Lightweight Progressive Multilevel Feature Collaborative Network for Remote Sensing Image Salient Object Detection

Bei Cheng, *Member, IEEE*, Zao Liu, Qingwang Wang, *Member, IEEE*, Tao Shen, *Senior Member, IEEE*, Chengbiao Fu, *Member, IEEE*, and Anhong Tian, *Member, IEEE*

***Abstract*— In recent years, numerous outstanding technologies have been proposed for salient object detection (SOD) in remote sensing images (RSIs), but most of them focus solely on improving performance while disregarding computational, thereby lacking portability and mobility. This article introduces a novel lightweight progressive multilevel feature collaborative network, termed LPMFCNet. This framework constructs progressive feature information through multilevel image content extraction and designs a multichannel interactive deep neural network with information fusion and filtering functions. First, a spatial detail enhancement module (SDEM) is devised to acquire distant feature information through intermediate branch expansion of receptive fields while preserving multiscale information extraction. Second, an advanced semantic interaction module (ASIM) is proposed to model distant dependency relationships between deep semantic features to better identify the positional information of salient objects. Finally, a multilevel feature collaboration module (MFCM) is designed to collaboratively utilize target features from a multilevel perspective, which fully mining deep-level semantic positional information while retaining target detail information. Extensive experimental comparisons are conducted on two remote sensing datasets with 17 advanced methods. Results demonstrate that the proposed method exhibits superior detection performance while maintaining lightweightness. The LPMFCNet only contains 3.26M parameters and runs 0.5G FLOPs for a 256 × 256 image.**

***Index Terms*— Advanced semantics, detail enhancement, lightweight salient object detection (SOD), multilevel collaboration, optical remote sensing image (RSI).**

Received 22 May 2024; revised 27 August 2024; accepted 23 October 2024. Date of publication 28 October 2024; date of current version 11 November 2024. This work was supported in part by Yunnan Province Fundamental Research Projects under Grant 202401AU070154, in part by the Open Fund of Yunnan Key Laboratory of Computer Technologies Application, and in part by the National Natural Science Foundation of China under Grant 42067029. *(Corresponding author: Tao Shen.)*

Bei Cheng is with the Faculty of Information Engineering and Automation, Kunming University of Science and Technology, Kunming 650500, China, and also with Yunnan Key Laboratory of Computer Technologies Application, Kunming 650500, China (e-mail: chengbei@kust.edu.cn).

Zao Liu, Qingwang Wang, Tao Shen, and Chengbiao Fu are with the Faculty of Information Engineering and Automation, Kunming University of Science and Technology, Kunming 650500, China (e-mail: liuzao@stu.kust.edu.cn; wangqingwang@kust.edu.cn; shentao@kust.edu.cn; fcb@kust.edu.cn).

Anhong Tian is with the Faculty of Information Engineering and Automation, Kunming University of Science and Technology, Kunming 650500, China, and also with the Faculty of Land Resource Engineering, Kunming University of Science and Technology, Kunming 650093, China (e-mail: tah@kust.edu.cn).

Digital Object Identifier 10.1109/TGRS.2024.3487244

I. Introduction

As an important image preprocessing step, salient object detection (SOD) has been successfully used in many image processing fields, such as object segmentation [1], object tracking [2], and image quality assessment [3]. SOD aims to identify the most visually appealing regions within an image and output them as pixel-level saliency maps. The rapid advancement of deep neural networks has facilitated notable achievements in SOD for natural scene images (NSI). Concurrently, with the enhancement of remote sensing image (RSI) acquisition capabilities, researchers have turned their attention to SOD in RSI. This attention stems from its potential to enhance information processing capabilities in diverse remote sensing tasks such as object detection [4], image fusion [5], and scene classification [6]. Different from NSI, RSI is typically captured from high altitudes, exhibiting significant scale variations of objects, larger image dimensions, and more complex background information. Directly applying SOD methods designed for NSI to RSI often yields unsatisfactory results. Consequently, achieving effective SOD for RSI poses a challenging task that warrants investigation.

In recent years, researchers have proposed numerous SOD methods based on deep learning, particularly in NSI, resulting in significant enhancements in detection accuracy [8], [9], [10], [11], [12], [13]. Optical RSI has been widely used in land planning [14], disaster monitoring [15], military reconnaissance [16], and other fields due to its large coverage, high resolution, rich content, and intuitive characteristics. Therefore, various SOD algorithms specifically for optical RSI have been developed. Among these, the encoder–decoder structure is widely employed, with additional ingenious strategies integrated to achieve improved performance. For instance, MFENet [17] simultaneously extracts global and semantic feature information, constructing a multiscale feature-enhanced RSI-SOD network, which can well generalize different scenes and object categories. AESINet [18] utilizes a differential perception mechanism to adaptively enhance the edge information of salient objects, while designing a multiscale feature module and a deep semantic interaction module to effectively capture multiscale information and mitigate the interference of shadow. MCCNet [19] makes full use of the foreground features, edge features, background features, and

1558-0644 © 2024 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission.
See https://www.ieee.org/publications/rights/index.html for more information.

Authorized licensed use limited to: Kunming Univ of Science and Tech. Downloaded on November 22,2024 at 07:21:10 UTC from IEEE Xplore. Restrictions apply.

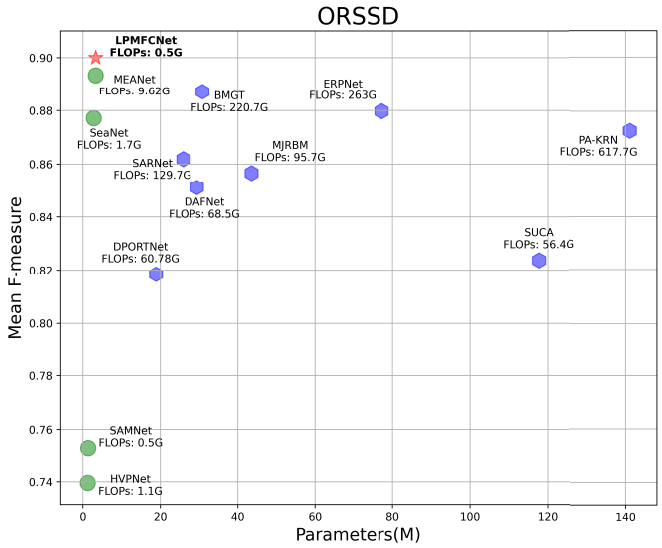

Fig. 1. Proposed method was compared to other methods using the ORSSD [7] dataset in terms of mean F-measure, FLOPs, and parameters. The red star represents the proposed method, while the green circles and blue hexagons represent lightweight and nonlightweight methods, respectively.

global image-level features of RSI, exploiting their complementary content to highlight prominent areas. ACCoNet [20] introduces a novel adjacent context coordination network to explore the coordination of adjacent features in the RSI-SOD encoder–decoder architecture, fully exploiting contextual information in RSI. RRNet [21] proposes a relational reasoning network with a parallel multiscale attention module to perform semantic relationship inference on high-level features, thereby promoting the generation of more complete detection results.

However, although existing methods have made certain progress in RSI-SOD, there are still some challenges in effectiveness. First, the above methods often have large amounts of parameters and calculations, making them unsuitable for deployment on mobile terminals. Fig. 1 plots some comparisons in terms of parameters and F-measure. As depicted, the majority of methods possess large parameter sizes, rendering them impractical for deployment on mobile devices. In actual scenarios, some applications have real-time requirements based on changes in user input, wherein lightweight models can exhibit faster response times when executed on mobile terminals. To address this issue, lightweight SOD methods have gradually emerged, such as CorrNet [22] and SeaNet [23], which significantly reduce parameter counts but still require improvement in detection performance. Therefore, this article proposes parameter counts but still requires improvement in detection performance. Second, the size of salient objects in RSI varies greatly, with object sizes ranging from a few pixels (e.g., cars) to hundreds (e.g., islands), resulting in the coexistence of multiple salient objects with significant scale disparities within the same scene. However, due to the shallowness and thinness of the backbone networks compared to conventional neural networks, lightweight backbones inevitably struggle to extract representative features. Consequently, when designing lightweight solutions for RSI-SOD, one of the challenges to be addressed is how to remedy the information loss within these shallow and thin backbones. Furthermore, since optical RSI captures scenes covering extensive areas, leading to cluttered backgrounds containing abundant background information besides salient objects. In this case, how to perceive the location of salient objects from complex scenes while maintaining the efficiency of the model is a significant challenge. Finally, since RSI significantly exhibits complex structures and topologies, this poses a new obstacle to accurately understanding the complete regions of salient objects while maintaining clear boundaries. Considering the different contributions of features at different levels to the SOD task, it is important to design multilevel feature fusion layers to fuse different levels of perceptual field information while simple addition or splicing may not obtain a satisfactory salience graph. Therefore, the fusion process should also consider parameter constraints and information compensation for lightweight backbones.

Driven by the above observations, this article proposes a novel lightweight progressive multilevel feature collaborative network (LPMFCNet), which aims to improve SOD performance while achieving lightweight. It consists of three key modules: spatial detail enhancement module (SDEM), advanced semantic interaction module (ASIM), and multilevel feature collaboration module (MFCM). As is known to all, multilevel learning can effectively utilize the complementary nature of features at different levels. We construct progressive feature information through multilevel image content extraction, designing a multichannel interactive deep neural network with information fusion and filtering functions. Specifically, SDEM expands the receiving field through the intermediate branch to obtain remote feature information while retaining the current information. It performs extended convolution on the two branches in parallel, which promotes the utilization of original features while mitigating the impact of losing local information, and strengthens the relevant details. The core of ASIM lies in long-range dependency capture based on nonlocal (NL) modules. By generating a rough saliency map, the guidance model can accurately identify the location of salient objects from the cluttered background. It utilizes depth-separable convolution blocks to extract deeper features while reducing the amount of computation. To ensure that the extracted target features encompass diverse target feature distributions, this article designs MFCM, which combines low-level features and deep-level features to jointly enhance mid-level features. It removes the influence of noise and nonimportant information through multilevel information fusion channels, and combines the two to achieve collaborative detection of targets.

The main contributions of this study can be summarized as follows.

1) This article conducts a hierarchical research on the feature extraction capability of MobileNet and proposes a novel LPMFCNet. This network improves the SOD accuracy in RSI while maintaining the model lightweightness, with a parameter count of 3.26M and 0.5G FLOPs for a 256×256 image.
2) The SDEM is proposed to integrate multiscale features from different branches, enhancing feature

Authorized licensed use limited to: Kunming Univ of Science and Tech. Downloaded on November 22,2024 at 07:21:10 UTC from IEEE Xplore. Restrictions apply.

representations of salient objects across various scales and quantities, thereby overcoming the limitations of traditional CNNs and expanded convolutions. Its mechanism involves extending receptive fields to capture distant feature information while preserving attentional cues.

3) The ASIM is proposed to model long-range dependencies between deep semantic features in optical RSI. This module generates rough saliency maps by designing continuous semantic connections across layers to guide high-level features in better identifying the positional information of salient objects.
4) The MFCM is proposed to collaboratively utilize target features from a multilevel perspective, enhancing intermediate features through the concurrent exploitation of low-level and deep-level features. Multilevel features are mapped to each other to capture diverse distributions of target features, thereby fully exploiting deep semantic positional information while retaining fine-grained texture details of the target.

II. Related Work

A. CNN Methods for SOD in NSI

Convolutional neural network (CNN) can automatically learn hierarchical aspects of images, enabling them to successfully differentiate salient objects from the background. Zhao et al. [24] demonstrated the efficacy of integrating edge information into the process of salient object recognition. By efficiently addressing the challenge of accurately defining the boundaries of objects, not only does the precision of saliency detection improve, but also provides crucial insight into the importance of edge information in visual perception. Chen et al. [25] incorporated a global context module to effectively capture global information in images. The module enhances the model's capacity for understanding the overall layout of scenes and the relationships between objects within them. In addition, it efficiently combines low-level appearance features, high-level semantic features, and global context features. This integration is a significant source of inspiration for our work. Zhou et al. [26] incorporated an interactive two-stream decoder that enables the examination of various cues, such as saliency and contour, and their significance. In addition, a novel adaptive contour loss method was developed to autonomously detect challenging samples during the learning process. The SUCA model, proposed by Li et al. [27], is a U-Net architecture that incorporates channel-wise attention through stacking. This method highlights the issue of feature mappings from different layers having varied roles in saliency detection. It also emphasizes that the significance of different feature channels within the same layer is not uniform when it comes to saliency. Xu et al. [28] utilized a knowledge review network to reintegrate the most effective feature mappings from all preceding layers. Lu et al. [29] introduced DCTNet for multimodal video SOD. This model captures the long-range dependencies of multimodal remote relationships, improving the feature representations of three modalities: RGB images, optical flow images, and depth maps. This mitigates the substantial loss of information along the top-down pathway and improves the precision of the final detection. While the approaches discussed above have shown impressive performance, their accuracy is obtained by sacrificing computing efficiency, and they are not well-suited for adapting to optical RSI.

B. CNN Methods for SOD in Optical RSI

SOD in optical RSI presents unique challenges compared to NSI, such as variable in object sizes, intricate backgrounds, and various imaging environments. In order to deal with these particular features, researchers have developed CNN structures that are customized to the unique attributes of RSI. Zhang et al. [30] introduced DAFNet, a complete framework specifically developed for identifying salient objects in optical RSI. The architecture incorporates a global context-aware attention module that dynamically perceives and utilizes global context information. It is crucial to effectively detect salient objects in RSI under complex backgrounds. Tu et al. [31] presented MJRBM, a method that integrates local cues generated from pyramid pooling with global information to create robust multiscale regional characteristics of salient objects and improve their borders. The architecture is designed to effectively integrate boundary and regional features, enhancing the ability to detect important objects in complex RSI. ERPNet, established by Zhou et al. [32], leverages edge information to guide the iterative localization of salient objects. This improves the model's capacity to precisely identify and locate objects, as well as to separate salient objects from intricate backgrounds typically encountered in RSI. Zeng et al. [18] invented ACCoNet, which includes a module for extracting features at multiple scales. This module can extract useful information from feature maps of the same level and different scales. In addition, they presented a more advanced approach to obtaining profound semantic information by utilizing graph convolutional networks. Luo and Liang [33] pioneered SEINet, which utilizes convolutional kernels with different shapes to efficiently handle the varied topology and directionality of salient objects. The described methodologies have yielded favorable outcomes in the analysis of RSI. Nevertheless, their considerable computational complexity makes them unsuitable for mobile devices. The restriction served as the impetus for the adoption of the streamlined methodology explored in this article.

C. CNN Methods for Object Detection and Image Segmentation in Optical RSI

The field of RSI analysis, a crucial modality for geographical information acquisition and environmental monitoring, encompasses various pivotal technologies, among which SOD, object detection, and image segmentation constitute the three fundamental tasks. Although they cater to distinct image analysis requirements, they frequently interconnect in practical applications, fostering a mutually reinforcing relationship. Specifically, SOD preempts by identifying potential regions for subsequent object detection and image segmentation, thereby curtailing computational costs. Conversely, object detection

Authorized licensed use limited to: Kunming Univ of Science and Tech. Downloaded on November 22,2024 at 07:21:10 UTC from IEEE Xplore. Restrictions apply.

and image segmentation endeavors to refine and analyze these salient objects, laying the groundwork for advanced geographical information analyses. In the realm of object detection, Li et al. [34] introduced dynamic adjustments to the extensive spatial receptive field within convolutional kernels, simulating the ranging contexts of diverse objects in remote sensing scenes with enhanced precision. On the other hand, Zhang et al. [35] presented an angle marker technology, facilitating the encoding, embedding, and learning of object-oriented orientational knowledge within remote sensing scenes. Their compact dual-path encoding architecture effectively captures both global long-range dependencies and local semantic relationships, thereby reinforcing inductive bias. Chen et al. [36] leveraged cue learning and SAM [37] in image segmentation, generating semantically discernible segmentation outcomes for RSIs. Meanwhile, Li et al. [38] crafted a synergistic attention module, adept at extracting channel affinities while preserving spatial nuances, and enhancing semantic segmentation capabilities for remote sensing imagery. This multitiered image analysis framework not only augments the efficiency and accuracy of RSI processing but also solidifies the technical foundation for intricate geospatial data interpretation.

D. Lightweight Methods for SOD

Although CNN methods have demonstrated impressive achievements in detecting salient objects, they frequently demand significant computational resources, which might be problematic for devices with limited capabilities. Therefore, there has been a rise in the development of lightweight SOD methods. Liu et al. [39] presented SAMNet, a model that incorporates a novel stereoscopic attention multiscale module. This module employs a stereoscopic attention method to dynamically combine features from various scales, efficiently capturing detailed information about salient objects while using little processing resources. In addition, Liu et al. [40] introduced HVPNet, which incorporate a hierarchical visual perception module to replicate the hierarchical perceptual learning observed in the visual cortex of primates. This module effectively learns contextual information at multiple scales by utilizing dense connections. Nevertheless, these methods fail to effectively incorporate the identified characteristics and are unable to precisely deduce the positional details of salient objects. The objective of this research is to develop streamlined and lightweight modules that are specifically designed for RSI.

Ensuring excellent performance while keeping a lightweight model is a challenging task due to the specific attributes of RSI. CorrNet, an innovative method proposed by Li et al. [22], aims to reduce the computational burden of the VGG-16 [41] backbone by constructing lightweight networks for feature extraction. This method seeks to address the memory and computational expenses that previous methods overlooked, while enhancing detection accuracy. Li et al. [23] presented SeaNet, a method that employs the lightweight MobileNet-V2 [42] for extracting features. SeaNet offers two lightweight modules: the dynamic semantic matching module for processing high-level features and the edge self-alignment module for processing low-level features. This method achieves effective identification and accurate positioning of salient objects in RSI. Liang and Luo [43] proposed MEANet to incorporates the multiscale edge-embedded attention module and the multilevel semantic guidance module, both of which are designed to have a low computational burden. Chai et al. [44] proposed a shallow CNN to achieve cloud detection in a more efficient way by removing the pooling/unpooling layers and normalization layers in the deep CNN and retaining only three convolutional layers. These techniques decrease the number of parameters and processing demands of the model while maintaining the same level of detection performance.

Building upon these studies, the current research focuses on exploring a streamlined and comprehensive network. The proposed LPMFCNet in this article differs from current methods by utilizing two adjacent and linked features to enhance refinement and extraction. This approach effectively integrates information and fully combines features from various levels. Its objective is to achieve an appropriate balance between the quantity of parameters and the level of performance.

III. Methodology

In this section, we will elaborate on the proposed lightweight architecture of LPMFCNet. In Section III-A, we depict the network overview. In Sections III-B–III-D, we introduce SDEM, ASIM, and MFCM, respectively. In Sections III-E and III-F, we present the decoder and loss function.

A. Network Overview

The proposed LPMFCNet is illustrated in Fig. 2, which comprises five primary components: the encoder, the SDEM, the ASIM, the MFCM, and the decoder. The network includes a variety of modules to extract and construct progressive feature information from multilevel image content.

The framework utilizes MobileNetV3 [45] as the encoder, specifically selecting MobileNetV3-Large as the backbone network for feature extraction. In particular, the first 15 bottleneck blocks are retained, but the last three convolutional layers and the final global average pooling layer are removed to make it suitable for remote sensing saliency detection. A residual connection links the bottleneck's input and output. In some residual connections, the squeeze-and-excitation [46] is applied. Nonlinear activation functions, such as ReLU and h-swish, are used depending on the layer index. ReLU is used in the first half of the network, while h-swish is used in the second half. The MobileNet V3-Large is divided into five blocks, namely, F_i, where $i \in \{1, 2, 3, 4, 5\}$. The output is denoted as $f_e^i \in \mathbb{R}^{c_i \times h_i \times w_i}$, where h_i and w_i represent $\left(256/2^i\right)$, and $c_i \in \{16,24,40,112,160\}$.

The SDEM is employed to extract extensive contextual features and multiscale feature information for the lower level features f_e^1 and f_e^2. This facilitates the perception of correlations between the two processing pathways and adapts to the varying sizes and quantities of salient objects. The output is denoted as f_{sdem}. The ASIM processes deep-level features f_e^4 and f_e^5 to capture semantic relations among them, obtaining

Authorized licensed use limited to: Kunming Univ of Science and Tech. Downloaded on November 22,2024 at 07:21:10 UTC from IEEE Xplore. Restrictions apply.

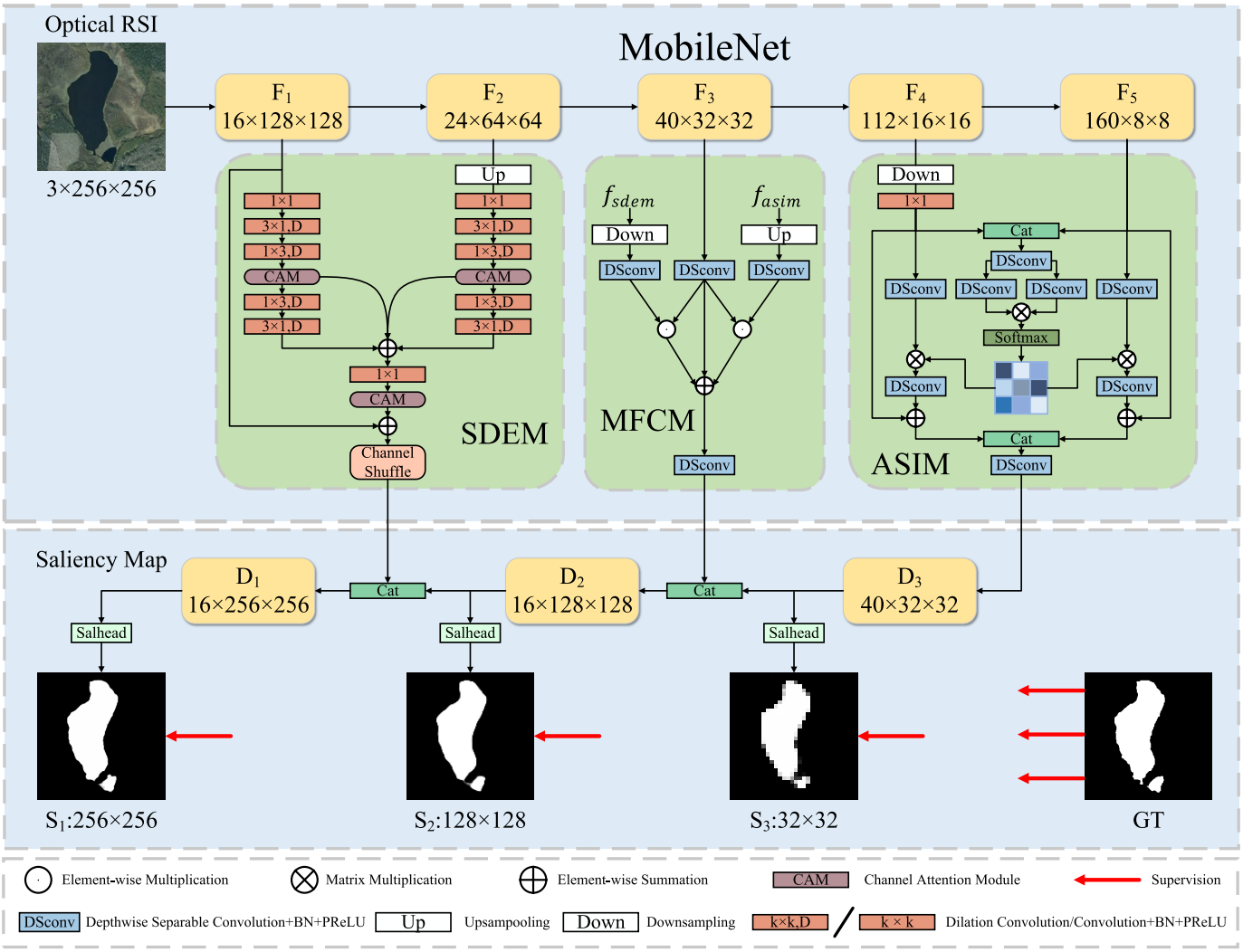

Fig. 2. Overall network structure. The article presents a network architecture based on the encoder–decoder structure, comprising five components: the encoder, the SDEM, the ASIM, the MFCM, and the decoder. The encoder extracts multilevel features, while the SDEM refines lower level features to improve the model's ability to capture details and multiscale objects. The ASIM focuses on enhancing the model's comprehension and representational capabilities by strengthening the expression of deep semantic information. The MFCM is designed to efficiently aggregate features from lower, middle, and deep layers, ensuring effective integration of information flow. Finally, the decoder uses these integrated features to produce a precise output, generating the final saliency map ($\mathbf{S}_1$). The intermediate saliency maps ($\mathbf{S}_2$) and ($\mathbf{S}_3$) produced by the decoder are subjected to deep supervision using the ground truth images.

coarse salient object location information. The resulting output is denoted as f_{asim}. After obtaining f_{sdem} and f_{asim}, they are integrated with the mid-level features f_e^3 through the MFCM. This allows the mid-level features to be guided by multiscale, fine-grained, and semantic information, resulting in complete and accurately located salient objects. The output is denoted as f_{mfcm}. Finally, three decoders, denoted as D_j where $j \in \{1, 2, 3\}$, gradually restore the resolution to obtain the final prominent object. A saliency head is added after each decoder to implement deep supervision [47]. This approach can improve gradient backpropagation and enhance the inference of salient objects while preventing overfitting. The detailed introduction of each module is shown below.

B. Spatial Detail Enhancement Module

Currently, the majority of the SOD methods in optical RSI use the backbone networks designed for image classification. However, this approach lacks a foundation for pixel-level precise predictions, which makes the model performance vulnerable to scale changes. To overcome this limitation, this article introduces SDEM, a lightweight approach that executes grouped dilated convolutions on two branches in parallel, thereby reducing the parameter count compared to existing methods [48], [49], [50]. This method expands the receptive field to capture distant feature information while preserving attention information via an intermediate branch. The SDEM utilizes channel attention module (CAM) [51] to promotes the utilization of original features while mitigating the impact of losing local information. The correlation and complementarity between the two pathways can be noticed due to the utilization of two distinct feature maps from shallow layers. This enables the model to acquire and incorporate information from many levels of characteristics, so capturing information at several scales and enhancing its ability to recognize targets.

Take f_e^1 and f_e^2 as inputs. The first step involves aligning the dimensions of these inputs by upsampling f_e^2 to match the

Authorized licensed use limited to: Kunming Univ of Science and Tech. Downloaded on November 22,2024 at 07:21:10 UTC from IEEE Xplore. Restrictions apply.

size of f_e^1. To decrease computational complexity, 1×1 point convolutions are employed to unify the channel numbers of two inputs, which are then halved to f_e^1 in the branch process. These adjusted inputs are denoted as $f_e^{1'}$ and $f_e^{2'}$, where $\{f_e^{1'}, f_e^{2'}\} \in \mathbb{R}^{(c_1/2)\times h_1 \times w_1}$.

What's more, SDEM employs an asymmetric convolution structure, which decomposes 2-D convolution into two 1-D operations. This enables independent feature learning in two directions, facilitating the capture of complex spatial relationships. After each convolution, batch normalization [52] and PReLU [53] operations are performed. This decomposes approach, which is based on MobileNet [54], reduces the number of model parameters while maintaining model performance. The two branches go through a 3×1 dilated convolution and a 1×3 dilated convolution in order

$$y_1 = f_{\mathrm{CAM}}\left(C_{1\times 3, r=(1,3)}\left(C_{3\times 1, r=(3,1)}\left(f_e^{1'}\right)\right)\right) \quad (1)$$

$$y_2 = f_{\mathrm{CAM}}\left(C_{1\times 3, r=(1,3)}\left(C_{3\times 1, r=(3,1)}\left(f_e^{2'}\right)\right)\right) \quad (2)$$

where r represents the dilation rate, $C_{m\times n}$ represents a convolution kernel of size $m \times n$.

Next, CAM is applied to prevent the potential loss of local information, followed by a series of 1×3 and 3×1 dilated convolutions. This process is expressed as

$$y_3 = C_{3\times 1, r=(3,1)}\left(C_{1\times 3, r=(1,3)}\left(f_{\mathrm{CAM}}(y_1)\right)\right) \quad (3)$$

$$y_4 = C_{3\times 1, r=(3,1)}\left(C_{1\times 3, r=(1,3)}\left(f_{\mathrm{CAM}}(y_2)\right)\right) \quad (4)$$

where f_{CAM} represents the channel attention.

The outputs of the four branches, y_1, y_2, y_3, and y_4, are combined and the channel number is restored to that of f_e^1 using a 1×1 point convolution. Subsequently, a CAM is applied again to enhance the integrated channel information. As f_e^1 contains the most detailed information at the lowest level, thus it is reintroduced through a residual connection. Furthermore, channel shuffle is employed to enhance the module's expressive capability. This process can be represented as follows:

$$f_{\mathrm{sdem}} = f_{\mathrm{shuffle}}\left(f_{\mathrm{CAM}}(C_{1\times 1}(y_1 + y_2 + y_3 + y_4)) + f_e^1\right) \quad (5)$$

where f_{shuffle} represents the channel shuffle, and f_{sdem} is the output of the SDEM.

SDEM integrates multiscale features from different branches, resulting in a comprehensive fusion of various feature types. This method is particularly effective in addressing the challenge posed by the significant size differences between them. It enhances the feature representation of salient objects across different scales and quantities, overcoming the limitations of traditional CNNs and dilated convolutions. Section IV will conduct ablation studies to validate the proposed SDEM's parameters and performance.

Fig. 3 shows the selected sections of the SDEM for feature map visualization. These portions comprise include each of the two branches passing through a 3×1 convolution, CAM, a second 3×1 convolution, and then the CAM following feature information fusion. Passing through CAM clearly distinguishes salient and nonsalient objects in y_1 and y_2. This reduces the likelihood that small objects will merge into the background. It can be seen that the final feature map FCA_{12} obtained after fusion clearly separates the significant targets with different scales.

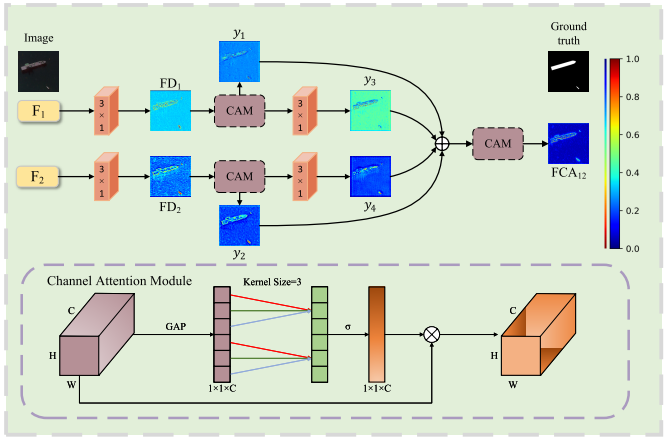

Fig. 3. Partial visualization of the SDEM. The outputs FD_1 and FD_2 are derived from the two branches that traverse the initial 3×1 convolution block. Subsequently, y_1 and y_2 represent the results of FD_1 and FD_2 passing through the initial CAM. Following this, y_3 and y_4 are the outcomes of y_1 and y_2 progressing through the subsequent 3×1 convolution block. The final output, FCA_{12}, is the result of the CAM after incorporating y_1, y_2, y_3, and y_4.

C. Advanced Semantic Interaction Module

Deep feature semantic information can help create coarse saliency maps and identify their locations. Therefore, this article proposes ASIM to model semantic information, which utilizes an NL module [55] to simulate long-range interdependence between deep semantic features in optical RSI. The proposed ASIM takes into account semantic connections that exist across several layers. By generating coarse saliency maps, it guides the recognition of salient object positions by other hierarchical features. To decrease the computational load and simplify the proposed ASIM, we extract deeper features using depthwise separable convolution blocks which include convolutions, batch normalization, and PReLU. Fig. 4 displays the specific implementation details of ASIM.

The ASIM takes f_e^4 and f_e^5 as inputs. Initially, f_e^4 is changed in terms of height and width via downsampling. Its channel count is then adjusted to match that of f_e^5 using a 1×1 point convolution, denoted as $f_e^{4'}$, where $\{f_e^{4'}, f_e^5\} \in \mathbb{R}^{c_5\times h_5 \times w_5}$. The features $f_e^{4'}$ and f_e^5 are then aggregated via channel concatenation, denoted as f_{cat}. Feature refinement is achieved by employing 3×3 depthwise separable convolution blocks, also referred to as DSconv. The semantic similarity matrix W_a is obtained by a two-step path. Each path undergoes DSconv to adjust channel counts and reshape the output. The output of first path is denoted as $f_m^4 \in \mathbb{R}^{(h_5\times w_5)\times (c_5/2)}$, and the output of second path is denoted as $f_m^5 \in \mathbb{R}^{(c_5/2)\times (h_5\times w_5)}$. Then, a query-key matching method is designed to extract the pairwise relationships of deep semantic information. This can be represented as

$$f_m^4 = \mathrm{Reshape}\left(\mathrm{DSconv}\left(\mathrm{cat}\left(f_e^{4'}, f_e^5\right)\right)\right) \quad (6)$$

$$f_m^5 = \mathrm{Reshape}\left(\mathrm{DSconv}\left(\mathrm{cat}\left(f_e^{4'}, f_e^5\right)\right)\right)^{\mathrm{T}} \quad (7)$$

$$W_a = \mathrm{Softmax}\left(f_m^4 \otimes f_m^5\right) \quad (8)$$

Authorized licensed use limited to: Kunming Univ of Science and Tech. Downloaded on November 22,2024 at 07:21:10 UTC from IEEE Xplore. Restrictions apply.

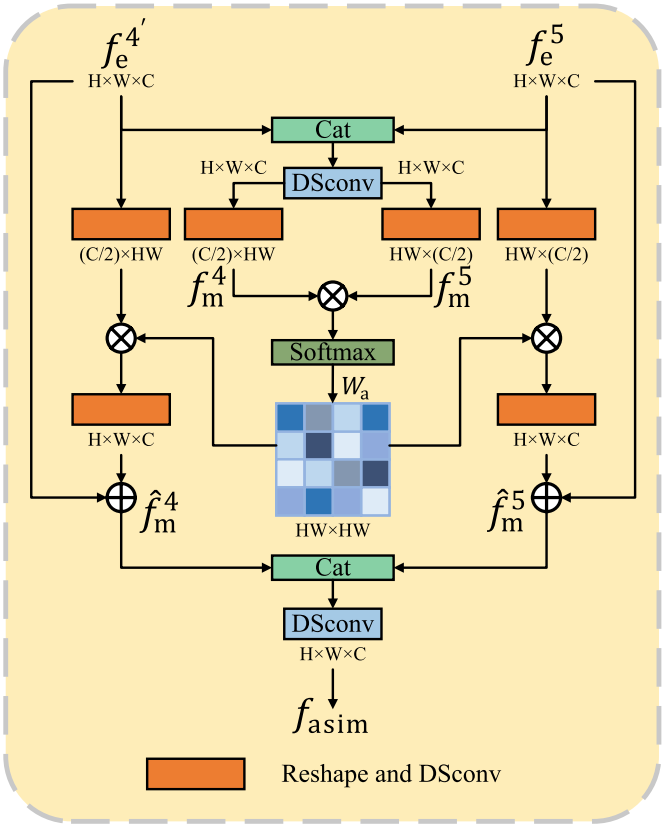

Fig. 4. Illustration of the ASIM. Analyze the semantic information present within the deep features of two adjacent layers.

where cat represents concatenation along the channel dimension. Reshape refers to the process of reshaping the tensor. T denotes the transpose, and $\otimes$ indicates matrix multiplication. And $W_a \in \mathbb{R}^{(h_5\times w_5)\times(h_5\times w_5)}$.

By modifying the dimensions of the channels $f_e^{4'}$ and f_e^5 and altering their shape, we obtain $f_r^{4'}$ and $f_r^{5'}$,where $\{f_r^{4'}, f_r^5\} \in \mathbb{R}^{(h_5\times w_5)\times(c_5/2)}$. A semantic dependency relationship is established by returning the semantic similarity matrix to the original features. Next, the semantically enhanced parts are aggregated and subjected to a refinement transformation. This procedure can be described as

$$\hat{f}_m^4 = \mathrm{DSconv}\left(W_a \otimes f_r^{4'}\right) + f_e^{4'} \tag{9}$$

$$\hat{f}_m^5 = \mathrm{DSconv}\left(W_a \otimes f_r^5\right) + f_e^5 \tag{10}$$

$$f_{\mathrm{asim}} = \mathrm{DSconv}\left(\mathrm{cat}\left(\hat{f}_m^4, \hat{f}_m^5\right)\right) \tag{11}$$

where f_{asim} is the output of the ASIM.

The procedure results in a combined representation of semantic information by merging deep feature maps f_e^4 and f_e^4. The feature f_{asim} retains the information from the original features and gathers global context information, enabling the network to precisely determine the position information of important targets.

D. Multilevel Feature Collaboration Module

Since RSI significantly exhibits complex structures and topologies, this poses a new obstacle to accurately understanding the complete area of a salient object while maintaining clear boundaries. Shallow features preserve rich details and background noise while maintaining possess clear boundaries, which are crucial for generating accurate saliency maps. On the other hand, deep features maintain consistent semantics and a clear background. The middle layer is characterized by both. Therefore, it is very important to design a multilevel feature fusion layer to fuse different levels of perceptual field information to aggregate contextual information. However, mixing information from multiple levels directly would invariably result in the emergence of redundant features. To address this issue, this article proposes an MFCM to combine both low-level and deep-level features for jointly enhance the intermediate-level features. It leverages the product of features to uncover crucial information that coexists across different features, facilitating collaborative target recognition. Subsequently, the sum of features is employed to comprehensively capture all information contained within them, ensuring no detail is overlooked, ultimately refining the object identification process.

Initially, the two inputs, f_{sdem} and f_{asim}, are aligned to f_e^3 by downsampling and upsampling.Then, their channel counts and dimensions are adjusted to match f_e^3 using DSconv. These adjusted inputs are denoted as $\hat{f}_{\mathrm{sdem}}$ and $\hat{f}_{\mathrm{asim}}$, where $\{\hat{f}_{\mathrm{sdem}}, \hat{f}_{\mathrm{asim}}\} \in \mathbb{R}^{c_3\times h_3\times w_3}$. f_e^3 also obtains $\hat{f}_e^3$ via DSconv. Due to the presence of a combination of detail information and semantic features, the middle layer is enhanced by merging it with $\hat{f}_{\mathrm{sdem}}$ and $\hat{f}_{\mathrm{asim}}$. This is achieved through element-wise multiplication to extract detail features from f_3 and $\hat{f}_{\mathrm{sdem}}$, as well as semantic features from f_3 and $\hat{f}_{\mathrm{asim}}$. The resulting information is then readded to the original features to balance detail and semantic information. Finally, the features are refined using DSconv. The process can be represented as follows:

$$f_{\mathrm{mfcm}} = \mathrm{DSconv}\left(\hat{f}_{\mathrm{sdem}} \odot \hat{f}_e^3 + \hat{f}_{\mathrm{asim}} \odot \hat{f}_e^3 + \hat{f}_e^3\right) \tag{12}$$

where $\odot$ represents element-wise multiplication, and f_{mfcm} is the output of the MFCM.

The feature maps f_{mfcm} produced by MFCM contain both deep semantic features and shallow detail features. This combination enhances the quality of edge information and enables more precise determination of target placement, irrespective of the scale.

E. Decoder

To accurately identify relevant objects, a model must effectively incorporate characteristics from several levels throughout the decoding stage. Although there is extensive literature on this topic [56], [57], [58], [59], [60], existing methods often design complex and redundant decoders without giving sufficient importance lightweight and efficient considerations. This article has developed a lightweight decoder that incorporates multilevel characteristics to ensure efficiency during deployment on mobile devices. We designed the decoder section of LPMFCNet, as depicted in Fig. 5, to include this method.

Initially, a 1×1 point convolution is applied to reduce the number of channels. During the decoding phase, multiscale

Authorized licensed use limited to: Kunming Univ of Science and Tech. Downloaded on November 22,2024 at 07:21:10 UTC from IEEE Xplore. Restrictions apply.

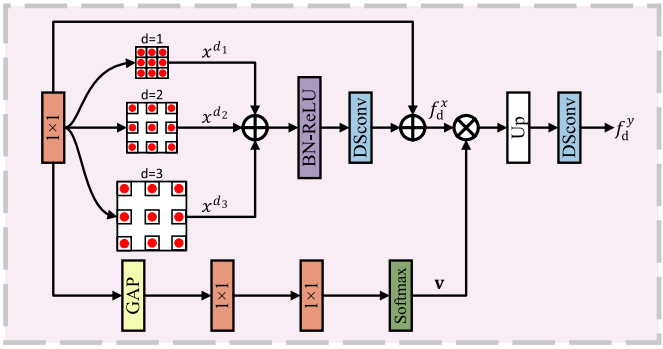

Fig. 5. Illustration of the decoder structure. d indicates the dilation rate, and GAP stands for global average pooling.

learning is performed to combine various levels of features that likewise possess multiscale representations. In order to accomplish lightweight multiscale learning, the process of multiscale fusion is carried out by utilizing parallel connections of DSconv with dilation rates of 1–3. The output, denoted as f_d^x, is optimized by using residual connections on the original features. Another path involves using a global average pooling layer to create a feature vector that can be used to recalibrate the combined multilevel features. This vector is subsequently processed by two fully connected layers to calculate the attention vector v. The resulting vector v is applied to f_d^x to produce the output f_d^y, which can be represented as

$$x^{d_1} = \mathrm{DSconv}_{d=1}(x) \tag{13}$$

$$x^{d_2} = \mathrm{DSconv}_{d=2}(x) \tag{14}$$

$$x^{d_3} = \mathrm{DSconv}_{d=3}(x) \tag{15}$$

$$f_d^x = \mathrm{DSconv}\left(\mathrm{ReLU}\left(\mathrm{BN}\left(x^{d_1} + x^{d_2} + x^{d_3}\right)\right)\right) + x \tag{16}$$

$$\mathbf{v} = \mathrm{Sigmoid}(\mathrm{FC}_2(\mathrm{ReLU}(\mathrm{FC}_1(\mathrm{GAP}(x))))) \tag{17}$$

$$f_d^y = \mathrm{DSconv}\left(\mathrm{Up}\left(\mathbf{v} \otimes f_d^x\right)\right) \tag{18}$$

where d represents the dilation rate, FC stands for Fully Connected layers, and UP denotes upsampling.

F. Loss Function

To improve the representation of features during the training phase, SalHeads are added after the three decoder blocks to produce saliency maps at three different resolutions. These are $\mathbf{S}_3 \in [0,1]^{1\times 32\times 32}$, $\mathbf{S}_2 \in [0,1]^{1\times 128\times 128}$, and $\mathbf{S}_1 \in [0,1]^{1\times 256\times 256}$, respectively. Ground truth maps are used to supervise $\mathbf{S}_3$, $\mathbf{S}_2$, and $\mathbf{S}_1$. Each SalHead comprises a dropout layer [61] and a 1×1 convolution layer. This article utilizes the saliency map $\mathbf{S}_1$, which is generated by the last Salhead, to predict the final saliency map.

The article employs a loss function that combines binary cross-entropy [62] (BCE) and intersection over union [62] (IoU) to train LPMFCNet. These loss functions can be represented as follows:

$$L_{\mathrm{bce}} = -\sum_{i=1}^{W\cdot H}\left[\mathbf{G}(i)\log(\mathbf{S}(i)) + (1-\mathbf{G}(i))\log(1-\mathbf{S}(i))\right] \tag{19}$$

$$L_{\mathrm{iou}} = 1 - \frac{\sum_{i=1}^{W\cdot H}\mathbf{S}(i)\cdot\mathbf{G}(i)}{\sum_{i=1}^{W\cdot H}[\mathbf{S}(i)+\mathbf{G}(i)-\mathbf{S}(i)\cdot\mathbf{G}(i)]} \tag{20}$$

where W and H denote the height and width of the image. $\mathbf{G}(i) \in \{0,1\}$ and $\mathbf{S}(i) \in \{0,1\}$ are the ground truth label and predicted saliency score of the ith pixel, respectively.

As a result, the total loss of LPMFCNet is defined as follows:

$$L_{\mathrm{total}} = \sum_{i=1}^{3}\left(L_{\mathrm{bce}}^i + L_{\mathrm{iou}}^i\right) \tag{21}$$

where L_{bce}^i and L_{iou}^i represent the BCE loss and IoU loss, respectively.

IV. Experiments

A. Experimental Setup

1) Dataset: This research compares the performance of LPMFCNet on two commonly used public datasets for remote sensing saliency object detection.

The first dataset, ORSSD [7], is publicly available and consists of 800 optical RSI with corresponding pixel-level annotations. It includes a variety of scenes, including ships, cars, planes, rivers, and islands. The training set comprises 600 images, while the test set contains 200 images.

EORSSD [30] is an expanded dataset based on ORSSD. It comprises 2000 optical RSI and corresponding pixel-level annotations, including more complex and challenging scenes. The dataset is divided into two sets: 1400 images for training and 600 images for testing.

2) Implementation Details: The LPMFCNet was implemented using the public PyTorch [63] toolbox on an NVIDIA GeForce RTX 4060Ti GPU. During the training phase, the size of the input images was resized to 256×256. In order to address the problem of overfitting, data augmentation techniques such as flipping and rotation were employed in this article. The network optimizer used was Adam [64], with a batch size of 8 and a learning rate of $1e^{-4}$. The learning rate declined by a factor of one-tenth per 30 epochs, and the training epoch is 50. The newly added convolutional layers and batch normalization layers are initialized using the normal distribution. Each dataset is trained on a separate training set and then tested on its own test set.

3) Evaluation Metrics: This article evaluates LPMFCNet using five commonly used quantitative evaluation metrics: S-measure (S_α, $\alpha = 0.5$) [65], E-measure (E_ξ) [66], F-measure (F_β, $\beta^2 = 0.3$) [67], mean absolute error ($\mathcal{M}$ or MAE) [68], and precision–recall (PR) curve [67], as well as F-measure curve and computational complexity metrics. The higher the first three metrics, the better, while the fourth metric should be lower. Furthermore, it is desirable for the PR curve and F-measure curve to approach the point (1, 1) in order to achieve superior outcomes. The computational complexity is determined by the quantity of parameters, measured in millions (M), and the number of floating-point operations (FLOPs), measured in billions (G). Lower values for these metrics indicate better performance.

Authorized licensed use limited to: Kunming Univ of Science and Tech. Downloaded on November 22,2024 at 07:21:10 UTC from IEEE Xplore. Restrictions apply.

The S_α evaluates the structural similarity between the real-valued saliency map and the binary ground truth. It considers object-aware (S_o) and region-aware (S_r) structure similarities

$$S_\alpha = \alpha \times S_o + (1-\alpha) \times S_r \tag{22}$$

where α is empirically set to 0.5. S_o is referred to as the object-aware structural similarity by comparing global distributions of foreground and background regions, and S_r as the region-aware structural similarity based on global structure information.

The E_ξ considers global means of the image and local pixel matching simultaneously

$$E_\xi = \frac{1}{W \times H} \sum_{i=1}^{W} \sum_{j=1}^{H} \phi_S(i, j) \tag{23}$$

where W and H represent the width and height of the image, respectively, ϕ_s is the enhanced alignment matrix, reflecting the correlation between saliency prediction map ($\boldsymbol{S} \in [0,1]^{W \times H}$) and binary ground-truth mask ($\boldsymbol{G} \in \{0,1\}^{W \times H}$) after subtracting their global means, respectively.

The F_β comprehensively considers both precision and recall by computing the weighted harmonic mean

$$F_\beta = \frac{\left(1+\beta^2\right) \times \text{Precision} \times \text{Recall}}{\beta^2 \times \text{Precision} + \text{Recall}} \tag{24}$$

where $\beta^2 = 0.3$ is set to 0.3 [67] to put more emphasis on precision. The metrics used in this study are maximum F-measure ($F_\beta^{\max}$) and mean F-measure (F_β^{mean}). The F_β^{mean} is calculated by binarizing the predicted saliency map with an adaptive threshold that is twice the mean value of the saliency prediction.

The MAE calculates the average absolute error of pixels between saliency prediction map $\boldsymbol{S}$ and binary ground-truth mask $\boldsymbol{G}$

$$\text{MAE} = \frac{1}{W \times H} \sum_{i=1}^{W} \sum_{j=1}^{H} |\boldsymbol{G}(i, j) - \boldsymbol{S}(i, j)|. \tag{25}$$

B. Performance Analysis

This article compares the proposed LPMFCNet with 17 other state-of-the-art (SOTA) methods, which include seven conventional saliency detection methods for NSI: PoolNet [69], EGNet [24], GCPA [25], ITSD [26], SUCA [27], PA-KRN [28], and DPORTNet [70]; as well as two lightweight saliency detection methods for NSI: SAMNet [39] and HVPNet [40]. Four conventional and one transformer saliency detection methods for optical RSI are compared to LPMFCNet. These methods included DAFNet [30], SARNet [71], MJRBM [31], ERPNet [32], as well as the transformer method BMGT [72]. Furthermore, the comparison also includes three lightweight methods: CorrNet [22], SeaNet [23], and MEANet [43]. The results for these methods are presented in Table I. All predicted saliency maps were assessed using identical metrics.

1) In Comparison to Deep Learning Methods for NSI: As can be seen from Table I that the LPMFCNet shows excellent performance and significant improvement results on two datasets. In comparison to the best-performing PA-KRN, LPMFCNet requires 43.1 times fewer parameters and 1235.4 times fewer FLOPs, while also making significant gains in other measures. Both ITSD and DPORTNet are clearly less effective than PA-KRN on two datasets. They prioritize minimizing the number of parameters and FLOPs to improve the model's overall efficiency when compared to lightweight approaches, there is still a gap. In contrast to SAMNet, ITSD exhibits a significantly higher number of parameters and FLOPs, with values approximately 12.84 and 109 times greater than those of SAMNet, respectively. Similarly, DPORTNe demonstrates a substantial increase in the number of parameters and FLOPs compared to SAMNet, with figures approximately 14.16 and 121.56 times higher, respectively. Our approach demonstrates superior performance compared to deep learning applied to NSI across multiple metrics, all while preserving the attributes of having a reduced parameters and FLOPs.

While lightweight methods for NSI have advantages in terms of the number of parameters and computational complexity, HVPNet has the lowest number of parameters among all compared models and SAMNet maintains the lowest FLOPs along with LPMFCNet among all compared models. However, HVPNet and SAMNet experience varying degrees of reduction in S_α, E_ξ^{mean}, $E_\xi^{\max}$, F_β^{mean}, $F_\beta^{\max}$, and MAE. This indicates that lightweight models exhibit limited generalization ability when applied to challenging scenes in RSI compared to NSI, highlighting the benefits of specialized methods. It can be seen that LPMFCNet achieves a harmonious combination of effectiveness and efficiency.

2) In Comparison to Deep Learning Methods for RSI: The LPMFCNet achieves better results compared to methods specifically designed for RSI, which maintains high accuracy while further reducing the number of parameters and computational complexity. In the ORSSD dataset, our model achieves the highest rankings in S_α, E_ξ^{mean}, $E_\xi^{\max}$, $F_\beta^{\max}$, and MAE, and the second highest ranking in F_β^{mean}. For EORSSD dataset, our model achieves the highest rank in S_α and E_ξ^{mean} and the second in $E_\xi^{\max}$ and $F_\beta^{\max}$, and third in F_β^{mean} and MAE. In EORSSD dataset, LPMFCNet performs worse than BMGT [72] in terms of $F_\beta^{\max}$ and MAE. This phenomenon occurs due to the utilization of BMGT, a model founded on the transformer architecture, which extensively investigates and preserves global semantic information, resulting in decreased error rates in intricate situations. However, LPMFCNet outperforms BMGT in other indicators, E_ξ^{mean} exhibits the highest level of lag in BMGT, amounting to 2.18%. Simultaneously, as evident from Table I, the lightweight model surpasses the majority of heavy models in performance. This is attributable to the fact that many of these heavier models rely on outdated network architectures, whereas BMGT employs a more advanced Transformer architecture, contributing to its commendable performance. The lightweight model, on the other hand, introduces a more innovative network architecture, meticulously redesigns the subsequent processing modules

Authorized licensed use limited to: Kunming Univ of Science and Tech. Downloaded on November 22,2024 at 07:21:10 UTC from IEEE Xplore. Restrictions apply.

TABLE I

COMPARES LPMFCNet WITH SOTA METHODS. THE TABLE USES A COLOR SCHEME TO INDICATE THE BEST PERFORMANCE, WITH RED INDICATING THE BEST, BLUE THE SECOND BEST, AND GREEN THE THIRD BEST. THE TABLE INCLUDES CNN-BASED METHODS FOR NSIS (C-NSIS), CNN-BASED METHODS FOR RSIS (C-RSIS), TRANSFORMER-BASED METHODS FOR RSIS (T-RSIS), LIGHTWEIGHT METHODS FOR NSIS (L-NSIS), AND LIGHTWEIGHT METHODS FOR RSIS (L-RSIS)

Methods	Type	Year	Param (M)↓	FLOPs (G)↓	ORSSD [7]						EORSSD [30]					
					$S_\alpha\uparrow$	$E_\xi^{mean}\uparrow$	$E_\xi^{max}\uparrow$	$F_\beta^{mean}\uparrow$	$F_\beta^{max}\uparrow$	$\mathcal{M}\downarrow$	$S_\alpha\uparrow$	$E_\xi^{mean}\uparrow$	$E_\xi^{max}\uparrow$	$F_\beta^{mean}\uparrow$	$F_\beta^{max}\uparrow$	$\mathcal{M}\downarrow$
PoolNet [69]	C-NSI	2019	53.63	123.4	0.8403	0.8650	0.9343	0.6999	0.7706	0.0358	0.8207	0.8193	0.9292	0.6406	0.7545	0.0210
EGNet [24]	C-NSI	2019	108.07	291.9	0.8721	0.9013	0.9731	0.7500	0.8332	0.0216	0.8601	0.8775	0.9570	0.6967	0.7880	0.0110
GCPA [25]	C-NSI	2020	67.06	54.3	0.9026	0.9341	0.9509	0.8433	0.8687	0.0168	0.8869	0.9167	0.9524	0.7905	0.8347	0.0102
ITSD [26]	C-NSI	2020	17.08	54.5	0.9050	0.9482	0.9601	0.8502	0.8735	0.0165	0.9050	0.9407	0.9556	0.8221	0.8523	0.0106
SUCA [27]	C-NSI	2021	117.71	56.4	0.8989	0.9400	0.9584	0.8237	0.8484	0.0145	0.8988	0.9277	0.9520	0.7949	0.8229	0.0097
PA-KRN [28]	C-NSI	2021	141.06	617.7	0.9239	0.9620	0.9680	0.8727	0.8890	0.0139	0.9192	0.9536	0.9616	0.8358	0.8639	0.0104
DPORTNet [70]	C-NSI	2022	18.84	60.78	0.8827	0.9139	0.9214	0.8184	0.8309	0.0220	0.8960	0.9116	0.9423	0.7937	0.8363	0.0150
DAFNet [30]	C-RSI	2021	29.35	68.5	0.9191	0.9539	0.9771	0.8511	0.8928	0.0113	0.9166	0.9291	0.9861	0.7845	0.8614	0.0060
SARNet [71]	C-RSI	2021	25.91	129.7	0.9134	0.9477	0.9557	0.8619	0.8850	0.0187	0.9240	0.9555	0.9620	0.8541	0.8719	0.0099
MJRBM [31]	C-RSI	2022	43.54	95.7	0.9204	0.9415	0.9623	0.8566	0.8842	0.0163	0.9197	0.9350	0.9646	0.8239	0.8656	0.0099
ERPNet [32]	C-RSI	2022	77.19	263	0.9352	0.9304	0.9738	0.8798	0.9036	0.0114	0.9252	0.9366	0.9665	0.8269	0.8743	0.0082
BMGT [72]	T-RSI	2023	30.8	220.7	0.9368	0.9609	0.9781	0.8872	0.9105	0.0092	0.9293	0.9483	0.9740	0.8410	0.8829	0.0058
SAMNet [39]	L-NSI	2021	1.33	0.5	0.8761	0.8818	0.9478	0.7531	0.8137	0.0217	0.8622	0.8700	0.9421	0.7214	0.7813	0.0132
HVPNet [40]	L-NSI	2021	1.23	1.1	0.8610	0.8717	0.9320	0.7396	0.7938	0.0225	0.8734	0.8721	0.9482	0.7377	0.8036	0.0110
CorrNet [22]	L-RSI	2022	4.09	21.1	0.9380	0.9746	0.9790	0.9002	0.9129	0.0098	0.9289	0.9646	0.9696	0.8620	0.8778	0.0083
SeaNet [23]	L-RSI	2023	2.76	1.7	0.9260	0.9722	0.9767	0.8772	0.8942	0.0105	0.9208	0.9651	0.9710	0.8519	0.8649	0.0073
MEANet [43]	L-RSI	2024	3.27	9.62	0.9340	0.9729	0.9768	0.8934	0.9033	0.0098	0.9282	0.9657	0.9708	0.8678	0.8766	0.0070
LPMFCNet	L-RSI	2024	3.26	0.5	0.9390	0.9761	0.9824	0.9000	0.9130	0.0086	0.9300	0.9701	0.9768	0.8589	0.8780	0.0066

tailored to features at varying levels, and incorporates postprocessing optimization techniques like pruning. These measures significantly diminish the computational requirements and parameter count, all while preserving the model's expressive capabilities.

In conclusion, the results indicate that LPMFCNet has remarkable efficiency and accuracy, particularly well-suited for tasks necessitating thorough and dependable analysis of remote sensing data. This is because of the utilization of progressive modules in LPMFCNet for feature extraction, tailored to the distinct characteristics of various levels of feature maps. In addition, targeted processing is conducted based on the specific attributes of RSI to enhance the overall performance.

When compared to the most recent lightweight methods for RSI, LPMFCNet scores second in terms of parameter count and first in FLOPs. Despite having fewer parameters, the performance of SeaNet has fallen dramatically. In contrast, the LPMFCNet described in this research outperforms SeaNet in all metrics with only a minor increase in parameters, owing to its use of a more advanced backbone network architecture and a compact module design. Among lightweight models, CorrNet boasts the largest number of parameters and computational complexity, yet it occupies only the second position in terms of the index ranking, with LPMFCNet topping the list in the first place. Specifically, in the ORSSD dataset, LPMFCNet outperforms CorrNet with a 0.1% increase in S_α, a 0.18% increase in E_ξ^{mean}, a 0.34% increase in E_ξ^{max}, a 0.01% increase in F_β^{max} and a 0.12% improvement in MAE.Similarly, on the EORSSD dataset, LPMFCNet achieves a 0.11% increase in S_α, a 0.55% increase in E_ξ^{mean}, a 0.72% increase in E_ξ^{max}, a 0.02% increase in F_β^{max} and a 0.04% improvement in MAE, further demonstrating its superior performance. Compared to method MEANet, LPMFCNet maintains a similar number of parameters while outperforming it comprehensively on the ORSSD dataset and maintaining an advantage in all metrics except for F_β^{mean} on the EORSSD dataset. This demonstrates that LPMFCNet achieves better performance while maintaining a lightweight design.

C. Qualitative Comparison

To further demonstrate the advantages of LPMFCNet, Fig. 6 presents visual comparisons against seven SOTA methods using selected images from the ORSSD and EORSSD datasets. The scenarios include large objects, multiple small objects, objects with shadows, complex geometric objects, and low-contrast objects. The second column of Fig. 6 displays the ground truth maps.

Authorized licensed use limited to: Kunming Univ of Science and Tech. Downloaded on November 22,2024 at 07:21:10 UTC from IEEE Xplore. Restrictions apply.

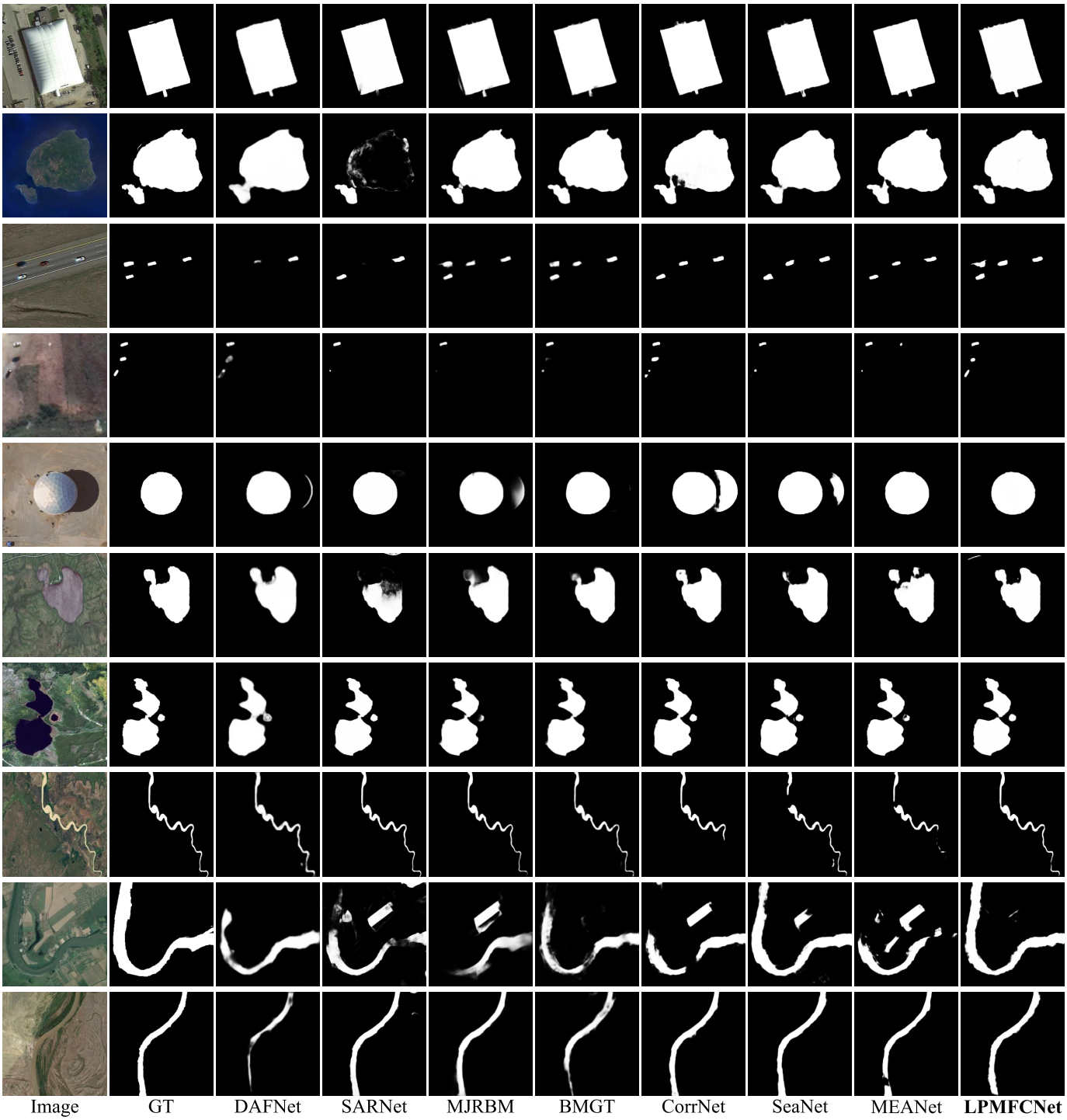

Fig. 6. Visual comparison with seven of the most advanced RSI-SOD networks.

The first and second lines in Fig. 6 depict scenarios with large objects. The scene in the first line shows a building with an orderly shape that all methods correctly identify as the salient object. However, the scene in the second line contains islands, and the complex boundary between the big and small islands leads to many methods connecting wrongly, including DAFNet, MJRBM, CorrNet, SeaNet, and MEANet. SARNet failed to accurately identify important land blocks. As can be seen from the figure, LPMFCNet precisely defines the boundaries between different objects.

The third and fourth lines in Fig. 6 show a plurality of vehicle targets with smaller sizes. In the third line, LPMFCNet correctly recognizes all notable objects, whereas other methods miss some targets. DAFNet and SARNet only identify two salient objects, and the three lightweight methods CorrNet, SeaNet, and MEANet find three salient objects. DAFNet successfully detected all the salient objects in the fourth line, but it was unable to capture specific details. In contrast, the other methods could not manage to identify all the salient objects. Conversely, LPMFCNet captures

Authorized licensed use limited to: Kunming Univ of Science and Tech. Downloaded on November 22,2024 at 07:21:10 UTC from IEEE Xplore. Restrictions apply.

both the details of small objects and identifies all salient objects.

The scene in the fifth line in Fig. 6 shows an object with a shadow. The presence of shadows interferes with the accurate capture of salient objects. DAFNet, MJRBM, CorrNet, and SeaNet are negatively impacted by the shadow's interference, resulting in high segmentation errors. In contrast, LPMFCNet, like specialized methods such as SARNet, BMGT, and MEANet, distinguishes perfectly between shadows and salient objects.

Lines six, seven, and eight of Fig. 6 depict scenes with complex geometric objects. Due to the complex shape of the target in the sixth line, SARNet, MJRBM, BMGT, SeaNet, and MEANet, all exhibit significant missing target structures. The scene in the seventh line is an intricate structure composed of lakes of different sizes and all methods can accurately locate large lakes. The target of large lakes can be identified by all methods. BMGT failed to accurately locate the target in small lakes, whereas DAFNet, MJRBM, and MEANet successfully identified the target but exhibited an incomplete spatial structure. In the eighth line, the large span and complex shape of the salient object make CorrNet, SeaNet, and MEANet unable to maintain the integrity of the target. However, LPMFCNet can adapt to objects with complex shapes and maintain their geometry and topological structure.

The ninth and tenth lines of Fig. 6 depict scenarios involving objects with low contrast. The complicated background in the ninth and tenth line's scene, exacerbated by the low contrast, greatly hinders the identification ability of salient objects. This interference leads to significant background noise in all other approaches, resulting in incomplete or excessive segmentation. In the ninth line scene, all methods are adversely affected by significant background interference, leading to the inability to correctly identify objects. The tenth line example shows incomplete segmentation of DAFNet, BMGT, and MEANet. On the other hand, LPMFCNet remains largely unaffected and generates the most accurate saliency maps.

Overall, the proposed LPMFCNet produces accurate and precise saliency maps due to the influence of its key modules. The ASIM provides precise localization capabilities, the SDEM offers segmentation capabilities for objects of varying scales, and the MFCM integrates high-level semantics with complex details. These features allow our model to properly identify salient objects, overcoming the limitations posed by RSI.

This study provides a comparative analysis of the PR and F-measure curves of LPMFCNet with SARNet, MJRBM, CorrNet, SeaNet, and MEANet on the ORSSD and EORSSD datasets, as shown in Fig. 7. The LPMFCNet is represented by the red curve. The results demonstrate that the proposed method, as shown by the red curve, exhibits superiority on the ORSSD dataset, since it is positioned in closer proximity to the top right corner. Our model exhibited moderate overall performance on the EORSSD dataset, primarily due to variations in F_β^{mean} and $F_\beta^{\max}$. Our approach appears to be particularly sensitive to threshold adjustments in this dataset, effectively balancing precision and recall under specific thresholds. Notably, $F_\beta^{\max}$ achieves impressive results, yet as thresholds shift, the model's performance fluctuates, leading to average performance metrics such as F_β^{mean}. Nevertheless, the proposed LPMFCNet maintains robust competitiveness in accurately detecting salient objects within RSI, bolstered by its optimal parameter size and computational efficiency.

TABLE II

PERFORMANCE METRICS OF SDEM WITH STANDARD CONVOLUTIONS AND VARYING DILATION RATES. THE COLOR RED SIGNIFIES THE HIGHEST LEVEL OF PERFORMANCE

Models	Param (M)↓	FLOPs (G)↓	ORSSD [7]			
			$S_\alpha \uparrow$	$E_\xi^{\max} \uparrow$	$F_\beta^{\max} \uparrow$	$\mathcal{M} \downarrow$
$SDEM_{\text{conv}}$	3.26	0.52	0.9389	0.9809	0.9086	0.0119
$SDEM_{\text{d=5}}$	3.26	0.503	0.9293	0.9717	0.8964	0.0114
$SDEM_{\text{d=7}}$	3.26	0.505	0.9302	0.9762	0.9004	0.0115
LPMFCNet	3.26	0.50	0.9390	0.9824	0.9130	0.0086

D. Ablation Studies

In order to showcase the efficacy of the suggested modules, we conducted comprehensive testing on the ORSSD dataset employing the subsequent methodologies: 1) justifying the need for implementing the decomposed convolution approach in SDEM and the logic for selecting the dilation rate; 2) assessing the efficacy of MFCM; and 3) evaluating the individual impact of each module. For each variant experiment, we rigorously retrain it using the same parameter settings and dataset as in Section IV-A.

1) Investigate the Rationality of Using Dilated Convolutions in SDEM: Replacing the extended convolution of two branches with ordinary convolution can omit the decomposition convolution operation and reduce the parameters of expansion rate and number of groups. The configuration is denoted as $\text{SDEM}_{\text{conv}}$. In order to determine the optimal parameters of SDEM, various expansion rates are tested in this section. The initial convolution kernel sizes of 3×1 and 1×3 were substituted with 5×1, 1×5, 7×1, and 1×7 correspondingly, denoted as $\text{SDEM}_{\text{d=5}}$ and $\text{SDEM}_{\text{d=7}}$. The results are presented in Table II. It was noted that the MAE increased from 0.0086 to 0.0119 when standard convolutions were employed. The cause of this rise is the incapacity to enlarge the receptive field. The extension is essential for obtaining additional supplementary multiscale information, which is vital for precise target localization. Consequently, there has been a significant decrease in the evaluation measures at the pixel level. Moreover, the lack of a decomposition strategy leads to a small rise in complexity. Nevertheless, overly enlarging the receptive field and the size of the convolution kernel will result in a drop in other metrics to different extents. The ineffectiveness of SDEM un catching salient objects of varying sizes may be attributed to an excessively broad receptive field and greater padding. Therefore, the convolution kernel sizes of SDEM in this article are set as 3×1 and 1×3 in SDEM. In addition, Table II shows that the SDEM proposed in this study improves performance without substantially increasing the parameter count, thereby

Authorized licensed use limited to: Kunming Univ of Science and Tech. Downloaded on November 22,2024 at 07:21:10 UTC from IEEE Xplore. Restrictions apply.

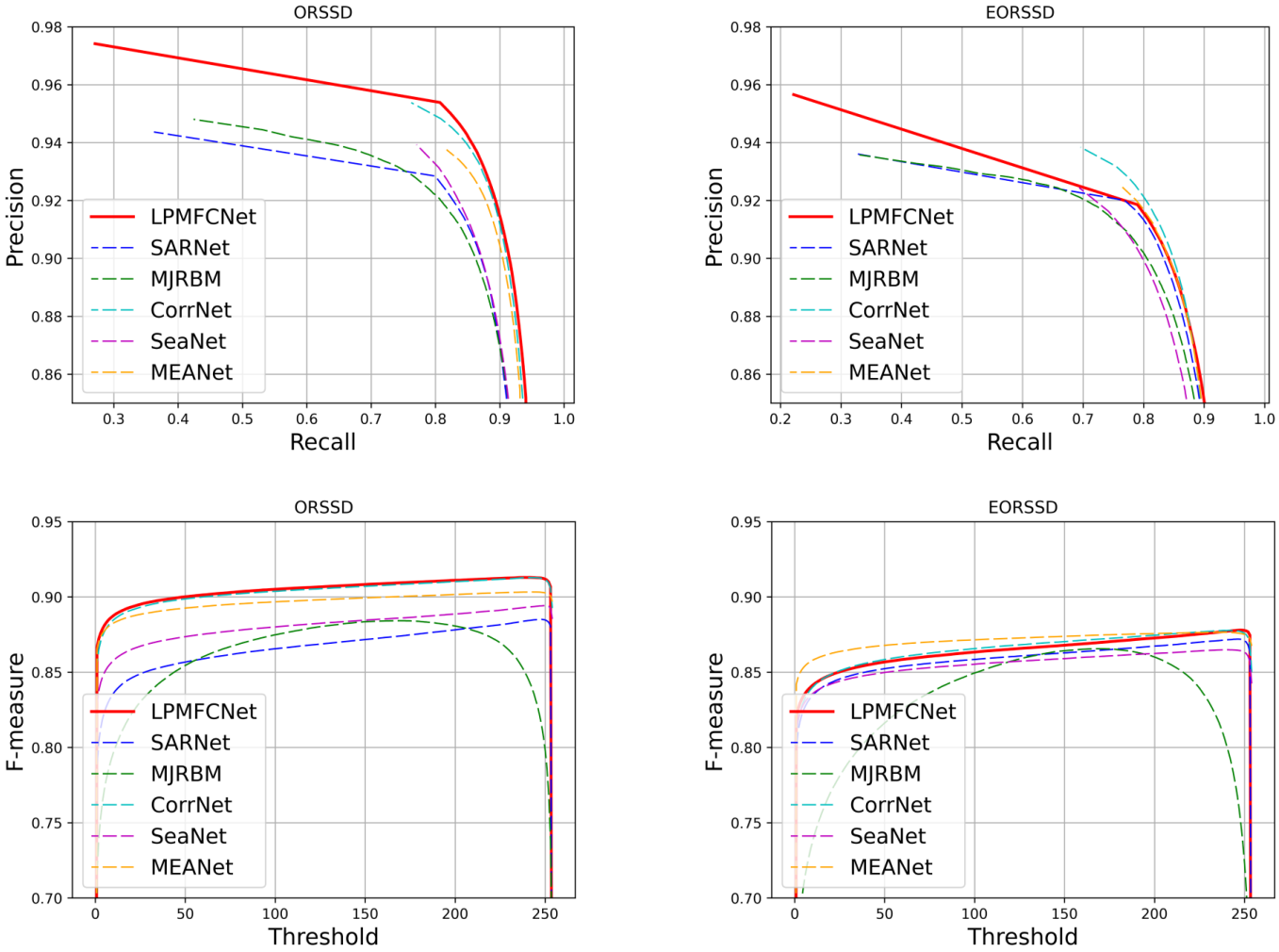

Fig. 7. PR curve (the first row) and F-measure curve (the second row) for the ORSSD and EORSSD datasets.

TABLE III

METRICS FOR MFCM AFTER EXCLUDING VARIOUS BRANCHES. THE COLOR RED SIGNIFIES THE HIGHEST LEVEL OF PERFORMANCE

No.	Param (M)↓	FLOPs (G)↓	ORSSD [7]			
			$S_\alpha \uparrow$	$E_\xi^{\max} \uparrow$	$F_\beta^{\max} \uparrow$	$\mathcal{M} \downarrow$
1	3.25	0.48	0.9332	0.9783	0.9009	0.0107
2	3.26	0.49	0.9354	0.9790	0.9047	0.0103
3	3.25	0.49	0.9336	0.9785	0.9051	0.0099
4	3.26	0.50	0.9390	0.9824	0.9130	0.0086

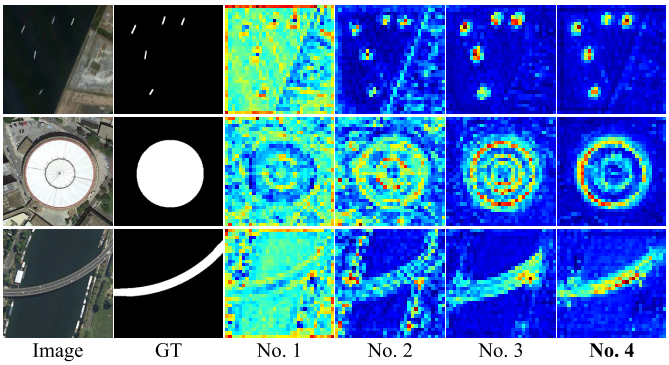

Fig. 8. Feature visualization of four conditions. The designations No.1–4 represent the w/o MFCM, w/o ASIM branch, w/o SDEM branch, and w/ MFCM, respectively.

showcasing the efficiency and robust performance of SDEM as a lightweight module.

2) Test the Effectiveness of MFCM: We assessed the performance of LPMFCNet in four configurations: the experiments were conducted under four conditions: 1) without the MFCM, namely, w/o MFCM; 2) without the f_{asim} input branch, namely, w/o ASIM branch; 3) without the f_{sdem} input branch, namely, w/o SDEM branch; and 4) with the complete MFCM, namely, w/ MFCM. The data presented in Table III demonstrates that the MFCM is characterized by its lightweight nature, as it utilizes a mere 0.014 million parameters. After analyzing the simulated visual representations of the results in various situations, it is evident from Table III and Fig. 8 that the disparities under varying circumstances are easily distinguishable. The first scenario corresponds to the output of the third layer in the backbone network. If the feature maps extracted from the backbone network are not

Authorized licensed use limited to: Kunming Univ of Science and Tech. Downloaded on November 22,2024 at 07:21:10 UTC from IEEE Xplore. Restrictions apply.

TABLE IV
ABLATION RESULTS OF EVALUATING THE INDIVIDUAL CONTRIBUTION OF EACH MODULE IN LPMFCNet. THE COLOR RED SIGNIFIES THE HIGHEST LEVEL OF PERFORMANCE

Baseline	SDEM	ASIM	Param (M)↓	FLOPs (G)↓	ORSSD [7]			
					$S_\alpha \uparrow$	$E_\xi^{\max} \uparrow$	$F_\beta^{\max} \uparrow$	$\mathcal{M} \downarrow$
✓			3.05	0.47	0.9289	0.9744	0.8954	0.0117
✓	✓		3.05	0.47	0.9308	0.9779	0.9001	0.0110
✓		✓	3.19	0.48	0.9301	0.9749	0.8982	0.0111
✓	✓	✓	3.25	0.48	0.9332	0.9783	0.9009	0.0107

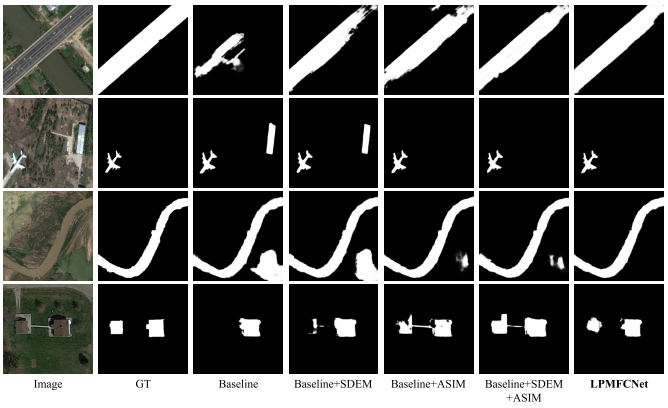

Fig. 9. Visualization examples that demonstrate the effectiveness of key modules.

processed, there will inevitably be a large number of irrelevant features. Feeding these features directly into the decoder will have a direct impact on the production of the final saliency map, leading to contamination. In the second scenario, When there is multiscale information and edge detail information but no semantic guidance, the pseudo-visualizations exhibited erroneous location localization. In comparison to the initial situation, there were enhancements of 0.22%, 0.07%, 0.38%, and 0.04% in S_α, $E_\xi^{\max}$, $F_\beta^{\max}$, and MAE, respectively. The third scenario incorporated semantic information but did not have the guidance of multiscale information and edge detail information, leading to the absence of edge details. In comparison to the initial situation, enhancements of 0.04%, 0.02%, 0.42%, and 0.08% were noted in S_α, $E_\xi^{\max}$, $F_\beta^{\max}$, and MAE, respectively. Although the increases in S_α and $E_\xi^{\max}$ were not as significant as in the second scenario, the accuracy of the position meant that the MAE was the lowest among the first three scenarios. The fourth scenario demonstrated superior metrics, indicating that the most favorable outcomes are achieved by combining multiscale and edge information at shallow layers, coupled with semantic information at deeper layers.

3) Investigate the Contribution of Each Module: Given that MFCM has already been tested in Section IV-D2, this section will now focus on the other modules. We designed four variants of LPMFCNet, which are presented in Table IV: 1) baseline; 2) baseline + SDEM; 3) baseline + ASIM; and 4) baseline + SDEM + ASIM. For ablation modules SDEM and ASIM with two inputs and one output, we selectively keep the up-sampling or down-sampling process. Subsequently, we employed convolutions to align the channels before combining them. The results are shown in Table IV. It can be seen that the addition of SDEM resulted in improvements of 0.19%, 0.35%, 0.47%, and 0.07% for S_α, $E_\xi^{\max}$, $F_\beta^{\max}$, and MAE, respectively. Similarly, the addition of ASIM resulted in improvements of 0.12%, 0.05%, 0.28%, and 0.06% for S_α, $E_\xi^{\max}$, $F_\beta^{\max}$, and MAE, respectively. Finally, the combined addition of SDEM and ASIM resulted in improvements of 0.43%, 0.39%, 0.55%, and 0.1% for S_α, $E_\xi^{\max}$, $F_\beta^{\max}$, and MAE, respectively. The results presented here illustrate the efficacy of the suggested modules for detecting salient objects in optical remote sensing.

From Table IV, it is clear that compared to the benchmark model, SDEM achieves performance improvement without an increase in either parameters or computational complexity. Conversely, ASIM experiences a larger increase in parameters due to the more complex matrix operations required for deep feature extraction, yet its performance enhancement is relatively modest when compared to SDEM. This is attributed to the paramount significance of scale difference in RSI.

The prediction maps generated by various ablation variants are depicted in Fig. 9. The introduction of SDEM improves the model's ability to recognize objects of different scales and clarifies and completes the boundaries of salient objects. ASIM plays a crucial role in achieving precise target localization due to its rich semantic information. The combination of SDEM and ASIM enhances the model's ability to accurately identify salient objects in complex scenes. The joint action of SDEM, ASIM, and MFCM produces the most distinctive saliency images, highlighting the importance of each module in optical remote sensing saliency detection.

V. CONCLUSION

In this article, a novel LPMFCNet is proposed to realize SOD in RSI. This framework constructs progressive feature information through multilevel image content extraction. For low-level features, SDEM is designed to enhance the feature representation of salient objects at different scales and numbers, which emphasizes original features and reduces the impact of missing local information. For high-level features, ASIM is proposed to model long-range dependencies between deep semantic features in optical RSI. In order to further utilize multilevel features, MFCM is designed to collaboratively utilize target features from a multilevel perspective, and use low-level features and deep-level features to jointly enhance mid-level features. LPMFCNet has only 3.26M parameters and 0.5G FLOPs, making it superior to most lightweight networks currently available. Although our LPMFCNet still lags behind some SOTA, complex architecture models in processing high-resolution RSIs with intricate backgrounds, this deficiency serves as a catalyst, driving our next research goal toward advancing in this direction.

Lightweight models boast vast applications, particularly on resource-limited devices, reducing complexity, speeding inference, and cutting deployment costs. They're vital for real-time

Authorized licensed use limited to: Kunming Univ of Science and Tech. Downloaded on November 22,2024 at 07:21:10 UTC from IEEE Xplore. Restrictions apply.

image processing on drones and processing remote sensing data for disaster response, military targeting, and maritime security. For example, detecting ships and pollution in ocean scenes, they bolster maritime security and eco-preservation. In summary, these models propel AI's feasibility and practicality in real-world use, fostering widespread adoption across industries.

REFERENCES

[1] M. Gao, F. Zheng, J. J. Q. Yu, C. Shan, G. Ding, and J. Han, "Deep learning for video object segmentation: A review," *Artif. Intell. Rev.*, vol. 56, no. 1, pp. 457–531, Jan. 2023.

[2] C. Fu et al., "Siamese object tracking for unmanned aerial vehicle: A review and comprehensive analysis," *Artif. Intell. Rev.*, vol. 56, no. S1, pp. 1417–1477, Oct. 2023, doi: 10.1007/s10462-023-10558-5.

[3] C. Chen et al., "TOPIQ: A top-down approach from semantics to distortions for image quality assessment," *IEEE Trans. Image Process.*, vol. 33, pp. 2404–2418, 2024.

[4] Y. Wang et al., "Remote sensing image super-resolution and object detection: Benchmark and state of the art," *Expert Syst. Appl.*, vol. 197, Jul. 2022, Art. no. 116793.

[5] Z. Wang, Y. Ma, and Y. Zhang, "Review of pixel-level remote sensing image fusion based on deep learning," *Inf. Fusion*, vol. 90, pp. 36–58, Feb. 2023.

[6] J. Wang, W. Li, M. Zhang, R. Tao, and J. Chanussot, "Remote-sensing scene classification via multistage self-guided separation network," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5615312.

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

[8] L. Li et al., "Discriminative co-saliency and background mining transformer for co-salient object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 7247–7256.

[9] P. Zheng et al., "GCoNet+: A stronger group collaborative co-salient object detector," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 9, pp. 10929–10946, Sep. 2023.

[10] C. Xie, C. Xia, M. Ma, Z. Zhao, X. Chen, and J. Li, "Pyramid grafting network for one-stage high resolution saliency detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 11707–11716.

[11] Z. Zhu, Z. Zhang, Z. Lin, X. Sun, and M.-M. Cheng, "Co-salient object detection with co-representation purification," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 7, pp. 8193–8205, Jul. 2023.

[12] Y. Wang, R. Wang, X. Fan, T. Wang, and X. He, "Pixels, regions, and objects: Multiple enhancement for salient object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 10031–10040.

[13] J.-J. Liu, Q. Hou, Z.-A. Liu, and M.-M. Cheng, "PoolNet+: Exploring the potential of pooling for salient object detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 1, pp. 887–904, Jan. 2023.

[14] A. A. Darem, A. A. Alhashmi, A. M. Almadani, A. K. Alanazi, and G. A. Sutantra, "Development of a map for land use and land cover classification of the northern border region using remote sensing and GIS," *Egyptian J. Remote Sens. Space Sci.*, vol. 26, no. 2, pp. 341–350, Aug. 2023.

[15] N. Casagli, E. Intrieri, V. Tofani, G. Gigli, and F. Raspini, "Landslide detection, monitoring and prediction with remote-sensing techniques," *Nature Rev. Earth Environ.*, vol. 4, no. 1, pp. 51–64, Jan. 2023.

[16] B. Zeng, S. Gao, Y. Xu, Z. Zhang, F. Li, and C. Wang, "Detection of military targets on ground and sea by UAVs with low-altitude oblique perspective," *Remote Sens.*, vol. 16, no. 7, p. 1288, Apr. 2024.

[17] Z. Wang, J. Guo, C. Zhang, and B. Wang, "Multiscale feature enhancement network for salient object detection in optical remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5634819.

[18] X. Zeng, M. Xu, Y. Hu, H. Tang, Y. Hu, and L. Nie, "Adaptive edge-aware semantic interaction network for salient object detection in optical remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5617416.

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

[20] G. Li, Z. Liu, D. Zeng, W. Lin, and H. Ling, "Adjacent context coordination network for salient object detection in optical remote sensing images," *IEEE Trans. Cybern.*, vol. 53, no. 1, pp. 526–538, Jan. 2022.

[21] R. Cong, Y. Zhang, L. Fang, J. Li, Y. Zhao, and S. Kwong, "RRNet: Relational reasoning network with parallel multiscale attention for salient object detection in optical remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–11, 2021.

[22] G. Li, Z. Liu, Z. Bai, W. Lin, and H. Ling, "Lightweight salient object detection in optical remote sensing images via feature correlation," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5617712.

[23] G. Li, Z. Liu, X. Zhang, and W. Lin, "Lightweight salient object detection in optical remote-sensing images via semantic matching and edge alignment," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, Jan. 2023, Art. no. 5601111.

[24] J. Zhao, J.-J. Liu, D.-P. Fan, Y. Cao, J. Yang, and M.-M. Cheng, "EGNet: Edge guidance network for salient object detection," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 8779–8788.

[25] Z. Chen, Q. Xu, and R. Cong, "Global context-aware progressive aggregation network for salient object detection," in *Proc. AAAI Conf. Artif. Intell.*, 2020, vol. 34, no. 7, pp. 10599–10606.

[26] 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. 9141–9150.

[27] 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.

[28] B. Xu, H. Liang, R. Liang, and P. Chen, "Locate globally, segment locally: A progressive architecture with knowledge review network for salient object detection," in *Proc. AAAI Conf. Artif. Intell.*, 2021, vol. 35, no. 4, pp. 3004–3012.

[29] Y. Lu, D. Min, K. Fu, and Q. Zhao, "Depth-cooperated trimodal network for video salient object detection," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Oct. 2022, pp. 116–120.

[30] 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, 2020.

[31] 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, 2021.

[32] X. Zhou et al., "Edge-guided recurrent positioning network for salient object detection in optical remote sensing images," *IEEE Trans. Cybern.*, vol. 53, no. 1, pp. 539–552, Jan. 2022.

[33] H. Luo and B. Liang, "Semantic-edge interactive network for salient object detection in optical remote sensing images," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 16, pp. 6980–6994, 2023.

[34] Y. Li, Q. Hou, Z. Zheng, M.-M. Cheng, J. Yang, and X. Li, "Large selective kernel network for remote sensing object detection," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2023, pp. 16794–16805.

[35] C. Zhang, J. Su, Y. Ju, K.-M. Lam, and Q. Wang, "Efficient inductive vision transformer for oriented object detection in remote sensing imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5616320.

[36] K. Chen et al., "RSPrompter: Learning to prompt for remote sensing instance segmentation based on visual foundation model," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, 2024, Art. no. 4701117.

[37] A. Kirillov et al., "Segment anything," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, Oct. 2023, pp. 4015–4026.

[38] X. Li et al., "A synergistical attention model for semantic segmentation of remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5400916.

[39] Y. Liu, X.-Y. Zhang, J.-W. Bian, L. Zhang, and M.-M. Cheng, "SAMNet: Stereoscopically attentive multi-scale network for lightweight salient object detection," *IEEE Trans. Image Process.*, vol. 30, pp. 3804–3814, 2021.

[40] 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, Sep. 2020.

Authorized licensed use limited to: Kunming Univ of Science and Tech. Downloaded on November 22,2024 at 07:21:10 UTC from IEEE Xplore. Restrictions apply.

[41] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*.
[42] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "MobileNetV2: Inverted residuals and linear bottlenecks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 4510–4520.
[43] B. Liang and H. Luo, "MEANet: An effective and lightweight solution for salient object detection in optical remote sensing images," *Expert Syst. Appl.*, vol. 238, Mar. 2024, Art. no. 121778.
[44] D. Chai, J. Huang, M. Wu, X. Yang, and R. Wang, "Remote sensing image cloud detection using a shallow convolutional neural network," *ISPRS J. Photogramm. Remote Sens.*, vol. 209, pp. 66–84, Mar. 2024.
[45] A. Howard et al., "Searching for MobileNetV3," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, Nov. 2019, pp. 1314–1324.
[46] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Oct. 2018, pp. 7132–7141.
[47] C.-Y. Lee, S. Xie, P. Gallagher, Z. Zhang, and Z. Tu, "Deeply-supervised nets," in *Proc. 18th Int. Conf. Artif. Intell. Statist.*, vol. 38, 2015, pp. 562–570.
[48] G. Gao, G. Xu, Y. Yu, J. Xie, J. Yang, and D. Yue, "MSCFNet: A lightweight network with multi-scale context fusion for real-time semantic segmentation," *IEEE Trans. Intell. Transp. Syst.*, vol. 23, no. 12, pp. 25489–25499, Dec. 2022.
[49] Y. Wang et al., "LEDNet: A lightweight encoder–decoder network for real-time semantic segmentation," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Jul. 2019, pp. 1860–1864.
[50] S. Mehta, M. Rastegari, A. Caspi, L. Shapiro, and H. Hajishirzi, "ESPNet: Efficient spatial pyramid of dilated convolutions for semantic segmentation," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Sep. 2018, pp. 552–568.
[51] Q. Wang, B. Wu, P. Zhu, P. Li, W. Zuo, and Q. Hu, "ECA-Net: Efficient channel attention for deep convolutional neural networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 11534–11542.
[52] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *Proc. Int. Conf. Mach. Learn.*, 2015, pp. 448–456.
[53] K. He, X. Zhang, S. Ren, and J. Sun, "Delving deep into rectifiers: Surpassing human-level performance on ImageNet classification," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 1026–1034.
[54] Z. Huang, H.-X. Chen, T. Zhou, Y.-Z. Yang, C.-Y. Wang, and B.-Y. Liu, "Contrast-weighted dictionary learning based saliency detection for VHR optical remote sensing images," *Pattern Recognit.*, vol. 113, May 2021, Art. no. 107757.
[55] X. Wang, R. Girshick, A. Gupta, and K. He, "Non-local neural networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7794–7803.
[56] J.-X. Zhao, Y. Cao, D.-P. Fan, M.-M. Cheng, X.-Y. Li, and L. Zhang, "Contrast prior and fluid pyramid integration for RGBD salient object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 3927–3936.
[57] C. Li, R. Cong, Y. Piao, Q. Xu, and C. C. Loy, "RGB-D salient object detection with cross-modality modulation and selection," in *Proc. 16th Eur. Conf. Comput. Vis.*, Glasgow, U.K. Cham, Switzerland: Springer, 2020, pp. 225–241.
[58] N. Liu, N. Zhang, and J. Han, "Learning selective self-mutual attention for RGB-D saliency detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 13756–13765.
[59] Y. Piao, W. Ji, J. Li, M. Zhang, and H. Lu, "Depth-induced multi-scale recurrent attention network for saliency detection," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 7254–7263.
[60] K. Fu, D.-P. Fan, G.-P. Ji, Q. Zhao, J. Shen, and C. Zhu, "Siamese network for RGB-D salient object detection and beyond," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 9, pp. 5541–5559, Sep. 2022.
[61] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *J. Mach. Learn. Res.*, vol. 15, no. 56, pp. 1929–1958, 2014.
[62] X. Qin, Z. Zhang, C. Huang, C. Gao, M. Dehghan, and M. Jagersand, "BASNet: Boundary-aware salient object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 7479–7489.
[63] A. Paszke et al., "Pytorch: An imperative style, high-performance deep learning library," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 32, 2019, pp. 1–12.
[64] M. Guo et al., "Attention mechanisms in computer vision: A survey," *Comput. Vis. Media*, vol. 8, no. 3, pp. 331–368, 2022.
[65] 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.
[66] 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. Joint Conf. Artif. Intell.*, Jul. 2018, pp. 698–704.
[67] 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.
[68] F. Perazzi, P. Krähenbühl, Y. Pritch, and A. Hornung, "Saliency filters: Contrast based filtering for salient region detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2012, pp. 733–740.
[69] 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. 3917–3926.
[70] Y. Liu, D. Zhang, N. Liu, S. Xu, and J. Han, "Disentangled capsule routing for fast part-object relational saliency," *IEEE Trans. Image Process.*, vol. 31, pp. 6719–6732, 2022.
[71] Z. Huang, H. Chen, B. Liu, and Z. Wang, "Semantic-guided attention refinement network for salient object detection in optical remote sensing images," *Remote Sens.*, vol. 13, no. 11, p. 2163, May 2021.
[72] K. Huang, C. Tian, and G. Li, "Bidirectional mutual guidance transformer for salient object detection in optical remote sensing images," *Int. J. Remote Sens.*, vol. 44, no. 13, pp. 4016–4033, Jul. 2023.

Bei Cheng (Member, IEEE) received the Ph.D. degree in electronic information science and technology from Chongqing University, Chongqing, China, in 2022.

Now, she joined Kunming University of Science and Technology, Kunming, China. Her research interests include machine learning and its application in remote sensing data analysis, salient object detection in remote sensing images, complex scene classification, object detection and recognition, multitype and multisource image object detection, and continuous frame image object recognition and tracking.

Zao Liu is currently pursuing the M.S. degree with the School of Information Engineering and Automation, Kunming University of Science and Technology, Kunming, China.

His research interests include computer vision and saliency detection.

Qingwang Wang (Member, IEEE) received the B.E. and Ph.D. degrees in electronics and information engineering and information and communication engineering from Harbin Institute of Technology, Harbin, China, in 2014 and 2020, respectively.

From 2020 to 2021, he worked as a Senior Engineer with Huawei Technology Company Ltd., Shenzhen, China, to study autonomous driving. Now, he joined Kunming University of Science and Technology, Kunming, China, as a high-level talent. His research interests include machine learning and its application to remote sensing data analysis, autonomous driving, and edge calculation. More specially, his studies currently focus on using kernel methods, deep learning, broad learning, and graph convolutional neural networks to extract information from RGB-T images, hyperspectral images, LiDAR data, and multispectral LiDAR point clouds.

Authorized licensed use limited to: Kunming Univ of Science and Tech. Downloaded on November 22,2024 at 07:21:10 UTC from IEEE Xplore. Restrictions apply.

Tao Shen (Senior Member, IEEE) received the Ph.D. degree from Illinois Institute of Technology, Chicago, IL, USA, in 2013.

He is a Professor and the Dean of the College of Information Engineering and Automation, Kunming University of Science and Technology, Kunming, China. He has published more than 20 papers in first-class SCI/EI and other internationally famous journals and top international conferences in relevant research fields. His research interests cover artificial intelligence, multisource remote sensing data processing, blockchain technology, and the Industrial Internet.

Chengbiao Fu (Member, IEEE) received the B.S. and M.S. degrees from Chongqing University of Posts and Telecommunications, Chongqing, China, in 2005 and 2009, respectively, and the Ph.D. degree from Kunming University of Science and Technology, Kunming, China, in 2021.

Since 2020, he has been an Associate Professor with Kunming University of Science and Technology. His research interests are in information identification and processing, hyperspectral remote sensing, and machine learning.

Anhong Tian (Member, IEEE) received the B.S. and M.S. degrees from Chongqing University of Posts and Telecommunications, Chongqing, China, in 2007 and 2010, respectively, and the Ph.D. degree from Kunming University of Science and Technology, Kunming, China, in 2020.

Since 2021, she has been a Professor with Kunming University of Science and Technology. Her research interests are in hyperspectral remote sensing, nonlinear systems, and artificial intelligence.

Authorized licensed use limited to: Kunming Univ of Science and Tech. Downloaded on November 22,2024 at 07:21:10 UTC from IEEE Xplore. Restrictions apply.