries and textures, which are not enough for evaluating a model. The dataset needs to contain more challenges, such as various sizes of objects and background interference.

Thus, we construct a more challenging dataset for evaluating ORSI SOD models. In order to construct the dataset, we initially collect 6020 images, most of which are from Google Earth, and a small part of which are from the existing ORS object detection datasets, such as the AID dataset [5], the DIOR dataset [39], the DOTA dataset [40], and the NWPU VHR-10 dataset [41]. These images were selected according to one of the following rules at least.

1) There are multiple salient objects in an image.
2) Large differences in the shape or size of the object in the same category.
3) Strong background interference.
4) The texture and the structure of the object in an image are complex.
5) The salient object is at the image boundary, not in the center.
6) The color contrast (the color ratio of the salient object and its surrounding region) is less than 0.6.
7) Due to some special weathers, the images taken by the aerial sensor are blurred.

To reduce label inconsistency, we ask six people to annotate salient objects in 6020 images and divide every two people into a group: one person in each group annotates the image, and the other person checks that. That is, each group individually

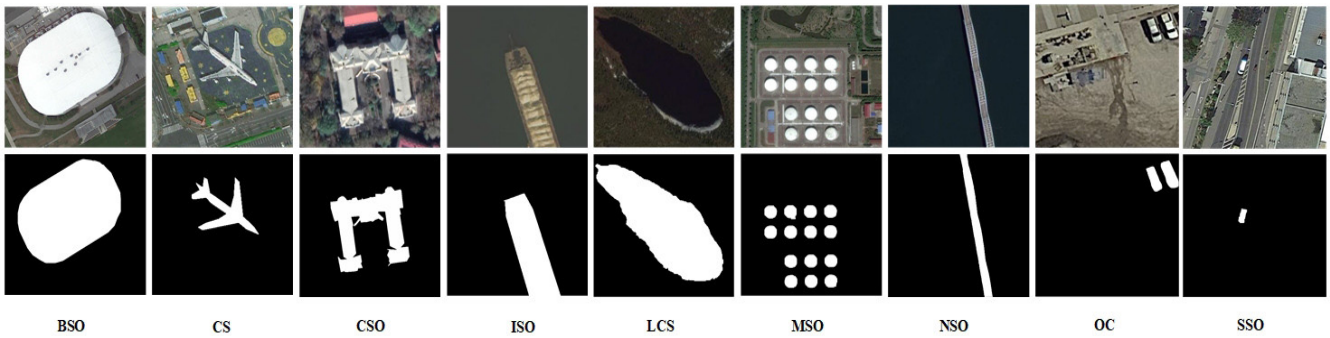


Fig. 5. Samples of several challenges in the ORSI-4199 dataset that we constructed. The first row shows the ORS images. The second row gives the pixelwise annotation.

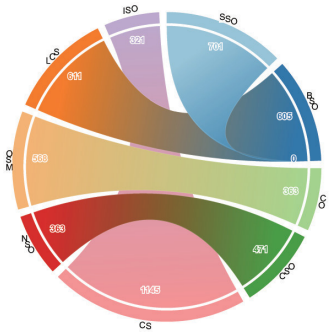


Fig. 6. Number of images belonging to each attribute in the proposed ORSI-4199 dataset.

annotates 6020 images, using a designed interactive annotation method. On average, each group takes 2–3 min to annotate an image. It takes six months to produce the dataset.

Assume that $a_x = 1$ in the image means that the object x is annotated as a salient one, and $a_x = 0$ indicates that it is not salient.

1) If there is no dispute between members of the same group about an annotated image, the number of annotations in the same position of the image is not less than 2.
2) If there is a dispute about an annotated image between members of a group, six people will annotate this image individually; finally, we will annotate it if the number of annotations in the same position is not less than 3.

If an image meets the above criteria, it will be selected for annotation; otherwise, it will be discarded. Excluding the images that do not meet the above rules, we finally retain 4199 images. Each image has its own attributes, which helps to objectively evaluate the model performance under different attributes. The challenging attributes of the dataset are summarized in detail, as shown in Table I.

Some sample images in our dataset are shown in Fig. 5. It is worth noting that an image may have multiple challenge attributes. For example, the image in the penultimate column not only has the attribute **OC** but also has **MSO**. The attribute of the image in the last column not only has **SSO** but also has **CS**. We also give the distribution of attributes in our dataset, as shown in Fig 6. **BSO** and **SSO** together account for about 31%, indicating that the shape and the size of salient objects in our dataset are very different. **CS** and **LCS** occupy approximately 42%, indicating that there are many cluttered backgrounds and blurred scenes, which are more consistent with the real-world scene.

TABLE I

SALIENT OBJECTS' ATTRIBUTES IN ORSIS AND THE CORRESPONDING DESCRIPTION IN ORSI-4199

Attribute	Description
BSO (big salient object)	The ratio between salient object and the entire image is not less than 0.5.
SSO (small salient object)	The ratio between salient object and the entire image is not larger than 0.3.
OC (off center)	The salient object is at the border of the image, not at the center.
CSO (complex salient object)	The salient object is complicated, mainly due to the texture, structure and complex boundaries of the object itself.
CS (complex scene)	Cluttered background region or blurred object.
NSO (narrow salient object)	The salient object is narrow and slender, such as the river and the road.
MSO (multiple salient objects)	The image has multiple salient objects.
LCS (low contrast scene)	The minimum chi-square distance between the color histograms of the salient object and its surrounding area is less than 0.6.
ISO (incomplete salient object)	The salient object is not complete.

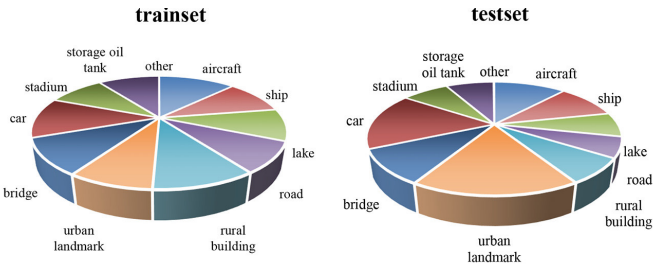


Fig. 7. Distribution of object category in the constructed training set and test set.

In the proposed dataset, most of the objects are aircraft, ship, rural building, lake, road, storage oil tank, urban landmark, car, stadium, and bridge. The specific distribution is shown in Fig. 7. It can be seen from the pie chart that the object categories in the training set are almost evenly distributed, while, in the test set, high-proportion categories are urban

landmarks, cars, and aircraft. In the test set, there are more challenges in these high-proportion categories, which is more beneficial to reflect the performance of a method. The proportion of each kind of object has been changed compared to that in the training set, which can reflect the generalization ability of the method.

V. Experiments

A. Dataset

We evaluate the performance of the proposed SOD method on the ORSSD dataset, the EORSSD dataset, and the ORSI-4199 dataset that we constructed.

1) *ORSSD [1]:* ORSSD contains 600 images with the corresponding GT.
2) *EORSSD [2]:* EORSSD is an extended version of ORSSD, which contains 2000 images with the corresponding GT.
3) *ORSI-4199:* We construct a new dataset with 4199 images for SOD in ORSI and also with pixelwise GT annotation of objects.

B. Implementation Details

We add data diversity to the EORSSD dataset and the ORSSD dataset according to authors' recommendations [1], [2]. We use flips and rotations to increase the diversity of the training set of the ORSI-4199 dataset. Each of input image is resized to 352 × 352. We use the Pytorch to build our model and use a Tesla P100-PCIE GPU for acceleration. The parameter initialization of our backbone is determined by VGG-16 [34] and ResNet-50 [35], and other convolutional layers are initialized using Pytorch's default settings. We do not apply any postprocessing to our model. The proposed model is trained by Adam's optimization method [47]. The batch size is set to 15, the learning rate is set as 10^{-4}, and the number of training is 100 epochs.

C. Evaluation Criteria

We use three measures to evaluate our method: precision–recall (PR) curve [48], F-measure [48], and mean absolute error (MAE) [49].

The precision and recall are calculated by the GT and the predicted saliency map. The PR curve is drawn by the precision and the recall. The vertical axis represents the precision, and the horizontal axis represents the recall. The PR curve is more closed to the coordinates (1, 1), the better the model performance will be.

F-measure is defined as the weighted harmonic mean of precision and recall, and F-measure is represented as

$$F_\beta = \frac{(1+\beta^2)\cdot \text{Precision}\cdot \text{Recall}}{\beta^2 \cdot \text{Precision}\cdot \text{Recall}} \tag{13}$$

where β^2 is set to 0.3 for emphasizing precision over recall as same as [48].

MAE is to measure the difference between predicted map S and GT G

$$\text{MAE} = \frac{1}{W\times H}\sum_{i=1}^{W}\sum_{j=1}^{H} \mid S(i,j) - G(i,j) \mid \tag{14}$$

TABLE II

Quantitative Evaluation. The Max F_β (Larger Is Better) and MAE (Smaller Is Better) of Different Methods on the ORSSD and ORSI-4199 Datasets. The Best Results Are in Red, Green, and Blue

	ORSI-4199		ORSSD		EORSSD	
Methods	max F_β	MAE	max F_β	MAE	max F_β	MAE
MJRBM-ResNet	0.8685	0.0372	0.9025	0.0145	0.8701	0.0099
MJRBM-VGG	0.8628	0.0373	0.9014	0.0141	0.8765	0.0095
LV-Net[1]	-	-	0.8414	0.0207	0.8051	0.0145
RAS [42]	0.7861	0.0671	0.8782	0.0159	0.8491	0.0105
ENFNet[23]	0.7764	0.0608	0.8758	0.0216	0.8506	0.0109
CPD[33]	0.8507	0.0421	0.8862	0.0151	0.8667	0.0107
BASNet[21]	0.8405	0.0454	0.8750	0.0155	0.8631	0.0112
EGNet[22]	0.8449	0.0440	0.8492	0.0283	0.8453	0.0136
PoolNet[43]	0.8312	0.0541	0.8606	0.0229	0.8341	0.0124
AADFNet[44]	0.8367	0.0488	0.8354	0.0235	0.7169	0.0676
R3Net[14]	0.8250	0.0401	0.8167	0.0241	0.7236	0.0394
PiCANet[45]	0.7173	0.0974	0.8519	0.0302	0.8081	0.0202
PFAN[46]	0.8203	0.0580	0.8344	0.0543	0.7740	0.0159
NLDF[18]	0.7723	0.0636	0.8537	0.0273	0.8148	0.0202

where W and H are the width and the height of GT. The smaller the MAE score is, the closer the prediction mask is to GT.

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

We compare our method with 12 state-of-the-art SOD methods on ORSSD, EORSSD, and ORSI-4199, including PiCANet [45], R3Net [14], PFAN [46], PoolNet [43], BASNet [21], EGNet [22], CPD [33], AADFNet [44], NLDF [18], LVNet [1], ENFNet [23], and RAS [42]. All results are generated by the codes provided by their authors. For a fair comparison, we use ORSSD, EORSSD, and ORSI-4199 datasets to retrain the deep learning-based SOD methods. LVNet [1] is the earliest ORSI SOD method. It does not provide the code but provides the test results on ORSSD and EORSSD.

1) Quantitative Comparison: Table II shows the results of our method and other state-of-the-art methods on three test datasets. It can be seen that our method outperforms the existing ORSI SOD methods. Compared with LV-Net [1], our method exceeds it by 5.71% and 31.40% on the ORSSD dataset with the max F_β and MAE metrics, respectively. The reason is that our method extracts object features with different scales accurately and eliminates the ambiguity of features caused by spatial inconsistency. In addition, BASNet [21], EGNet [22], PoolNet [43], and ENFNet [23] use boundary prior to guide the generation of salient objects. We can see that our method also surpasses them significantly on the complex ORSI-4199 dataset ([2.23%, 17.87%], [1.79%, 15.23%], [3.16%, 31.05%], and [8.64%, 38.65%]), as we jointly learn boundary features and regional features, which can effectively avoid the nonideal prediction caused by wrong boundaries. In Table II, our method still gets the best results on the

TABLE III

ATTRIBUTE-BASED PERFORMANCE ON THE CHALLENGING ORSI-4199 DATASET. THE AVERAGE SALIENT-OBJECT PERFORMANCE IS PRESENTED IN LAST ROW. THE TOP THREE MAX F_β'S ARE DISPLAYED IN RED, GREEN, AND BLUE

Attr	AADFNet[44]	ENFNet[23]	RAS[42]	BASNet[21]	CPD[33]	EGNet[22]	PFAN[46]	PICANet[45]	PoolNet[43]	R3Net[14]	NLDF[18]	MJRBM
BSO	0.9147	0.8687	0.8209	0.9043	0.9158	0.9139	0.8969	0.7908	0.8833	0.9029	0.8808	0.9166
CS	0.8821	0.8232	0.8132	0.8794	0.8887	0.8734	0.8615	0.7553	0.8603	0.8663	0.8255	0.8961
CSO	0.8780	0.8196	0.8106	0.8736	0.8831	0.8705	0.8568	0.7512	0.8539	0.8629	0.8214	0.8899
ISO	0.9040	0.8348	0.8179	0.8883	0.8928	0.9017	0.8739	0.7631	0.8643	0.9142	0.8600	0.8937
LCO	0.7603	0.7135	0.7358	0.7859	0.7820	0.7808	0.7620	0.6428	0.7605	0.7555	0.7115	0.7916
MSO	0.8211	0.7792	0.8142	0.8181	0.8524	0.8372	0.8064	0.7846	0.8324	0.8072	0.7736	0.8585
NSO	0.8256	0.6959	0.6904	0.8261	0.8384	0.8479	0.8442	0.5827	0.8253	0.8449	0.6796	0.8748
OC	0.7547	0.7401	0.7642	0.8104	0.8156	0.7897	0.7675	0.7479	0.8060	0.7643	0.7151	0.8422
SSO	0.7539	0.6972	0.7573	0.7645	0.7874	0.7681	0.7338	0.7145	0.7755	0.7152	0.6921	0.8091
Avg	0.8327	0.7747	0.7805	0.8390	0.8507	0.8426	0.8226	0.7259	0.8291	0.8259	0.7733	0.8636

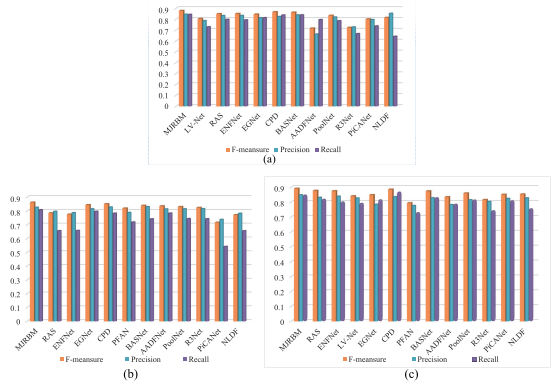


Fig. 8. Comparison of mean F_β, precision, and recall among different methods on ORSSD, EORSSD, and ORSI-4199 datasets. (a) Results on the EORSSD dataset. (b) Results on the ORSI-4199 dataset. (c) Results on the ORSSD dataset.

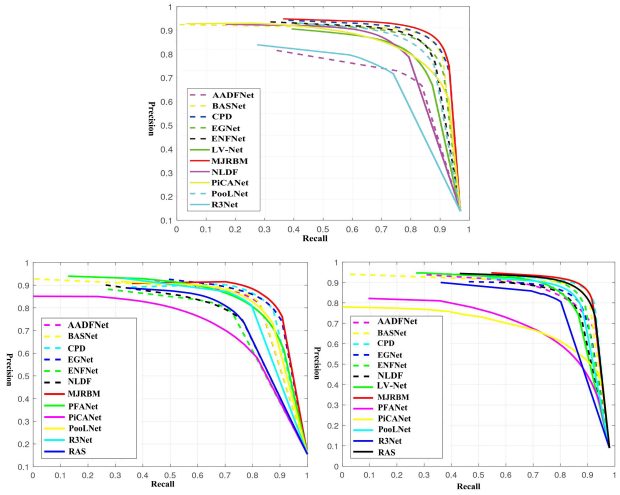


Fig. 9. First row shows the PR curves of different methods on the EORSSD dataset. The second row left shows PR curves of different methods on the ORSI-4199 dataset. The second right shows PR curves of different methods on the ORSSD dataset.

EORSSD dataset. Compared with the second and third places, our MAE metric increases 9.5% and 11.21%, respectively. The most important is that the max F_β and MAE metrics have significant increases ([7.14%, 34.48%]) compared with LV-Net [1]. In addition, in order to verify that our proposed workflow is effective, even based on other popular backbone networks, as an example, we take that ResNet-50 [50] replaces VGG-16 [34] and do the experiments, as shown in Table II. We find that the performance of the network based on ResNet-50 [50] is close to the performance with VGG-16 [34] on these three datasets. The reason is given as follows. Although the ResNet-50 [50] is better at extracting abundant features than VGG-16 [34], it has a deeper architecture which results in lower resolution of feature maps. Therefore, the extracted salient region features from ResNet-50 [50] cannot capture tiny objects or boundaries very well. In addition, inaccurate boundary features bring wrong guidance information to the network and then lead to inaccurate saliency predictions. Anyway, our network flow with different backbones can be superior to other methods. The PR curves, precision, and recall of the abovementioned methods on three test datasets are shown in Figs. 8 and 9. It can be seen from Fig. 8 that our model not only has high precision and recall on three datasets but also has small gaps between the two evaluation indicators. This result means that the saliency map that we predict always shows consistent confidence in the salient region. In PR curves (see Fig. 9), it can be seen that our model is closer to the (1, 1) coordinate, meaning that our model performs much better than other models.

2) Attributes-Based Performance on ORSI-4199: In the challenging ORSI-4199 dataset, each image has attributes that reflect the typical challenges in the remote sensing scene. These annotations help to study the advantages and disadvantages of SOD methods. Table III shows the scores of the proposed model and the state-of-the-art models. We can see that our model ranks first on seven of the nine attributes. In addition, our model also ranks first on average. These results show that the proposed model outperforms existing models in most challenging situations. Although the scores that we obtained on the one attribute are lower than R3Net [14], it applied the postprocessing methods to refine their salient maps.

3) Qualitative Evaluation: Several representative examples are shown in Fig. 10. These examples are from various scenes, including small salient objects (first, second, fifth, and sixth rows), narrow salient objects (second and fifth rows),

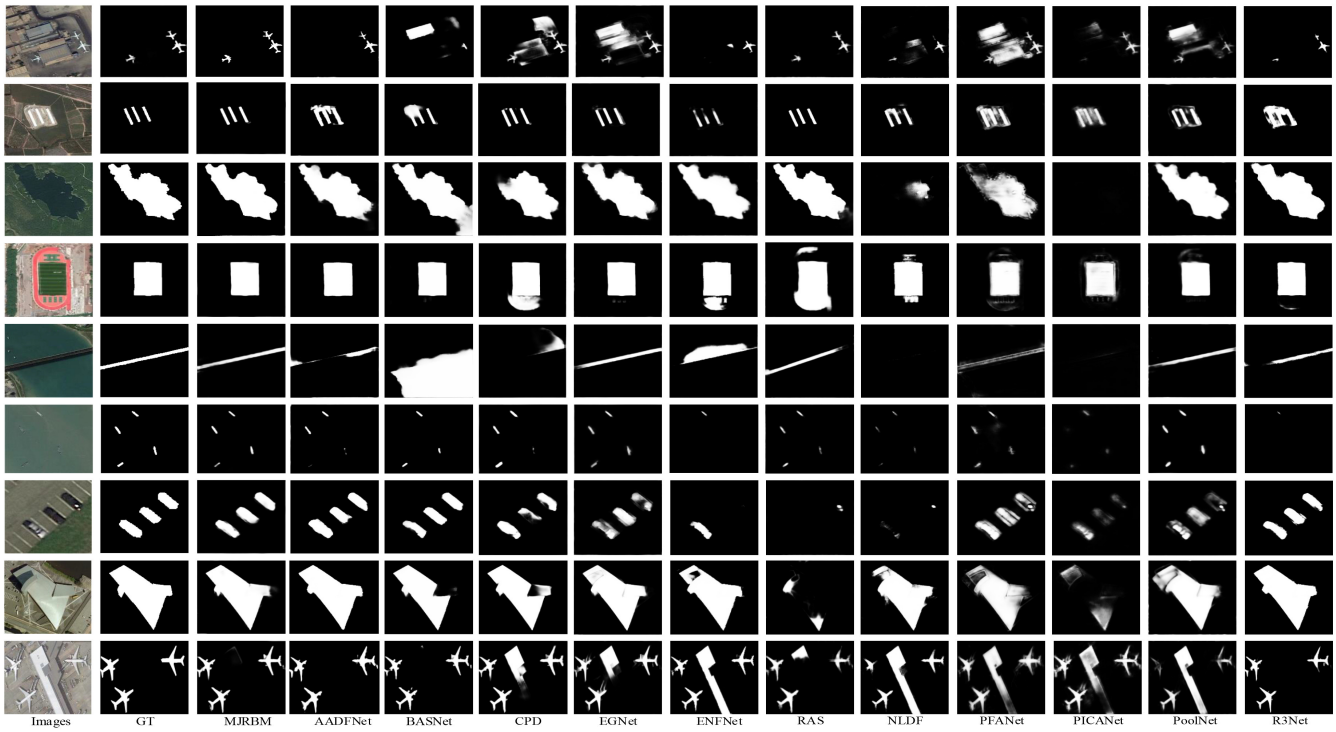


Fig. 10. Visual comparisons between our results and state-of-the-art methods. Our results perform best for images with various challenges. (GT: ground truth.)

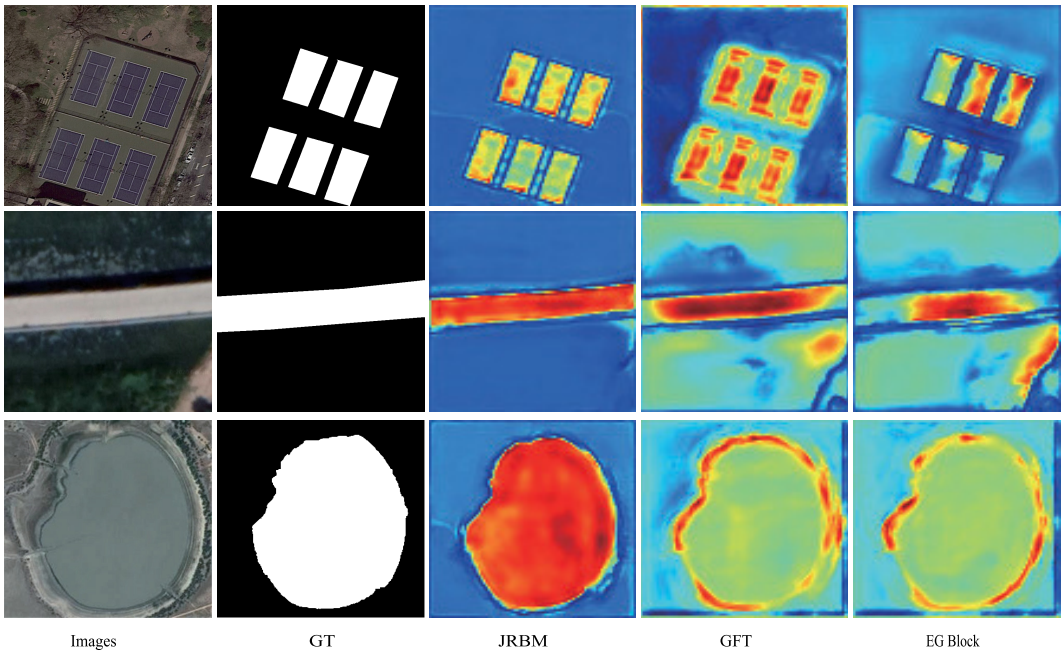


Fig. 11. Visual comparison of feature maps obtained by different modules. The first and second columns are the image and corresponding GT, respectively. The last three columns are the visualization of the feature map obtained by our proposed module (JRBM), GFT [37], and the EG Block [23], respectively.

low contrast scenes (third, fifth, and sixth rows), complex scenes (fourth and ninth rows), and big salient objects (fourth and eighth rows). When objects have large scale difference, AADFNet [44], ENFNet [23], and EGNet [22] can only detect large objects, while small objects are lost. Even though other methods (CPD [33] and BASNet [21]) can locate the object, they cannot get a refined prediction map. In addition, in complex scenarios, most models cannot predict accurate results because they are influenced by too much noise. In contrast, our method not only accurately locates the salient target but also obtains better boundary prediction, which benefits from the proposed two modules.

E. Ablation Study

1) Module Comparison: We adopt the EG Block [23] and GFT [37] to compare with our proposed JRBM. The EG Block and GFT use boundary information as a prior to guiding the generation of salient region features. As shown

TABLE IV

RESULTS GENERATED WITH THE EG BLOCK [23], GFT [37], OR JRBM MODULES ON THE ORSI-4199 DATASET

	max F_β	MAE
EG Block [23]	0.8520	0.0423
GFT [37]	0.8535	0.0412
JRBM	0.8628	0.0373

TABLE V

INFLUENCE OF DIFFERENT HYPERPARAMETERS IN THE LOSS FUNCTION ON THE PREDICTION MAPS. THE SCORE OF MAX F_β AND MAE IN OUR METHOD WHEN ω AND η ARE GIVEN DIFFERENT VALUES. THE BEST RESULT IS SHOWN IN RED. THE TEST DATASETS ARE EORSSD AND ORSSD

	EORSSD		ORSSD	
ω=1	max F_β	MAE	max F_β	MAE
η =0.9	0.8700	0.0100	0.8851	0.0151
η=0.8	0.8718	0.0109	0.8940	0.0164
η=0.7	0.8758	0.0095	0.9014	0.0141
η=0.6	0.8697	0.0101	0.8861	0.0166
η=1	max F_β	MAE	max F_β	MAE
ω=0.9	0.8736	0.0102	0.8960	0.0150
ω=0.8	0.8720	0.0098	0.8914	0.0146
ω=0.7	0.8740	0.0105	0.8927	0.0170
ω=0.6	0.8718	0.0110	0.8859	0.0180
η=1, ω=1	0.8723	0.0098	0.8943	0.0151

in Table IV, JRBM outperforms the EG Block [23] with the increase in 1.08% and 11.82% in terms of the max F_β and MAE, respectively. Similarly, compared with GFT [37], max F_β and MAE increase by 0.93% and 9.47%, respectively. Obtained from Table IV, GFT [37] can improve the accuracy than the EG Block [23]. The EG Block [23] extracts edges from the original images and then uses edges as an input to help generate accurate salient regions. However, remote sensing images contain redundant texture information, and the method with direct guidance produces some wrong information, which leads to unsatisfactory results. Different from the EG Block [23], GFT [37] adopts an adaptive method to improve the features of salient regions with boundary features. This method still has limitations because the generation of boundary features is relatively stable, and the wrong boundary cannot be modified during the network training, resulting in the limited enhancement for features of salient regions. The JRBM module uses a bidirectional feature transformation to collaboratively optimize boundary features and region features, which can modify the incorrect feature in time. In addition, to show the effect of JRBM more intuitively, we visualize the features in Fig. 11. It can be seen that JRBM can learn clearer boundaries and more accurate salient regions through joint learning for boundaries and regions.

2) Loss Function Analysis: We add the hyperparameter ω to the coarse prediction branch corresponding to subitem L_c and the hyperparameter η to boundary prediction branch corresponding to subitem L_e in the total loss function. We set up two sets of experiments: 1) when $\omega = 1$ remains unchanged, we assign different values to η and 2) when $\eta = 1$ remains unchanged, we assign different values to ω. It can be seen from Table V that, when $\omega = 1$ and $\eta = 0.7$, the network can get the best performance. For our network, the parameters of the coarse prediction branch and final prediction branch are not shared, the coarse prediction branch is to obtain a rough predicted map to refine features in the encoder of the final prediction branch, and the final prediction branch helps the coarse prediction branch to focus on the salient object. By giving $\omega = 1$ to the subloss item of the coarse prediction branch, our model can uniformly highlight the salient objects while suppressing the interference. The boundary prediction branch is to obtain boundary features so that the region features and boundary features are jointly optimized. The adding weight $\eta = 0.7$ to the boundary loss subitem (L_e) is helpful to optimize the learning process and avoid excessive weight from affecting the optimization direction. This shows that reasonable boundary optimization can improve salient features and ultimately get more accurate predicted maps.

TABLE VI

QUANTITATIVE EVALUATION FOR ABLATION STUDIES ON THE ORSI-4199 AND ORSSD DATASETS. "JRBM" IS THE JOINT REGION AND BOUNDARY MODULE, "AGGREGATION" IS THE AGGREGATION MODULE, AND "HAM" IS THE HIERARCHICAL ATTENTION MODULE

	ORSI-4199		ORSSD	
	max F_β	MAE	max F_β	MAE
Baseline	0.8221	0.0511	0.8164	0.0456
JRBM	0.8442	0.0432	0.8448	0.0378
HAM	0.8377	0.0424	0.8508	0.0273
HAM+aggregation	0.8498	0.0412	0.8777	0.0213
JRBM+aggregation	0.8501	0.0427	0.8618	0.0218
RFB [3]+JRBM+aggregation	0.8568	0.0410	0.8862	0.0190
MJRBM	0.8628	0.0373	0.9014	0.0141

3) Module Analysis: In order to show the effect of each module on the overall network, we perform the following ablation studies.

1) *Baseline:* We use the network architecture proposed in [33], and the decoder uses 3 × 3 convolution instead.
2) *JRBM:* We only use the JRBM in the network decoder.
3) *HAM:* We only use the HAM in the network decoder.
4) *HAM + aggregation:* The HAM and the aggregation module are used in the network decoder.
5) *JRBM + aggregation:* The JRBM and the aggregation module are used in the network decoder.
6) *RFB + JRBM + aggregation:* The HAM in the proposed network is replaced by the RFB [3] module, and the other modules remain.
7) *MJRBM:* It means the proposed complete network.

Table VI shows the benefit of each module quantitatively. We take the VGG-16 backbone with encoder framework [33] as the baseline. As shown in Table VI, when JRBM is added to baseline, the performance has been significantly improved by 2.84% and 17.11%, respectively, on max F_β and MAE on

the ORSSD [1] dataset. These data prove that JRBM can learn accurate boundaries of the salient object, and there is a small difference between false negatives and false positives. When adding HAM to the baseline, max F_β and MAE increased by 3.44% and 40.13%, respectively, which also confirms that this approach improves missing detection for salient objects. Adding an aggregation module can also help refine the final predicted map. In order to verify that HAM is better than RFB module [3], we add a comparison for them and find that max F_β and MAE increased by 1.23% and 25.26%, respectively. This comparison experiment proves the spatial inconsistency might be caused by the aggregation of branches of multiple receptive fields, which ultimately causes missing some parts in the predicted map. The HAM module can avoid suboptimal results caused by spatial inconsistency. On the ORSI-4199 dataset, it also shows that the module proposed in this article is helpful for accurate SOD detection.

VI. Conclusion

In this article, we propose a novel multiscale joint region and boundary model for ORSI SOD. We adopt the HAM to extract multiscale information and increase the correlation between channels, which is conducive to detecting salient objects with different scales. Then, we fuse local cues and global information to extract boundary features. We design a JRBM to embed the boundary features into the multiscale regional features. The module uses the jointly learning method to simultaneously optimize the boundary features and regional features. Finally, the proposed model can obtain complete predicted maps with accurate boundaries. Moreover, we construct an ORSI dataset with attribute annotations. Experimental results demonstrate that our proposed ORSI SOD network improves the performance on available datasets and outperforms the state-of-the-art methods under different evaluation metrics.

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] S. Liu, D. Huang, and A. Wang, "Receptive field block net for accurate and fast object detection," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 385–400.

[4] J. Fu *et al.*, "Dual attention network for scene segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 3146–3154.

[5] C. Dong, J. Liu, and F. Xu, "Ship detection in optical remote sensing images based on saliency and a rotation-invariant descriptor," *Remote Sens.*, vol. 10, no. 3, p. 400, Mar. 2018.

[6] Z. Xiao, Y. Gong, Y. Long, D. Li, X. Wang, and H. Liu, "Airport detection based on a multiscale fusion feature for optical remote sensing images," *IEEE Geosci. Remote Sens. Lett.*, vol. 14, no. 9, pp. 1469–1473, Sep. 2017.

[7] M.-M. Cheng, N. J. Mitra, X. Huang, P. H. S. Torr, and S.-M. Hu, "Global contrast based salient region detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 3, pp. 569–582, Mar. 2015.

[8] C. Yang, L. Zhang, H. Lu, X. Ruan, and M.-H. Yang, "Saliency detection via graph-based manifold ranking," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2013, pp. 3166–3173.

[9] H. Jiang, J. Wang, Z. Yuan, Y. Wu, N. Zheng, and S. Li, "Salient object detection: A discriminative regional feature integration approach," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2013, pp. 2083–2090.

[10] A. Borji, M.-M. Cheng, H. Jiang, and J. Li, "Salient object detection: A benchmark," *IEEE Trans. Image Process.*, vol. 24, no. 12, pp. 5706–5722, Dec. 2015.

[11] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 3431–3440.

[12] P. Zhang, D. Wang, H. Lu, H. Wang, and X. Ruan, "Amulet: Aggregating multi-level convolutional features for salient object detection," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 202–211.

[13] L. Zhang, J. Dai, H. Lu, Y. He, and G. Wang, "A bi-directional message passing model for salient object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 1741–1750.

[14] Z. Deng *et al.*, "R^3Net: Recurrent residual refinement network for saliency detection," in *Proc. 27th Int. Joint Conf. Artif. Intell.*, Jul. 2018, pp. 684–690.

[15] W. Wang, J. Shen, M.-M. Cheng, and L. Shao, "An iterative and cooperative top-down and bottom-up inference network for salient object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 5968–5977.

[16] B. Wang, Q. Chen, M. Zhou, Z. Zhang, X. Jin, and K. Gai, "Progressive feature polishing network for salient object detection," in *Proc. AAAI Conf. Artif. Intell.*, 2020, pp. 12128–12135.

[17] J. Wei, S. Wang, and Q. Huang, "F^3Net: Fusion, feedback and focus for salient object detection," in *Proc. AAAI Conf. Artif. Intell.*, 2020, pp. 12321–12328.

[18] 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. 6609–6617.

[19] Y. Zhuge, G. Yang, P. Zhang, and H. Lu, "Boundary-guided feature aggregation network for salient object detection," *IEEE Signal Process. Lett.*, vol. 25, no. 12, pp. 1800–1804, Dec. 2018.

[20] J. Su, J. Li, Y. Zhang, C. Xia, and Y. Tian, "Selectivity or invariance: Boundary-aware salient object detection," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 3799–3808.

[21] X. Qin, Z. Zhang, C. Huang, C. Gao, M. Dehghan, and M. Jagersand, "BASNet: Boundary-aware salient object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 7479–7489.

[22] J. 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. (ICCV)*, Oct. 2019, pp. 8779–8788.

[23] Z. Tu, Y. Ma, C. Li, J. Tang, and B. Luo, "Edge-guided non-local fully convolutional network for salient object detection," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 31, no. 2, pp. 582–593, Feb. 2021.

[24] D. Zhao, J. Wang, J. Shi, and Z. Jiang, "Sparsity-guided saliency detection for remote sensing images," *J. Appl. Remote Sens.*, vol. 9, no. 1, Sep. 2015, Art. no. 095055.

[25] 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, Nov. 2020.

[26] Y. Li, Y. Zhang, X. Huang, and A. L. Yuille, "Deep networks under scene-level supervision for multi-class geospatial object detection from remote sensing images," *ISPRS J. Photogramm. Remote Sens.*, vol. 146, pp. 182–196, Dec. 2018.

[27] Y. Li, W. Chen, Y. Zhang, C. Tao, R. Xiao, and Y. Tan, "Accurate cloud detection in high-resolution remote sensing imagery by weakly supervised deep learning," *Remote Sens. Environ.*, vol. 250, Dec. 2020, Art. no. 112045.

[28] P. Du, H. Zhang, and H. Ma, "Classifier refinement for weakly supervised object detection with class-specific activation map," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2019, pp. 3367–3371.

[29] Y. Zhang, X. Liu, Y. Zhang, X. Ling, and X. Huang, "Automatic and unsupervised water body extraction based on spectral-spatial features using GF-1 satellite imagery," *IEEE Geosci. Remote Sens. Lett.*, vol. 16, no. 6, pp. 927–931, Jun. 2019.

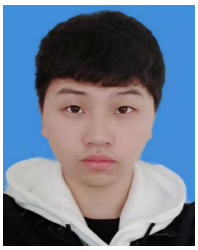
[30] W. Feng, H. Sui, J. Tu, W. Huang, and K. Sun, "A novel change detection approach based on visual saliency and random forest from multi-temporal high-resolution remote-sensing images," *Int. J. Remote Sens.*, vol. 39, no. 22, pp. 7998–8021, Nov. 2018.

[31] H. Li, X. Yu, and X. Wang, "A saliency-based method for SAR target detection," in *Proc. IEEE Int. Geosci. Remote Sens. Symp. (IGARSS)*, Jul. 2018, pp. 2837–2840.

[32] G. Hu, Z. Yang, J. Han, L. Huang, J. Gong, and N. Xiong, "Aircraft detection in remote sensing images based on saliency and convolution neural network," *EURASIP J. Wireless Commun. Netw.*, vol. 2018, no. 1, pp. 1–16, Dec. 2018.
[33] Z. Wu, L. Su, and Q. Huang, "Cascaded partial decoder for fast and accurate salient object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 3907–3916.
[34] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*. [Online]. Available: http://arxiv.org/abs/1409.1556
[35] K. He, X. Zhang, S. Ren, and J. Sun, "Spatial pyramid pooling in deep convolutional networks for visual recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 9, pp. 1904–1916, Sep. 2015.
[36] D.-P. Fan, Y. Zhai, A. Borji, J. Yang, and L. Shao, "BBS-Net: RGB-D salient object detection with a bifurcated backbone strategy network," in *Proc. Eur. Conf. Comput. Vis.* 2020, pp. 275–292.
[37] C. Li, W. Xia, Y. Yan, B. Luo, and J. Tang, "Segmenting objects in day and night: Edge-conditioned CNN for thermal image semantic segmentation," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 32, no. 7, pp. 3069–3082, Jul. 2020.
[38] Z. Wu, L. Su, and Q. Huang, "Stacked cross refinement network for edge-aware salient object detection," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 7264–7273.
[39] K. Li, G. Wan, G. Cheng, L. Meng, and J. Han, "Object detection in optical remote sensing images: A survey and a new benchmark," *ISPRS J. Photogramm. Remote Sens.*, vol. 159, pp. 296–307, Jan. 2020.
[40] G.-S. Xia *et al.*, "DOTA: A large-scale dataset for object detection in aerial images," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 3974–3983.
[41] G. Cheng, P. Zhou, and J. Han, "Learning rotation-invariant convolutional neural networks for object detection in VHR optical remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 54, no. 12, pp. 7405–7415, Dec. 2016.
[42] S. Chen, X. Tan, B. Wang, H. Lu, X. Hu, and Y. Fu, "Reverse attention-based residual network for salient object detection," *IEEE Trans. Image Process.*, vol. 29, pp. 3763–3776, 2020.
[43] 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. (CVPR)*, Jun. 2019, pp. 3917–3926.
[44] L. Zhu *et al.*, "Aggregating attentional dilated features for salient object detection," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 30, no. 10, pp. 3358–3371, Oct. 2020.
[45] N. Liu, J. Han, and M.-H. Yang, "PiCANet: Learning pixel-wise contextual attention for saliency detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 3089–3098.
[46] T. Zhao and X. Wu, "Pyramid feature attention network for saliency detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 3085–3094.
[47] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*. [Online]. Available: http://arxiv.org/abs/1412.6980
[48] R. Achanta, S. Hemami, F. Estrada, and S. Susstrunk, "Frequency-tuned salient region detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 1597–1604.
[49] F. Perazzi, P. Krähenbühl, Y. Pritch, and A. Hornung, "Saliency filters: Contrast based filtering for salient region detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2012, pp. 733–740.
[50] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

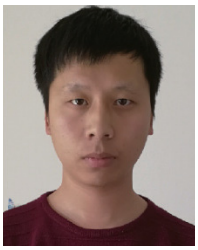
Zhengzheng Tu received the M.S. and Ph.D. degrees from the School of Computer Science and Technology, Anhui University, Hefei, China, in 2007 and 2015, respectively.

She is currently an Associate Professor with the School of Computer Science and Technology, Anhui University. Her research interests include computer vision and deep learning.



Chao Wang received the B.S. degree from the Anhui Institute of Information Technology, Wuhu, China, in 2019. He is currently pursuing the M.S. degree with the School of Computer Science and Technology, Anhui University, Hefei, China.

His research is optical remote sense image (ORSI) SOD based on deep learning.



Chenglong Li received the M.S. and Ph.D. degrees from the School of Computer Science and Technology, Anhui University, Hefei, China, in 2013 and 2016, respectively.

From 2014 to 2015, he was a Visiting Student with the School of Data and Computer Science, Sun Yat-sen University, Guangzhou, China. He was a Post-Doctoral Research Fellow with the National Laboratory of Pattern Recognition (NLPR), Center for Research on Intelligent Perception and Computing (CRIPAC), Institute of Automation, Chinese Academy of Sciences (CASIA), Beijing, China. He is currently an Associate Professor with the School of Computer Science and Technology, Anhui University. His research interests include computer vision and deep learning.

Dr. Li was a recipient of the ACM Hefei Doctoral Dissertation Award in 2016.

Minghao Fan received the Ph.D. degree in mechanical and electronic control engineering from the State Key Laboratory of Fluid Power and Mechatronic Systems of Zhejiang University, Hangzhou, China.

He is currently a Professorate Senior Engineer with the State Grid Anhui Electric Power Research Institute, Hefei, China. His research interests include electric fire and safety protection, and fire safety issues in energy utilization.

Haifeng Zhao received the B.Eng. degree in electrical engineering and the Ph.D. degree in computer science from Anhui University, Hefei, China, in 1995 and 2006, respectively.

He is currently a Professor with the School of Computer Science and Technology, Anhui University. His research interests include medical image processing, pattern recognition, and computer vision.

Bin Luo (Senior Member, IEEE) received the B.Eng. degree in electronics and the M.Eng. degree in computer science from Anhui University, Hefei, China, in 1984 and 1991, respectively, and the Ph.D. degree in computer science from the University of York, York, U.K., in 2002.

He is currently a Professor with Anhui University. He has authored over 200 articles in journals, edited books, and refereed conferences. His research interests include random graph-based pattern recognition, image and graph matching, graph spectral analysis, and video analysis.

Dr. Luo is also the Chair of the IEEE Hefei Subsection. He has served as a Peer Reviewer for international academic journals, such as the IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, *Pattern Recognition*, *Pattern Recognition Letters*, the *International Journal of Pattern Recognition and Artificial Intelligence*, *Knowledge and Information Systems*, and *Neurocomputing*.