Adjacent Context Coordination Network for Salient Object Detection in Optical Remote Sensing Images

Gongyang Li, Zhi Liu, *Senior Member, IEEE*, Dan Zeng, *Senior Member, IEEE*
Weisi Lin, *Fellow, IEEE*, and Haibin Ling, *Senior Member, IEEE*

***Abstract*—Salient object detection (SOD) in optical remote sensing images (RSIs), or *RSI-SOD*, is an emerging topic in understanding optical RSIs. However, due to the difference between optical RSIs and natural scene images (NSIs), directly applying NSI-SOD methods to optical RSIs fails to achieve satisfactory results. In this paper, we propose a novel Adjacent Context Coordination Network (ACCoNet) to explore the coordination of adjacent features in an encoder-decoder architecture for RSI-SOD. Specifically, ACCoNet consists of three parts: an encoder, Adjacent Context Coordination Modules (ACCoMs), and a decoder. As the key component of ACCoNet, ACCoM activates the salient regions of output features of the encoder and transmits them to the decoder. ACCoM contains a local branch and two adjacent branches to coordinate the multi-level features simultaneously. The local branch highlights the salient regions in an adaptive way, while the adjacent branches introduce global information of adjacent levels to enhance salient regions. Additionally, to extend the capabilities of the classic decoder block (*i.e.*, several cascaded convolutional layers), we extend it with two bifurcations and propose a Bifurcation-Aggregation Block to capture the contextual information in the decoder. Extensive experiments on two benchmark datasets demonstrate that the proposed ACCoNet outperforms 22 state-of-the-art methods under nine evaluation metrics, and runs up to 81 *fps* on a single NVIDIA Titan X GPU. The code and results of our method are available at https://github.com/MathLee/ACCoNet.**

***Index Terms*—Optical remote sensing images, salient object detection, adjacent context coordination, bifurcation-aggregation block.**

I. Introduction

SALIENT object detection (SOD) aims at distinguishing and highlighting visually attractive objects/regions in a scene, which has been extended from natural scene images (NSIs) [1]–[3] to videos [4], image groups [5], RGB-D images [6], *etc*. It has many applications, such as object segmentation [7], [8], object tracking [9], [10], quality assessment [11], [12], hyperspectral image classification [13], *etc*. Recently, SOD has been extended to optical remote sensing images (RSIs) [14]–[22], and has produced encouraging results. For conciseness, in the rest of the paper, we use *RSI-SOD* for the task of SOD in optical RSIs.

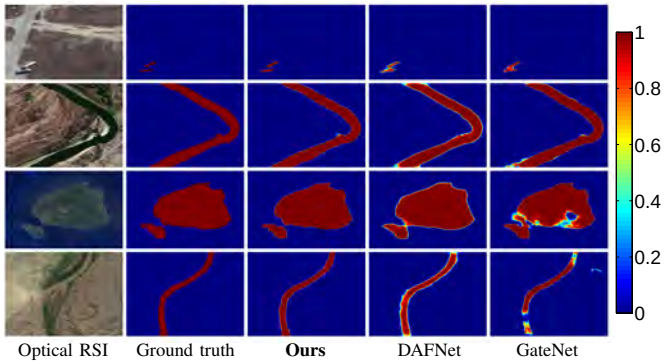


Fig. 1. Saliency maps produced by our method and two state-of-the-art methods DAFNet [20] and GateNet [23] on optical RSIs.

During the past decades, SOD in NSIs [1]–[3], or *NSI-SOD* in short, has made a remarkable progress, especially when armed with deep learning techniques such as convolutional neural network (CNN) [24]. Naturally, researchers will consider applying the mature NSI-SOD solutions to RSI-SOD. However, there are significant differences in shooting devices, scenes and view orientations between NSIs and optical RSIs, resulting in differences in their resolutions, object types and object scales [18], [20]. Consequently, direct migration of NSI-SOD solutions to RSI-SOD often leads to unsatisfactory performance. As shown in the last column of Fig. 1, GateNet [23], which is a representative CNN-based NSI-SOD method and retrained on optical RSIs, cannot highlight salient objects in optical RSIs completely.

The existing specialized methods for RSI-SOD can be divided into traditional methods and CNN-based methods. Traditional methods rely heavily on specific handcrafted features based on classical principles, such as color information content [14], sparse representation [15], saliency feature analysis [16], and self-adaptive multiple feature fusion [17]. They usually fail in complex scenes of optical RSIs. CNN-based methods focus on exploring effective feature interaction strategies to overcome the complex topology and unique scenes of optical RSIs. The nested network [18] fuses multi-resolution features; the parallel down-up fusion network [19] focuses on the cross-path interaction, which is from low-level path/features to high-level path/features, between two adjacent features; and the dense attention fluid network (DAFNet) [20] transfers shallow-layer attention cues of low-level features,

Gongyang Li, Zhi Liu, and Dan Zeng are with Shanghai Institute for Advanced Communication and Data Science, Shanghai University, Shanghai 200444, China, and School of Communication and Information Engineering, Shanghai University, Shanghai 200444, China. Dan Zeng is also with the Key Laboratory of Specialty Fiber Optics and Optical Access Networks, Joint International Research Laboratory of Specialty Fiber Optics and Advanced Communication, Shanghai University, Shanghai 200444, China (email: ligongyang@shu.edu.cn; liuzhisjtu@163.com; dzeng@shu.edu.cn).

Weisi Lin is with the School of Computer Science and Engineering, Nanyang Technological University, Singapore 639798 (e-mail: wslin@ntu.edu.sg).

Haibin Ling is with the Department of Computer Science, Stony Brook University, Stony Brook, NY 11794 USA (email: hling@cs.stonybrook.edu).

Corresponding author: Zhi Liu.

which capture edge and texture information, to deep layers, *i.e.*, high-level features, which capture semantic and object location information. However, the influence of high-level features on low-level features is ignored, the coverage in feature interaction is insufficient, and the cascade structure of decoder blocks is plain, which may lead to incomplete exploration of the contextual information in optical RSIs. As shown in the penultimate column of Fig. 1, the saliency maps of DAFNet [20], which is currently the best specialized method, lose sharp boundaries and finer details.

Inspired by the above observations, in this paper, we propose a novel specialized solution for RSI-SOD, namely *Adjacent Context Coordination Network* (ACCoNet), which focuses on coordinating adjacent features and capturing contextual information to adapt to diverse object types and object sizes in optical RSIs. Our key idea is to comprehensively explore the contextual information contained in adjacent features, expand the coverage of feature interaction, and improve the context capture capability of plain decoder blocks. Specifically, we consider features processing of three adjacent blocks (*i.e.*, the current, previous and subsequent blocks) in the backbone in a special module. This way, the previous and subsequent features can provide comprehensive global auxiliary information to the current features. Besides, we introduce bifurcations into plain decoder blocks to capture multi-scale context and increase the feature diversity.

In particular, we implement our ACCoNet in an encoder-decoder architecture. ACCoNet is composed of an *Adjacent Context Coordination Module* (ACCoM) for three adjacent features and a *Bifurcation-Aggregation Block* (BAB) for the decoder. ACCoM consists of three branches, one for local information and the other two for adjacent context. Specifically, the local branch is responsible for modulating and enhancing current features in an adaptive manner, while the other two adjacent branches are responsible for assisting current features with the previous and subsequent features through the previous-to-current and subsequent-to-current interactions. For BAB, we put a bifurcation after each cascaded convolutional layer, and then aggregate these bifurcations to capture diverse contexts. In this way, our ACCoNet achieves the best performance as compared with 22 state-of-the-art methods (an average S_α of 93.64%, an average max F_β of 89.93% and an average max E_ξ of 97.62% on two datasets), and generates the most accurate saliency maps, as exemplified in the middle column of Fig. 1.

Our main contributions are summarized as follows:

- We explore the coordination of adjacent features in an encoder-decoder architecture for RSI-SOD, and propose a novel *Adjacent Context Coordination Network* (ACCoNet), which effectively promotes the interaction of adjacent features for comprehensive coordination and fully captures contextual information, outperforming previous methods on public benchmarks.
- We propose an *Adjacent Context Coordination Module* (ACCoM) to coordinate cross-scale interactions in the feature embedding provided by the encoder and to deliver the valuable information to the decoder.
- We extend the cascade structure of classic decoder blocks to the bifurcation-aggregation structure, and propose a *Bifurcation-Aggregation Block* (BAB) to capture the multi-scale contextual information in the decoder.

The remaining parts of this paper are organized as follows. In Sec. II, we summarize the related works of NSI-SOD and RSI-SOD. In Sec. III, we elaborate our ACCoNet. In Sec. IV, we present the experiments and ablation studies of our ACCoNet. Finally, the conclusion is drawn in Sec. V.

II. Related Work

In this section, we review the classic works of NSI-SOD and RSI-SOD, including traditional methods and CNN-based methods.

A. Salient Object Detection in NSIs

1) Traditional NSI-SOD Methods. Salient object detection starts with natural scene images [25], and a lot of traditional methods [1] have investigated hand-crafted features for NSI-SOD. Traditional NSI-SOD methods can been divided into three categories: unsupervised methods [25]–[34], semi-supervised methods [35], and supervised methods [36]. Numerous principles and technologies have been proposed for unsupervised methods, such as center-surround differences [25], the maximal entropy random walk [26], the saliency tree [27], the regularized random walks ranking [28], [29], directional information [30], the high-dimensional color transform [31], the sparse graph [32], the structured matrix decomposition [33], the hybrid sparse learning [34], *etc*. Compared with unsupervised methods, there are relatively fewer semi-supervised and supervised methods in traditional methods. Zhou *et al*. [35] first utilized a boundary homogeneity model to generate pseudo labels. Then based on a linear feedback control system model, they presented an iterative semi-supervised learning framework to establish relationships between control states and saliency map. Liang *et al*. [36] trained a support vector machine to select features through the supervised learning, which removes redundant features and speeds up model learning. Wang *et al*. [37] presented a supervised multiple-instance learning framework for saliency detection, which incorporates a set of low-, mid-, and high-level features to comprehensively predict the scores of salient regions.

2) CNN-based NSI-SOD Methods. Different from traditional methods, most CNN-based NSI-SOD methods [2], [3] are based on supervised learning, and they greatly improve the detection accuracy. A large number of well-known strategies of feature processing have been proposed, such as the multi-level and multi-scale feature interaction [38], [39], the feature suppress and balance [23], the sparse and dense labeling aggregation [40], the edge-aware feature fusion [41], [42], the global context-aware aggregation [43], [44]. In addition, many popular mechanisms in the deep learning community are applied to NSI-SOD, such as the deep supervision [45], [46], the recurrent mechanism [47], [48], the attention mechanism [44], [49]–[51], the generative adversarial learning [52], and the adversarial attack [53]. Differently, Li *et al*. [54] focused on

the detection speed and proposed the depthwise nonlocal network, which achieves competitive performance using a single CPU thread. Liu *et al*. [55] explored a lightweight architecture for NSI-SOD and imitated the primate visual cortex in their network via hierarchical visual perception learning.

The above NSI-SOD methods have a great influence on RSI-SOD methods. However, due to the essential differences between NSIs and optical RSIs, these RSI-SOD methods have made specific modifications to the hand-crafted features or the CNN feature processing strategies of original NSI-SOD methods.

B. Salient Object Detection in Optical RSIs

As an emerging field, the SOD in optical RSIs, *i.e*., RSI-SOD, has attracted more and more attention. Zhang *et al*. [14] first performed the color information content analysis on the input optical RSI to get the saliency scores of each color component, and then they constructed the saliency map based on these saliency scores. Zhao *et al*. [15] obtained low-level features via the global cues and background prior, and the sparse representation was introduced to transform low-level features to high-level features for saliency map integration. In [16], Zhang *et al*. combined the super-pixel segmentation and statistical saliency feature analysis for RSI-SOD. Zhang *et al*. [17] fused the features of color, intensity, texture and global contrast adaptively based on the low-rank matrix recovery to generate the saliency map. Faur *et al*. [56] combined the mean-shift-based segmentation and the rate distortion-based optimization together for salient remote sensing image segmentation.

Different from the above traditional RSI-SOD methods, CNN-based RSI-SOD solutions explore the unique characteristics from optical RSI data, and have made a promising progress. Li *et al*. [18] constructed a challenging dataset for RSI-SOD. They proposed an LV-shaped network, where the L-shaped two-stream pyramid module receives input images of five resolutions and the V-shaped nested connections structure infers salient objects based on multi-resolution features. In [19], Li *et al*. designed five parallel paths with dense connections, which exploit the in-path and cross-path information contained in two adjacent features to detect diversely scaled salient objects in optical RSIs. Zhang *et al*. [20] first established shallow-to-deep connections between different levels through dense attention fluid structure, and then they exploited global-context information to achieve feature alignment and reinforcement. Zhang *et al*. [21] combined the weakly and fully supervised learning for RSI-SOD. They obtained pseudo annotations based on a classification network and the gradient-weighted class activation mapping to train the feedback saliency analysis network. Tu *et al*. [57] proposed a multiscale joint region and boundary model for RSI-SOD. Following [18], Zhou *et al*. [58] proposed a three inputs based edge-aware feature integration network.

Aside from the above studies, there are some works on tasks related to RSI-SOD, such as airport detection [59], building extraction [60], residential areas extraction [61], ship detection [62], oil tank detection [63], [64], and region-of-interest detection/extraction [65]–[68]. These methods show good performance in specific scenes of optical RSIs, but may not generalize well to various optical RSI scenes, resulting in poor performance in RSI-SOD.

As we know, the salient objects in optical RSIs usually have complex geometry structures, variable sizes and uncertain quantities, and are often accompanied with occlusion, shadows and abnormal illumination. The specialized methods mentioned above put forward meaningful solutions to the characteristics of optical RSIs. However, we believe that the contextual information in optical RSIs needs to be further explored, which is important to overcome these challenging scenes. We thoroughly explore the contextual information in both encoder and decoder of our ACCoNet. Concretely, the previous-to-current and subsequent-to-current feature interactions are established among three adjacent blocks in the encoder, and the cascade structure is updated to the bifurcation-aggregation structure in the decoder.

III. Methodology

In this section, we elaborate the proposed Adjacent Context Coordination Network (ACCoNet). In Sec. III-A, we clarify the network overview and motivation of our ACCoNet. In Sec. III-B, we present our Adjacent Context Coordination Module (ACCoM) in detail. In Sec. III-C, we give the detailed formulas of our Bifurcation-Aggregation Block (BAB). In Sec. III-D, we introduce the loss function.

A. Network Overview and Motivation

The proposed ACCoNet is based on the encoder-decoder architecture, which has shown outstanding ability in pixel-level prediction tasks, such as semantic segmentation [69], medical image segmentation [70], NSI-SOD [23], [43] and RGB-D SOD [71]–[73]. As shown in Fig. 2, ACCoNet consists of an encoder network, several ACCoM components, and a decoder network with BABs.

1) Encoder Network. Following [69], [71]–[73], we adopt the plain VGG-16 [74] as our basic encoder network, where the last max-pooling layer and three fully connected layers are truncated. As shown at the top of Fig. 2, our encoder network consists of five blocks, denoted by E^t ($t \in \{1,2,3,4,5\}$ is the block index), and we adopt the feature map of the last convolutional layer of each block, *i.e*., *conv1-2*, *conv2-2*, *conv3-3*, *conv4-3* and *conv5-3*, denoted by $\boldsymbol{f}_{\mathrm{e}}^t \in \mathbb{R}^{h_t \times w_t \times c_t}$ ($c_{t=\{1,2,3,4,5\}} = \{64, 128, 256, 512, 512\}$), for subsequent processing. The input size of our encoder network is $256 \times 256 \times 3$, so $h_t = \frac{256}{2^{t-1}}$ and $w_t = \frac{256}{2^{t-1}}$.

2) Adjacent Context Coordination Module. Contextual information is crucial for RSI-SOD. It exists not only in one feature level, but also in features at adjacent levels. Using convolutional layers with different convolution kernels in parallel is a popular strategy to capture local and global contents within one feature level. This is conducive to capturing salient objects with variable sizes or uncertain quantities in optical RSIs. And introducing feature interaction among features at adjacent levels is an effective strategy to capture cross-level contextual complementary information. This is effective for refining the details and determining the location of salient

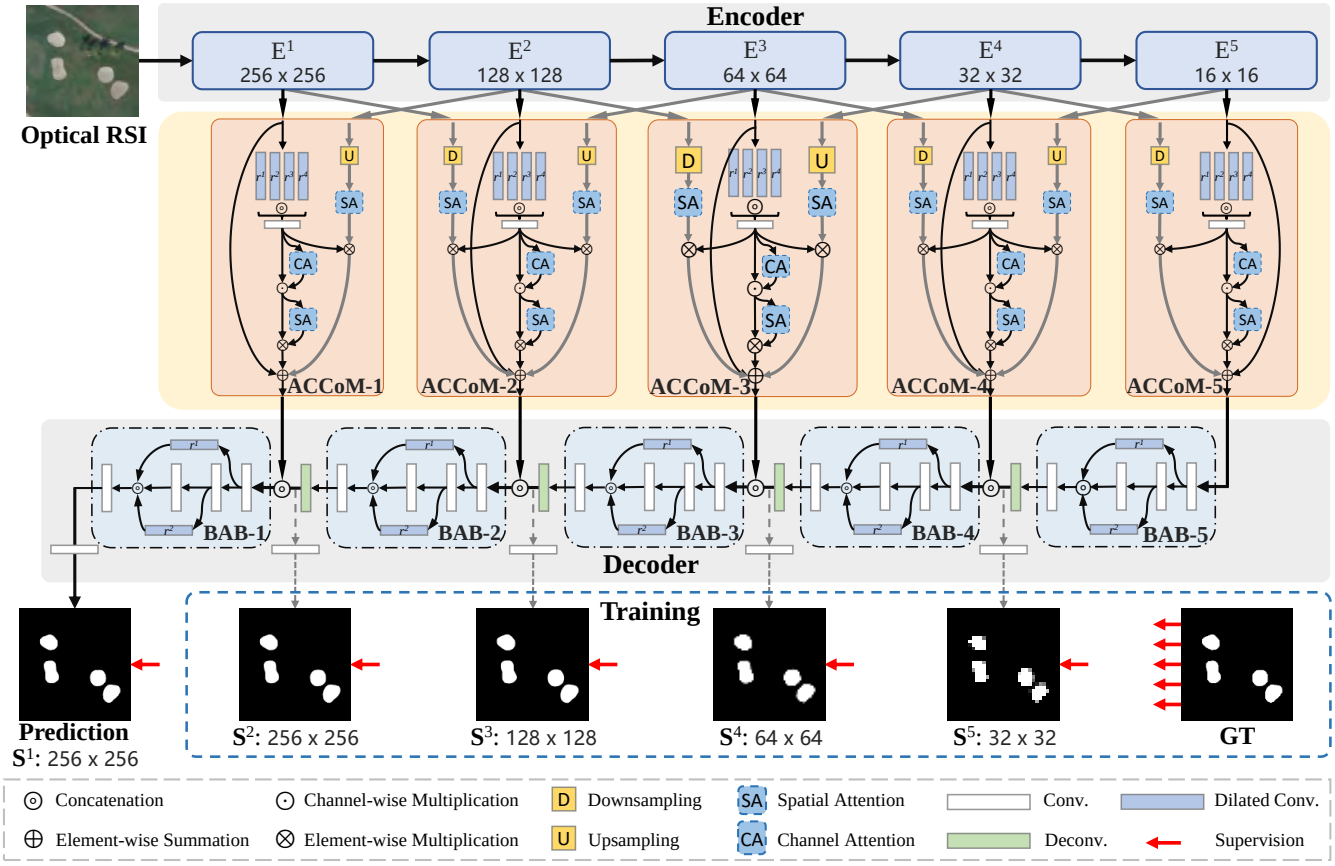


Fig. 2. Pipeline of the proposed ACCoNet, which is comprised of three key parts: the encoder network, the Adjacent Context Coordination Module (ACCoM) and the decoder network with Bifurcation-Aggregation Blocks (BABs). First, the encoder network extracts the basic features at five scales. Then, these basic features are fed to five ACCoMs to coordinate the feature activation. Finally, the output contextual features of ACCoM are transmitted to the decoder, which employs BABs to further capture contextual information, for inferring the salient objects. Notably, in the training phase, we adopt the deep supervision strategy, and attach the pixel-level supervision to each decoder block. GT denotes ground truth.

objects in optical RSIs. Thus motivated, we explore the above two kinds of contextual information with these two mentioned strategies. Since high-level features provide a lot of semantic clues and low-level features provide a lot of fine details, we coordinate cross-scale features from the current, previous and subsequent blocks.

In practice, we design three branches (*i.e.*, one local branch and two adjacent branches) in ACCoM. The local branch is based on the first strategy. Moreover, it is equipped with the attention mechanism for further feature modulation in an adaptive way. The two adjacent branches are based on the second strategy, and consist of the previous-to-current branch and the subsequent-to-current branch. Since that the previous and subsequent features are different in scale from the current features, the two adjacent branches provide cross-scale information via two spatial attention maps to align salient regions twice. Comprehensive coordination enables the proposed ACCoM to transmit valuable contextual information to the decoder. Notably, as shown in Fig. 2, for ACCoM-1 and ACCoM-5, due to their special position, we can only make one adjacent branch in them. We present ACCoM in detail in Sec. III-B, and assess its effectiveness in Sec. IV-C.

3) Bifurcation-Aggregation Block. The decoder network is in charge of inferring the salient objects. Generally, the classic decoder network [69], [70] is comprised of five plain decoder blocks, in which the convolutional layers are cascaded. However, the inference ability of the cascade structure depends more on the features transmitted by the encoder, and the cascade structure is not sensitive to the unique scenes of optical RSIs, which may damage the inference accuracy of salient objects of the decoder network. As previously mentioned, contextual information is crucial for RSI-SOD, so we further explore them in the decoder. We introduce dilated convolutions [75] as bifurcations after the first two cascaded convolutional layers, and then aggregate the information from the two bifurcations and the original one via the concatenation-convolution operation. In this way, the bifurcation-aggregation structure enriches the topology of the decoder through two dilated convolutions, expands the receptive field of features and captures rich contextual information, which is beneficial for inferring the salient objects. We present BAB in detail in Sec. III-C, and show its ablation studies in Sec. IV-C.

B. Adjacent Context Coordination Module

Adjacent Context Coordination Module is the key component in ACCoNet. It connects the encoder and the decoder, and its details are illustrated in Fig. 2. There are usually three branches in ACCoM (*e.g.*, ACCoM-2, ACCoM-3 and

ACCoM-4): one local branch (the middle one in ACCoM) and two adjacent branches (the left and right ones in ACCoM). While ACCoM-1 and ACCoM-5 only contain two branches: one local branch and one adjacent branch. Thus, we generally define the processing of ACCoM as $\mathrm{F}(\cdot)$, which is formulated as follows:

$$\boldsymbol{f}_{\mathrm{accom}}^{t} = \begin{cases} \mathrm{F}(\boldsymbol{f}_{\mathrm{e}}^{t}, \boldsymbol{f}_{\mathrm{e}}^{t+1}), & t = 1 \\ \mathrm{F}(\boldsymbol{f}_{\mathrm{e}}^{t-1}, \boldsymbol{f}_{\mathrm{e}}^{t}, \boldsymbol{f}_{\mathrm{e}}^{t+1}), & t = 2, 3, 4 \\ \mathrm{F}(\boldsymbol{f}_{\mathrm{e}}^{t-1}, \boldsymbol{f}_{\mathrm{e}}^{t}), & t = 5, \end{cases} \tag{1}$$

where $\boldsymbol{f}_{\mathrm{accom}}^{t} \in \mathbb{R}^{h_t \times w_t \times c_t}$ is the output feature of ACCoM-*t*, and $\boldsymbol{f}_{\mathrm{e}}^{t-1}$, $\boldsymbol{f}_{\mathrm{e}}^{t}$ and $\boldsymbol{f}_{\mathrm{e}}^{t+1}$ are the previous, current and subsequent features, respectively.

1) Local Branch. The local branch operates on the current features $\boldsymbol{f}_{\mathrm{e}}^{t} \in \mathbb{R}^{h_t \times w_t \times c_t}$, and contains two main operations. First, we apply four dilated convolutions [75] (rather than normal convolutional layers) with different dilation rates in parallel to $\boldsymbol{f}_{\mathrm{e}}^{t}$, which is defined as follows:

$$\boldsymbol{f}_{\mathrm{dc}}^{t,i} = \mathrm{DConv}_{\sigma}(\boldsymbol{f}_{\mathrm{e}}^{t}; \mathbf{W}_{3\times3}^{t,i}, r^i), \ i \in \{1, 2, 3, 4\}, \tag{2}$$

where $\boldsymbol{f}_{\mathrm{dc}}^{t,i} \in \mathbb{R}^{h_t \times w_t \times c_t}$ is the output feature of each dilated convolution, $\mathrm{DConv}_{\sigma}(*;*,*)$ is the dilated convolution with Batch Normalization (BN) [76] and ReLU activation function σ, $\mathbf{W}_{3\times3}^{t,i}$ is the parameters with 3×3 kernel, and $r^i = i$ is the dilation rate. This can effectively traverse regions of different sizes in $\boldsymbol{f}_{\mathrm{e}}^{t}$.

Then, we summarize these output features using the concatenation-convolution operation, obtaining features with rich contextual cues, *i.e.*, $\boldsymbol{f}_{\mathrm{c}}^{t} \in \mathbb{R}^{h_t \times w_t \times c_t}$, which is defined as follows:

$$\boldsymbol{f}_{\mathrm{c}}^{t} = \mathrm{Conv}_{\sigma}\big(\mathrm{Concat}\big(\boldsymbol{f}_{\mathrm{dc}}^{t,1}, \boldsymbol{f}_{\mathrm{dc}}^{t,2}, \boldsymbol{f}_{\mathrm{dc}}^{t,3}, \boldsymbol{f}_{\mathrm{dc}}^{t,4}\big); \mathbf{W}_{3\times3}^{t}\big), \tag{3}$$

where $\mathrm{Concat}(\cdot)$ is the cross-channel concatenation, and $\mathrm{Conv}_{\sigma}(*;*)$ is the normal convolutional layer with BN and ReLU activation function. The subsequent operations in ACCoM are based on $\boldsymbol{f}_{\mathrm{c}}^{t}$.

However, the summary operation is relatively straightforward, resulting in some redundant information in $\boldsymbol{f}_{\mathrm{c}}^{t}$. We adopt the subtle channel attention (CA) and spatial attention (SA) [77], [78] to further purify $\boldsymbol{f}_{\mathrm{c}}^{t}$ in an adaptive manner, which is formulated as follows:

$$\boldsymbol{f}_{\mathrm{loc}}^{t} = \mathrm{SA}(\mathrm{CA}(\boldsymbol{f}_{\mathrm{c}}^{t}) \odot \boldsymbol{f}_{\mathrm{c}}^{t}) \otimes \boldsymbol{f}_{\mathrm{c}}^{t}, \tag{4}$$

where $\boldsymbol{f}_{\mathrm{loc}}^{t} \in \mathbb{R}^{h_t \times w_t \times c_t}$ is the output feature of the local branch, $\odot$ is the channel-wise multiplication, and $\otimes$ is the element-wise multiplication. Specifically, we implement CA with a spatial-wise global max pooling (GMP), a fully connected layer with ReLU activation function and a fully connected layer with sigmoid activation function; and we implement SA with a channel-wise GMP and a convolutional layer with sigmoid activation function. Such an adaptive modulation process selects valuable contents from $\boldsymbol{f}_{\mathrm{c}}^{t}$.

2) Adjacent Branch(es). The adjacent branches contribute two types of assistance to $\boldsymbol{f}_{\mathrm{c}}^{t}$. The first one is the previous-to-current branch, which can be computed as:

$$\boldsymbol{f}_{\mathrm{pc}}^{t} = \mathrm{SA}(\mathrm{Down}(\boldsymbol{f}_{\mathrm{e}}^{t-1})) \otimes \boldsymbol{f}_{\mathrm{c}}^{t}, \quad t = 2, 3, 4, 5, \tag{5}$$

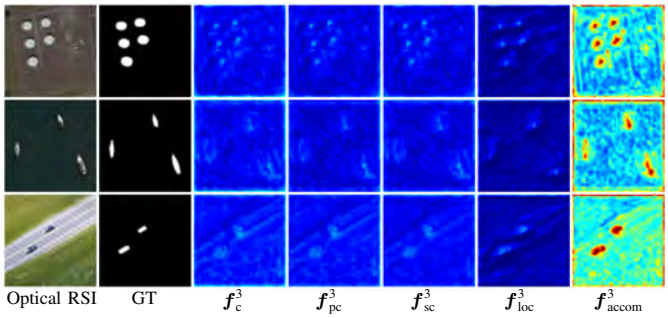


Fig. 3. Feature visualization of each branch in ACCoM-3. Please zoom-in for viewing details.

where $\boldsymbol{f}_{\mathrm{pc}}^{t} \in \mathbb{R}^{h_t \times w_t \times c_t}$ is the output feature of the previous-to-current branch, and $\mathrm{Down}(\cdot)$ is the $2\times$ downsampling implemented by max-pooling. This branch brings alignment information with fine details to $\boldsymbol{f}_{\mathrm{c}}^{t}$.

The second one is the subsequent-to-current branch, which can be computed as:

$$\boldsymbol{f}_{\mathrm{sc}}^{t} = \mathrm{SA}(\mathrm{Up}(\boldsymbol{f}_{\mathrm{e}}^{(t+1)})) \otimes \boldsymbol{f}_{\mathrm{c}}^{t}, \quad t = 1, 2, 3, 4, \tag{6}$$

where $\boldsymbol{f}_{\mathrm{sc}}^{t} \in \mathbb{R}^{h_t \times w_t \times c_t}$ is the output feature of the subsequent-to-current branch, and $\mathrm{Up}(\cdot)$ is the $2\times$ upsampling implemented by bilinear interpolation. This branch brings alignment information with object location to $\boldsymbol{f}_{\mathrm{c}}^{t}$.

3) Branches Integration. After the above effective coordination, we integrate the output features of these three (or two) branches with the original current features as follows:

$$\boldsymbol{f}_{\mathrm{accom}}^{t} = \begin{cases} \boldsymbol{f}_{\mathrm{loc}}^{t} \oplus \boldsymbol{f}_{\mathrm{sc}}^{t} \oplus \boldsymbol{f}_{\mathrm{e}}^{t}, & t = 1 \\ \boldsymbol{f}_{\mathrm{loc}}^{t} \oplus (\boldsymbol{f}_{\mathrm{pc}}^{t} \oplus \boldsymbol{f}_{\mathrm{sc}}^{t}) \oplus \boldsymbol{f}_{\mathrm{e}}^{t}, & t = 2, 3, 4 \\ \boldsymbol{f}_{\mathrm{loc}}^{t} \oplus \boldsymbol{f}_{\mathrm{pc}}^{t} \oplus \boldsymbol{f}_{\mathrm{e}}^{t}, & t = 5, \end{cases} \tag{7}$$

where $\oplus$ is the element-wise summation and the original current features are regarded as the basic content. In summary, $\boldsymbol{f}_{\mathrm{e}}^{t}$ is coordinated by various contents, which greatly enhances the robustness and stability of $\boldsymbol{f}_{\mathrm{accom}}^{t}$.

In Fig. 3, we visualize features in ACCoM-3. It shows that, with all branches (*i.e.*, $\boldsymbol{f}_{\mathrm{loc}}^{3}$, $\boldsymbol{f}_{\mathrm{pc}}^{3}$ and $\boldsymbol{f}_{\mathrm{sc}}^{3}$) working together, ACCoM accurately activates each salient region through comprehensive coordination, making the salient objects in $\boldsymbol{f}_{\mathrm{accom}}^{3}$ more obvious than those in $\boldsymbol{f}_{\mathrm{c}}^{3}$.

C. Bifurcation-Aggregation Block

Bifurcation-Aggregation Block is the basic unit of the decoder. It processes the features from the current ACCoM and the previous BAB, and finally infers the mask of salient objects. We define the processing of BAB as $\mathrm{B}(\cdot)$, which is formulated as follows:

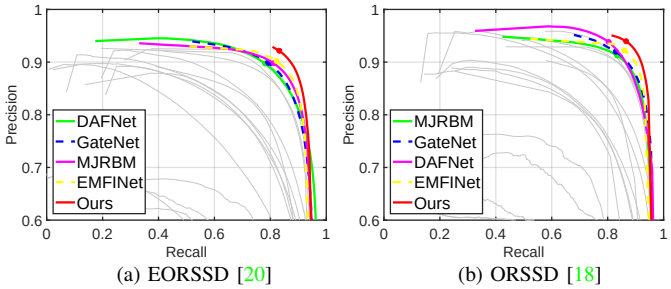
$$\boldsymbol{f}_{\mathrm{bab}}^{t} = \begin{cases} \mathrm{B}(\boldsymbol{f}_{\mathrm{accom}}^{t}, \mathrm{Deconv}(\boldsymbol{f}_{\mathrm{bab}}^{t+1})), & t = 1, 2, 3, 4 \\ \mathrm{B}(\boldsymbol{f}_{\mathrm{accom}}^{t}), & t = 5, \end{cases} \tag{8}$$

where $\boldsymbol{f}_{\mathrm{bab}}^{t}$ is the output feature of BAB-*t*, and $\mathrm{Deconv}(\cdot)$ is the deconvolution layer with BN and ReLU activation function.

For convenience, we define the features generated by the three cascaded convolutional layers in BAB-*t* as $\boldsymbol{f}_{\mathrm{b\text{-}c}}^{t,l}$ ($l \in$

TABLE I
DETAILED PARAMETERS OF TWO BIFURCATIONS (*i.e.*, DILATED CONVOLUTIONS) IN BAB, INCLUDING KERNEL SIZE, CHANNEL NUMBER, DILATION RATE AND THE SIZE OF OUTPUT FEATURE. FOR INSTANCE, $(3 \times 3, 64, 64)$ DENOTES THAT THE KERNEL SIZE IS 3×3, THE INPUT CHANNEL NUMBER IS 64, AND THE OUTPUT CHANNEL NUMBER IS 64.

Aspects	Dilated conv.	r^1	r^2	Output size
BAB-1	$(3 \times 3, 64, 64)$	5	3	$[256 \times 256 \times 64]$
BAB-2	$(3 \times 3, 128, 128)$	5	3	$[128 \times 128 \times 128]$
BAB-3	$(3 \times 3, 256, 256)$	5	3	$[64 \times 64 \times 256]$
BAB-4	$(3 \times 3, 512, 512)$	3	2	$[32 \times 32 \times 512]$
BAB-5	$(3 \times 3, 512, 512)$	3	2	$[16 \times 16 \times 512]$



(a) EORSSD [20] (b) ORSSD [18]

Fig. 4. Quantitative performance comparison on PR curve in two datasets. The top five methods are shown in color. Please zoom-in for details.

$\{1, 2, 3\}$). So the output feature of two bifurcations (*i.e.*, $\boldsymbol{f}_{\text{bif}}^{t,l}$) can be computed as:

$$\boldsymbol{f}_{\text{bif}}^{t,l} = \text{DConv}_{\sigma}(\boldsymbol{f}_{\text{b-c}}^{t,l}; \mathbf{W}_{3\times3}^{t,l}, r^l), \quad l = 1, 2, \tag{9}$$

in which we adopt the dilated convolution to expand the receptive field and capture contextual cues from $\boldsymbol{f}_{\text{accom}}^{t}$. In practice, considering the difference in feature resolution of each BAB, we set different dilation rates of bifurcations for different BABs. The detailed parameters are shown in Tab. I.

Then, we adopt the concatenation-convolution operation to aggregate these two bifurcations and the original $\boldsymbol{f}_{\text{b-c}}^{t,3}$ as:

$$\boldsymbol{f}_{\text{bab}}^{t} = \text{Conv}_{\sigma}(\text{Concat}(\boldsymbol{f}_{\text{bif}}^{t,1}, \boldsymbol{f}_{\text{bif}}^{t,2}, \boldsymbol{f}_{\text{b-c}}^{t,3}); \mathbf{W}_{3\times3}^{t}). \tag{10}$$

This way, BAB further scans regions with different sizes based on $\boldsymbol{f}_{\text{accom}}^{t}$ at the inference stage, which can be well adapted to the characteristics of changes in the shape, size, and quantity of salient objects in optical RSIs.

D. Loss Function

As shown at the bottom of Fig. 2, in the training phase, we attach the pixel-level supervision to each decoder block (*i.e.*, the deep supervision strategy [79]) for quick convergency. Specifically, we arrange a convolutional layer after BAB-t to generate the intermediate/final saliency map, denoted as $\mathbf{S}^t$. For $\mathbf{S}^t$, we combine the pixel-level binary cross-entropy (BCE) loss with the map-level intersection-over-union (IoU) loss [73], [80] for comprehensive and complementary content enhancement. We formulate the total loss function $\mathbb{L}$ as:

$$\mathbb{L} = \sum_{t=1}^{5} \left(\text{L}_{\text{bce}}^{t}(\text{Up}(\mathbf{S}^t), \mathbf{GT}) + \text{L}_{\text{iou}}^{t}(\text{Up}(\mathbf{S}^t), \mathbf{GT}) \right), \tag{11}$$

where $\text{L}_{\text{bce}}^{t}(\cdot, \cdot)$ is the BCE loss, $\text{L}_{\text{iou}}^{t}(\cdot, \cdot)$ is the IoU loss, and $\mathbf{GT}$ is the ground truth. In this way, the deep supervision strategy with hybrid losses not only stabilizes our ACCoNet training process, but also improves the detection accuracy.

IV. EXPERIMENTAL RESULTS

A. Experimental Protocol

1) Datasets. We evaluate the proposed method on two recently proposed datasets for RSI-SOD.

ORSSD [18] is the first publicly available dataset for RSI-SOD, collected from the Google Earth and some existing RSI datasets. It contains 800 optical RSIs and provides corresponding pixel-wise annotation for each image. Among these optical RSIs, 600 images are used as training set and the remaining 200 images as testing set.

EORSSD [20] is the largest public dataset for RSI-SOD. It extends the original ORSSD dataset to 2,000 images with corresponding pixel-wise GTs. Among these, 1,400 images are used as training set and 600 images as testing set.

2) Network Training Details. We implement the proposed ACCoNet by PyTorch [81] with an NVIDIA Titan X GPU. In the training and testing phases, the input optical RSIs are resized into 256×256. We adopt the parameters of the pre-trained VGG-16 model [74] to initialize the parameters of the encoder network in our ACCoNet, while the parameters of all other newly added layers are initialized by the normal distribution [82]. We set the initial learning rate to $1e^{-4}$, and it will be divided by 10 after 30 epochs. Due to the limitation of GPU memory, we set the batch size to 6. We use the Adam optimizer [83] for network optimization. For data augmentation, we adopt the flipping and rotation, producing seven additional variants of the original training data. Specifically, on the EORSSD dataset [20], we train our ACCoNet with 11,200 augmented pairs for 39 epochs. On the ORSSD dataset [18], we train our ACCoNet with 4,800 augmented pairs for 54 epochs.

3) Evaluation Metrics. We adopt nine widely used evaluation metrics, including S-measure (S_α, α = 0.5) [84], maximum, mean and adaptive F-measure (F_β, β^2 = 0.3) [85], maximum, mean and adaptive E-measure (E_ξ) [86], Mean Absolute Error (MAE, $\mathcal{M}$) and Precision-Recall (PR) curve, to comprehensively measure the performance of our ACCoNet and other compared methods. Specifically, **S-measure** simultaneously measures the region-aware and object-aware structural similarity. **F-measure** is the weighted harmonic mean of precision and recall, and we pay more attention to precision in the paper. **E-measure** jointly considers the local pixel-level match information and the global image-level statistics. **MAE** evaluates the average pixel-level errors. **PR curve** presents the correlation between precision and recall. The evaluation tool[1] provided by Fan *et al.* [6] is adopted by us for convenient evaluation.

[1] http://dpfan.net/d3netbenchmark/

TABLE II
QUANTITATIVE COMPARISON OF OUR METHOD AND OTHER 22 STATE-OF-THE-ART METHODS, INCLUDING FIVE TRADITIONAL NSI-SOD METHODS, TEN CNN-BASED NSI-SOD METHODS, AND SEVEN RSI-SOD METHODS, ON TWO POPULAR DATASETS IN TERMS OF S-MEASURE, MAXIMUM, MEAN AND ADAPTIVE F-MEASURE, MAXIMUM, MEAN AND ADAPTIVE E-MEASURE, AND MAE. WE ALSO REPORT THE FRAMES PER SECOND (FPS) OF ALL METHODS. ↑ AND ↓ INDICATE LARGER AND SMALLER IS BETTER, RESPECTIVELY. THE TOP THREE RESULTS ARE MARKED IN **RED**, **BLUE** AND **GREEN**, RESPECTIVELY. † MEANS DEEP LEARNING BASED METHOD. FOR SIMPLICITY, R3. IS R3NET, POOL. IS POOLNET, EG. IS EGNET, MI. IS MINET, GATE. IS GATENET, LV. IS LVNET, DAF. IS DAFNET, MJRB. IS MJRBM, EMFI. IS EMFINET, AND ACCO. IS ACCONET.

		Traditional NSI-SOD Methods					CNN-based NSI-SOD Methods										RSI-SOD Methods							
		RRWR	HDCT	DSG	SMD	RCRR	DSS†	RADF†	R3.†	PFAN†	Pool.†	EG.†	GCPA†	MI.†	ITSD†	Gate.†	VOS	CMC	SMFF	LV.†	DAF.†	MJRB.†	EMFI.†	**ACCo.†**
		2015	2016	2017	2017	2018	2017	2018	2018	2019	2019	2019	2020	2020	2020	2020	2018	2019	2019	2019	2021	2022	2022	2022
	Metric	[28]	[31]	[32]	[33]	[29]	[45]	[47]	[48]	[49]	[43]	[41]	[44]	[39]	[42]	[23]	[59]	[63]	[17]	[18]	[20]	[57]	[58]	Ours
	FPS ↑	0.3	7	0.6	–	0.3	8	7	2	–	–	–	23	12	16	25	–	–	–	1.4	26	32	25	81
EORSSD [20]	S_α ↑	.5992	.5971	.6420	.7101	.6007	.7868	.8179	.8184	.8348	.8207	.8601	.8869	.9040	.9050	.9114	.5082	.5798	.5401	.8630	**.9166**	**.9197**	**.9290**	**.9290**
	max F_β ↑	.3993	.5407	.5232	.5884	.3995	.6849	.7446	.7498	.7454	.7545	.7880	.8347	.8344	.8523	.8566	.2765	.3268	.5176	.7794	.8614	**.8656**	**.8720**	**.8837**
	mean F_β ↑	.3686	.4018	.4597	.5473	.3685	.5801	.6582	.6302	.6766	.6406	.6967	.7905	.8174	.8221	.8228	.2107	.2692	.2992	.7328	.7845	**.8239**	**.8486**	**.8552**
	adp F_β ↑	.3344	.2658	.4012	.4081	.3347	.4597	.4933	.4165	.5471	.4611	.5379	.6723	**.7705**	.7421	.7109	.1836	.2007	.2083	.6284	.6427	.7066	**.7984**	**.7969**
	max E_ξ ↑	.6894	.7861	.7260	.7697	.6882	.9186	.9140	.9483	.9266	.9292	.9570	.9524	.9442	.9556	.9610	.5982	.6803	.7744	.9254	**.9861**	.9646	**.9711**	**.9727**
	mean E_ξ ↑	.5943	.6376	.6594	.7286	.5946	.7631	.8567	.8294	.8638	.8193	.8775	.9167	.9346	**.9407**	.9385	.4886	.5894	.5197	.8801	.9291	.9350	**.9604**	**.9653**
	adp E_ξ ↑	.5639	.5192	.6188	.6416	.5636	.6933	.7261	.6462	.7738	.6836	.7566	.8647	**.9243**	.9103	.8909	.4767	.4890	.5014	.8445	.8446	.8897	**.9501**	**.9450**
	$\mathcal{M}$ ↓	.1677	.1088	.1246	.0771	.1644	.0186	.0168	.0171	.0160	.0210	.0110	.0102	.0093	.0106	.0095	.2096	.1057	.1434	.0146	**.0060**	.0099	**.0084**	**.0074**
ORSSD [18]	S_α ↑	.6835	.6197	.7195	.7640	.6849	.8262	.8259	.8141	.8613	.8403	.8721	.9026	.9040	.9050	.9186	.5366	.6033	.5312	.8815	.9191	**.9204**	**.9366**	**.9437**
	max F_β ↑	.5590	.5257	.6238	.6692	.5591	.7467	.7619	.7456	.8131	.7706	.8332	.8687	.8761	.8735	.8871	.3471	.3913	.4417	.8263	**.8928**	.8842	**.9002**	**.9149**
	mean F_β ↑	.5125	.4235	.5747	.6214	.5126	.6962	.6856	.7383	.7308	.6999	.7500	.8433	.8574	.8502	**.8679**	.2717	.3454	.2684	.7995	.8511	.8566	**.8856**	**.8971**
	adp F_β ↑	.4874	.3722	.5657	.5568	.4876	.6206	.5730	.7379	.6722	.6166	.6452	.7861	**.8251**	.8068	.8229	.2633	.3108	.2496	.7506	.7876	.8022	**.8617**	**.8806**
	max E_ξ ↑	.7649	.7719	.7912	.8230	.7651	.8860	.9130	.8913	.9519	.9343	.9731	.9509	.9545	.9601	.9664	.6514	.7064	.7402	.9456	**.9771**	.9623	**.9737**	**.9796**
	mean E_ξ ↑	.7017	.6495	.7337	.7745	.7021	.8362	.8298	.8681	.8553	.8650	.9013	.9341	.9454	.9482	.9538	.5352	.6417	.4920	.9259	**.9539**	.9415	**.9671**	**.9754**
	adp E_ξ ↑	.6949	.6291	.7532	.7682	.6950	.8085	.7678	.8887	.8504	.8124	.8226	.9205	.9423	.9335	**.9428**	.5826	.5996	.5676	.9195	.9360	.9328	**.9663**	**.9721**
	$\mathcal{M}$ ↓	.1324	.1309	.1041	.0715	.1277	.0363	.0382	.0399	.0243	.0358	.0216	.0168	.0144	.0165	.0137	.2151	.1267	.1854	.0207	**.0113**	.0163	**.0109**	**.0088**

B. Comparison with State-of-the-arts

1) Comparison Methods. Following the two popular RSI-SOD benchmarks [18], [20], we compare our method with 22 state-of-the-art NSI-SOD and RSI-SOD methods for a comprehensive evaluation. Concretely, these compared methods include five traditional NSI-SOD methods (RRWR [28], HDCT [31], DSG [32], SMD [33], and RCRR [29]), ten CNN-based NSI-SOD methods (DSS [45], RADF [47], R3Net [48], PFAN [49], PoolNet [43], EGNet [41], GCPA [44], MINet [39], ITSD [42], and GateNet [23]), three traditional RSI-SOD methods (VOS [59], CMC [63], and SMFF [17]), and four recent CNN-based RSI-SOD methods (LVNet [18], DAFNet [20], MJRBM [57], and EMFINet [58]). Notably, except for GCPA [44], MINet [39], ITSD [42] and GateNet [23], the saliency maps of all the other compared methods are provided by Zhang *et al.* [20][2] and/or by the authors. Following [18], [20], we fine-tune GCPA [44], MINet [39], ITSD [42] and GateNet [23] with their default hyperparameter settings using the same training data as our method on the two datasets.

2) Quantitative Comparison on EORSSD. We present the quantitative comparison of EORSSD [20] in terms of S_α, F_β, E_ξ and $\mathcal{M}$ in the upper part of Tab. II. Among the eight metrics in Tab. II, our method ranks first in four metrics and second in other four metrics. Overall, on the EORSSD dataset, our method performs the best among all compared methods. EMFINet [58] is the best among the seven existing RSI-SOD methods, and GateNet [23] is the best among existing NSI-SOD methods. In comparison to EMFINet, our method performs marginally lower in terms of adp F_β and adp E_ξ, but surpasses EMFINet by 1.17% on max F_β. Compared with GateNet, our method greatly outperforms it by 2.71%, 3.24%, 5.41% and 8.60% on max F_β, mean F_β, adp E_ξ and adp F_β, respectively. In addition, we show the PR curve in Fig. 4(a), and our method is better than all compared methods.

3) Quantitative Comparison on ORSSD. The quantitative comparison of ORSSD [18] on eight metrics is shown at the bottom part of Tab. II, and the PR curve is shown in Fig. 4(b). Our method consistently outperforms all compared methods among all nine quantitative metrics. Notably, compared with the second best method, the performance gain of our method reaches 1.89% on adp F_β, 1.47% on max F_β, and 1.15% on mean F_β. Among all the compared methods, ours is the only method whose $\mathcal{M}$ is lower than 0.0100, *i.e.*, 0.0088.

According to the quantitative comparison on the two datasets, our method is the best method for RSI-SOD. In addition, comparing the specialized RSI-SOD methods and the NSI-SOD methods in the same period, we can find that the specialized methods are better than the NSI-SOD methods, which indicates that the development of specialized methods is necessary and urgent.

4) Visual Comparison. We show some qualitative results in Fig. 5, including several representative and challenging scenes of optical RSIs, such as object with shadows, tiny object, multiple objects, multiple tiny objects, and irregular geometry structure.

For the first scene, shadows are usually connected with

[2]https://github.com/rmcong/DAFNet_TIP20

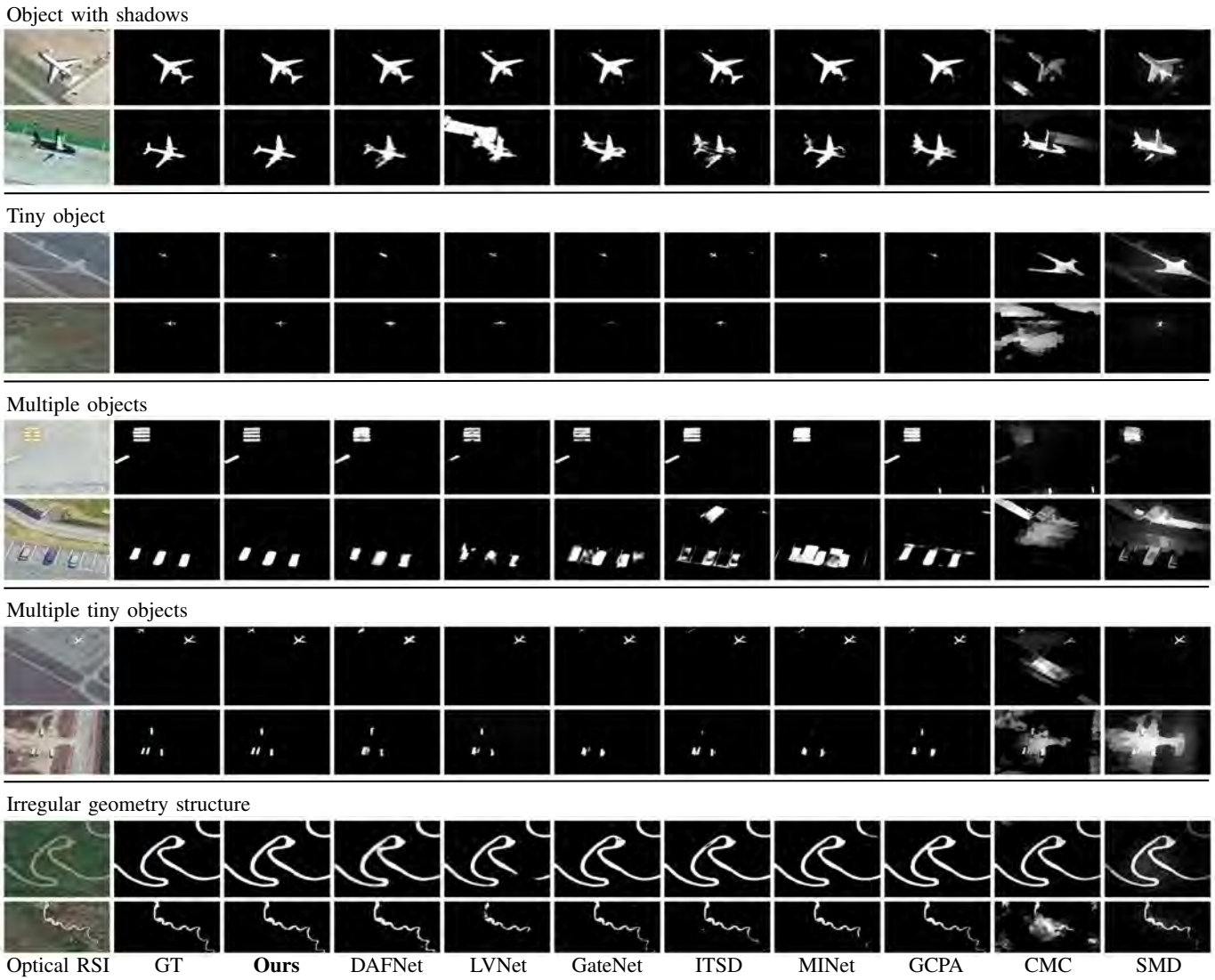


Fig. 5. Visual comparisons with eight representative state-of-the-art methods, including two CNN-based RSI-SOD methods (DAFNet [20] and LVNet [18]), four CNN-based NSI-SOD methods (GateNet [23], ITSD [42], MINet [39] and GCPA [44]), one traditional RSI-SOD method (CMC [63]), and one traditional NSI-SOD method (SMD [33]). Please zoom-in for the best view, especially for tiny object and multiple tiny objects.

salient objects, which will interfere with the detection and highlight inaccurate regions on the saliency map. We can clearly observe that in the second example, LVNet, GateNet, ITSD, CMC and SMD are in this dilemma, but our method can highlight the plane more accurately.

The second scene is unique to optical RSIs and is different from the scene with the small object in NSIs. In this scene, optical RSIs contain much smaller object, *i.e.*, the tiny object. Such extreme scene invalidates traditional methods and two CNN-based NSI-SOD methods, *i.e.*, CMC, SMD, MINet and GCPA. The first two methods detect wrong objects in the first example, and the latter two methods fail to detect any objects in the second example. Besides, other methods can only roughly determine the location of the tiny object but the details cannot be described well. Our method can capture the tiny object with fine details.

Scene with multiple objects has always been the difficulty of the SOD task. In the first example, MINet misses an object. Although other methods detect all objects, objects are

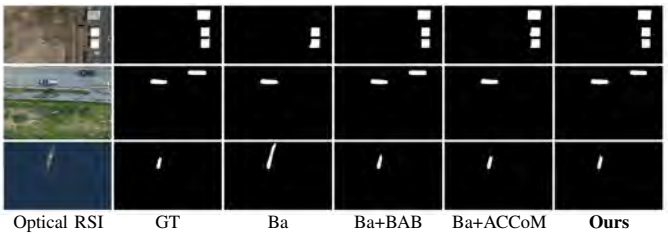


Fig. 6. Visual examples of ablation studies. "Ba" represents the basic encoder-decoder network. Please zoom-in for details.

incomplete. In the second example, due to the complexity of the scene, GateNet, ITSD, MINet, CMC and SMD incorrectly detect more regions. On the contrary, our method locates all objects finely without any redundant regions.

The fourth scene is a combination of the second and third scenes, which puts forward higher requirements for the SOD method. All representative compared methods appear to miss the detection of some objects, while our method distinguishes

TABLE III
ABLATION ANALYSES ON MEASURING THE OVERALL CONTRIBUTIONS OF ACCOM AND BAB IN ACCONET. BASELINE IS THE ENCODER-DECODER NETWORK. THE BEST RESULT IN EACH COLUMN IS **BOLD**.

No.	Baseline	ACCoM	BAB	EORSSD [20]			ORSSD [18]		
				maxF_β	$\mathcal{M}$	maxE_ξ	maxF_β	$\mathcal{M}$	maxE_ξ
1	✓			.8642	.0093	.9547	.8832	.0138	.9566
2	✓	✓		.8819	.0076	.9673	.9117	.0098	.9766
3	✓		✓	.8777	.0086	.9655	.8987	.0131	.9661
4	✓	✓	✓	**.8837**	**.0074**	**.9727**	**.9149**	**.0088**	**.9796**

TABLE IV
ABLATION RESULTS ON CONFIRMING THE EFFECTIVENESS OF TWO TYPES OF BRANCHES IN ACCOM AND THE RATIONALITY OF THE DILATED CONVOLUTION BASED BIFURCATIONS IN BAB. THE BEST RESULT IN EACH COLUMN IS **BOLD**.

Models	EORSSD [20]			ORSSD [18]		
	maxF_β ↑	$\mathcal{M}$ ↓	maxE_ξ ↑	maxF_β ↑	$\mathcal{M}$ ↓	maxE_ξ ↑
ACCoNet (**Ours**)	**.8837**	**.0074**	**.9727**	**.9149**	**.0088**	**.9796**
w/o LB	.8800	.0079	.9681	.9029	.0113	.9691
w/o AB	.8830	.0075	.9704	.9072	.0108	.9739
w/ DC	.8831	.0075	**.9727**	.9136	.0093	.9790
w/ NC	.8834	**.0074**	.9716	.9144	.0090	.9783

w/o LB: ACCoM without local branch. *w/o AB*: ACCoM without adjacent branches.

w/ DC: two bifurcations of BAB are direct connection operations.

w/ NC: two bifurcations of BAB are normal convolutional layers.

all tiny objects.

The last scene refers specifically to the river. Rivers often have very complex irregular geometric structures and span the entire image. They have different widths in different positions, which is not friendly to some methods, causing LVNet, ITSD and MINet to detect only part of the river. Thanks to our method thoroughly explores the contextual information in both encoder and decoder, which is particularly advantageous for variable object scales, object shapes and object quantities in optical RSIs, our method can overcome the above common and complex scenes in optical RSIs.

5) Speed Comparison. In Tab. II, we report the speed of 15 compared methods and ours[3]. Our method reaches a fast processing speed of 81 *fps* on a GPU, which ranks first among 16 compared methods and is more than three times that of the second best method EFMINet (*i.e.*, 25 *fps*). Based on the above comprehensive comparison, our method shows remarkable detection accuracy and astonishing speed.

C. Ablation Studies

In this subsection, we conduct thorough ablation studies on EORSSD [20] and ORSSD [18] to investigate the impact of the two vital components in our method. Specifically, we analyze 1) the overall contributions of ACCoM and BAB in ACCoNet, 2) the effectiveness of two types of branches in ACCoM, 3) the rationality of the dilated convolution based bifurcations in BAB, 4) the complementarity between BCE and IoU in loss function, and 5) the flexibility of our method. For each variant, we strictly modify only one part at a time and retrain the variant on the two datasets using the same training settings as in Sec. IV-A.

1. The overall contributions of ACCoM and BAB in ACCoNet. As shown in Tab. III, to measure the overall contributions of the proposed ACCoM and BAB to ACCoNet, we offer three variants: 1) the encoder-decoder network (*i.e.*, "Baseline"), 2) the baseline network with only ACCoMs (*i.e.*, "Baseline+ACCoM"), and 3) the baseline network with only BABs (*i.e.*, "Baseline+ACCoM"). Besides, the complete ACCoNet is "Baseline+ACCoM+BAB". We report the quantitative results in Tab. III.

[3]The speed of RRWR, HDCT, DSG, RCRR, DSS, RADF, R3Net and LVNet are borrowed from [18], the speed of GCPA, MINet, ITSD, GateNet and MJRBM is obtained by our test, and the speed of DAFNet and EFMINet is obtained from original papers.

On the EORSSD dataset, we observe that "Baseline" only achieves 86.42% on max F_β, 0.0093 on $\mathcal{M}$ and 95.47% on max E_ξ. ACCoM increases "Baseline" by 1.76%, 0.0017 and 1.26% on these three metrics respectively, while BAB increases "Baseline" by 1.35%, 0.0007 and 1.08% on these three metrics respectively. With the joint cooperation of ACCoM and BAB, our complete ACCoNet improves "Baseline" by 1.95%, 0.0017 and 1.80% on these three metrics respectively. The trends on the ORSSD dataset are the same as that on the EORSSD dataset. Notably, our complete ACCoNet improves "Baseline" by 3.17%, 0.0050 and 2.30% on max F_β, $\mathcal{M}$ and max E_ξ, respectively, which more clearly validates the effectiveness of each proposed module.

Additionally, we show saliency maps of these three variants and our method in Fig. 6. In the first and second examples, "Ba" (*i.e.*, "Baseline") misses an object. In the first example, both ACCoM and BAB complete the missing object. Differently, in the second example, only BAB completes the missing object. This means that as long as ACCoM or BAB can complete the missing object, the complete ACCoNet (*i.e.*, "Ours") can get accurate saliency maps. In the third example, "Ba" mistakenly highlights the background region. BAB suppresses part of the background and ACCoM suppresses more background, resulting in a satisfactory saliency map of "Ours". The above quantitative and qualitative analysis confirms that both ACCoM and BAB are important for ACCoNet, and the contextual information explored by these two modules is conducive to the detection of salient objects in optical RSIs.

2. The effectiveness of two types of branches in ACCoM. To investigate the effectiveness of two types of branches in ACCoM, we provide two variants: 1) removing the local branch in ACCoM (*i.e.*, *w/o LB*) and 2) removing the adjacent branches in ACCoM (*i.e.*, *w/o AB*). The ablation results are reported in the third and fourth rows of Tab. IV.

We discover that the performances of *w/o LB* and *w/o AB* are worse than ours, which demonstrates that these two types of branches are effective. Concretely, on the ORSSD dataset, the performance of *w/o LB* is degraded, *e.g.*, max F_β: 91.49% → 90.29%, $\mathcal{M}$: 0.0088 → 0.0113, max E_ξ: 97.96% → 96.91%,

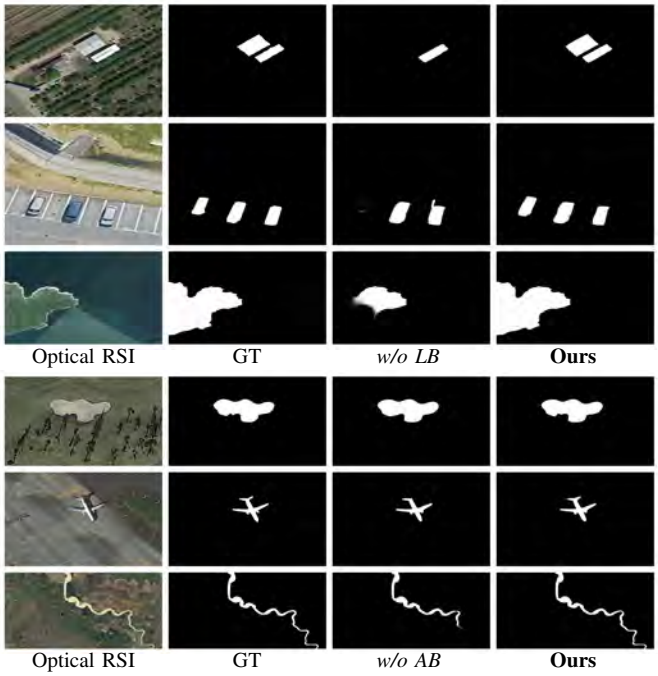


Fig. 7. Visual examples of two variants, *w/o LB* and *w/o AB*.

while the performance of *w/o AB* drops slightly, *e.g.*, max F_β: 91.49% $\rightarrow$ 90.72%, $\mathcal{M}$: 0.0088 $\rightarrow$ 0.0108, max E_ξ: 97.96% $\rightarrow$ 97.39%. The same trend is observed on the EORSSD dataset. The reason is that the feature modulation of adjacent branches is based on $\boldsymbol{f}_{\mathrm{c}}^t$, which belongs to the local branch. If we remove the local branch, the global assistance provided by two adjacent features will act on $\boldsymbol{f}_{\mathrm{e}}^t$, which cannot exert the maximum effect of global assistance. Thus, we conclude that the local branch is the core of ACCoM.

Specifically, in Fig. 7, we show saliency maps of these two variants and our complete method to visually evaluate the role of the local branch and the adjacent branches. As shown in the first three examples of Fig. 7, the saliency maps of *w/o LB* miss objects in the case of multiple salient objects (the first two examples), and cannot detect the complete object in the case of large salient object (the third one). This is because after removing the local branch, the location information of salient objects will be reduced, resulting in two types of missed detections. Differently, for the saliency maps of *w/o AB*, the salient objects are basically located accurately, but the details are not perfectly outlined, such as the regions occluded by the tree (the fourth one), the airplane tail (the fifth one), and the slender river (the last one). After removing the adjacent branches, the cross-level contextual complementary information is discarded, causing the damage of the salient object details. In summary, the local branch is good for scenes with multiple salient objects and large salient object, while the adjacent branches are good for scenes containing salient objects with fine details.

3. The rationality of the dilated convolution based bifurcations in BAB. To validate the rationality of the dilated convolution based bifurcations in BAB, we conduct two variants: 1) replacing dilated convolutions by direct connection operations (*i.e.*, *w/ DC*) and 2) replacing dilated convolutions by normal convolutional layers (*i.e.*, *w/ NC*). The ablation results are reported in the last two rows of Tab. IV.

TABLE V
ABLATION STUDY ON EVALUATING THE COMPLEMENTARITY BETWEEN BCE AND IOU IN LOSS FUNCTION. THE BEST RESULT IN EACH COLUMN IS **BOLD**.

No.	BCE	IoU	EORSSD [20]			ORSSD [18]		
			maxF_β ↑	$\mathcal{M}$ ↓	maxE_ξ ↑	maxF_β ↑	$\mathcal{M}$ ↓	maxE_ξ ↑
1	✓		.8731	.0085	.9666	.9018	.0117	.9703
2		✓	.8801	.0081	.9711	.9027	.0105	.9747
3	✓	✓	**.8837**	**.0074**	**.9727**	**.9149**	**.0088**	**.9796**

TABLE VI
PERFORMANCE ON DIFFERENT ENCODER BACKBONES OF OUR ACCONET.

Models	EORSSD [20]			ORSSD [18]		
	maxF_β ↑	$\mathcal{M}$ ↓	maxE_ξ ↑	maxF_β ↑	$\mathcal{M}$ ↓	maxE_ξ ↑
ACCoNet-VGG	.8837	.0074	.9727	.9149	.0088	.9796
ACCoNet-ResNet	.8821	.0067	.9759	.9149	.0087	.9819

In general, we find that the performance gap between these two variants and our original BAB is small. However, with direct connection operations, BAB cannot fully demonstrate its ability to capture contextual information, which leads to performance degradation, *e.g.*, max F_β: 88.31% (*w/ DC*) *v.s.* 88.37% (Ours) on the EORSSD and 91.36% (*w/ DC*) *v.s.* 91.49% (Ours) on the ORSSD. The normal convolutional layers slightly improve the ability of BAB compared to direct connection operations, *e.g.*, max F_β: 88.31% (*w/ DC*) $\rightarrow$ 88.34% (*w/ NC*) on the EORSSD and 91.36% (*w/ DC*) $\rightarrow$ 91.44% (*w/ NC*) on the ORSSD. In summary, the dilated convolution based bifurcations can capture better various contextual information with different receptive fields in the decoder.

4. The complementarity between BCE and IoU in loss function. To prove the complementarity between BCE and IoU in loss function, we provide two variants: 1) training our method with only BCE loss and 2) training our method with only IoU loss. We report the quantitative results in Tab. V.

As shown in Tab. V, training our ACCoNet with only BCE loss or IoU loss can achieve promising performance, but the performance of these two variants is worse than that of our complete loss function. This is because BCE loss is a pixel-level supervision, and IoU loss is a map-level supervision. The two losses train the network from different aspects, and they can complement each other. Combining the two losses to train our method together is conducive to keeping the completeness of salient objects. This composite loss function is popular in the field of SOD [58], [73], [80].

5. The flexibility of our method. To demonstrate the flexibility of our method, we provide a variant, namely ACCoNet-ResNet, which adopts ResNet-50 [87] as the encoder backbone, and report the performance in Tab. VI. As shown in Tab. VI, with the more powerful encoder backbone ResNet-

50, the performance of ACCoNet-ResNet is improved on most evaluation metrics as compared with our original method, *i.e.*, ACCoNet-VGG in Tab. VI, whose encoder backbone is VGG-16. We can conclude that our method shows strong adaptability to different encoder backbones.

V. Conclusion

In this paper, we investigate the contextual knowledge in an encoder-decoder architecture and proposed an effective ACCoNet for RSI-SOD. We believe that the contextual information is beneficial to tackle variable object scales, object shapes and object quantities in RSI-SOD. In the encoder, we propose the Adjacent Context Coordination Module (ACCoM) to coordinate the adjacent features (*i.e.*, the current, previous and subsequent features) and explore adjacent information for salient regions activation. In the decoder, we propose the Bifurcation-Aggregation Block (BAB) to capture the multi-scale contents for salient regions inference. Both ACCoMs and BABs learn contextual information to improve the representation of salient objects. In particular, we employ the deep supervision with hybrid losses to stabilize the network training. Extensive experiments, including quantitative, visual and speed comparisons and ablation studies, demonstrate that the proposed method is superior to 22 relevant state-of-the-art methods, and the two proposed modules contribute significantly to performance.

References

[1] 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.

[2] A. Borji, M.-M. Cheng, Q. Hou, H. Jiang, and J. Li, "Salient object detection: A survey," *Comput. Vis. Media*, vol. 5, no. 2, pp. 117–150, Jun. 2019.

[3] W. Wang, Q. Lai, H. Fu, J. Shen, H. Ling, and R. Yang, "Salient object detection in the deep learning era: An in-depth survey," *IEEE Trans. Pattern Anal. Mach. Intell.*, 2021. [Online]. Available: 10.1109/TPAMI.2021.3051099

[4] W. Wang, J. Shen, J. Xie, M.-M. Cheng, H. Ling, and A. Borji, "Revisiting video saliency prediction in the deep learning era," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 1, pp. 220–237, Jan. 2021.

[5] R. Cong, J. Lei, H. Fu, M.-M. Cheng, W. Lin, and Q. Huang, "Review of visual saliency detection with comprehensive information," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 29, no. 10, pp. 2941–2959, Oct. 2019.

[6] D.-P. Fan, Z. Lin, Z. Zhang, M. Zhu, and M.-M. Cheng, "Rethinking RGB-D salient object detection: Models, data sets, and large-scale benchmarks," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 32, no. 5, pp. 2075–2089, 2021.

[7] G. Li, Z. Liu, R. Shi, and W. Wei, "Constrained fixation point based segmentation via deep neural network," *Neurocomputing*, vol. 368, pp. 180–187, Nov. 2019.

[8] G. Li, Z. Liu, R. Shi, Z. Hu, W. Wei, Y. Wu, M. Huang, and H. Ling, "Personal fixations-based object segmentation with object localization and boundary preservation," *IEEE Trans. Image Process.*, vol. 30, pp. 1461–1475, Jan. 2021.

[9] S. Hong, T. You, S. Kwak, and B. Han, "Online tracking by learning discriminative saliency map with convolutional neural network," in *Proc. ICML*, Jul. 2015, pp. 597–606.

[10] 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, p. 107130, Apr. 2020.

[11] K. Gu, S. Wang, H. Yang, W. Lin, G. Zhai, X. Yang, and W. Zhang, "Saliency-guided quality assessment of screen content images," *IEEE Trans. Multimedia*, vol. 18, no. 6, pp. 1098–1110, Jun. 2016.

[12] S. Yang, Q. Jiang, W. Lin, and Y. Wang, "SGDNet: An end-to-end saliency-guided deep neural network for no-reference image quality assessment," in *Proc. ACM MM*, Oct. 2019, pp. 1383–1391.

[13] Q. Wang, J. Lin, and Y. Yuan, "Salient band selection for hyperspectral image classification via manifold ranking," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 27, no. 6, pp. 1279–1289, Jun. 2016.

[14] L. Zhang, S. Wang, and X. Li, "Salient region detection in remote sensing images based on color information content," in *Proc. IEEE IGARSS*, Jul. 2015, pp. 1877–1880.

[15] 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, pp. 1–14, Sept. 2015.

[16] L. Zhang, Y. Wang, and Y. Sun, "Salient target detection based on the combination of super-pixel and statistical saliency feature analysis for remote sensing images," in *Proc. IEEE ICIP*, Oct. 2018, pp. 2336–2340.

[17] 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, May 2019.

[18] 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.

[19] C. Li, R. Cong, C. Guo, H. Li, C. Zhang, F. Zheng, and Y. Zhao, "A parallel down-up fusion network for salient object detection in optical remote sensing images," *Neurocomputing*, vol. 415, pp. 411–120, Nov. 2020.

[20] Q. Zhang, R. Cong, C. Li, M.-M. Cheng, Y. Fang, X. Cao, Y. Zhao, and S. Kwong, "Dense attention fluid network for salient object detection in optical remote sensing images," *IEEE Trans. Image Process.*, vol. 30, pp. 1305–1317, 2021.

[21] 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, 2021.

[22] G. Li, Z. Liu, W. Lin, and H. Ling, "Multi-content complementation network for salient object detection in optical remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–13, 2022.

[23] X. Zhao, Y. Pang, L. Zhang, H. Lu, and L. Zhang, "Suppress and balance: A simple gated network for salient object detection," in *Proc. ECCV*, Aug. 2020, pp. 35–51.

[24] Y. LeCun, B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, and L. D. Jackel, "Backpropagation applied to handwritten zip code recognition," *Neural Comput.*, vol. 1, no. 4, pp. 541–551, Dec. 1989.

[25] L. Itti, C. Koch, and E. Niebur, "A model of saliency-based visual attention for rapid scene analysis," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 20, no. 11, pp. 1254–1259, Nov. 1998.

[26] J.-G. Yu, J. Zhao, J. Tian, and Y. Tan, "Maximal entropy random walk for region-based visual saliency," *IEEE Trans. Cybern.*, vol. 44, no. 9, pp. 1661–1672, Sept. 2014.

[27] Z. Liu, W. Zou, and O. Le Meur, "Saliency tree: A novel saliency detection framework," *IEEE Trans. Image Process.*, vol. 23, no. 5, pp. 1937–1952, May 2014.

[28] C. Li, Y. Yuan, W. Cai, Y. Xia, and D. D. Feng, "Robust saliency detection via regularized random walks ranking," in *Proc. IEEE CVPR*, Jun. 2015, pp. 2710–2717.

[29] 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.

[30] 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.

[31] 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.

[32] 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.

[33] 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.

[34] S. Wang, S. Yang, M. Wang, and L. Jiao, "New contour cue-based hybrid sparse learning for salient object detection," *IEEE Trans. Cybern.*, 2019. [Online]. Available: 10.1109/TCYB.2018.2881482

[35] 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.

[36] M. Liang and X. Hu, "Feature selection in supervised saliency prediction," *IEEE Trans. Cybern.*, vol. 45, no. 5, pp. 914–926, May 2015.

[37] 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.

[38] M. Huang, Z. Liu, L. Ye, X. Zhou, and Y. Wang, "Saliency detection via multi-level integration and multi-scale fusion neural networks," *Neurocomputing*, vol. 364, pp. 310–321, Oct. 2019.

[39] Y. Pang, X. Zhao, L. Zhang, and H. Lu, "Multi-scale interactive network for salient object detection," in *Proc. IEEE CVPR*, Jun. 2020, pp. 9410–9419.

[40] 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, 2021.

[41] 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 ICCV*, Oct. 2019, pp. 8779–8788.

[42] H. Zhou, X. Xie, J.-H. Lai, Z. Chen, and L. Yang, "Interactive two-stream decoder for accurate and fast saliency detection," in *Proc. IEEE CVPR*, Jun. 2020, pp. 9138–9147.

[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 CVPR*, Jun. 2019, pp. 3912–3921.

[44] Z. Chen, Q. Xu, R. Cong, and Q. Huang, "Global context-aware progressive aggregation network for salient object detection," in *Proc. AAAI*, Feb. 2020, pp. 10 599–10 606.

[45] Q. Hou, M.-M. Cheng, X. Hu, A. Borji, Z. Tu, and P. Torr, "Deeply supervised salient object detection with short connections," in *Proc. IEEE CVPR*, Jul. 2017, pp. 5300–5309.

[46] Y. Liu, M.-M. Cheng, X.-Y. Zhang, G.-Y. Nie, and M. Wang, "DNA: Deeply supervised nonlinear aggregation for salient object detection," *IEEE Trans. Cybern.*, 2021. [Online]. Available: 10.1109/TCYB.2021.3051350

[47] X. Hu, L. Zhu, J. Qin, C.-W. Fu, and P.-A. Heng, "Recurrently aggregating deep features for salient object detection," in *Proc. AAAI*, Feb. 2018, pp. 6943–6950.

[48] Z. Deng, X. Hu, L. Zhu, X. Xu, J. Qin, G. Han, and P.-A. Heng, "R^3Net: Recurrent residual refinement network for saliency detection," in *Proc. IJCAI*, Jul. 2018, pp. 684–690.

[49] T. Zhao and X. Wu, "Pyramid feature attention network for saliency detection," in *Proc. IEEE CVPR*, Jun. 2019, pp. 3080–3089.

[50] J. Li, Z. Pan, Q. Liu, Y. Cui, and Y. Sun, "Complementarity-aware attention network for salient object detection," *IEEE Trans. Cybern.*, vol. 52, no. 2, pp. 873–886, 2022.

[51] S. Chen, B. Wang, X. Tan, and X. Hu, "Embedding attention and residual network for accurate salient object detection," *IEEE Trans. Cybern.*, vol. 50, no. 5, pp. 2050–2062, May 2020.

[52] Y. Wu, Z. Liu, and X. Zhou, "Saliency detection using adversarial learning networks," *J. Vis. Commun. Image Represent.*, vol. 67, p. 102761, Feb. 2020.

[53] H. Li, G. Li, and Y. Yu, "ROSA: Robust salient object detection against adversarial attacks," *IEEE Trans. Cybern.*, vol. 50, no. 11, pp. 4835–4847, Nov. 2020.

[54] 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, 2021.

[55] Y. Liu, Y.-C. Gu, X.-Y. Zhang, W. Wang, and M.-M. Cheng, "Lightweight salient object detection via hierarchical visual perception learning," *IEEE Trans. Cybern.*, vol. 51, no. 9, pp. 4439–4449, 2021.

[56] 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.

[57] Z. Tu, C. Wang, C. Li, M. Fan, H. Zhao, and B. Luo, "ORSI salient object detection via multiscale joint region and boundary model," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–13, 2022.

[58] X. Zhou, K. Shen, Z. Liu, C. Gong, J. Zhang, and C. Yan, "Edge-aware multiscale feature integration network for salient object detection in optical remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–15, 2022.

[59] 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.

[60] E. Li, S. Xu, W. Meng, and X. Zhang, "Building extraction from remotely sensed images by integrating saliency cue," *IEEE J. Sel. Top. Appl. Earth Observ. Remote Sens.*, vol. 10, no. 3, pp. 906–919, Mar. 2017.

[61] L. Zhang, A. Li, Z. Zhang, and K. Yang, "Global and local saliency analysis for the extraction of residential areas in high-spatial-resolution remote sensing image," *IEEE Trans. Geosci. Remote Sens.*, vol. 54, no. 7, pp. 3750–3763, Jul. 2016.

[62] C. Dong, J. Liu, F. Xu, and C. Liu, "Ship detection from optical remote sensing images using multi-scale analysis and Fourier HOG descriptor," *Remote Sens.*, vol. 11, no. 13, pp. 1–19, Jun. 2019.

[63] 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, pp. 1–18, May 2019.

[64] M. Jing, D. Zhao, M. Zhou, Y. Gao, Z. Jiang, and Z. Shi, "Unsupervised oil tank detection by shape-guide saliency model," *IEEE Geosci. Remote Sens. Lett.*, vol. 16, no. 3, pp. 477–481, Mar. 2019.

[65] L. Zhang and K. Yang, "Region-of-interest extraction based on frequency domain analysis and salient region detection for remote sensing image," *IEEE Geosci. Remote Sens. Lett.*, vol. 11, no. 5, pp. 916–920, May 2014.

[66] L. Ma, B. Du, H. Chen, and N. Q. Soomro, "Region-of-interest detection via superpixel-to-pixel saliency analysis for remote sensing image," *IEEE Geosci. Remote Sens. Lett.*, vol. 13, no. 12, pp. 1752–1756, Dec. 2016.

[67] T. Li, J. Zhang, X. Lu, and Y. Zhang, "SDBD: A hierarchical region-of-interest detection approach in large-scale remote sensing image," *IEEE Geosci. Remote Sens. Lett.*, vol. 14, no. 5, pp. 699–703, May 2017.

[68] G. Liu, L. Qi, Y. Tie, and L. Ma, "Region-of-interest detection based on statistical distinctiveness for panchromatic remote sensing images," *IEEE Geosci. Remote Sens. Lett.*, vol. 16, no. 2, pp. 271–275, Feb. 2019.

[69] V. Badrinarayanan, A. Kendall, and R. Cipolla, "SegNet: A deep convolutional encoder-decoder architecture for image segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 12, pp. 2481–2495, Dec. 2017.

[70] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. MICCAI*, Oct. 2015, pp. 234–241.

[71] G. Li, Z. Liu, and H. Ling, "ICNet: Information conversion network for RGB-D based salient object detection," *IEEE Trans. Image Process.*, vol. 29, pp. 4873–4884, Mar. 2020.

[72] G. Li, Z. Liu, L. Ye, Y. Wang, and H. Ling, "Cross-modal weighting network for RGB-D salient object detection," in *Proc. ECCV*, Aug. 2020, pp. 665–681.

[73] G. Li, Z. Liu, M. Chen, Z. Bai, W. Lin, and H. Ling, "Hierarchical alternate interaction network for RGB-D salient object detection," *IEEE Trans. Image Process.*, vol. 30, pp. 3528–3542, Mar. 2021.

[74] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *ICLR*, May 2015, pp. 1–14.

[75] F. Yu and V. Koltun, "Multi-scale context aggregation by dilated convolutions," in *ICLR*, May 2016, pp. 1–13.

[76] S. Ioffe and C. Szegedy, "Batch Normalization: Accelerating deep network training by reducing internal covariate shift," in *Proc. ICML*, vol. 37, Jul. 2015, pp. 448–456.

[77] J. Hu, L. Shen, S. Albanie, G. Sun, and E. Wu, "Squeeze-and-excitation networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 8, pp. 2011–2023, Aug. 2020.

[78] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "CBAM: Convolutional block attention module," in *Proc. ECCV*, Sept. 2018, pp. 3–19.

[79] S. Xie and Z. Tu, "Holistically-nested edge detection," in *Proc. IEEE ICCV*, Dec. 2015, pp. 1395–1403.

[80] X. Qin, Z. Zhang, C. Huang, C. Gao, M. Dehghan, and M. Jagersand, "BASNet: Boundary-aware salient object detection," in *Proc. IEEE CVPR*, Jun. 2019, pp. 7479–7489.

[81] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Kopf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, and S. Chintala, "PyTorch: An imperative style, high-performance deep learning library," in *Proc. NeurIPS*, Dec. 2019, pp. 8024–8035.

[82] K. He, X. Zhang, S. Ren, and J. Sun, "Delving deep into rectifiers: Surpassing human-level performance on ImageNet classification," in *Proc. IEEE ICCV*, Dec. 2015, pp. 1026–1034.

[83] D. P. Kingma and J. L. Ba, "Adam: A method for stochastic optimization," in *ICLR*, May 2015, pp. 1–15.

[84] 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 ICCV*, Oct. 2017, pp. 4548–4557.

[85] R. Achanta, S. Hemami, F. Estrada, and S. Susstrunk, "Frequency-tuned salient region detection," in *Proc. IEEE CVPR*, Jun. 2009, pp. 1597–1604.

[86] D.-P. Fan, C. Gong, Y. Cao, B. Ren, M.-M. Cheng, and A. Borji, "Enhanced-alignment measure for binary foreground map evaluation," in *Proc. IJCAI*, Jul. 2018, pp. 698–704.
[87] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE CVPR*, Jun. 2016, pp. 770–778.