Multi-Content Complementation Network for Salient Object Detection in Optical Remote Sensing Images

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

***Abstract*—In the computer vision community, great progresses have been achieved in salient object detection from natural scene images (NSI-SOD); by contrast, salient object detection in optical remote sensing images (RSI-SOD) remains to be a challenging emerging topic. The unique characteristics of optical RSIs, such as scales, illuminations, and imaging orientations, bring significant differences between NSI-SOD and RSI-SOD. In this article, we propose a novel multi-content complementation network (MCCNet) to explore the complementarity of multiple content for RSI-SOD. Specifically, MCCNet is based on the general encoder–decoder architecture, and contains a novel key component named multi-content complementation module (MCCM), which bridges the encoder and the decoder. In MCCM, we consider multiple types of features that are critical to RSI-SOD, including foreground features, edge features, background features, and global image-level features, and exploit the content complementarity between them to highlight salient regions over various scales in RSI features through the attention mechanism. Besides, we comprehensively introduce pixel-level, map-level, and metric-aware losses in the training phase. Extensive experiments on two popular datasets demonstrate that the proposed MCCNet outperforms 23 state-of-the-art methods, including both NSI-SOD and RSI-SOD methods. The code and results of our method are available at https://github.com/MathLee/MCCNet.**

***Index Terms*—Background, edge, multi-content complementation, optical remote sensing images, salient object detection (SOD).**

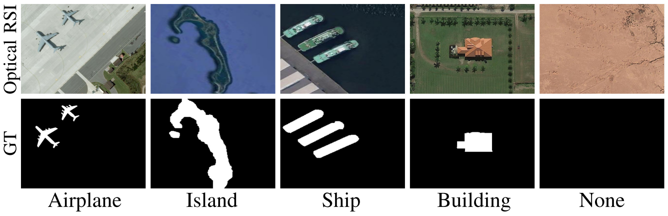


Fig. 1. Representative example scenes in the RSI-SOD task. “None” means there is no salient object in this scene. GT is the ground truth.

I. Introduction

Visual attention mechanism aims to capture the most attractive regions in a scene, and plays an important role in the human visual system. In computer vision, efforts have been devoted to model this mechanism and can be generally divided into two important topics: *fixation prediction* and *salient object detection*. The former predicts visual saliency degree of regions, while the latter highlights salient object regions. In this article, we focus on salient object detection (SOD) [1]–[4], which has shown successful applications in various computer vision tasks, such as object segmentation [5], [6], image quality assessment [7], [8], image retargeting [9], *etc*. And different from the classic SOD in natural scene images (NSI-SOD), we are dedicated to SOD in optical remote sensing images (RSI-SOD) [10], [11]. Specifically, the optical RSIs refer to color images photographed by satellites and aerial sensors in the range of 400–760 nm [10], [12], and have only three optical bands (RGB), which are different from hyperspectral images that include more spectral bands information [13]. RSI-SOD aims at highlighting airplanes, islands, ships, buildings, and rivers, which attract humans' attention, at the pixel level in the optical RSI.

Convolutional neural networks (CNNs) [14] significantly stimulate NSI-SOD [2] and greatly improve the detection accuracy. Recently, as many thought-provoking ideas and techniques, such as multilevel/scale fusion [15], edge guidance/preservation [16], [17], attention [18], [19], complementary losses [20], [21], *etc*., are introduced into NSI-SOD; NSI-SOD has become more mature. However, there are big differences between the acquisition of NSIs and optical RSIs. Optical RSIs are photographed by satellite and aerial sensors, so the object types, scales, illuminations, imaging orientations, and backgrounds of optical RSIs are fundamentally different from NSIs. Some representative scenes in RSI-SOD task are shown in Fig. 1. The last scene of Fig. 1 is special, there is no salient object. Thus, directly applying NSI-SOD methods to optical RSIs may be inappropriate.

Manuscript received August 29, 2021; revised November 3, 2021; accepted November 24, 2021. Date of publication November 30, 2021; date of current version February 21, 2022. This work was supported in part by the National Natural Science Foundation of China under Grant 62171269, in part by the China Scholarship Council under Grant 202006890079, and in part by the Singapore Ministry of Education Tier-2 Fund MOE2016-T2-2-057(S). *(Corresponding author: Zhi Liu.)*

Gongyang Li is with Shanghai Institute for Advanced Communication and Data Science, Shanghai University, Shanghai 200444, China, also with the School of Communication and Information Engineering, Shanghai University, Shanghai 200444, China, and also with the School of Computer Science and Engineering, Nanyang Technological University, Singapore 639798 (e-mail: ligongyang@shu.edu.cn).

Zhi Liu is with Shanghai Institute for Advanced Communication and Data Science, Shanghai University, Shanghai 200444, China, and also with the School of Communication and Information Engineering, Shanghai University, Shanghai 200444, China (e-mail: liuzhisjtu@163.com).

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 (e-mail: hling@cs.stonybrook.edu).

Digital Object Identifier 10.1109/TGRS.2021.3131221

However, as an emerging topic of saliency detection, RSI-SOD solutions are heavily inspired by NSI-SOD ones, especially the CNN-based ones. Concretely, as a pioneer work in RSI-SOD, LVNet [10] fuses multiresolution inputs in a nested structure to perceive objects of different sizes. PDFNet [22] integrates five-scale features from five branches for comprehensive detection. DAFNet [11] not only employs the salient edge map as the additional supervision, but also performs attention in a dense fluid manner. Similar to [10], EMFINet [23] adopts optical RSIs with three different resolutions as inputs, but different from [11], it employs edge supervision to generate features with edge-aware constraint and introduces a hybrid loss to infer salient objects with shape boundaries. These specialized CNN-based RSI-SOD methods are based on the characteristics of optical RSI to propose effective solutions and obtain promising performance.

Motivated by the above observations, we expand the advantages of NSI-SOD methods [16], [17], [19] and propose a novel *multi-content complementation module* (MCCM) to adapt to the characteristics of optical RSIs. Specifically, we first integrate the foreground content into our MCCM. Similar to [11], [16], [17], [23], we introduce edge content, but the difference is that we employ edge supervision to produce an edge attention map for edge activation in features. For RSI-SOD, we believe that in addition to foreground and edge, the background [19] is also important. Here, we consider the complex background content of optical RSIs. The above three kinds of content cover local information in detail. Inspired by [24], we incorporate global image-level content for comprehensive content complementation. In this way, our MCCM captures both local and global content simultaneously, which is effective for accurately perceiving salient regions and distinguishing cluttered background regions.

Moreover, to improve the robustness of our MCCM, we implement MCCM at multiple feature scales. We deploy MCCM in an encoder–decoder network, which is a general backbone for NSI-SOD, and propose a simple yet effective *multi-content complementation network* (MCCNet) for RSI-SOD. Benefiting from the progressive inference procedure in the backbone, our MCCNet can highlight salient regions with various scales and object types and flexibly adapt to the challenging scenes of optical RSIs. In addition, following [20], [21], we construct a comprehensive loss function to efficiently train our MCCNet.

Our main contributions are summarized as follows.

1) We propose a MCCM to explore the complementarity of multiple content in features of optical RSIs for salient regions perception. In MCCM, the local content, i.e., foreground, edge, and background, and the global image-level content are simultaneously exploited.
2) We embed MCCM on multiple feature scales in an encoder–decoder network, and propose an effective and efficient MCCNet for RSI-SOD, which runs at a fast inference speed of 95 frames/s on a single GPU. MCCNet perfectly combines the feature complementation ability of MCCM and the inference ability of the basic network.
3) We conduct comprehensive experiments on two benchmark RSI-SOD datasets. The experimental results demonstrate that the proposed MCCNet is superior to 23 state-of-the-art methods under various evaluation metrics, and the effectiveness of the proposed MCCM is also verified.

The rest of this article is organized as follows. Section II reviews the related work of NSI-SOD and RSI-SOD, Section III presents our MCCNet in detail, Section IV elaborates experiments and ablation studies, and Section V draws the conclusion.

II. Related Work

In this section, we first summarize the works of NSI-SOD, and then elaborate RSI-SOD methods. For each topic, we introduce both traditional and CNN-based methods.

A. Salient Object Detection in Natural Scene Images

As a pioneer in saliency detection, Itti *et al.* [25] proposed the first computational visual attention model for NSIs, which is the cornerstone of other traditional work. Liu *et al.* [26] proposed an unsupervised method based on kernel density estimation. Liu *et al.* [27] proposed the saliency tree framework based on salient region merging and salient node selection. The regularized random walks ranking was proposed in [28], and Yuan *et al.* [29] further combined it with reversion correction. Meanwhile, Zou *et al.* [30] jointly handled the SOD and object segmentation, effectively exploring the complementary cues of the two tasks. Kim *et al.* [31] extended the high-dimensional color transform-based SOD method with a local learning-based method. Zhou *et al.* [32] integrated the diffusion results of foreground and background into the final saliency map. Peng *et al.* [33] applied the structured matrix decomposition to NSI-SOD. Although traditional methods do not achieve impressive performance, they provide numerous valuable and thought-provoking solutions to NSI-SOD.

The CNN-based NSI-SOD methods [2] break through the performance bottleneck of traditional methods [1] and promote NSI-SOD to a new era. For instance, Hou *et al.* [34] implemented the deep supervision at multiple side-output layers for NSI-SOD. Many subsequent methods [16], [19], [35]–[39] have applied the deep supervision scheme to NSI-SOD. Zhang *et al.* [15] fused features over different scales to extract multiscale information, while Pang *et al.* [40] integrated features of three adjacent levels. Zhao *et al.* [41] proposed a gated dual branch control interference between different levels of features. Edge/boundary cues were maturely used in NSI-SOD in various ways. Wang *et al.* [17] directly extracted edge region from image, and sent it into the backbone network together with the image and superpixel region. Differently, Wu *et al.* [42] used the Sobel operator to obtain the edge label as additional edge supervision. Zhao *et al.* [16] captured the salient edge from the ground truth and used it to force network learn edge features for one-to-one guidance. Moreover, Liu *et al.* [18] learned pixel-wise local and global attention to facilitate detection. Chen *et al.* [19] introduced the background information through the proposed reverse attention. For supervision, in addition to the popular

BCE loss, Ma *et al.* [43] introduced the IoU loss, Qin *et al.* [20] further introduced the SSIM loss, and Xu *et al.* [39] introduced common losses of fixation prediction to NSI-SOD. Zhao *et al.* [21] further proposed a metric-aware F-measure loss based on the popular evaluation metric F-measure [44]. Besides, the global context-aware aggregation [36], [45] and the recurrent mechanism [46], [47] have also been widely explored.

Though existing NSI-SOD methods cannot be directly applied to optical RSIs, they still provide important references for RSI-SOD. Our method incorporates some advantages of these NSI-SOD methods, such as deep supervision, complementary losses, and edge information, to adapt to the particularity of optical RSIs.

B. Salient Object Detection in Optical Remote Sensing Images

Remote sensing image processing has been popular in the past decade. Hong *et al.* [49] proposed a general multimodal deep learning (MDL) framework, which consists of Ex-Net and Fu-Net, for pixel-level remote sensing image classification. They introduced and developed five fusion architectures, including early fusion, middle fusion, late fusion, en-de fusion, and cross fusion, in the MDL framework. In [50], graph convolutional networks were introduced into hyperspectral image classification. To address the shortage of identifying materials in cross-modality remote sensing data, X-ModalNet [51], a semisupervised deep cross-modal framework, was proposed for classification in remote sensing data. Moreover, Hong *et al.* [52] proposed an augmented linear mixing model to address spectral variability for hyperspectral unmixing. More in-depth analysis can be found in [13], which elaborates on the interpretable hyperspectral artificial intelligence.

In addition to the above popular tasks of remote sensing image processing, there are some tasks similar to RSI-SOD, such as airport detection [53], ship detection [54]–[56], oil tank detection [57], [58], building extraction [59], residential areas extraction [60], [61], and object detection from aerial images [62]. In fact, these object detection/extraction tasks mostly focus on specific scenes and objects, such as airport, ship, oil tank, building, and residential area. By contrast, the RSI-SOD task involves all these scenarios, and is hence more general and challenging.

In particular, RSI-SOD task aims at extracting the most attractive objects in optical RSIs, and considers the subjective initiative of human more than the region-of-interest extraction task [63]–[66]. Here, we first introduce some traditional RSI-SOD methods. Faur *et al.* [67] regarded RSI-SOD as a data information compression task, and proposed a rate-distortion measure-based method. Based on the global and background information, Zhao *et al.* [68] proposed a sparse representation-based saliency computation method. Zhang *et al.* [69]–[71] proposed a series of unsupervised methods: in [69], the saliency map was constructed based on color information content; in [70], the statistical saliency feature map and the information saliency feature map were fused for final saliency map; and in [71], a low-rank matrix recovery-based self-adaptive multiple feature fusion method was proposed.

Compared with traditional RSI-SOD methods, CNN-based RSI-SOD methods provide more powerful solutions for complex optical RSIs. In [10] and [11], two challenging datasets of RSI-SOD were constructed. And Li *et al.* [10] extracted multiscale features directly from five different resolution optical RSIs in a two-stream pyramid module, and further perceived objects of different sizes in a V-shaped module with nested connections. Following NSI-SOD methods such as [16], [42], Zhang *et al.* [11] constructed a multitask architecture, which predicts saliency map and salient edge map simultaneously, to sharpen the object boundaries. Furthermore, they proposed a cascaded pyramid attention module to solve the problem of object scale changes. Following [10], Zhou *et al.* [23] extracted multiscale features from three different resolution optical RSIs, and captured edge features using edge supervision for edge preservation. And they adopted the hybrid loss, including the pixel-level BCE loss, patch-level SSIM loss, and map-level IoU loss [20], to facilitate the training. Different from [10], [23], Li *et al.* [22] only extracted features from an optical RSI, but efficiently captured multiresolution information by integrating five different levels of features for inference. Due to lack of optical RSIs data, Zhang and Ma [72] introduced the weakly supervised learning into RSI-SOD. They first generated pseudo labels with auxiliary images in a classification network, and then constructed a deep but lightweight feedback saliency analysis network to progressively refine saliency map. We can find the impression of CNN-based NSI-SOD methods among the above-specialized CNN-based RSI-SOD methods, but these specialized methods are uniquely constructed according to the characteristics of optical RSIs.

The above-mentioned previous arts suggest that the foreground prior, edge cue, and background cue play an important role in SOD. However, they have been exploited independently in RSI-SOD. In our MCCM, we not only explore the complementation among three kinds of content (*i.e.*, foreground, edge, and background), but also introduce the global (image-level) content, which is more comprehensive than [11] and [23]. Moreover, we lay out MCCM with five feature scales in our MCCNet with one network input, which is more efficient than [10] and [23].

III. Proposed Method

In this section, we detail the proposed MCCNet. In Section III-A, we present the network overview of our MCCNet. In Section III-B, we elaborate our MCCM. In Section III-C, we clarify the comprehensive loss function.

A. Network Overview

As depicted in Fig. 2, our MCCNet is built on the encoder–decoder architecture, which is friendly to pixel-level image segmentation [73], [74] and various SOD tasks [41], [45], [75], [76], and comprises three key parts: encoder network, five MCCM components, and decoder network.

For the encoder network, we adapt the popular VGG-16 [48] for basic feature extraction. Different from the original VGG-16 structure for image classification task, we delete the

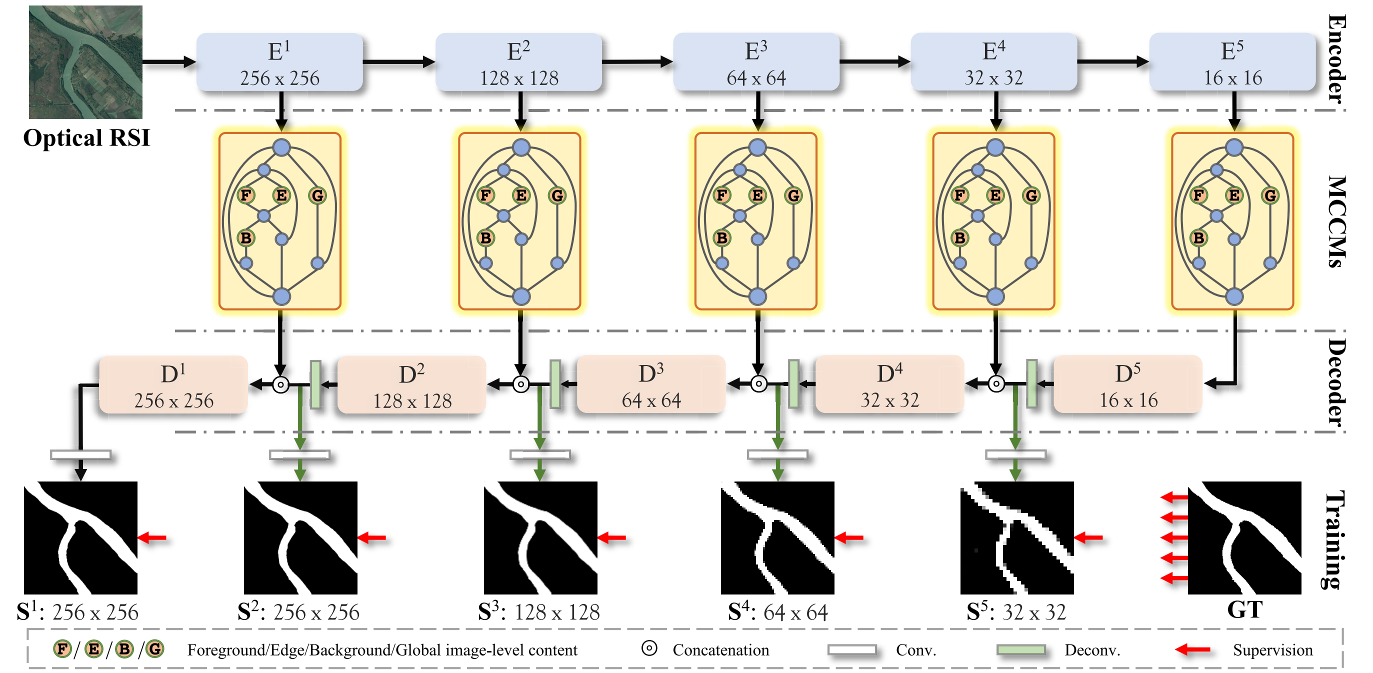


Fig. 2. Overall framework of the proposed MCCNet, which is based on the general encoder–decoder architecture. We first extract the basic features using the classic VGG-16 [48] from an optical RSI with a size of 256 × 256 × 3. Then, we model the complementary information between foreground features, edge features, background features, and global image-level features in the pivotal MCCM. Finally, we progressively infer salient objects using multiscale features output from five MCCMs in the decoder. In the training phase, we employ the comprehensive supervision to each decoder block, including the pixel-level BCE loss, map-level IoU loss, and metric-aware F-m loss.

last four layers, including one max-pooling layer and three fully connected layers, for our pixel-level RSI-SOD task, and denote the remaining five convolution blocks as E^t, where t is the block index and belongs to $\{1, 2, 3, 4, 5\}$. For $t = 1, 2$, E^t contains two convolutional layers; for $t = 3, 4, 5$, E^t contains three convolutional layers. For the input optical RSI $\mathbf{I} \in \mathbb{R}^{256\times256\times3}$, the extracted features of each convolution block are denoted as $\boldsymbol{f}_{\mathrm{e}}^t \in \mathbb{R}^{h_t\times w_t\times c_t}$, where h_t is $(256/2^{t-1})$, w_t is $(256/2^{t-1})$, and c_t belongs to $\{64, 128, 256, 512, 512\}$. Then, the basic features $\boldsymbol{f}_{\mathrm{e}}^t$ of five levels will be fed to the corresponding MCCM to produce $\boldsymbol{f}_{\mathrm{mccm}}^t \in \mathbb{R}^{h_t\times w_t\times c_t}$. In MCCM, we generate the foreground, edge, background, and global image-level features from the source features $\boldsymbol{f}_{\mathrm{e}}^t$, and explore the complementarity between them. Since MCCMs are deployed in five levels, they can capture multiscale complementary information, which is beneficial to RSI-SOD. Finally, our decoder network infers salient objects based on $\boldsymbol{f}_{\mathrm{mccm}}^t$ in a progressive resolution restoration manner. Our decoder network also contains five blocks, denoted by D^t, whose structure corresponds to that of E^t, *i.e.*, D^t contains two convolutional layers for $t = 1, 2$, and D^t contains three convolutional layers for $t = 3, 4, 5$. Between the two decoder blocks, we use a deconvolutional layer to restore the resolution. In addition to the classic binary cross-entropy (BCE) loss, we introduce intersection-over-union (IoU) loss and F-measure (F-m) loss as auxiliary losses for each decoder block to comprehensively supervise the network training.

B. Multi-Content Complementation Module

In the studies of saliency detection, foreground assumption is often regarded as the prior in traditional NSI-SOD methods [28], [32], and provides effective guidance to find obvious salient regions. It can be explored in deep learning through the attention mechanism [77]. The background information is another important cue in NSI-SOD to determine the nonsalient regions, and is modeled by the reverse attention [19]. The edge information is also widely used to complete the salient objects. Since the scenes of RSI-SOD are typically more complicated than that of NSI-SOD, it is insufficient to consider the above three kinds of content in RSI-SOD independently. According to the above motivation, we propose the MCCM, so as to address RSI-SOD based on the complementation of foreground, background, and edge. In addition, we integrate the global image-level content into MCCM, which is indispensable for RSI-SOD.

We illustrate the structure of MCCM in Fig. 3. In MCCM, the input features are $\boldsymbol{f}_{\mathrm{e}}^t$ from E^t. We regard $\boldsymbol{f}_{\mathrm{e}}^t$ as source features, which generate all four kinds of content. In the following, we elaborate MCCM based on these four kinds of content.

1) Foreground and Edge: As shown in Fig. 3, foreground and edge features are extracted in parallel. Considering that $\boldsymbol{f}_{\mathrm{e}}^t$ of VGG-16 is relatively rough, we first perform the channel attention [77] on $\boldsymbol{f}_{\mathrm{e}}^t$ to reduce redundant information and purify $\boldsymbol{f}_{\mathrm{e}}^t$ as follows:

$$\boldsymbol{f}_{\mathrm{ca}}^t = \mathrm{CA}\big(\boldsymbol{f}_{\mathrm{e}}^t\big) \odot \boldsymbol{f}_{\mathrm{e}}^t \tag{1}$$

where $\boldsymbol{f}_{\mathrm{ca}}^t \in \mathbb{R}^{h_t\times w_t\times c_t}$ denotes the purified features, $\mathrm{CA}(\cdot)$ is the channel attention,[1] and $\odot$ is the channel-wise multiplication.

[1] Channel attention is implemented by a spatial global max pooling (GMP_s), two fully connected layers, and a sigmoid activation function.

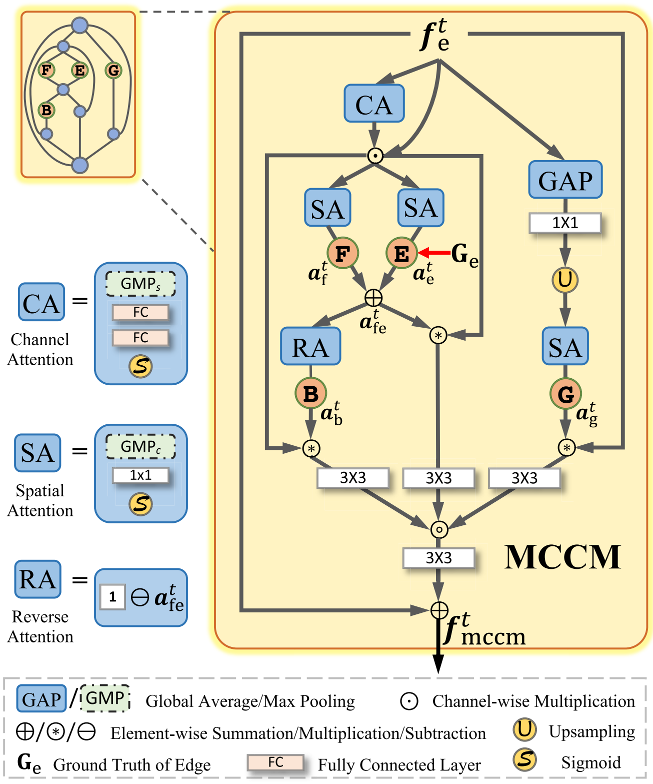


Fig. 3. Illustration of the MCCM.

Then, we obtain the foreground map and edge map, denoted by $\{a_{\mathrm{f}}^t, a_{\mathrm{e}}^t\} \in [0,1]^{h_t \times w_t \times 1}$, through the spatial attention [77] at the same time, which can be computed as

$$a_{\mathrm{f}}^t = \mathrm{SA}\left(f_{\mathrm{ca}}^t\right) \quad (2)$$
$$a_{\mathrm{e}}^t = \mathrm{SA}\left(f_{\mathrm{ca}}^t\right) \quad (3)$$

where $\mathrm{SA}(\cdot)$ is the spatial attention.[2] Notably, the foreground map is generated in an adaptive way, while the edge map is generated in a learning way, *i.e.*, under the supervision of the ground truth of edge in the training phase.

Since both foreground map and edge map are correlated with the salient regions and can complement each other, we aggregate them together using the element-wise summation, and obtain the foreground-edge map, denoted by $a_{\mathrm{fe}}^t \in [0,2]^{h_t \times w_t \times 1}$. And we adopt the foreground-edge map to highlight the salient regions at feature level as follows:

$$f_{\mathrm{fe}}^t = a_{\mathrm{fe}}^t \circledast f_{\mathrm{ca}}^t \quad (4)$$

where $f_{\mathrm{fe}}^t \in \mathbb{R}^{h_t \times w_t \times c_t}$ is the foreground-edge features and $\circledast$ is the element-wise multiplication. The way we merge the foreground map and edge map is different from those of using edge information in SOD [10], [16], [20], [23]. In particular, our method explicitly explores the complementary information between these two maps, and is therefore more effective.

2) Background: The generation of background map is closely related to the foreground-edge map. Following [19], we obtain the background map, denoted by $a_{\mathrm{b}}^t \in [-1,1]^{h_t \times w_t \times 1}$, through the reverse attention. In the background map a_{b}^t, we redefine the concept of background and foreground, *i.e.*, the background is defined as 1, and the foreground is defined as −1. And we also adopt the background map to highlight the nonsalient regions at feature level. The process is written as

$$a_{\mathrm{b}}^t = \mathbf{1} \ominus a_{\mathrm{fe}}^t \quad (5)$$
$$f_{\mathrm{b}}^t = a_{\mathrm{b}}^t \circledast f_{\mathrm{ca}}^t \quad (6)$$

where $\mathbf{1}$ is a matrix with size $h_t \times w_t \times 1$, where all elements are 1, $\ominus$ is the element-wise subtraction, and $f_{\mathrm{b}}^t \in \mathbb{R}^{h_t \times w_t \times c_t}$ are the background features. We can find that the background features are based on the foreground-edge features, but they are the opposite of the foreground-edge features. In the subsequent processing of MCCM, we will merge them at channel level, and implicitly extract the complementary information between them.

3) Global Image-Level Content: In fact, whether it is foreground-edge features or background features, they contain local information, which benefits to complete the detail and boundary of salient objects. And inspired by [24], we introduce the global image-level content to capture the overall tone of source features in our MCCM.

Concretely, following [24], we apply spatial-wise global average pooling on f_{e}^t to extremely compress global distribution information into pixels and get the basic image-level features, and perform a 1 × 1 convolutional layer for feature smoothing. Then, we reconstruct the image-level content to the same size as the original f_{e}^t using upsampling with bilinear interpolation. Such a rough operation will lose a lot of detailed information, but the reconstructed features can reflect the overall tone of source features. Different from [24], which directly integrates the image-level content at channel level through concatenation, we compress the reconstructed image-level content into an elegant response map, namely the global image-level map $a_{\mathrm{g}}^t \in [0,1]^{h_t \times w_t \times 1}$, via the spatial attention. The entire process is formulated as follows:

$$a_{\mathrm{g}}^t = \mathrm{SA}\left(\mathrm{up}\left(\mathrm{conv}_{1\times1}\left(\mathrm{GAP}_s\left(f_{\mathrm{e}}^t\right)\right)\right)\right) \quad (7)$$

where $\mathrm{GAP}_s(\cdot)$ is the spatial global average pooling, $\mathrm{conv}_{1\times1}(\cdot)$ is the 1×1 convolutional layer, and $\mathrm{up}(\cdot)$ is the upsampling operation. We adopt a_{g}^t to reflect the overall tone at feature level as follows:

$$f_{\mathrm{g}}^t = a_{\mathrm{g}}^t \circledast f_{\mathrm{e}}^t \quad (8)$$

where $f_{\mathrm{g}}^t \in \mathbb{R}^{h_t \times w_t \times c_t}$ is the global image-level feature.

4) Multi-Content Aggregation: Through the above thorough operations, we obtain features of four kinds of content, *i.e.*, f_{fe}^t, f_{b}^t, and f_{g}^t, and further polish them using the 3 × 3 convolutional layer, obtaining $\hat{f}_{\mathrm{fe}}^t$, $\hat{f}_{\mathrm{b}}^t$, and $\hat{f}_{\mathrm{g}}^t$. Then, we aggregate them using the adaptive concatenation–convolution operation. Besides, we adopt a short connection to retain the original content to generate the output features of MCCM $f_{\mathrm{mccm}}^t \in \mathbb{R}^{h_t \times w_t \times c_t}$. The entire aggregation process is written as

$$f_{\mathrm{mccm}}^t = \mathrm{conv}_{3\times3}\left(\hat{f}_{\mathrm{fe}}^t \circledcirc \hat{f}_{\mathrm{b}}^t \circledcirc \hat{f}_{\mathrm{g}}^t\right) \oplus f_{\mathrm{e}}^t \quad (9)$$

[2] Spatial attention is implemented by a GMP_c, a convolutional layer and a sigmoid activation function.

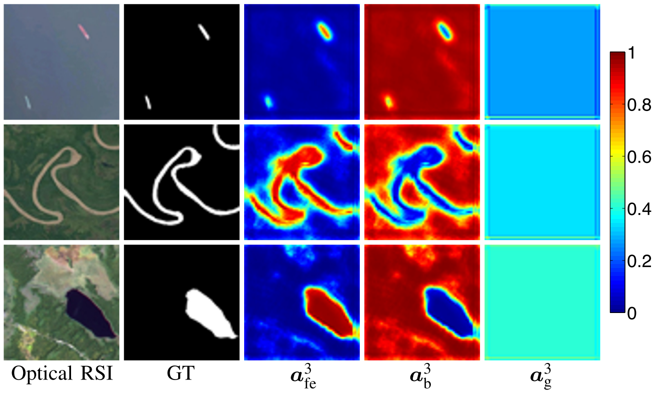


Fig. 4. Feature visualization of components in MCCM attached to E^3. Please zoom-in for viewing details.

where ⊚ is the cross-channel concatenation and ⊕ is the element-wise summation. In summary, $\boldsymbol{f}_{\text{mccm}}^t$ comprises the essence of foreground, edge, background, global image-level content, and original content, which makes MCCM an indispensable part of MCCNet. The MCCMs equipped with five feature levels assist MCCNet to adapt to the complex scenes and changeable objects in optical RSIs.

In Fig. 4, we visualize three maps of four kinds of content in MCCM attached to E^3, *i.e.*, $\boldsymbol{a}_{\text{fe}}^3$, $\boldsymbol{a}_{\text{b}}^3$, and $\boldsymbol{a}_{\text{g}}^3$. With the combination of $\boldsymbol{a}_{\text{f}}^3$ and $\boldsymbol{a}_{\text{e}}^3$, $\boldsymbol{a}_{\text{fe}}^3$ can clearly highlight the salient regions, regardless of whether they have a tiny size, a large size, multiple objects, or a complex topology, and meanwhile $\boldsymbol{a}_{\text{b}}^3$ makes the nonsalient regions obvious. And $\boldsymbol{a}_{\text{g}}^3$ provides the specific basic tone of source features of each optical RSI.

C. Comprehensive Loss Function

For a successful CNN-based model, an effective architecture and several well-designed modules are necessary. In addition, a good training strategy can improve model performance without extra model parameters. As shown in Fig. 2, we adopt the widely used deep supervision [34], [78] in the training phase to monitor the intermediate saliency maps of different sizes, which forces features to learn the characteristics of salient regions of different sizes. And inspired by the successful usage of hybrid and complementary loss in SOD [20], [23], [79], we adopt the classic pixel-level BCE loss and map-level IoU loss in our loss function. We also include the metric-aware F-m loss [21] in our loss function to further facilitate the network training. Thus, we construct a comprehensive loss function $\mathbb{L}_{\text{s}}^t$ to supervise the predicted saliency map $\mathbf{S}^t$ of D^t, which can be formulated as

$$\mathbb{L}_{\text{s}}^t = \ell_{\text{bce}}(\text{up}(\mathbf{S}^t), \mathbf{G}) + \ell_{\text{iou}}(\text{up}(\mathbf{S}^t), \mathbf{G}) + \ell_{\text{fm}}(\text{up}(\mathbf{S}^t), \mathbf{G}) \tag{10}$$

where $\mathbf{G}$ is the ground truth, and $\ell_{\text{bce}}(\cdot)$, $\ell_{\text{iou}}(\cdot)$, and $\ell_{\text{fm}}(\cdot)$ are BCE loss, IoU loss, and F-m loss, respectively. They can be computed as follows:

$$\ell_{\text{bce}} = -\sum_{i=1}^{W\cdot H}[\mathbf{G}(i)\log(\mathbf{S}(i)) + (1-\mathbf{G}(i))\log(1-\mathbf{S}(i))] \tag{11}$$

$$\ell_{\text{iou}} = 1 - \frac{\sum_{i=1}^{W\cdot H}\mathbf{S}(i)\cdot\mathbf{G}(i)}{\sum_{i=1}^{W\cdot H}[\mathbf{S}(i)+\mathbf{G}(i)-\mathbf{S}(i)\cdot\mathbf{G}(i)]} \tag{12}$$

$$\ell_{\text{fm}} = 1 - \frac{(1+\beta^2)\cdot P(\mathbf{S},\mathbf{G})\cdot R((\mathbf{S},\mathbf{G}))}{\beta^2\cdot P(\mathbf{S},\mathbf{G}) + R(\mathbf{S},\mathbf{G})} \tag{13}$$

where $\mathbf{G}(i) \in \{0,1\}$ and $\mathbf{S}(i) \in [0,1]$ are the ground truth label and predicted saliency score of the ith pixel, respectively, β^2 is 0.3, $P = (\text{TP}/(\text{TP}+\text{FP}))$, $R = (\text{TP}/(\text{TP}+\text{FN}))$, $\text{TP}(\mathbf{S},\mathbf{G}) = \sum_{i=1}^{W\cdot H}\mathbf{S}(i)\cdot\mathbf{G}(i)$, $\text{FP}(\mathbf{S},\mathbf{G}) = \sum_{i=1}^{W\cdot H}\mathbf{S}(i)\cdot(1-\mathbf{G}(i))$, and $\text{FN}(\mathbf{S},\mathbf{G}) = \sum_{i=1}^{W\cdot H}(1-\mathbf{S}(i))\cdot\mathbf{G}(i)$. Our comprehensive loss function helps our MCCNet better adapt to the special scenes of optical RSIs.

Besides, in the t-th MCCM, we generate the edge map $\boldsymbol{a}_{\text{e}}^t$ by a model learned from minimizing with the edge loss $\mathbb{L}_{\text{e}}^t$, which can be formulated as

$$\mathbb{L}_{\text{e}}^t = \ell_{\text{bce}}\left(\text{up}\left(\boldsymbol{a}_{\text{e}}^t\right), \mathbf{G}_{\text{e}}\right) \tag{14}$$

where $\mathbf{G}_{\text{e}}$ is the ground truth of edge, generated in the same way as [16]. Therefore, the total loss $\mathbb{L}_{\text{total}}$ of our MCCNet in the training phase can be expressed as

$$\mathbb{L}_{\text{total}} = \sum_{t=1}^{5}\left(\mathbb{L}_{\text{s}}^t + \mathbb{L}_{\text{e}}^t\right). \tag{15}$$

IV. Experiments

A. Experimental Protocol

1) Datasets: We train and evaluate our method and the compared methods on two public RSI-SOD datasets.

ORSSD [10] consists of 800 optical RSIs with corresponding pixel-wise ground truths, including multiple scenes, such as ships, cars, airplanes, playgrounds, rivers, and islands. We adopt 600 images with their ground truths for training and the remaining 200 images for testing.

EORSSD [11] expands the ORSSD dataset to include more complex and challenging scenes, resulting in 2000 optical RSIs with corresponding pixel-wise annotation, which is the largest available RSI-SOD dataset. We adopt 1400 images with their ground truths for training and the remaining 600 images for testing.

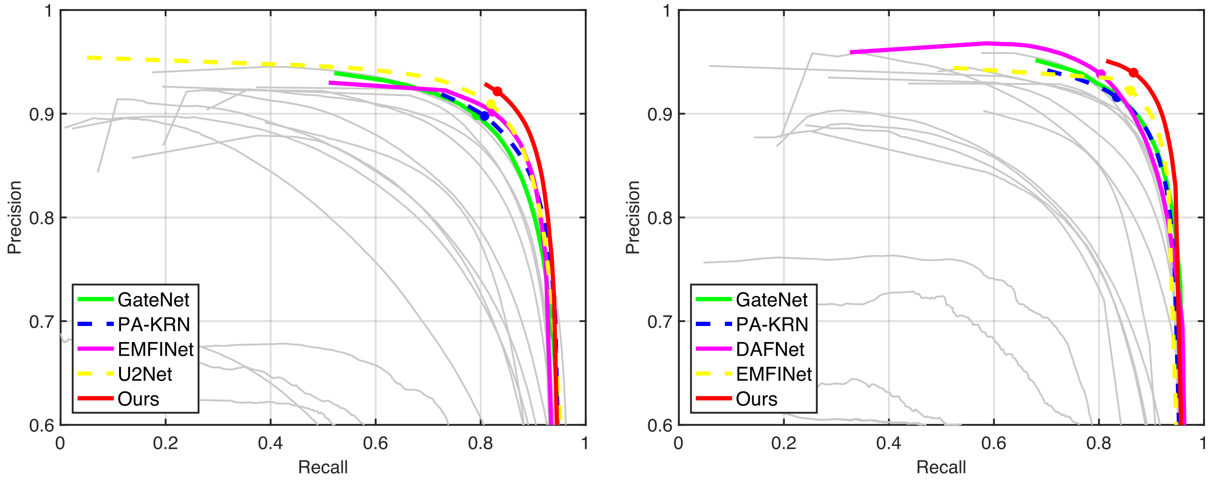
2) Implementation Details: We conduct experiments of our proposed MCCNet on PyTorch [80] platform with an NVIDIA Titan X GPU (12 GB memory). During the network training, each optical RSI is resized to 256 × 256 and augmented by flipping and rotation, producing seven additional training samples. We initialize the encoder network of our MCCNet and other newly added convolutional layers by the pre-trained VGG-16 model [48] and the normal distribution [81], respectively. We utilize the Adam optimizer [82] for network optimization with the batch size 8 and the initial learning rate $1e^{-4}$, which will be divided by 10 after 30 epochs. Following [11], [23], for the EORSSD dataset [11], we train

TABLE I

QUANTITATIVE RESULTS ON TWO RSI-SOD DATASETS, INCLUDING EORSSD AND ORSSD. THERE ARE 23 STATE-OF-THE-ART METHODS, INCLUDING FIVE TRADITIONAL NSI-SOD METHODS, 12 CNN-BASED NSI-SOD METHODS, 3 TRADITIONAL RSI-SOD METHODS, AND 3 CNN-BASED RSI-SOD METHODS. ↑/↓ MEANS A LARGER/SMALLER SCORE IS BETTER. THE TOP THREE RESULTS ARE HIGHLIGHTED IN **RED**, **BLUE**, AND **GREEN**, RESPECTIVELY

Methods	Type	Speed	#Param	FLOPs	EORSSD [11]								ORSSD [10]							
		(*fps*)↑	(M)↓	(G)↓	$S_\alpha \uparrow$	$F_\beta^{\max} \uparrow$	$F_\beta^{\text{mean}} \uparrow$	$F_\beta^{\text{adp}} \uparrow$	$E_\xi^{\max} \uparrow$	$E_\xi^{\text{mean}} \uparrow$	$E_\xi^{\text{adp}} \uparrow$	$\mathcal{M} \downarrow$	$S_\alpha \uparrow$	$F_\beta^{\max} \uparrow$	$F_\beta^{\text{mean}} \uparrow$	$F_\beta^{\text{adp}} \uparrow$	$E_\xi^{\max} \uparrow$	$E_\xi^{\text{mean}} \uparrow$	$E_\xi^{\text{adp}} \uparrow$	$\mathcal{M} \downarrow$
RRWR$_{15}$ [28]	T.N.	0.3	-	-	.5992	.3993	.3686	.3344	.6894	.5943	.5639	.1677	.6835	.5590	.5125	.4874	.7649	.7017	.6949	.1324
HDCT$_{16}$ [31]	T.N.	7	-	-	.5971	.5407	.4018	.2658	.7861	.6376	.5192	.1088	.6197	.5257	.4235	.3722	.7719	.6495	.6291	.1309
DSG$_{17}$ [32]	T.N.	0.6	-	-	.6420	.5232	.4597	.4012	.7260	.6594	.6188	.1246	.7195	.6238	.5747	.5657	.7912	.7337	.7532	.1041
SMD$_{17}$ [33]	T.N.	-	-	-	.7101	.5884	.5473	.4081	.7697	.7286	.6416	.0771	.7640	.6692	.6214	.5568	.8230	.7745	.7682	.0715
RCRR$_{18}$ [29]	T.N.	0.3	-	-	.6007	.3995	.3685	.3347	.6882	.5946	.5636	.1644	.6849	.5591	.5126	.4876	.7651	.7021	.6950	.1277
DSS$_{17}$ [34]	C.N.	8	62.23	114.6	.7868	.6849	.5801	.4597	.9186	.7631	.6933	.0186	.8262	.7467	.6962	.6206	.8860	.8362	.8085	.0363
RADF$_{18}$ [46]	C.N.	7	62.54	214.2	.8179	.7446	.6582	.4933	.9140	.8567	.7162	.0168	.8259	.7619	.6856	.5730	.9130	.8298	.7678	.0382
R3Net$_{18}$ [47]	C.N.	2	56.16	47.5	.8184	.7498	.6302	.4165	.9483	.8294	.6462	.0171	.8141	.7456	.7383	.7379	.8913	.8681	.8887	.0399
EGNet$_{19}$ [16]	C.N.	9	108.07	291.9	.8601	.7880	.6967	.5379	.9570	.8775	.7566	.0110	.8721	.8332	.7500	.6452	.9731	.9013	.8226	.0216
PoolNet$_{19}$ [45]	C.N.	25	53.63	123.4	.8207	.7545	.6406	.4611	.9292	.8193	.6836	.0210	.8403	.7706	.6999	.6166	.9343	.8650	.8124	.0358
GCPA$_{20}$ [36]	C.N.	23	67.06	54.3	.8869	.8347	.7905	.6723	.9524	.9167	.8647	.0102	.9026	.8687	.8433	.7861	.9509	.9341	.9205	.0168
ITSD$_{20}$ [37]	C.N.	16	17.08	54.5	.9050	.8523	.8221	.7421	.9556	.9407	.9103	.0106	.9050	.8735	.8502	.8068	.9601	.9482	.9335	.0165
MINet$_{20}$ [40]	C.N.	12	47.56	146.3	.9040	.8344	.8174	.7705	.9442	.9346	.9243	.0093	.9040	.8761	.8574	.8251	.9545	.9454	.9423	.0144
GateNet$_{20}$ [41]	C.N.	25	100.02	108.3	.9114	.8566	.8228	.7109	.9610	.9385	.8909	.0095	.9186	.8871	.8679	.8229	.9664	.9538	.9428	.0137
U2Net$_{20}$ [35]	C.N.	25	44.01	376.2	**.9199**	**.8732**	.8329	.7221	.9649	.9373	.8989	**.0076**	.9162	.8738	.8492	.8038	.9539	.9387	.9326	.0166
SUCA$_{21}$ [38]	C.N.	24	117.71	56.4	.8988	.8229	.7949	.7260	.9520	.9277	.9082	.0097	.8989	.8484	.8237	.7748	.9584	.9400	.9194	.0145
PA-KRN$_{21}$ [39]	C.N.	16	141.06	617.7	.9192	.8639	**.8358**	**.7993**	.9616	**.9536**	**.9416**	.0104	**.9239**	.8890	**.8727**	**.8548**	.9680	**.9620**	**.9579**	.0139
VOS$_{18}$ [53]	T.R.	-	-	-	.5082	.2765	.2107	.1836	.5982	.4886	.4767	.2096	.5366	.3471	.2717	.2633	.6514	.5352	.5826	.2151
CMC$_{19}$ [57]	T.R.	-	-	-	.5798	.3268	.2692	.2007	.6803	5894	.4890	.1057	.6033	.3913	.3454	.3108	.7064	.6417	.5996	.1267
SMFF$_{19}$ [71]	T.R.	-	-	-	.5401	.5176	.2992	.2083	.7744	.5197	.5014	.1434	.5312	.4417	.2684	.2496	.7402	.4920	.5676	.1854
LVNet$_{19}$ [10]	C.R.	1.4	-	-	.8630	.7794	.7328	.6284	.9254	.8801	.8445	.0146	.8815	.8263	.7995	.7506	.9456	.9259	.9195	.0207
DAFNet$_{21}$ [11]	C.R.	26	29.35	68.5	.9166	.8614	.7845	.6427	**.9861**	.9291	.8446	**.0060**	.9191	**.8928**	.8511	.7876	**.9771**	.9539	.9360	**.0113**
EMFINet$_{21}$ [23]	C.R.	25	107.26	480.9	**.9290**	**.8720**	**.8486**	**.7984**	**.9711**	**.9604**	**.9501**	.0084	**.9366**	**.9002**	**.8856**	**.8617**	**.9737**	**.9671**	**.9663**	**.0109**
Ours	C.R.	95	67.65	112.8	**.9327**	**.8904**	**.8604**	**.8137**	**.9755**	**.9685**	**.9538**	**.0066**	**.9437**	**.9155**	**.9054**	**.8957**	**.9800**	**.9758**	**.9735**	**.0087**

T.N.: Traditional NSI-SOD method; C.N.: CNN-based NSI-SOD method; T.R.: Traditional RSI-SOD method; C.R.: CNN-based RSI-SOD method.



(a) (b)

Fig. 5. Quantitative comparison in terms of PR curve on two datasets for RSI-SOD, *i.e.*, EORSSD and ORSSD. We show the top five methods in color. (a) EORSSD [11]. (b) ORSSD [10].

our MCCNet with 1400 original optical RSI and GT pairs and their 9800 augmented samples for 39 epochs. While for the ORSSD dataset [10], we train our MCCNet with 600 original optical RSI and GT pairs and their 4200 augmented samples for 34 epochs.

3) Evaluation Metrics: We use five evaluation metrics to evaluate our method and other compared methods: S-measure (S_α, $\alpha = 0.5$) [83] is responsible for evaluating the structural similarity at object-aware and region-aware levels. F-measure (F_β) [44] is the weighted harmonic average of precision and recall. We set β^2 to 0.3 to emphasize the precision over recall, and adopt its maximum, mean, and adaptive forms (*i.e.*, $F_\beta^{\max}$, F_β^{mean}, and F_β^{adp}) for comprehensive measure. E-measure (E_ξ) [84] simultaneously captures the local match information at pixel-level and the global statistics at image-level. We report its maximum, mean, and adaptive values (*i.e.*, $E_\xi^{\max}$, E_ξ^{mean}, and E_ξ^{adp}). Mean Absolute Error (MAE, $\mathcal{M}$) evaluates the average pixel-level difference. Precision-recall (PR) curve plots different combinations of precision and recall with the threshold ranging from 0 to 255.

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

We compare our proposed method with 23 state-of-the-art SOD methods, which can be divided into four categories. The first one contains the traditional NSI-SOD methods, including RRWR [28], HDCT [31], DSG [32], SMD [33], and RCRR [29]. The second one includes the CNN-based NSI-SOD methods, including DSS [34], RADF [46], R3Net [47], EGNet [16], PoolNet [45], GCPA [36], ITSD [37], MINet [40], GateNet [41], U2Net [35], SUCA [38], and PA-KRN [39]. The third one includes the traditional RSI-SOD methods, including VOS [53], CMC [57], and SMFF [71]. The last one contains the CNN-based RSI-SOD methods, including LVNet [10], DAFNet [11], and EMFINet [23]. For a fair comparison, we use the saliency maps provided by the available public RSI-SOD benchmarks [10], [11] and/or by the authors. Specifically, we retrain seven recent CNN-based NSI-SOD methods, *i.e.*, GCPA, ITSD, MINet, GateNet, U2Net, SUCA, and PA-KRN, on EORSSD and ORSSD datasets with their default settings using the same training data as our method.

1) Quantitative Comparison: In Table I, we report the quantitative comparison results of our method and all compared methods on two RSI-SOD datasets in terms of S_α, $F_\beta^{\max}$, F_β^{mean}, F_β^{adp}, $E_\xi^{\max}$, E_ξ^{mean}, E_ξ^{adp}, and $\mathcal{M}$, among which the higher the first seven evaluation metrics, the better, while the last one is opposite to them.

Overall, our method shows excellent performance as compared with all four categories of methods on EORSSD and ORSSD. Specifically, on the EORSSD dataset, our method is weaker than DAFNet in terms of $F_\beta^{\max}$ and $\mathcal{M}$, but is significantly better than DAFNet on the other six metrics, *e.g.*, F_β^{adp}: 0.8137 (Ours) versus 0.6427 (DAFNet), and E_ξ^{adp}: 0.9538 (Ours) versus 0.8446 (DAFNet). While on the ORSSD dataset, our method fully surpasses DAFNet in all eight metrics, *e.g.*, S_α: 0.9437 (Ours) versus 0.9191 (DAFNet), and $\mathcal{M}$: 0.0087 (Ours) versus 0.0113 (DAFNet). Compared with the latest CNN-based RSI-SOD method EMFINet, our method consistently outperforms it on both datasets, *e.g.*, $F_\beta^{\max}$: 2.11% and 1.70% better than it and $\mathcal{M}$: 21.43% and 20.18% lower than it on EORSSD and ORSSD, respectively. In comparison to the eight traditional methods, including NSI- and RSI-SOD, our method is a lot ahead of them. Even though the CNN-based NSI-SOD methods are retrained on optical RSI data, their performance is generally lower than that of specialized RSI-SOD methods, which illustrates the urgency and necessity of proposing specialized solutions. In addition, we present PR curves in Fig. 5. We observe that the specialized CNN-based methods show great strength, and the PR curve of our method is closer to the upper right corner than all compared methods, which is consistent with the remarkable quantitative results of our method on two datasets in Table I.

2) Computational Complexity Comparison: We measure the computational complexity from three aspects, including inference speed (without I/O time), network parameters and FLOPs, which are captured from the available public RSI-SOD benchmarks [10], [11] and our retraining, and report them in Table I. Overall, we can find that most CNN-based methods run in real time (23–25 frames/s), especially our method reaches an astonishing inference speed of 95 frames/s, which is friendly to practical applications. The network parameters and FLOPs of our method are at the midstream level. However, compared with the second best method EMFINet, our network parameters and FLOPs are much smaller, *e.g.*, #Param: 67.65M (Ours) versus 107.26M (EMFINet), and FLOPs: 112.8M (Ours) versus 480.9M (EMFINet). From the above quantitative comparison and computational complexity comparison, we can conclude that our method is effective and efficient.

3) Visual Comparison: We show some representative visual examples of optical RSIs in Fig. 6, including airplane, car, building, river, and pool. We summarize the five specific scenes as follows: 1) three cases of airplane: general airplane, airplane with shadows, and airplane with interferences; 2) three cases of car: multiple cars, multiple tiny cars, and cars with complex background; 3) three cases of building: general building, multiple buildings, and building with inconsistent colors; 4) two cases of river: river with irregular topology and river with low contrast; and 5) two cases of pool: pool with complex geometry and pool with interferences. The above cases include challenging scenes of optical RSIs, such as multiple objects, tiny objects, object with interferences, object with shadows, complex background, low contrast, and complex topology. We present the saliency maps of the representative methods in four categories, including EMFINet, DAFNet, LVNet, PA-KRN, U2Net, SMFF, and SMD.

Obviously, the two traditional methods SMFF and SMD are often confused by the unique scenes of optical RSIs. The two recent CNN-based NSI-SOD methods PA-KRN and U2Net often show weak inadaptability to optical RSIs data. The three CNN-based RSI-SOD methods can overcome difficult scenes to a certain extent, but generate saliency maps with flaws. Using the complementarity of four kinds of content, *i.e.*, foreground, edge, background, and global image-level content, our method can accurately locate salient objects and outline fine details, which shows strong adaptability and robustness in these scenes.

C. Ablation Studies

Here, we conduct comprehensive experiments to evaluate the effectiveness of important components of our MCCNet on EORSSD and ORSSD datasets. In particular, we investigate 1) the individual contribution of each content in MCCM; 2) the necessity of merging the original content in MCCM; and 3) the effectiveness of our comprehensive loss function.

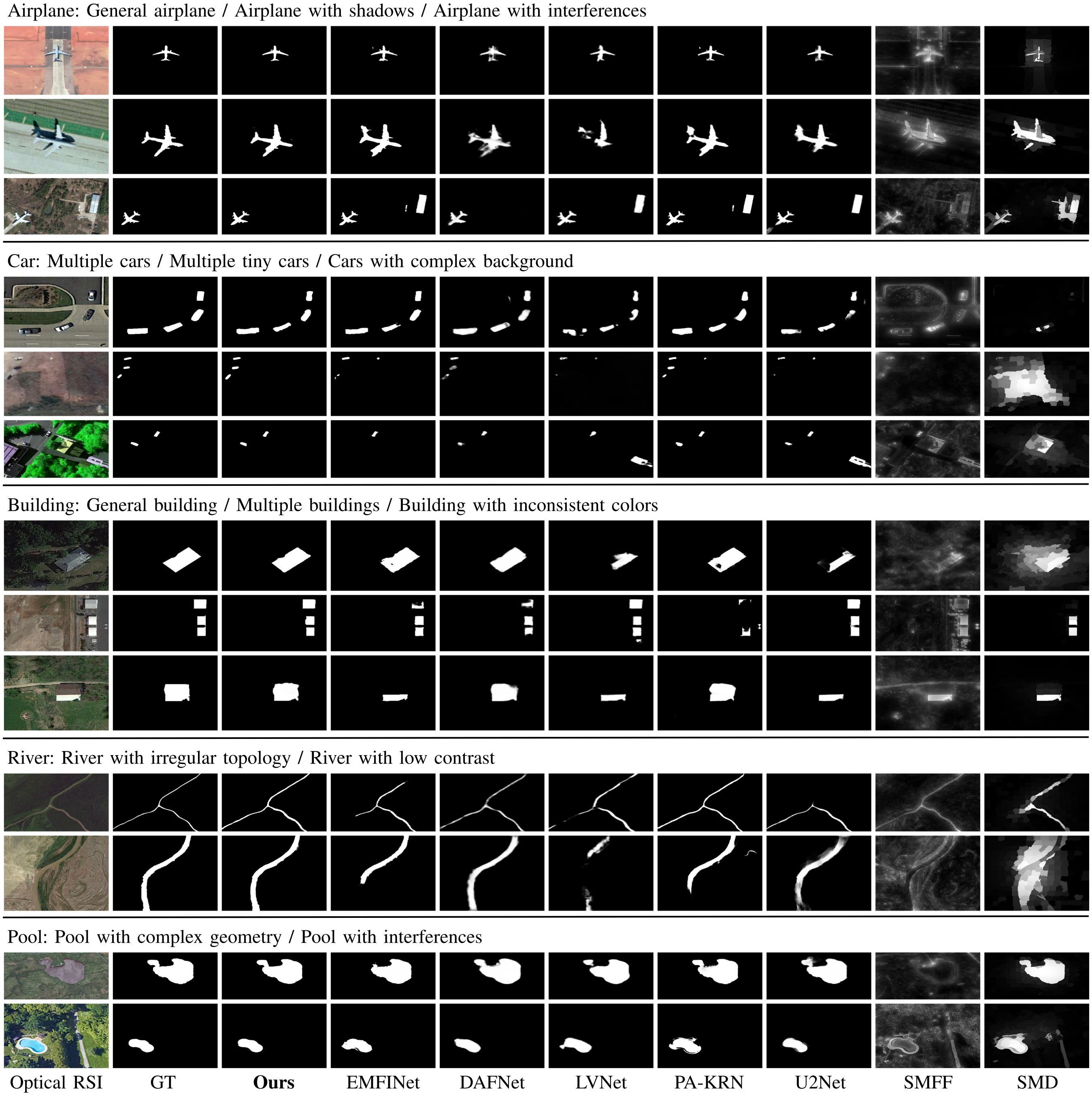

Airplane: General airplane / Airplane with shadows / Airplane with interferences
Car: Multiple cars / Multiple tiny cars / Cars with complex background
Building: General building / Multiple buildings / Building with inconsistent colors
River: River with irregular topology / River with low contrast
Pool: Pool with complex geometry / Pool with interferences
Optical RSI
GT
Ours
EMFINet
DAFNet
LVNet
PA-KRN
U2Net
SMFF
SMD

Fig. 6. Visual comparisons with seven representative state-of-the-art methods, including three CNN-based RSI-SOD methods (EMFINet [23], DAFNet [11] and LVNet [10]), two CNN-based NSI-SOD methods (PA-KRN [39] and U2Net [35]), one traditional RSI-SOD method (SMFF [71]), and one traditional NSI-SOD method (SMD [33]), on various scenes. Please zoom-in for the best view.

For each variant experiment, we rigorously retrain it with the same parameter settings and datasets as in Section IV-A.

1) Individual Contribution of Each Content in MCCM: To evaluate the individual contribution of each content, *i.e.*, foreground (FG), edge (EG), background (BG), and global image-level content (GIC), in MCCM, we first provide four progressive variants of MCCM on the upper part of Table II: 1) Baseline, which is the encoder-decoder network with skip connections (*i.e.*, replacing MCCM with skip connection in MCCNet); 2) Baseline + FG; 3) Baseline + FG + EG; and 4) Baseline + FG + EG + BG. For an intuitive understanding of variants, we illustrate three variants (No. 2–No. 4) in Fig. 7.

Based on the quantitative results in Table II, we observe consistently the increasing trends of performance on both datasets. FG greatly activates the potential of the network, and based on it, by additionally exploring the complementation of EG and BG, the performance takes off further. Notably, although GIC only provides the specific basic tone of source features, as expressed in Section III-B and Fig. 4, it increases

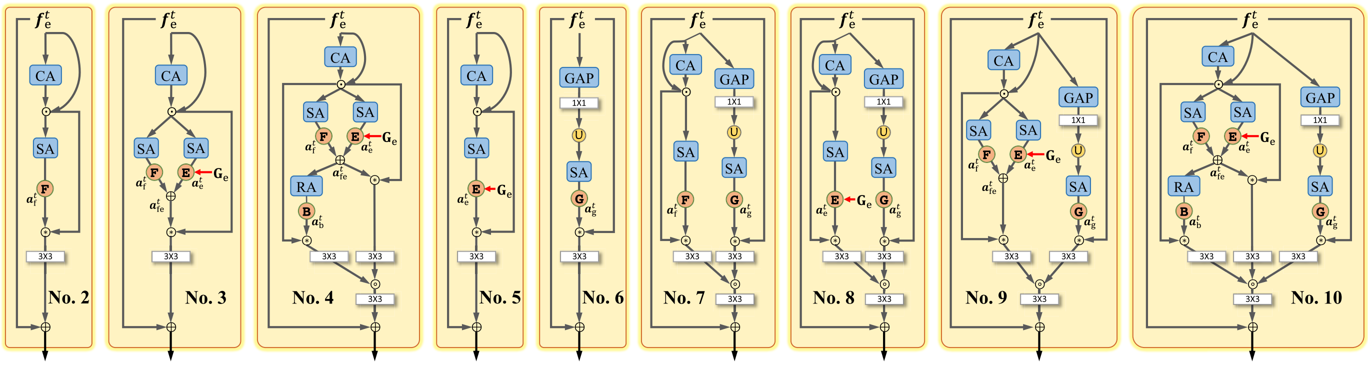


Fig. 7. Structures of eight MCCM variants (*i.e.*, No. 2–No. 9) and the full MCCM (*i.e.*, No. 10). Please zoom-in for viewing details.

TABLE II

ABLATION STUDY ON EVALUATING THE INDIVIDUAL CONTRIBUTION OF EACH CONTENT IN MCCM. BASELINE IS THE ENCODER–DECODER NETWORK WITH SKIP CONNECTIONS. FG, EG, BG, AND GIC MEAN FOREGROUND, EDGE, BACKGROUND, AND GLOBAL IMAGE-LEVEL CONTENT, RESPECTIVELY. THE BEST RESULT IN EACH COLUMN IS **BOLD**

No.	Baseline	FG	EG	BG	GIC	EORSSD [11]		ORSSD [10]	
						$F_\beta^{\max}$ ↑	$E_\xi^{\max}$ ↑	$F_\beta^{\max}$ ↑	$E_\xi^{\max}$ ↑
1	✓					.8687	.9585	.8878	.9608
2	✓	✓				.8829	.9703	.9043	.9711
3	✓	✓	✓			.8856	.9727	.9066	.9742
4	✓	✓	✓	✓		.8870	.9747	.9096	.9774
5	✓		✓			.8842	.9713	.9047	.9718
6	✓				✓	.8740	.9623	.9014	.9689
7	✓	✓			✓	.8855	.9728	.9116	.9755
8	✓		✓		✓	.8864	.9736	.9114	.9746
9	✓	✓	✓		✓	.8896	.9747	.9127	.9776
10	✓	✓	✓	✓	✓	**.8904**	**.9755**	**.9155**	**.9800**

$F_\beta^{\max}$ from 0.8879 to 0.8904 on the EORSSD dataset and from 0.9096 to 0.9155 on the ORSSD dataset. In summary, our full MCCM improves "Baseline" by 2.17% and 1.70% on $F_\beta^{\max}$ and $E_\xi^{\max}$, respectively, on the EORSSD dataset. The performance improvement is more significant on the ORSSD dataset, that is, our full MCCM improves "Baseline" by 2.77% and 1.92% on $F_\beta^{\max}$ and $E_\xi^{\max}$, respectively.

Moreover, to comprehensively analyze the different combinations of four kinds of content, we provide another five variants of MCCM on the middle part of Table II: 5) Baseline + EG; 6) Baseline + GIC; 7) Baseline + FG + GIC; 8) Baseline + EG + GIC; and 9) Baseline + FG + EG + GIC. We also illustrate these five variants (No. 5–No. 9) in Fig. 7. Notably, since BG is jointly defined by FG and EG according to Eq. (5), BG cannot be obtained by the single FG or the single EG.

By comparing the fifth and sixth variants with the first one, we can clearly observe the contributions of EG and GIC. By comparing the seventh and eighth variants with the sixth one, we can find that based on GIC, FE or EG can further promote the performance. When we abandon BG in MCCM, the performance of "Baseline + FG + EG + GIC" is adversely affected. Through the comprehensive comparison of all nine variants and our complete MCCM, we can conclude that each content in MCCM contributes to the final excellent performance and the proposed MCCM is effective.

TABLE III

ABLATION STUDY ON PROVING THE NECESSITY OF MERGING THE ORIGINAL CONTENT IN MCCM. THE BEST RESULT IN EACH COLUMN IS **BOLD**

Models	EORSSD [11]		ORSSD [10]	
	$F_\beta^{\max}$ ↑	$E_\xi^{\max}$ ↑	$F_\beta^{\max}$ ↑	$E_\xi^{\max}$ ↑
w/o original content	.8859	.9729	.9120	.9763
w/ original content (**Ours**)	**.8904**	**.9755**	**.9155**	**.9800**

2) Necessity of Merging the Original Content in MCCM: To prove the necessity of merging the original content in MCCM, we provide a variant that removes the original content in (9), *i.e.*, *w/o original content*. As summarized in Table III, we can observe that the performance degradation occurs in *w/o original content*, *e.g.*, $F_\beta^{\max}$ and $E_\xi^{\max}$ are reduced by 0.0045 and 0.0026 on the EORSSD dataset, and 0.0035 and 0.0037 on the ORSSD dataset. This demonstrates that the basic information of salient regions provided by the original content is necessary for a better performance.

3) Effectiveness of Our Comprehensive Loss Function: To validate the effectiveness of our comprehensive loss function, we provide three loss variants for network training: 1) the single BCE loss; 2) BCE loss with IoU loss; and 3) BCE loss with F-m loss. We report the quantitative results in Table IV.

Our MCCNet trained with the single BCE loss achieves acceptable performance, *e.g.*, $F_\beta^{\max}$: 0.8747 and $E_\xi^{\max}$: 0.9705 on the EORSSD dataset, and $F_\beta^{\max}$: 0.9094 and $E_\xi^{\max}$: 0.9728 on the ORSSD dataset. And with the assistance of IoU loss or F-m loss, the performance is further improved by about 0.0018–0.0127 in $F_\beta^{\max}$ and 0.0018–0.0038 in $E_\xi^{\max}$. Integrating the three losses together to train our MCCNet, our MCCNet achieves the best performance, which increases the simplest variant by 0.0158 in $F_\beta^{\max}$ on the EORSSD dataset and 0.0072 in $E_\xi^{\max}$ on the ORSSD dataset. In general, the

TABLE IV

ABLATION STUDY ON EVALUATING THE EFFECTIVENESS OF OUR COMPREHENSIVE LOSS FUNCTION. BCE, IOU AND F-M REPRESENT BCE LOSS, IOU LOSS AND F-MEASURE LOSS, RESPECTIVELY. THE BEST RESULT IN EACH COLUMN IS **BOLD**

No.	BCE	IoU	F-m	EORSSD [11]		ORSSD [10]	
				$F_{\beta}^{\max} \uparrow$	$E_{\xi}^{\max} \uparrow$	$F_{\beta}^{\max} \uparrow$	$E_{\xi}^{\max} \uparrow$
1	✓			.8746	.9705	.9094	.9728
2	✓	✓		.8873	.9726	.9136	.9766
3	✓		✓	.8857	.9723	.9112	.9760
4	✓	✓	✓	**.8904**	**.9755**	**.9155**	**.9800**

above analysis clearly verifies the effectiveness of our comprehensive loss function, and training network with suitable losses is efficient to boost our method without additional parameters.

V. CONCLUSION

In this article, we propose an effective MCCM to model the complementarity of multiple content, including foreground, edge, background, and global image-level content, for optical RSI data. In MCCM, the foreground map and edge map are directly integrated to complete salient regions, and then each content complements others in an adaptive manner. Moreover, we equip MCCM to the encoder–decoder network, and propose a full solution, namely MCCNet, for RSI-SOD. MCCMs on five feature sizes can highlight salient regions well, and successfully address the variable object scales/types/quantities of optical RSI data. Finally, a comprehensive loss function is employed in the training phase to boost performance. We conduct extensive experiments on two public RSI-SOD datasets. The experimental results demonstrate the superiority of the proposed MCCNet as well as the effectiveness of the proposed MCCM. Moreover, the fast inference speed of 95 frames/s is extremely conducive to applying our MCCNet in practical applications.

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] 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.*, early access, Jan. 12, 2021, doi: 10.1109/TPAMI.2021.3051099.

[3] 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.

[4] 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.

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

[6] G. Li *et al.*, "Personal fixations-based object segmentation with object localization and boundary preservation," *IEEE Trans. Image Process.*, vol. 30, pp. 1461–1475, 2021.

[7] K. Gu *et al.*, "Saliency-guided quality assessment of screen content images," *IEEE Trans. Multimedia*, vol. 18, no. 6, pp. 1098–1110, Jun. 2016.

[8] 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. 27th ACM Int. Conf. Multimedia*, Oct. 2019, pp. 1383–1391.

[9] Y. Fang, Z. Chen, W. Lin, and C.-W. Lin, "Saliency detection in the compressed domain for adaptive image retargeting," *IEEE Trans. Image Process.*, vol. 21, no. 9, pp. 3888–3901, Sep. 2012.

[10] 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.

[11] 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.

[12] R. Cong *et al.*, "RRNet: Relational reasoning network with parallel multi-scale attention for salient object detection in optical remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, early access, Nov. 1, 2021, doi: 10.1109/TGRS.2021.3123984.

[13] D. Hong *et al.*, "Interpretable hyperspectral artificial intelligence: When nonconvex modeling meets hyperspectral remote sensing," *IEEE Geosci. Remote Sens. Mag.*, vol. 9, no. 2, pp. 52–87, Jun. 2021.

[14] Y. LeCun *et al.*, "Backpropagation applied to handwritten zip code recognition," *Neural Comput.*, vol. 1, no. 4, pp. 541–551, Dec. 1989.

[15] 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.

[16] J.-X. Zhao, J.-J. Liu, D.-P. Fan, Y. Cao, J. Yang, and M.-M. Cheng, "EGNet: Edge guidance network for salient object detection," in *Proc. IEEE ICCV*, Oct. 2019, pp. 8779–8788.

[17] X. Wang, H. Ma, X. Chen, and S. You, "Edge preserving and multi-scale contextual neural network for salient object detection," *IEEE Trans. Image Process.*, vol. 27, no. 1, pp. 121–134, Jan. 2018.

[18] N. Liu, J. Han, and M.-H. Yang, "PiCANet: Pixel-wise contextual attention learning for accurate saliency detection," *IEEE Trans. Image Process.*, vol. 29, pp. 6438–6451, 2020.

[19] 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.

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

[21] K. Zhao, S. Gao, W. Wang, and M.-M. Cheng, "Optimizing the F-measure for threshold-free salient object detection," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 8849–8857.

[22] C. Li *et al.*, "A parallel down-up fusion network for salient object detection in optical remote sensing images," *Neurocomputing*, vol. 415, pp. 120–411, Nov. 2020.

[23] 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.*, early access, Jul. 5, 2021, doi: 10.1109/TGRS.2021.3091312.

[24] L.-C. Chen, G. Papandreou, F. Schroff, and H. Adam, "Rethinking atrous convolution for semantic image segmentation," 2017, *arXiv:1706.05587*.

[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] Z. Liu, R. Shi, L. Shen, Y. Xue, K. N. Ngan, and Z. Zhang, "Unsupervised salient object segmentation based on kernel density estimation and two-phase graph cut," *IEEE Trans. Multimedia*, vol. 14, no. 4, pp. 1275–1289, Aug. 2012.

[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 Conf. Comput. Vis. Pattern Recognit. (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] W. Zou, Z. Liu, K. Kpalma, J. Ronsin, Y. Zhao, and N. Komodakis, "Unsupervised joint salient region detection and object segmentation," *IEEE Trans. Image Process.*, vol. 24, no. 11, pp. 3858–3873, Nov. 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] Q. Hou, M.-M. Cheng, X. Hu, A. Borji, Z. Tu, and P. H. Torr, "Deeply supervised salient object detection with short connections," in *Proc. Comput. Vis. Pattern Recognit.*, Jul. 2017, pp. 5300–5309.

[35] X. Qin, Z. Zhang, C. Huang, M. Dehghan, O. R. Zaiane, and M. Jagersand, "U^2-net: Going deeper with nested U-structure for salient object detection," *Pattern Recognit.*, vol. 106, Oct. 2020, Art. no. 107404.

[36] 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. 10599–10606.

[37] 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/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 9138–9147.

[38] J. Li, Z. Pan, Q. Liu, and Z. Wang, "Stacked U-shape network with channel-wise attention for salient object detection," *IEEE Trans. Multimedia*, vol. 23, pp. 1397–1409, 2021.

[39] B. Xu, H. Liang, R. Liang, and P. Chen, "Locate globally, segment locally: A progressive architecture with knowledge review network for salient object detection," in *Proc. AAAI*, Feb. 2021, pp. 3004–3012.

[40] Y. Pang, X. Zhao, L. Zhang, and H. Lu, "Multi-scale interactive network for salient object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 9410–9419.

[41] 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.

[42] Z. Wu, L. Su, and Q. Huang, "Decomposition and completion network for salient object detection," *IEEE Trans. Image Process.*, vol. 30, pp. 6226–6239, 2021.

[43] M. Ma, C. Xia, and J. Li, "Pyramidal feature shrinking for salient object detection," in *Proc. AAAI*, Feb. 2021, pp. 2311–2318.

[44] 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.

[45] 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.*, Jun. 2019, pp. 3912–3921.

[46] 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.

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

[48] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. 3rd Int. Conf. Learn. Represent. (ICLR)*, 2015, pp. 1–14.

[49] D. Hong, L. Gao, N. Yokoya, J. Yao, J. Chanussot, Q. Du, and B. Zhang, "More diverse means better: Multimodal deep learning meets remote-sensing imagery classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 5, pp. 4340–4354, May 2021.

[50] D. Hong, L. Gao, J. Yao, B. Zhang, A. Plaza, and J. Chanussot, "Graph convolutional networks for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 7, pp. 5966–5978, Jul. 2021.

[51] D. Hong, N. Yokoya, G.-S. Xia, J. Chanussot, and X. X. Zhu, "X-ModalNet: A semi-supervised deep cross-modal network for classification of remote sensing data," *ISPRS J. Photogramm. Remote Sens.*, vol. 167, pp. 12–23, Sep. 2020.

[52] D. Hong, N. Yokoya, J. Chanussot, and X. X. Zhu, "An augmented linear mixing model to address spectral variability for hyperspectral unmixing," *IEEE Trans. Image Process.*, vol. 28, no. 4, pp. 1923–1938, Apr. 2019.

[53] 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.

[54] Q. Li, L. Mou, Q. Liu, Y. Wang, and X. X. Zhu, "HSF-Net: Multiscale deep feature embedding for ship detection in optical remote sensing imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 12, pp. 7147–7161, Dec. 2018.

[55] 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.

[56] H. Chen, T. Gao, W. Chen, Y. Zhang, and J. Zhao, "Contour refinement and EG-GHT-based inshore ship detection in optical remote sensing image," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 11, pp. 8458–8478, Nov. 2019.

[57] 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.

[58] 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.

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

[60] 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.

[61] D. Peng, H. Guan, Y. Zang, and L. Bruzzone, "Full-level domain adaptation for building extraction in very-high-resolution optical remote-sensing images," *IEEE Trans. Geosci. Remote Sens.*, early access, Jul. 13, 2021, doi: 10.1109/TGRS.2021.3093004.

[62] P. Zhu *et al.*, "Detection and tracking meet drones challenge," 2020, *arXiv:2001.06303*.

[63] 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.

[64] 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.

[65] 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.

[66] 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.

[67] 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.

[68] 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, Sep. 2015.

[69] L. Zhang, S. Wang, and X. Li, "Salient region detection in remote sensing images based on color information content," in *Proc. IEEE Int. Geosci. Remote Sens. Symp. (IGARSS)*, Jul. 2015, pp. 1877–1880.

[70] 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. 25th IEEE Int. Conf. Image Process. (ICIP)*, Oct. 2018, pp. 2336–2340.

[71] 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.

[72] L. Zhang and J. Ma, "Salient object detection based on progressively supervised learning for remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 11, pp. 9682–9696, Nov. 2021, doi: 10.1109/TGRS.2020.3045708.

[73] 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.

[74] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. Med. Image Comput. Comput.-Assisted Intervent.*, Oct. 2015, pp. 234–241.

[75] 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, 2020.

[76] 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.

[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. Xie and Z. Tu, "Holistically-nested edge detection," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2015, pp. 1395–1403.

[79] 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, 2021.

[80] A. Paszke *et al.*, "PyTorch: An imperative style, high-performance deep learning library," in *Proc. NeurIPS*, Dec. 2019, pp. 8024–8035.

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

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

[83] D.-P. Fan, M.-M. Cheng, Y. Liu, T. Li, and A. Borji, "Structure-measure: A new way to evaluate foreground maps," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 4548–4557.

[84] 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. 27th Int. Conf. Artif. Intell.*, Jul. 2018, pp. 698–704.

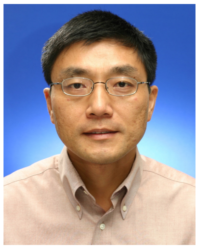
Gongyang Li received the B.E. degree from Shanghai Normal University, Shanghai, China, in 2016. He is currently pursuing the Ph.D. degree with the School of Communication and Information Engineering, Shanghai University, Shanghai.

His research interests include image/video object segmentation and saliency detection.

Zhi Liu (Senior Member, IEEE) received the B.E. and M.E. degrees from Tianjin University, Tianjin, China, in 1999 and 2002, respectively, and the Ph.D. degree from the Institute of Image Processing and Pattern Recognition, Shanghai Jiao Tong University, Shanghai, China, in 2005.

He is currently a Professor with the School of Communication and Information Engineering, Shanghai University, Shanghai. From August 2012 to August 2014, he was a Visiting Researcher with the SIROCCO Team, IRISA/INRIA-Rennes, France, with the support by EU FP7 Marie Curie Actions. He has authored or coauthored more than 200 refereed technical articles in international journals and conferences. His research interests include image/video processing, machine learning, computer vision, and multimedia communication.

Dr. Liu was a TPC Member/Session Chair in ICIP 2017, PCM 2016, VCIP 2016, ICME 2014, WIAMIS 2013. He co-organized special sessions on visual attention, saliency models, and applications at WIAMIS 2013 and ICME 2014. He is an Area Editor of *Signal Processing: Image Communication* and served as a Guest Editor for the special issue on *Recent Advances in Saliency Models, Applications, and Evaluations* in *Signal Processing: Image Communication*.



Weisi Lin (Fellow, IEEE) received the Ph.D. degree from King's College London, London, U.K., in 1993.

He is currently a Professor with the School of Computer Science and Engineering, Nanyang Technological University, Singapore. His research interests include image processing, visual quality evaluation, and perception-inspired signal modeling, with more than 340 refereed articles published in international journals and conferences.

Dr. Lin is a fellow of the Institution of Engineering Technology and an Honorary Fellow of the Singapore Institute of Engineering Technologists. He has been on the Editorial Board of the IEEE TRANSACTIONS ON IMAGE PROCESSING, the IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY, the IEEE TRANSACTIONS ON MULTIMEDIA, the IEEE SIGNAL PROCESSING LETTERS, and *Journal of Visual Communication and Image Representation*. He has been elected as APSIPA (2012/2013) Distinguished Lecturer. He was a Technical-Program Chair for Pacific-Rim Conference on Multimedia 2012, the IEEE International Conference on Multimedia and Expo 2013, and the International Workshop on Quality of Multimedia Experience 2014.

Haibin Ling (Senior Member, IEEE) received the B.S. and M.S. degrees from Peking University in 1997 and 2000, respectively, and the Ph.D. degree from the University of Maryland, College Park, in 2006.

From 2000 to 2001, he was an Assistant Researcher with Microsoft Research Asia. From 2006 to 2007, he worked as a Post-Doctoral Scientist with the University of California at Los Angeles, Los Angeles, CA, USA. In 2007, he joined Siemens Corporate Research as a Research Scientist; then, from 2008 to 2019, he worked as a Faculty Member of the Department of Computer Sciences at Temple University. In fall 2019, he joined Stony Brook University as a SUNY Empire Innovation Professor with the Department of Computer Science. His research interests include computer vision, augmented reality, medical image analysis, and human computer interaction.

Dr. Ling received the Best Student Paper Award at ACM UIST in 2003, the NSF CAREER Award in 2014, the Yahoo Faculty Research and Engagement Program Award in 2019, and the Amazon AWS Machine Learning Research Award in 2019. He serves as an Associate Editor for several journals including the IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE (PAMI), *Pattern Recognition (PR)*, and *Computer Vision and Image Understanding (CVIU)*. He has served as Area Chairs for CVPR (2014, 2016, 2019, and 2020) and ECCV (2020).