Edge-Aware Multiscale Feature Integration Network for Salient Object Detection in Optical Remote Sensing Images

Xiaofei Zhou, Kunye Shen, Zhi Liu, *Senior Member, IEEE*, Chen Gong, *Member, IEEE*, Jiyong Zhang, and Chenggang Yan

***Abstract*— The optical remote sensing images (RSIs) show various spatial resolutions and cluttered background, where salient objects with different scales, types, and orientations are presented in diverse RSI scenes. Therefore, it is inappropriate to directly extend cutting-edge saliency detection methods for conventional RGB images to optical RSIs. Besides, the existing saliency models targeting RSIs often render imperfect saliency maps, where some of them are with coarse boundary details. To solve this problem, this article attempts to introduce the edge information to precisely detect salient objects in RSIs. Accordingly, we propose an edge-aware multiscale feature integration network (EMFI-Net) for salient object detection by conducting multiscale feature integration under the explicit and implicit assistance of salient edge cues. Specifically, our network contains two parts including the encoder and decoder. First, the encoder extracts multiscale deep features from three RSIs with different resolutions, where the high-level deep semantic features from three RSIs are integrated using a cascaded feature fusion module. Second, the encoder explicitly enriches the multiscale deep features by integrating the salient edge cues extracted by a salient edge extraction module. Meanwhile, we also implicitly deploy an edge-aware constraint to the supervision of the saliency map prediction by introducing a hybrid loss function. Finally, the decoder integrates the enriched multiscale deep features in a coarse-to-fine way, yielding a high-quality saliency map. The experiments conducted on two public optical RSI datasets clearly prove the effectiveness and superiority of the proposed EMFI-Net against the state-of-the-art saliency models.**

***Index Terms*— Multiscale deep features, optical remote sensing images, salient edge cues, salient object detection.**

Manuscript received December 4, 2020; revised April 6, 2021 and May 31, 2021; accepted June 15, 2021. This work was supported in part by the National Key Research and Development Program of China under Grant 2020YFB1406604; in part by the National Natural Science Foundation of China under Grant 61901145, Grant 61973162, Grant 61931008, Grant 61671196, Grant 62071415, Grant 62001146, Grant 61701149, Grant 61801157, Grant 61971268, Grant 61901150, and Grant 61972123; in part by the Fundamental Research Funds for the Central Universities under Grant 30920032202 and Grant 30921013114; in part by the "Young Elite Scientists Sponsorship Program" by CAST under Grant 2018QNRC001; in part by the Hong Kong Scholars Program under Grant XJ2019036; in part by the Zhejiang Province Nature Science Foundation of China under Grant LR17F030006; and in part by the 111 Project under Grant D17019. *(Corresponding author: Chenggang Yan.)*

This work did not involve human subjects or animals in its research.

Xiaofei Zhou, Kunye Shen, Jiyong Zhang, and Chenggang Yan are with the School of Automation, Hangzhou Dianzi University, Hangzhou 310018, China (e-mail: zxforchid@outlook.com; kunyeshen@outlook.com; jzhang@hdu.edu.cn; cgyan@hdu.edu.cn).

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

Chen Gong is with the PCA Laboratory, Key Laboratory of Intelligent Perception and Systems for High-Dimensional Information of Ministry of Education, School of Computer Science and Engineering, Nanjing University of Science and Technology, Nanjing 210094, China (e-mail: chen.gong@njust.edu.cn).

Digital Object Identifier 10.1109/TGRS.2021.3091312

I. Introduction

VISION system of humans tries to locate the most visually distinctive regions based on the visual attention mechanism [1], [2], which is the foundation of salient object detection. Recently, salient object detection has received wide attention around the world because of its successful applications in many research areas such as image/video segmentation [3]–[5], image/video compression [6], [7], image editing [8], image quality assessment [9], retargeting [10], visual categorization [11], and so on.

In 1998–2020, the main efforts of salient object detection mainly go through the computation of center-surround difference [12], [13], constructing feature-driven machine-learning system [14], [15], and building deep learning-based framework [16], [17]. Especially, the deep learning-based saliency models significantly elevate the performance of salient object detection. Obviously, the existing saliency models can be applied to conventional RGB images (natural scenes) [18], [19], RGB-D images [20]–[24], RGB-T images [25], videos [3], [26], [27], light-field images [28], and optical remote sensing images [29], [30]. Among them, because of the crucial roles in the military, agriculture, and disaster relief, optical remote sensing images have attracted an increasing attention recently. Here, we should note that the optical RSIs used by this article are different from the hyperspectral remote sensing images [31]–[36], which have more spectrum bands and try to acquire the spectrum of each pixel. However, on the one hand, there are only a small number of related [29], [30], [37]–[43] major works in performing salient object detection on optical RSIs, where their performance will degrade on some challenging scenes. On the other hand, the optical RSIs are photographed by the high-angle satellite, and thus they often show various scene patterns, as shown in Fig. 1. Generally, these patterns mainly include illumination variations, objects of various sizes, multiple salient objects, different object types, cluttered backgrounds, complex texture

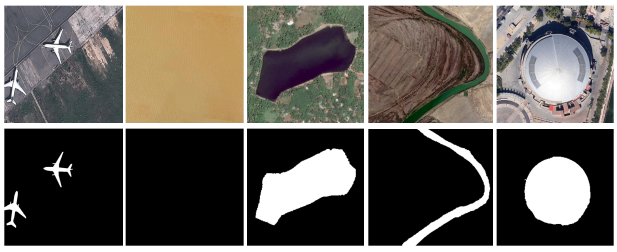


Fig. 1. Examples of optical RSIs: (top row) optical RSIs and (bottom row) GT.

structures, and even no salient objects. This also poses the barrier in achieving an encouraging performance by directly applying existing natural scene saliency models.

Motivated by the aforementioned descriptions about salient object detection in optical RSIs, we propose a novel edge-aware multiscale feature integration network (EMFI-Net) shown in Fig. 2, which is an encoder-decoder architecture network. To be specific, first, taking the scale diversity of salient objects in optical RSIs into consideration, we deploy three convolutional branches with the same structure and use three images of different resolutions as input. After that, the cascaded feature fusion module, marked in purple dotted box shown in Fig. 2, is deployed to combine the high-level features from three images with different resolutions progressively. Following this way, we can obtain the multiscale deep features, namely the multilevel deep features $\{\mathbf{F}_4^1, \mathbf{F}_3^1, \mathbf{F}_2^1, \mathbf{F}_1^1\}$ from the original image and the high-level deep semantic feature $\{\tilde{\mathbf{F}}_5^1\}$ from the cascaded feature fusion module shown in Fig. 2. In this process, the rich representational deep features not only depict the local detail of salient objects but also present the global context of salient objects.

Second, through a deep analysis, we find that most existing optical RSI saliency models (see [29], [30], [39]) ignore the effect of fine edges or boundary details in depicting salient objects, where the predicted saliency maps are often with low-quality boundary details. Meanwhile, to tackle the coarse boundary problem, the existing natural image saliency models (see [17], [44], [45]) either only deploy edge information to the stage of saliency inference or just embed it to loss functions. Differently, we try to sufficiently utilize the salient edge cues in both ways. On the one hand, we employ an edge module to combine the low- and high-level deep features, yielding the rich salient edge features, which are used to enrich the multiscale deep features explicitly. On the other hand, in the loss computation, we introduce the hybrid loss [44] containing edge-aware constraint to implicitly inject the fine edge information to saliency maps. Lastly, the decoder, i.e. the deep feature aggregation module equipped with a set of convolution and up-sampling operations, integrates the enriched multiscale deep features in a coarse-to-fine way, yielding high-quality saliency maps with complete structure, distinct details, and accurate boundaries.

Overall, our main contributions can be summarized as follows.

1) We propose a novel optical RSIs saliency model, namely EMFI-Net, which is an encoder-decoder architecture network including multiscale feature exaction, salient edge digging and integration, and deep feature aggregation.
2) The multiscale deep features from three images with different resolutions present salient objects with diverse scales in terms of local details and global context, and the salient edge cues endow the deep features with accurate boundary information in explicit and implicit ways.

The remaining of this article is organized as follows. The related works on salient object detection are reviewed in Section II. Section III gives a detailed description of the proposed EMFI-Net. In Section IV, comprehensive experiments and the detailed analysis are presented. Finally, the conclusion for this work is detailed in Section V.

II. Related Works

In past decades, various theories have been applied to build saliency models, and we have fortunately witnessed the booming research progress in salient object detection. In this part, we will first give a brief introduction of salient object detection in natural scene images, and then review some saliency models targeting optical remote sensing images.

A. Saliency Models for Natural Scene Images

The pioneering saliency model [1] proposed the well-known center-surround difference mechanism to locate salient objects. Following this way, in [12], saliency was defined as the difference between the current region and other regions. Meanwhile, some other theories are also applied to saliency computation. For example, in [46], the boundary connectivity-based saliency map depicted the background probability of each region. In [47], based on the boundary prior information, the saliency reversion correction method together with the regularized random walk ranking model were used to obtain high-quality saliency maps. Zhou *et al.* [48] applied the compactness concept to acquire two initial saliency maps. In recent years, many classical efforts are constructed based on traditional machine-learning methods. For example, Liu *et al.* [49] employed the conditional random field to fuse multiple saliency maps. Jiang *et al.* [14] utilized the random forest regressor to map the multiple region-level features to saliency scores. In [15], the Adaboost algorithm was exploited to perform an unsupervised saliency computation process. The multiple instance-learning theory together with a simple-to-complex optimization method were employed by Huang *et al.* [50] to predict saliency maps.

The deep-learning techniques have also been successfully applied to elevate the performance of salient object detection. For example, Li and Yu [16] employed three convolutional branches on three different resolution images to extract multiscale deep features, which are mapped to saliency values using a shallow neural network. Hou *et al.* [51] combined the multilevel deep features to generate saliency maps using the holistically nested edge detector architecture. Differently,

our network contains both kinds of multiscale deep features, namely the multilevel deep features from the original input image and the high-level deep semantic feature from the multiresolution input images. Interestingly, the recurrent structure has become popular in some deep saliency models. A recurrent residual refinement network [18] was designed to locate salient objects. Hu *et al.* [52] employed the multilevel deep features integrated from different layers to refine each layer's deep features in a recurrent way. By incorporating the saliency prior knowledge, Wang *et al.* [53] deployed a recurrent architecture to generate reliable saliency maps by iteratively correcting the previous errors. A cascaded partial decoder [54] was presented to perform fast and accurate saliency detection by discarding shallower layers' features and employing effective attention maps. However, it is very time consuming and laborious to prepare pixel-wise annotations for the network training. Therefore, Zhang *et al.* [55] proposed a supervision synthesis scheme-based framework to learn deep saliency model, which can be trained without human annotation. Successively, Han *et al.* [56] further proposed a weakly supervised learning framework to explore the object segmentation and category-specific 3D shape reconstruction.

Besides, some researchers pay attention to the local details and boundary quality of saliency maps. For instance, Liu *et al.* [17] proposed two pooling based modules to provide spatial contextual information and promote the fusion of multilevel deep features, respectively. Qin *et al.* [44] proposed an end-to-end predict-refine network to obtain accurate saliency maps by implicitly exploiting boundary information. After that, Qin *et al.* [57] designed a two-level nested U-shaped network equipped with Residual U-blocks to pop out salient objects. Zhao *et al.* [45] proposed an edge-aware network (i.e., EGNet) to explore the complementarity between salient edge and salient objects. The differences between EGNet and our model mainly focus on the details of the generation and the usage of edge features. First, in our model, the high-level features employed to acquire edge information are generated using a cascaded feature fusion module, which aggregates high-level features from three different resolution images. Therefore, the high-level features adopted by our model present more effective global context information than EGNet, which adopts the output of encoder's last convolutional layer as the high-level features. Second, the edge features in EGNet are just deployed to enrich the multiscale deep features, which are used to perform saliency prediction separately. Differently, the edge-enhanced multiscale deep features of our model are integrated into the final saliency map using the decoder in a progressive way, where the edge information will flow across different decoder blocks. Wu *et al.* [58] designed the stacked cross refinement network by simultaneously elevating the salient object and edge features. Li *et al.* [23] utilized the saliency-guided position-edge attention module to remit the edge blur problem. The differences between our model and [23] can be summarized as two aspects. First, our model employs the low- and high-level deep features to generate edge features, whereas in [23], five edge maps are generated by using each level of RGB-D features, modulated features, and up-sampled features. Second, in our model, the edge feature is concatenated with every level of deep feature. Differently, in [23], edge maps are used as an attention map, which are integrated with deep features by conducting element-wise multiplication. In [24], an attention-steered interweave fusion network was proposed to progressively integrate cross-modal and cross-level deep features, where the side outputs supervision was also employed. Particularly, in [24], they adopt deep supervision to three side outputs, which are employed as the feature selectors to weigh the features in the same convolutional block. Differently, our model deploys the supervision to more side outputs, which is only treated as supervision signals.

Compared with the existing saliency models, which target natural scene images and are unsuitable for directly detecting salient objects in optical RSIs, our model makes some special designs by sufficiently taking into account the complex scene patterns of optical RSIs. For example, the multiscale deep semantic features are extracted from three images with different resolutions, and this gives a more effective representation for the salient objects with different scales. After that, the cascaded feature fusion attempts to integrate the multiscale high-level semantic features, which further gives powerful global context information for diverse scene patterns. Particularly, compared with the edge-based saliency models, our model adequately utilizes the edge information in explicit and implicit ways, which gives precise detection for salient objects in optical RSIs.

B. Saliency Models for Optical RSIs

Although many efforts have been devoted to salient object detection in nature scene images, the research on optical RSIs saliency models is insufficient. There are only a small number of prior works on salient object detection in optical RSIs. For example, Zhao *et al.* [39] proposed a sparsity-guided saliency model to perform saliency map integration through the acquisition of the global and background cues. Ma *et al.* [40] presented a superpixel-to-pixel saliency model to detect regions of interest by using the texture and color features. Besides, there are also some works aiming at locating special salient targets. For instance, a two-step saliency estimation method [37] was proposed to locate buildings, where a probabilistic model was used to aggregate each building's multiple saliency cues. Zhang *et al.* [41] integrated the vision-oriented and knowledge-oriented saliency maps to accurately locate airports. In [38], the color features-based and the radial symmetric circle-based feature maps are combined to detect the oil tank.

Furthermore, the recently published works have pushed forward the progress of this area to some degree. For example, Zhang *et al.* [43] proposed a self-adaptively feature fusion model to fuse multiple saliency cues including the color, intensity, texture, and global contrast using the low-rank matrix recovery method. Dong *et al.* [59] designed the multiscale pyramid architecture to generate saliency maps, which were further used to conduct graph-based segmentation. Li *et al.* [29] proposed an end-to-end LV-Net, which consists of a two-stream pyramid module and an encoder-decoder

module, to detect salient objects in optical RSIs. Meanwhile, this work provides a public optical RSIs dataset. A parallel down-up fusion network proposed by Li *et al.* [30] sufficiently utilized the in/cross-path information and the multiresolution features to detect salient objects in optical RSIs. In [60], an attention guided feature-learning architecture was deployed to perform salient object detection, and the authors also built a large-scale optical RSIs dataset. In addition, there are also some models targeting object detection in remote sensing images. In [61], a dynamic curriculum learning strategy was employed to progressively learn the object detectors, where the difficulty of training examples is ranked with the entropy-based image difficulty measure criterion. To deal with the challenging scene such as rotation variations and appearance ambiguity, Li *et al.* [62] designed the region proposal network (RPN) and local-contextual feature fusion network to extract the proposals and locate the geospatial objects.

Among the existing saliency models aiming at processing optical RSIs, some of the previous works only treat the saliency detection as an auxiliary unit for related vision tasks. Some other previous works including the traditional models and the deep learning-based models are less effective in some challenging scenes, where they either fail to sufficiently utilize the edge information or ignore the effect of edge information. By contrast, our model pays more attention on the edge information. This way not only enhances the multiscale deep features but also acts as constraints for establishing supervision. Through the multiscale feature integration under the explicit and implicit assistance of salient edge cues, the generated saliency maps are endowed with complete objects, accurate boundaries, and distinct local details.

III. The Proposed Method

In this section, the architecture of the proposed EMFI-Net is first introduced in Section III-A. Then, some key operations such as the multiscale feature extraction, salient edge extraction and integration, and deep feature aggregation will be described in Sections III-B–III-D, respectively.

A. Overall Architecture

The architecture of the proposed EMFI-Net is shown in Fig. 2, which is an encoder-decoder structure network consisting of multiscale feature extraction, salient edge extraction and integration, and deep feature aggregation. To be specific, three images $\{\mathbf{I}^i\}_{i=1}^3$ of different resolutions are first passed into the multiscale feature extraction module, in which three parallel convolutional branches with the same structure followed by a cascaded feature fusion module try to relieve the obstacle caused by the scale diversity of salient objects. Following this way, we can obtain the multiscale deep features $\{\mathbf{F}_1^1, \mathbf{F}_2^1, \mathbf{F}_3^1, \mathbf{F}_4^1, \tilde{\mathbf{F}}_5^1\}$, where the multiresolution high-level deep features $\{\mathbf{F}_5^i\}_{i=1}^3$ can be obtained from three different resolution images and they are fused to generate the high-level deep semantic feature $\tilde{\mathbf{F}}_5^1$ using the cascaded feature fusion module. Then, the low-level deep feature $\{\mathbf{F}_2^1\}$ and the high-level deep feature $\{\tilde{\mathbf{F}}_5^1\}$ are combined to acquire the salient edge cues $\{\mathbf{E}_i\}_{i=1}^5$ using an edge module, i.e., the salient edge extraction module shown in Fig. 2. After that, the multiscale deep features $\{\mathbf{F}_1^1, \mathbf{F}_2^1, \mathbf{F}_3^1, \mathbf{F}_4^1, \tilde{\mathbf{F}}_5^1\}$ are tamped by integrating with the salient edge cues, yielding the enriched multiscale deep features $\{\mathbf{F}_j^E\}_{j=1}^5$. Next, the decoder, namely the deep feature aggregation module, progressively integrates the multiscale deep features using a set of convolution and up-sampling operations, yielding the high-quality saliency map $\mathbf{S}$ with complete structure, accurate boundary, and distinct details. Besides, we also introduce the hybrid loss to implicitly endow the saliency maps with well-defined boundaries. In the following, we will give a detailed description for each of the components.

B. Multiscale Feature Extraction

Salient objects in optical RSIs usually show various sizes, which include both tiny object like ship on the ocean and large objects like stadium roof. This phenomenon will lead to the performance degradation of saliency models. To confront the scale diversity problem, we deploy the multiscale feature extraction module, which contains three parallel convolutional branches with the same structure and a cascaded feature fusion module.

Formally, we first downsample the original image $\mathbf{I}^1$ by factors 2 and 4, generating other two images $\{\mathbf{I}^2, \mathbf{I}^3\}$. Then, the three images $\{\mathbf{I}^1, \mathbf{I}^2, \mathbf{I}^3\}$ with different resolutions are passed to three parallel convolutional branches with the same structure, as shown in Fig. 2. Concretely, each convolutional branch containing five convolutional blocks Conv-Bi ($i = 1, \ldots, 5$) is constructed based on ResNet-34 [63], which embeds the residual learning to each pair of 3×3 convolutional layers by using shortcut connections. It should be noted that the convolutional branch in EMFI-Net is slightly different from that in ResNet-34. To be specific, the convolutional layer (kernel size $= 7 \times 7$, channel $= 64$, stride $= 2$) in "conv1" of ResNet-34 is replaced with a convolutional layer with 3×3 kernel size, 64 channels, and 1 stride. In addition, the max pooling layer is abolished after "conv1." Therefore, in EMFI-Net, we set the "conv1" and "conv2_x" as the first convolutional block Conv-B1. After that, Conv-B2, Conv-B3, and Conv-B4 adopt the "conv3_x," "conv4_x," and "conv5_x" of ResNet-34, respectively. Moreover, to further enlarge the receptive field of our network, we deploy a max pooling layer (kernel size $= 2 \times 2$, stride $= 2$, padding $= 0$) and three basic res-blocks (channel $= 512$) after Conv-B4, and these layers constitute the convolutional block Conv-B5 of EMFI-Net. Following this architecture, we can obtain high-level deep semantic features $\{\mathbf{F}_5^1, \mathbf{F}_5^2, \mathbf{F}_5^3\}$ from three RSIs $\{\mathbf{I}^k\}_{k=1}^3$ and multilevel deep features $\{\mathbf{F}_1^1, \mathbf{F}_2^1, \mathbf{F}_3^1, \mathbf{F}_4^1\}$ from the original image $\mathbf{I}^1$. Subsequently, to aggregate the three deep semantic features, we deploy a cascaded feature fusion module shown in purple dotted box of Fig. 2.

Specifically, at first, the deep feature $\mathbf{F}_5^3$ is passed into a convolutional block Conv (i.e., three convolutional layers), yielding the enhanced deep feature $\tilde{\mathbf{F}}_5^3$, namely

$$\tilde{\mathbf{F}}_5^3 = \mathrm{Conv}\left(\mathbf{F}_5^3\right) \tag{1}$$

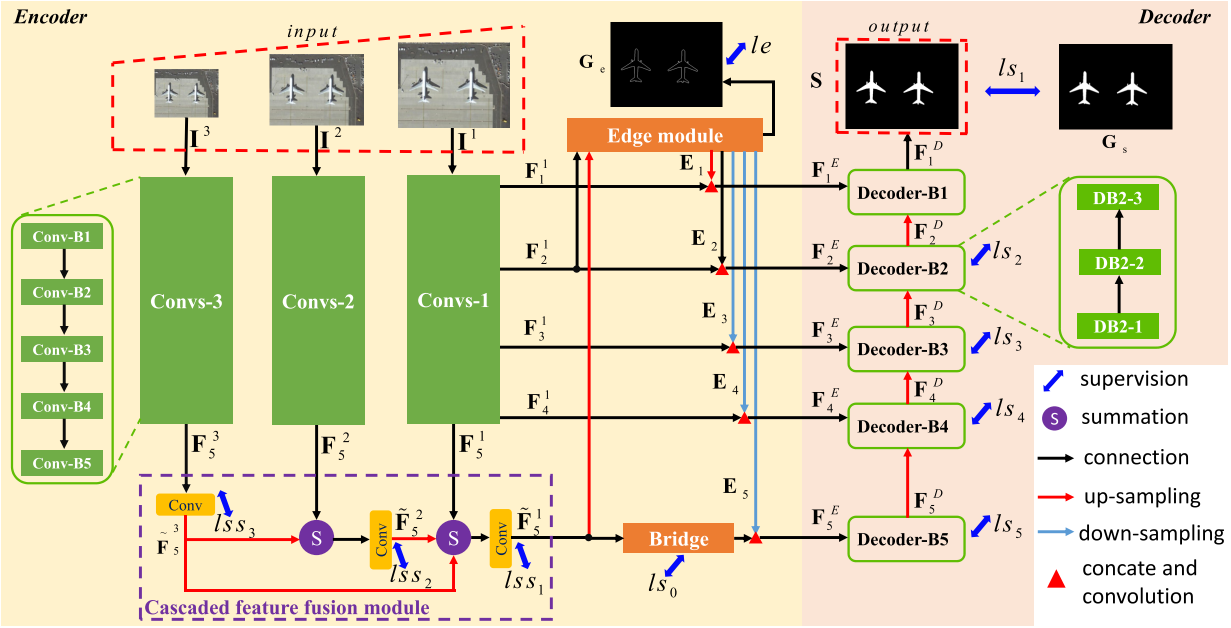


Fig. 2. Architecture of the proposed EMFI-Net: the input are three different resolution optical remote sensing images $\{\mathbf{I}^1, \mathbf{I}^2, \mathbf{I}^3\}$, and the output is the saliency map $\mathbf{S}$, which are all marked in dotted red boxes. The overall network consists of the encoder and decoder. To be specific, the encoder is used to extract multiscale deep features $\{\tilde{\mathbf{F}}_5^1, \mathbf{F}_4^1, \mathbf{F}_3^1, \mathbf{F}_2^1, \mathbf{F}_1^1\}$ and obtain salient edge cues $\{\mathbf{E}_i\}_{i=1}^5$, which are further applied to enhance the deep features (yielding the enriched deep features $\{\mathbf{F}_i^E\}_{i=1}^5$). The decoder is employed to aggregate the multiscale deep features in a coarse-to-fine way, and we can obtain the high-quality saliency map $\mathbf{S}$ for the original image $\mathbf{I}^1$. Here, the supervision (blue arrows) of edge module, decoder blocks, bridge module, and three convolutional branches of encoder is indicated by le, $\{\text{ls}_i\}_{i=1}^5$, ls_0, and $\{\text{lss}\}_{i=1}^3$, respectively.

where each convolutional layer (kernel size $= 3 \times 3$, stride $= 1$) in Conv is followed by a batch normalization (BN) layer and a ReLU layer.

Second, the enhanced deep feature $\tilde{\mathbf{F}}_5^3$ is up-sampled to the same size as $\mathbf{F}_5^2$ using bilinear interpolation. Then, both deep features are combined and sent into a convolutional block Conv. The process can be defined as

$$\tilde{\mathbf{F}}_5^2 = \text{Conv}\left(\text{up}_{\times 2}\left(\tilde{\mathbf{F}}_5^3\right) + \mathbf{F}_5^2\right) \tag{2}$$

where "$\text{up}_{\times 2}(.)$" denotes $2\times$ up-sampling operation by performing the bilinear interpolation and "$+$" is the element-wise summation operation.

Lastly, the deep semantic feature $\mathbf{F}_5^1$ together with the two enhanced deep features $\tilde{\mathbf{F}}_5^2$ and $\tilde{\mathbf{F}}_5^3$ are also sent into a convolution block Conv, yielding the enhanced deep feature $\tilde{\mathbf{F}}_5^1$, which can be formulated as

$$\tilde{\mathbf{F}}_5^1 = \text{Conv}\left(\text{up}_{\times 4}\left(\tilde{\mathbf{F}}_5^3\right) + \text{up}_{\times 2}\left(\tilde{\mathbf{F}}_5^2\right) + \mathbf{F}_5^1\right) \tag{3}$$

where "$\text{up}_{\times 4}(.)$" refers to $4\times$ up-sampling operation by using the bilinear interpolation.

Following this way, we can aggregate the three high-level semantic deep features $\{\mathbf{F}_5^i\}_{i=1}^3$, generating the deep feature $\tilde{\mathbf{F}}_5^1$. In the following, the multiscale (or multilevel) deep features $\{\mathbf{F}_1^1, \mathbf{F}_2^1, \mathbf{F}_3^1, \mathbf{F}_4^1, \tilde{\mathbf{F}}_5^1\}$ will be improved by the salient edge module.

C. Salient Edge Extraction and Integration

Through a thorough review of existing optical RSI saliency models, we can find that the usage of boundary information is insufficient in current works. Therefore, we try to take full advantage of salient edge cues in both explicit and implicit ways, which enable our model to generate clear boundary details.

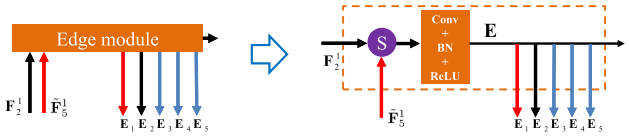


Fig. 3. Illustration of the edge module, where S means pixel-wise summation, the red line denotes up-sampling operation, the black line means connectivity, and the blue line refers to down-sampling operation.

Formally, the shallow layer features are able to depict rich spatial details such as edge information, whereas the deep layer features try to capture the semantic knowledge. Here, similar as the saliency model [45] for natural scene images, we first combine the low- and high-level deep features $\mathbf{F}_2^1$ and $\tilde{\mathbf{F}}_5^1$ to generate the salient edge cue $\mathbf{E}$. According to Figs. 2 and 3, the whole process can be formulated as

$$\mathbf{E} = f\left(\mathbf{F}_2^1 + \text{up}_{\times 8}\left(\tilde{\mathbf{F}}_5^1\right)\right) \tag{4}$$

where "$\text{up}_{\times 8}(.)$" means $8\times$ up-sampling, i.e. bilinear interpolation shown in red line, and f denotes the function of the combination of convolution layer, the BN layer and the ReLU layer, as shown in Fig. 3.

Then, we try to deploy the salient edge cue $\mathbf{E}$ to enrich multiscale deep features $\{\mathbf{F}_1^1, \mathbf{F}_2^1, \mathbf{F}_3^1, \mathbf{F}_4^1, \tilde{\mathbf{F}}_5^1\}$. Concretely, the salient edge cue $\mathbf{E}$ is first resized to the same size as the multiscale deep features using up-sampling or down-sampling operation, yielding the corresponding salient

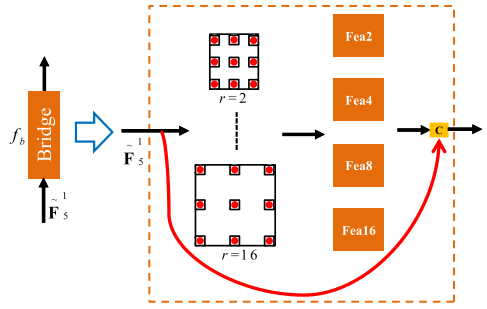


Fig. 4. Illustration of the bridge module, where f_b represents the function of bridge module and "C" denotes the concatenation operation.

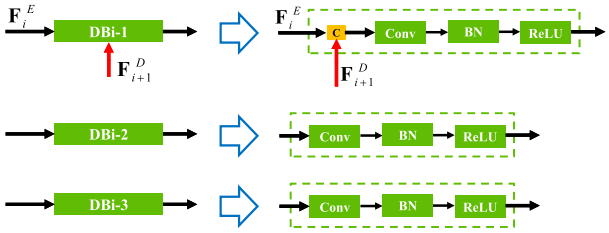


Fig. 5. Illustration of the ith decoder block "Decoder-Bi," where the red line denotes the up-sampling operation and "C" refers to the concatenation operation.

edge cues $\{\mathbf{E}_j\}_{j=1}^{5}$. Here, the jth edge cue $\mathbf{E}_j$ is with the same spatial size as the jth deep feature $\mathbf{F}_j^1$, and the corresponding sampling rate sr_j is set to $2^{|j-2|}$. Then, $\mathbf{E}_j$ is delivered to integrate with the multiscale deep features, namely

$$\mathbf{F}_j^E = \begin{cases} \mathrm{Conv}\left(\left[\mathbf{F}_j^1, \mathbf{E}_j\right]\right) & j = 1, 2, 3, 4 \\ \mathrm{Conv}\left(\left[f_b\left(\tilde{\mathbf{F}}_j^1\right), \mathbf{E}_j\right]\right) & j = 5 \end{cases} \tag{5}$$

where $\mathbf{F}_j^E$ denotes the enriched deep feature, Conv means the convolution operation after concatenation operation "$[\cdot,\cdot]$," and $f_b(\cdot)$ represents the function of bridge module shown in Fig. 2. Particularly, to enlarge the receptive field and capture the powerful global context, we add the bridge module shown in Fig. 2. To be specific, similar to the Atrous spatial pyramid pooling (ASPP) [64], the bridge module shown in Fig. 4 first deploys four parallel dilated convolution with dilation rates $r = \{2, 4, 8, 16\}$, and then the input of bridge module $\tilde{\mathbf{F}}_5^1$ and the obtained four deep features {Fea2, Fea4, Fea8, Fea16} are combined using the concatenation operation.

Meanwhile, to depict the salient edge cue $\mathbf{E}$ accurately, we employ the supervision to guide the salient edge extraction. Concretely, the salient edge cue $\mathbf{E}$ is first processed by a convolutional layer (kernel size $= 1 \times 1$, stride $= 1$, padding $=$ 0, and channel $= 1$), a bilinear interpolation layer ($2\times$ up-sampling), and a sigmoid function. This operation is denoted by $f_c(\cdot)$ and is used to obtain the salient edge map. Then, we adopt the cross-entropy loss to realize the supervision, namely

$$\begin{aligned} \mathrm{le} = -\sum_{i=1}^{W\times H} \{&\mathbf{GT}_{e_+}(i)\log(f_c(\mathbf{E})(i)) \\ &+ \mathbf{GT}_{e_-}(i)\log(1 - f_c(\mathbf{E})(i))\} \end{aligned} \tag{6}$$

where $\mathbf{GT}_{e_+}$ and $\mathbf{GT}_{e_-}$ refer to salient edge pixels and background pixels, respectively. Here, we generate the ground truth (GT) of salient edge $\mathbf{GT}_e$ by following [45]. Specifically, the gradient magnitude of salient object GT $\mathbf{GT}_s$ is first computed, and then we set the value of the pixels with nonzero gradient amplitude to 1. Therefore, the values of salient edge pixels $\mathbf{GT}_{e_+}$ are 1 and the values of background pixels $\mathbf{GT}_{e_-}$ are 0.

Lastly, except for the explicit usage of salient edge cues in the proposed EMFI-Net, the implicit utilization of edge information is also applied to further model the proposed EMFI-Net with rich edge details. Specifically, in the computation of loss functions (this will be detailed in Section III-D) shown in Fig. 2, we introduce the hybrid loss [44], which is defined as

$$\begin{cases} \mathrm{ls} = l_{\mathrm{bce}} + l_{\mathrm{ssim}} + l_{\mathrm{iou}} \\ l_{\mathrm{bce}} = -\sum_{i=1}^{W\times H} \{\mathbf{GT}_{s_+}(i)\log(\mathbf{S}(i)) \\ \qquad\qquad + \mathbf{GT}_{s_-}(i)\log(1 - \mathbf{S}(i))\} \\ l_{\mathrm{ssim}} = 1 - \dfrac{(2u_p u_{\mathrm{gt}} + \Omega_1)(2\delta_{p,\mathrm{gt}} + \Omega_2)}{\left(u_p^2 + u_{\mathrm{gt}}^2 + \Omega_1\right)\left(\delta_p^2 + \delta_{\mathrm{gt}}^2 + \Omega_2\right)} \\ l_{\mathrm{iou}} = 1 - \dfrac{\sum_{i=1}^{W\times H} \mathbf{S}(i)\mathbf{GT}_s(i)}{\sum_{i=1}^{W\times H}[\mathbf{S}(i) + \mathbf{GT}_s(i) - \mathbf{S}(i)\mathbf{GT}_s(i)]} \end{cases} \tag{7}$$

where l_{ssim}, l_{bce}, and l_{iou} refer to SSIM loss [65], BCE loss [66], and IoU loss [67], respectively. Besides, in the computation of l_{ssim}, u_p, and u_{gt}, δ_p and δ_{gt} denote the mean and the standard deviations of patch regions $\boldsymbol{p} = (p_1, \ldots, p_i, \ldots, p_N)$ and $\mathbf{gt} = (\mathrm{gt}_1, \ldots, \mathrm{gt}_i, \ldots, \mathrm{gt}_N)$, which are cropped from the generated saliency map $\mathbf{S}$ and the GT $\mathbf{GT}_s$. Here, N is the patch size, and p_i and gt_i are the ith pixel values of $\boldsymbol{p}$ and $\mathbf{gt}$, $\delta_{p,\mathrm{gt}}$ is the covariance of $\boldsymbol{p}$ and $\mathbf{gt}$, and Ω_1 and Ω_2 are usually set to 10^{-4} and 9×10^{-4}, respectively. In addition, among them, BCE loss aims to give a smooth gradient for each pixel, IoU loss pays more attention to salient regions, while the SSIM loss pays more attention to boundary pixels by incorporating the neighboring pixels' effects on them. Based on the hybrid loss, we deploy the implicit modeling of salient edges, where the predicted saliency maps can be enhanced with clear and accurate boundary details.

D. Deep Feature Aggregation

With the enriched multiscale deep features $\{\mathbf{F}_1^E, \mathbf{F}_2^E, \mathbf{F}_3^E, \mathbf{F}_4^E, \mathbf{F}_5^E\}$, the following crucial issue is how to effectively make a fusion for these features. Here, our model, which adopts the encoding-decoding architecture, treats the deep feature fusion as a decoding process and tries to integrate multiscale deep features in a progressive way.

Formally, according to Fig. 2, the decoder, namely deep feature aggregation module, contains five decoder blocks "Decoder-Bi" ($i = 1, \ldots, 5$), in which each of them consists of three convolutional blocks, i.e. DBi-1, DBi-2, DBi-3. For each convolutional block shown in Fig. 5, it contains a convolutional layer, a BN layer, and a ReLU layer. Correspondingly,

according to Fig. 2, the decoding process can be defined as

$$\mathbf{F}_i^D = \begin{cases} f_{\mathrm{di}}\left(\mathbf{F}_i^E\right) & i = 5 \\ f_{\mathrm{di}}\left(\left[\mathbf{F}_i^E, \mathrm{up}_{\times 2}\left(\mathbf{F}_{i+1}^D\right)\right]\right) & i = 1, 2, 3, 4 \end{cases} \tag{8}$$

where $\mathbf{F}_i^D$ denotes the output of the ith decoder block, f_{di} denotes the function of the ith decoder block "Decoder-Bi," and "$\mathrm{up}_{\times 2}(.)$" denotes the $2\times$ up-sampling operation by executing the bilinear interpolation, which is marked in red line shown in Fig. 5. Finally, to obtain the saliency map $\mathbf{S}$ shown in Fig. 2, the output of "Decoder-B1" $\mathbf{F}_1^D$ is further processed by a convolutional layer (kernel size $= 1 \times 1$, stride $= 1$, padding $= 0$, and channel $= 1$) and a sigmoid function. Here, to present each decoder block in a simple and convenient way, we do not draw this convolutional layer in Fig. 2.

Besides, the deeply supervised architecture has been successfully deployed by some saliency models [45], [51], [68]. Inspired by this, we also add the deep supervision to all decoder blocks by using the hybrid loss [44], namely $\{\mathrm{ls}_i\}_{i=1}^5$ shown in Fig. 2, where the side output of each decoder block can be generated by using a convolutional layer (kernel size $= 1 \times 1$, stride $= 1$, padding $= 0$, and channel $= 1$), an up-sampling layer (bilinear interpolation), and a sigmoid function. Here, the up-sampling layer is used to resize the output of the convolutional layer to the same size as the original input image $\mathbf{I}^1$. Meanwhile, we also employ the hybrid loss shown in Eq. 7 to form supervision for training the multiscale feature extraction module (i.e., the side outputs of the three parallel branches) and the bridge module shown in Fig. 2. Therefore, the total loss $\mathcal{L}$ of the proposed EMFI-Net can be formulated as

$$\mathcal{L} = \mathrm{le} + \sum_{i=1}^{3} \mathrm{lss}_i + \sum_{i=0}^{5} \mathrm{ls}_i \tag{9}$$

where ls_0 denotes the supervision of bridge module and $\{\mathrm{lss}_i\}_{i=1}^3$ denote the supervision of multiscale feature extraction module, as depicted in Fig. 2. Notice that the side outputs of multiscale feature extraction module are also generated using a convolutional layer (kernel size $= 1 \times 1$, stride $= 1$, padding $= 0$, and channel $= 1$), an up-sampling layer (bilinear interpolation), and a sigmoid function.

Following this way, the enriched multiscale deep features containing low-level deep spatial details and the high-level deep semantic information are aggregated in a coarse-to-fine way, and we can obtain the high-quality saliency map with complete structure, accurate boundary, and distinct details indicated in Fig. 2.

IV. Experimental Results

In this section, we first present the public optical remote sensing image dataset and the implementation details in Section IV-A. Second, the evaluation metrics are detailed in Section IV-B. Third, in Section IV-C, we will make some comparisons between the proposed EMFI-Net and the state-of-the-art saliency models. Fourth, the ablation study will be presented in Section IV-D. Lastly, in Section IV-E, we present the failure cases and analysis.

A. Datasets and Implementation

To comprehensively validate our model, we conduct extensive comparisons on two public benchmark optical remote sensing image datasets, namely ORSSD dataset [29] and EORSSD dataset [60]. Specifically, the ORSSD dataset contains 800 images, where 600 images are treated as the training set and 200 images are used for testing. This dataset exhibits various spatial resolutions, numerous object scales and types, cluttered background, and so on. Meanwhile, as an extension of the ORSSD, EORSSD contains 2000 images, where 1400 images are used for training and 600 images are adopted for testing. Notably, each image in the ORSSD and EORSSD datasets is furnished with pixel-wise annotation.

Here, according to [29], [30], [60], we employ the same training set, namely 600 images in ORSSD dataset and 1400 images in EORSSD dataset, to train our model. Besides, we also adopt 200 images in ORSSD dataset and 600 images in EORSSD dataset to constitute the test set. Furthermore, to train the proposed EMFI-Net, the training set is augmented by performing rotation with angles 90°, 180°, and 270° and conducting mirror reflection on those images. By following this way, the training set of ORSSD and EORSSD contains totally 4800 examples and 11 200 examples, respectively. In addition, each training image is resized to 256×256 during the training phase.

Our model is implemented with PyTorch on a PC with an Intel(R) Core(TM) i9-9900X 3.50 GHz CPU, 32 GB RAM, and an NVIDIA GTX 2080Ti GPU, in which parts of the encoder are initialized by using ResNet-34 [63] and the remaining parts are initialized by Xavier [69]. Besides, the proposed EMFI-Net can be trained in an end-to-end manner, and the Adam algorithm [70] is adopted to optimize the network, where the initial learning rate, batch size, and maximum epoch number are set to 10^{-4}, 4, and 130, respectively.

B. Evaluation Metrics

To quantitatively make a comparison for different saliency models on ORSSD and EORSSD datasets, we adopt the following evaluation metrics including precision-recall (PR) curve, F-measure curve, max F-measure (maxF), S-measure (S) [71], max E-measure (maxE) [72], and mean absolute error (MAE).

Precision and Recall is a standard metric to evaluate the model performance, where we totally compute 256 pairs of average precision value and average recall value over all saliency maps using the thresholds ranging from 0 to 255. Here, we plot the PR curve, where the vertical and horizontal axes denote precision value and recall value, respectively. Particularly, the closer the PR curve is to the coordinates (1, 1), the better the performance of the saliency model is.

F-measure is regarded as a comprehensive metric, which can be obtained by performing the weighted harmonic average on precision and recall values, namely

$$F_\beta = \frac{(1+\beta^2)\mathrm{Precision} \times \mathrm{Recall}}{\beta^2 \mathrm{Precision} + \mathrm{Recall}} \tag{10}$$

where β^2 is set to 0.3 to give more emphasis on precision as suggested in [73]. In this article, we report the max F-measure

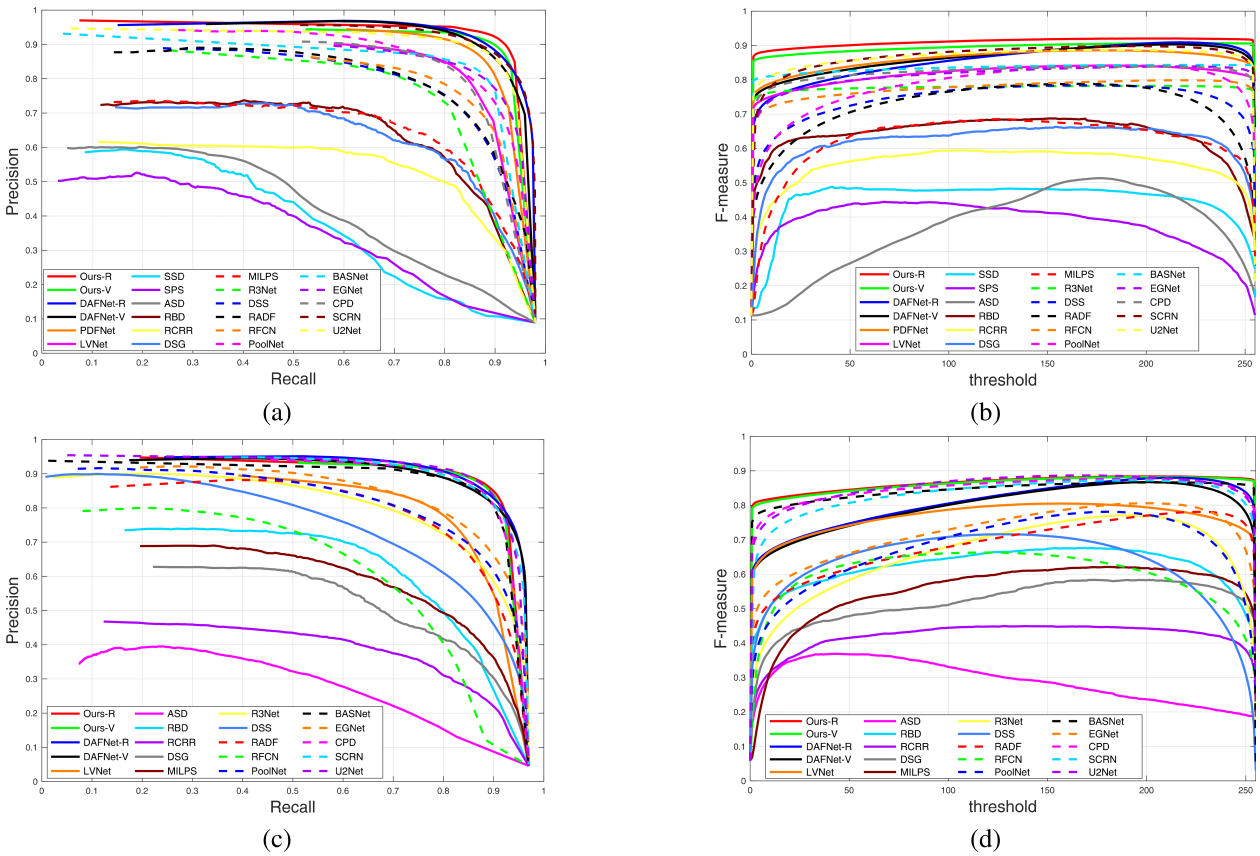


Fig. 6. (better viewed in color) Quantitative evaluation of different saliency models. (a) PR curves. (b) F-measure curves on ORSSD dataset. (c) PR curves. (d) F-measure curves on EORSSD dataset.

and show the F-measure curve simultaneously. Specially, for the F-measure curve, it can be drawn based on the pair of F score and threshold ([0, 255]), where each F score is computed by using (10) under each threshold. And thus, the larger the coordinate area covered by the F-measure curve, the better performance of the saliency model is.

MAE presents the absolute pixel-wise difference between the saliency map $\mathbf{S}$ and its corresponding GT $\mathbf{GT}_s$, which can be written as

$$\mathrm{MAE} = \frac{1}{W \times H} \sum_{i=1}^{W \times H} |\mathbf{S}(i) - \mathbf{GT}_s(i)| \tag{11}$$

where W and H represent the width and height of the saliency map, respectively.

S-measure measures the structural similarity of saliency maps, which simultaneously incorporates the region similarity (S_r) and the object similarity (S_o). The definition is presented as

$$S = \alpha \times S_o + (1 - \alpha) \times S_r \tag{12}$$

where α is set to 0.5 as suggested in [71].

E-measure evaluates the similarity between the predicted saliency maps and the GT by incorporating the local pixel saliency value and the image-level mean saliency value simultaneously. According to [72], the E-measure is formulated as

$$\xi = \frac{2\varphi_{\mathrm{GT}}(x, y) \circ \varphi_{\mathrm{FM}}(x, y)}{\varphi_{\mathrm{GT}}(x, y) \circ \varphi_{\mathrm{GT}}(x, y) + \varphi_{\mathrm{FM}}(x, y) \circ \varphi_{\mathrm{FM}}(x, y)}$$
$$E = \frac{1}{W \times H} \sum_{x=1}^{W} \sum_{y=1}^{H} f(\xi) \tag{13}$$

where $f(\cdot)$ is a convex function and $\circ$ refers to the Hadamard product. Besides, the alignment matrix ξ is constructed on the bias matrices φ_{GT} and φ_{FM}, which can be regarded as the centering operation on GT and binary saliency map, respectively.

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

In this section, we compare our model EMFI-Net denoted as "Ours" with 20 state-of-the-art saliency models on ORSSD and EORSSD datasets, in which there are six saliency models for optical RSIs (i.e., DAFNet [60], PDFNet [30], LVNet [29], SSD [39], SPS [40], ASD [41]), four traditional saliency models for natural scene RGB images (i.e., RBD [46], RCRR [47], DSG [48], MILPS [50]), and ten deep learning-based saliency models for natural scene RGB images (i.e., R3Net [18], DSS [51], RADF [52], RFCN [53], PoolNet [17], CPD [54], BASNet [44], EGNet [45], SCRN [58], andU2Net [57]). Meanwhile, for a fair comparison, saliency maps are generated by running the codes released or provided by the authors. All networks are trained from scratch on the training set of

TABLE I

QUANTITATIVE COMPARISON RESULTS OF S-MEASURE, MAX F-MEASURE, MAX E-MEASURE, AND MAE ON THE ORSSD AND EORSSD DATASETS. HERE, "↑" (↓) MEANS THAT THE LARGER (SMALLER) THE BETTER. THE BEST THREE RESULTS IN EACH ROW ARE MARKED IN RED, GREEN, AND BLUE, RESPECTIVELY

	ORSSD Dataset				EORSSD Dataset			
	$S\uparrow$	$F_\beta\uparrow$	$E_\beta\uparrow$	$MAE\downarrow$	$S\uparrow$	$F_\beta\uparrow$	$E_\beta\uparrow$	$MAE\downarrow$
DAFNet-V [60]	**0.9191**	0.8928	**0.9771**	0.0113	0.9166	0.8612	**0.9859**	**0.0060**
DAFNet-R [60]	0.9188	**0.8999**	**0.9821**	**0.0106**	**0.9184**	**0.8734**	**0.9815**	**0.0053**
PDFNet [30]	0.9112	0.8726	0.9608	0.0149	-	-	-	-
LVNet [29]	0.8815	0.8263	0.9456	0.0207	0.8644	0.7824	0.9279	0.0145
SSD [39]	0.5838	0.4460	0.7052	0.1126	-	-	-	-
SPS [40]	0.5758	0.3820	0.6472	0.1233	-	-	-	-
ASD [41]	0.5477	0.4701	0.7448	0.2119	0.5662	0.3054	0.6464	0.0563
RBD [46]	0.7662	0.6579	0.8501	0.0626	0.7409	0.6338	0.8238	0.0465
RCRR [47]	0.6849	0.5591	0.7651	0.1277	0.6013	0.4021	0.6898	0.1644
DSG [48]	0.7195	0.6238	0.7912	0.1041	0.6428	0.5259	0.7276	0.1246
MILPS [50]	0.7361	0.6519	0.8265	0.0913	0.6679	0.5749	0.7786	0.0958
R3Net [18]	0.8141	0.7456	0.8913	0.0399	0.8192	0.7516	0.9500	0.0171
DSS [51]	0.8262	0.7467	0.8860	0.0363	0.7874	0.6868	0.9205	0.0186
RADF [52]	0.8259	0.7619	0.9130	0.0382	0.8188	0.7473	0.9162	0.0168
RFCN [53]	0.8437	0.7742	0.9157	0.0293	0.7573	0.6290	0.8605	0.0236
PoolNet [17]	0.8551	0.8229	0.9368	0.0293	0.8217	0.7575	0.9318	0.0210
BASNet [44]	0.8963	0.8282	0.9346	0.0204	0.9191	0.8544	0.9567	0.0089
EGNet [45]	0.8774	0.8187	0.9165	0.0308	0.8601	0.7880	0.9570	0.0110
CPD [54]	0.8627	0.8033	0.9115	0.0297	0.9174	0.8627	0.9611	0.0114
SCRN [58]	0.9061	0.8846	0.9647	0.0157	0.8944	0.8544	0.9601	0.0142
U2Net [57]	0.9162	0.8738	0.9539	0.0166	0.9199	0.8732	0.9649	0.0076
Ours-V	**0.9366**	**0.9002**	0.9737	**0.0109**	**0.9290**	0.8720	0.9711	0.0084
Ours-R	**0.9432**	**0.9155**	**0.9813**	**0.0095**	**0.9319**	**0.8742**	**0.9712**	**0.0075**

the employed datasets. In the following, we will present the quantitative comparison as well as the qualitative comparison. Here, the results of SSD [39], SPS [40], and PDFNet [30] on EORSSD dataset are not provided by the authors, and thus the corresponding quantitative and qualitative results are not presented in this article.

1) Quantitative Comparison: To conduct quantitative evaluations on ORSSD and EORSSD datasets, we first present PR curves and F-measure curves in Fig. 6. Here, similar as the recently published RSIs saliency model DAFNet [60], we also provide the results of the proposed EMFI-Net under the backbone of VGG-16 [74] and ResNet-34, i.e., Ours-V and Ours-R. As shown in Fig. 6, we can find that our model performs better than other saliency models in terms of the PR curve and F-measure curves on the ORSSD and EORSSD datasets, where the PR curve of our model is the closest one to the coordinates (1, 1) and the area below the F-measure curve by our model is also the largest one.

Besides, to give a more intuitive presentation for different models, we provide Table I to show the comparison results in terms of S-measure (S), max F-measure (maxF), max E-measure (maxE), and mean absolute error (MAE). It can be seen that the performance of deep learning-based models such as DAFNet [60], PDFNet [30], and U2Net [57] is significantly better than the traditional saliency models including traditional saliency models for natural scene RGB images and RSIs saliency models. Particularly, compared with the top-level saliency models targeting natural scene images such as U2Net [57], CPD [54], and EGNet [45], the RSIs saliency models such as DAFNet [60] and PDFNet [30] acquire a better performance. This confirms the necessity of devising a RSIs saliency model exclusively. In addition, benefitting from the design of our network, our model (Ours-V and Ours-R) performs better than other models on both datasets except for the top-ranking RSI saliency model DAFNet [60]. Concretely, compared with one of the top-level models U2Net [57] on the ORSSD dataset, our approach (Ours-R) elevates the performance by 2.9%, 4.8%, and 2.9%, in terms of S-measure, max F-measure, and max E-measure, respectively. It reduces the MAE by 42.7%. On the EORSSD dataset, our model (Ours-R) also promotes the performance by 1.3%, 0.1%, and 0.7%, in terms of S-measure, max F-measure, and max E-measure, respectively, whereas the MAE is reduced by 1.3%. Moreover, compared with DAFNet [60], our model (Ours-V and Ours-R) is superior in terms of three metrics including S-measure, max F-measure, and MAE on the ORSSD dataset. On the EORSSD dataset, our model performs better than DAFNet [60] in terms of two metrics including S-measure and max F-measure. Therefore, through Fig. 6 and Table I, we can clearly observe the superiority and effectiveness of our model.

Furthermore, to evaluate the computational efficiency of different models, we present the model size (MB) and the average running time (seconds per image) of different models performed on the test set of ORSSD dataset, as summarized in Table II. Here, because the codes of PDFNet [30], SSD [39], SPS [40], and ASD [41] are not provided by the authors, we are unable to provide the average running time of them. Besides, the average running time and the model size of our model and DAFNet [60] are obtained under the ResNet backbone. As presented in Table II, the deep learning-based models are more efficient than the traditional saliency models. Especially, our model takes about 0.04 s for a 256 $\times$ 256 image, which is prominent among all models. However, compared with the top-level models, the model size of our network is slightly large. Therefore, we can say that there is still a large room for further improving the computational efficiency (especially the model size) of our model in future work.

2) Qualitative Comparison: To qualitatively make a comparison for all saliency models, some visual results on the ORSSD dataset and the EORSSD dataset are presented in Figs. 7 and 8, respectively. Here, each figure provides five examples selected from the corresponding dataset, and the results of the proposed EMFI-Net and DAFNet [60] are generated under the ResNet backbone. It can be found that the prediction results of our model shown in Figs. 7(c) and 8(c) are more complete and accurate than other models. Overall, the main advantages of our model lie in three aspects:

a) Superiority in the scenarios with multiple and small objects: In the first and second examples of Fig. 7, the traditional models [e.g., SPS [40] and RCRR [47] shown in Fig. 7(h) and (k)] are completely unable to detect salient objects, whereas the deep learning-based models [e.g., DAFNet [60], LVNet [29], and U2Net [57] shown in Fig. 7(d), (f), and (w)] either provide incomplete detection or falsely highlight background regions. By contrast, our model can successfully detect the two salient ships and two aircrafts from the above two examples. Similarly, in Fig. 8, for the

TABLE II

COMPARISON OF THE MODEL SIZE (MB) AND THE AVERAGE RUNNING TIME (SECONDS PER IMAGE) ON THE TEST SET OF ORSSD DATASET. NOTE THAT "M" DENOTES MATLAB, "C" DENOTES CAFFE, "T" DENOTES TENSORFLOW, AND "P" DENOTES PYTORCH

	DAFNet [60]	PDFNet [30]	LVNet [29]	SSD [39]	SPS [40]	ASD [41]	RBD [46]	RCRR [47]	DSG [48]	MILPS [50]	R3Net [18]	DSS [51]	RADF [52]	RFCN [53]	PoolNet [17]	BASNet [44]	EGNet [45]	CPD [54]	SCRN [58]	U2Net [57]	**EMFI-Net Ours**
Code	P	-	T	-	-	-	M	M	M	M	P	P	P	C	P	P	P	P	P	P	P
Model size	112	-	207	-	-	-	-	-	-	-	142	248	248	744	260	332	426	112	97	168	377
Time	0.04	-	0.74	-	-	-	0.62	3.14	1.57	26.34	0.48	0.12	0.15	1.10	0.01	0.02	0.03	0.02	0.04	0.04	0.04

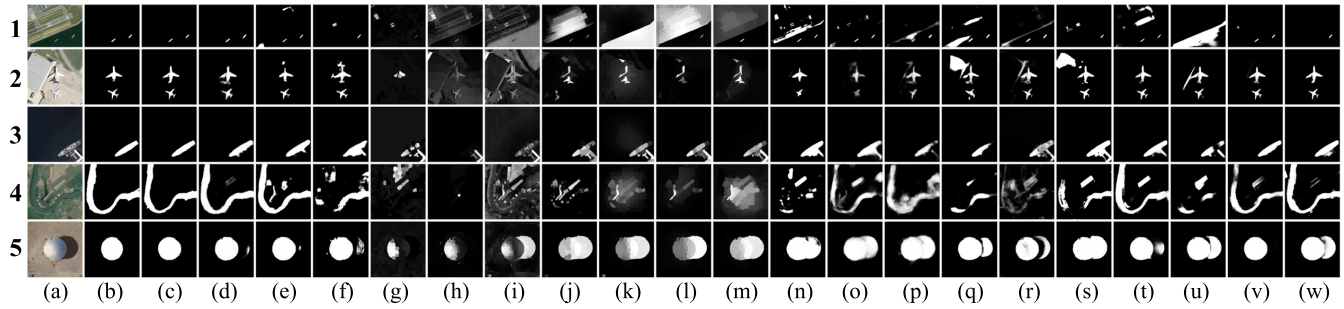


Fig. 7. Visual comparison of different saliency models on several challenging optical RSIs of ORSSD dataset. (a) Optical RSIs. (b) GT. (c) Ours. (d) DAFNet [60]. (e) PDFNet [30]. (f) LVNet [29]. (g) SSD [39]. (h) SPS [40]. (i) ASD [41]. (j) RBD [46]. (k) RCRR [47]. (l) DSG [48]. (m) MILPS [50]. (n) R3Net [18]. (o) DSS [51]. (p) RADF [52]. (q) RFCN [53]. (r) PoolNet [17]. (s) BASNet [44]. (t) EGNet [45]. (u) CPD [54]. (v) SCRN [58]. (w) U2Net [57].

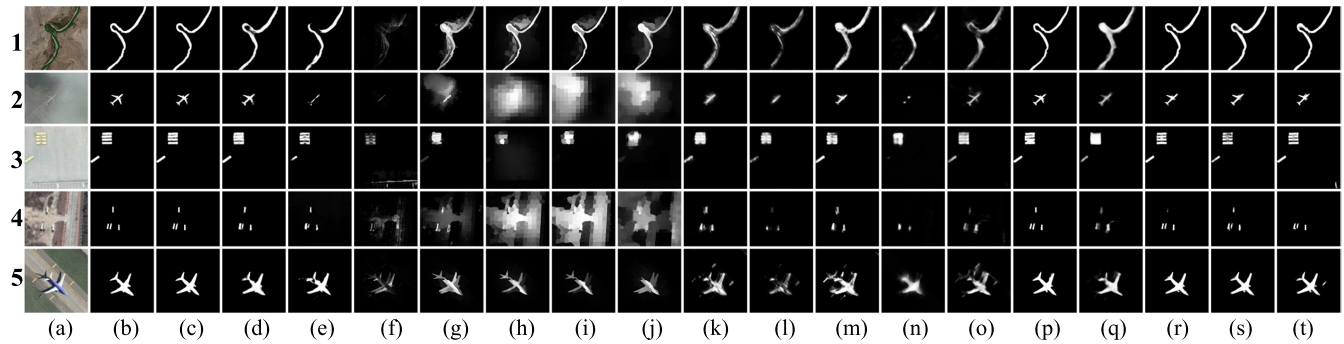
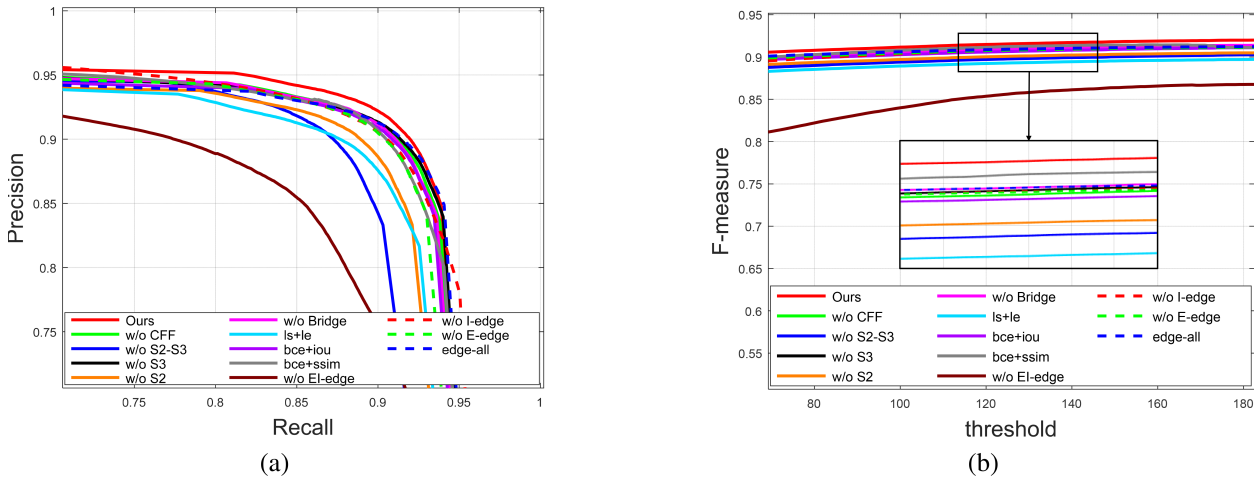


Fig. 8. Visual comparison of different saliency models on several challenging optical RSIs of EORSSD dataset. (a) Optical RSIs. (b) GT. (c) Ours. (d) DAFNet [60]. (e) LVNet [29]. (f) ASD [41]. (g) RBD [46]. (h) RCRR [47]. (i) DSG [48]. (j) MILPS [50]. (k) R3Net [18]. (l) DSS [51]. (m) RADF [52]. (n) RFCN [53]. (o) PoolNet [17]. (p) BASNet [44]. (q) EGNet [45]. (r) CPD [54]. (s) SCRN [58]. (t) U2Net [57].

third and fourth examples, the top-level deep learning-based models such as DAFNet [60], LVNet [29], and U2Net [57] either mistakenly highlight background regions or cannot distinguish salient objects clearly, as shown in Fig. 8(d), (e), and (t). By contrast, our model shown in Fig. 8(c) can pop-out all salient objects completely and clearly.

b) Superiority in cluttered and complex scenes: In the fourth and fifth examples of Fig. 7, no matter the traditional models [e.g., ASD [41] and RCRR [47] presented in Fig. 7(i) and (k)] or the deep learning-based models [e.g., DAFNet [60], LVNet [29], and U2Net [57] shown in Fig. 7(d), (f), and (w)] all failed, namely falsely highlighting background regions. By contrast, our model can provide complete and accurate inference for salient objects. Similarly, in the first and second examples of Fig. 8, some top-level deep learning-based models [e.g., DAFNet [60], DSS [51], and PoolNet [17] depicted in Fig. 8(d), (l), and (o)] either incorrectly detect the background regions or fail to completely pop-out salient objects. In stark contrast, our model shown in Fig. 8(c) can still successfully highlight the salient objects, where our results are with clear boundaries, especially on the river turning area and the airplane wing.

c) Superiority in highlighting of salient objects and suppression of background: In the third example of Fig. 7, some traditional models [e.g., SSD [39] and DSG [48] shown in Fig. 7(g) and (l)] and deep learning-based models [e.g., LVNet [29] and U2Net [57] shown in Fig. 7(f) and (w)] falsely highlight the wharf area. By contrast, our model shown in Fig. 7(c) performs better than other models, where the wharf area is effectively suppressed and the salient ship is highlighted completely. Similarly, in thefifth example of Fig. 8, the traditional models [e.g., ASD [41], RBD [46], and



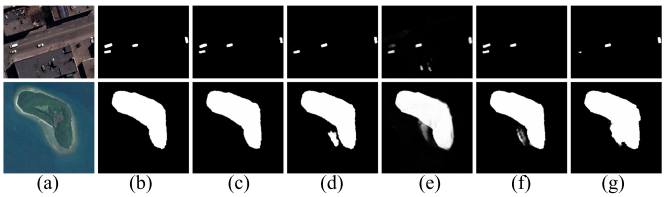
(a) (b)

Fig. 9. (better viewed in color) Quantitative evaluation for the ablation study. (a) PR curves. (b) F-measure curves.

TABLE III

ABLATION STUDIES ARE PERFORMED ON ORSSD DATASET, WHERE THE BEST RESULT IN EACH COLUMN IS MARKED IN RED BOLD FACE. NOTABLY, "↑" (↓) MEANS THAT THE LARGER (SMALLER) THE RESULT, THE BETTER IS

	ls+le	bce+iou	bce+ssim	edge-all	w/o EI-edge	w/o I-edge	w/o E-edge	w/o CFF	w/o S2-S3	w/o S3	w/o S2	w/o Bridge	**Ours**
$S\uparrow$	0.9230	0.9396	0.9345	0.9392	0.8631	0.9377	0.9300	0.9317	0.9195	0.9225	0.9287	0.9388	**0.9432**
$maxF\uparrow$	0.8866	0.9053	0.9042	0.9071	08459	0.9016	0.9013	0.9069	0.8806	0.8977	0.8946	0.9053	**0.9155**
$maxE\uparrow$	0.9646	0.9745	0.9733	0.9780	0.9313	0.9692	0.9683	0.9697	0.9523	0.9561	0.9645	0.9725	**0.9813**
$MAE\downarrow$	0.0128	0.0098	0.0104	0.0105	0.0299	0.0124	0.0146	0.0170	0.0181	0.0194	0.0144	0.0103	**0.0095**



(a) (b) (c) (d) (e) (f) (g)

Fig. 10. Qualitative comparisons of several variants of the proposed EMFI-Net. (a) Optical RSIs. (b) GT. (c) Ours. (d) Edge-all. (e) w/o EI-edge. (f) w/o I-edge. (g) w/o E-edge.

RCRR [47] presented in Fig. 8(f), (g), and (h)] and deep learning-based models [e.g., LVNet [29], RFCN [53], and EGNet [45] shown in Fig. 8(e), (n), and (q)] cannot detect the airplane completely, and even introduce some background regions. By contrast, our model shown in Fig. 8(c) can still generate a more complete saliency map, which is with more accurate boundaries.

In summary, through the aforementioned quantitative and qualitative comparisons, we can firmly demonstrate the effectiveness and superiority of the proposed EMFI-Net, namely our model can perform dense and precise salient object detection on optical RSIs.

D. Ablation Studies

In this part, we conduct comprehensive experiments to validate each key component of our model on ORSSD dataset. The experiments contain quantitative comparisons shown in Fig. 9 and Table III, and qualitative comparisons shown in Figs. 10–12. Here, the results of our model (denoted as "Ours") are obtained based on the ResNet backbone.

1) Validation of Edge Information: To demonstrate the effectiveness of edge cues, we design three variants including EMFI-Net without the explicit and implicit usage of salient edge cues, EMFI-Net without the implicit usage of edge information, and EMFI-Net without the explicit usage of salient edge cues, which are dubbed as "w/o EI-edge," "w/o I-edge," and "w/o E-edge," respectively. We also utilize all multiscale deep features to generate the salient edge cue, which is denoted as "edge-all." The results presented in Fig. 9 and Table III signify that our model performs better than w/o EI-edge, w/o I-edge, and w/o E-edge. We can also find that w/o E-edge and w/o I-edge perform better than w/o EI-edge. This firmly demonstrates the importance of edge information in our model, and further validates the necessity of the sufficient usage of edge information. Besides, it can be seen that our model performs better than edge-all, and this demonstrates the rationality of the design of our edge extraction module. In addition, as shown in the visual comparisons of Fig. 10, our model can detect the four white cars in the top example more completely, and give more accurate boundaries of the green island in the bottom example when compared with the four variants.

2) Validation of Multiscale Deep Features: To prove the effectiveness our multiscale strategy, we devise four variants, namely our model without the second- and third-scale

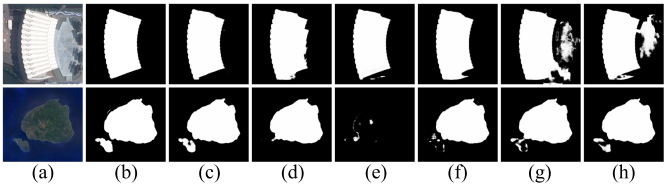


Fig. 11. Qualitative comparisons of several variants of the proposed EMFI-Net. (a) Optical RSIs. (b) GT. (c) Ours. (d) w/o CFF. (e) S2-S3. (f) w/o S3. (g) w/o S2. (h) w/o bridge.

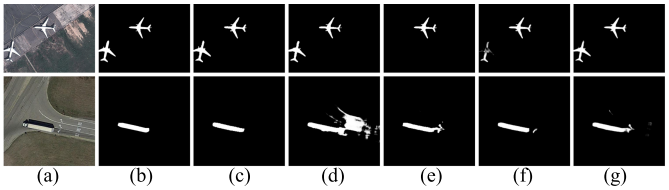


Fig. 12. Qualitative comparisons of several variants of the proposed EMFI-Net. (a) Optical RSIs. (b) GT. (c) Ours. (d) ls + le. (e) bce + iou. (f) bce + ssim. (g) bce.

branches, our model without the third-scale branch, and our model without the second-scale branch, which are denoted as "w/o S2-S3," "w/o S3," and "w/o S2," respectively. Moreover, to validate the effectiveness of the cascaded feature fusion module, we employ a simple concatenation operation to replace it, namely "w/o CFF." In addition, to demonstrate the effectiveness of bridge module, we define our network without the bridge module as "w/o Bridge." According to Fig. 9 and Table III, we can find that our model performs best when compared with the five variants. This demonstrates the rationality of the designed multiscale strategy with cascaded feature fusion module. Meanwhile, this also validates the effectiveness of the adopted bridge module. Besides, as shown in Fig. 11, the five variants either give more concern on background regions or lose salient objects, whereas our model is capable of giving more complete and accurate prediction for the building and the two islands.

3) Validation of Loss Functions: To demonstrate the effectiveness of our loss functions, we first define three variants, namely 1) our model is only equipped with BCE loss and IoU loss (bce + iou); 2) our model only adopts BCE loss and SSIM loss (bce + ssim); and 3) our model only employs BCE loss (bce). Actually, "bce" is our model without implicit usage of edge information, namely "w/o I-edge." Besides, we use "ls + le" to denote our network shown in Fig. 2, which only utilizes the BCE-based edge loss (i.e., le) and the output loss (i.e., ls_1). The results presented in Fig. 9 and Table III indicate that our model achieves the best performance when compared with the four variants. Similarly, in Fig. 12, the four variants cannot detect the two airplanes of the top example and the truck of the bottom example completely and accurately. By contrast, our model provides more complete and precise results, especially the two engines of each airplane can be precisely highlighted. Besides, we also provide the training time of these models, where our model, bce + iou, bce + ssim, bce (i.e., w/o I-edge) and ls + le take about 30.4 h (hours), 31.7 h, 36.7 h, 26 h, and 31 h, respectively. This indicates that different loss terms have little effect on training time. Therefore, through the above comparison results, we can demonstrate the rationality of the deployment of our loss functions.

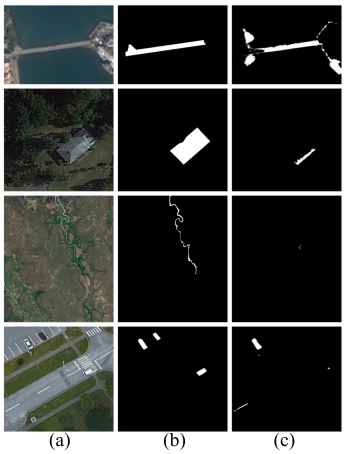


Fig. 13. Some failure examples. (a) Optical RSIs. (b) GT. (c) Saliency maps generated by the proposed EMFI-Net.

E. Failure Cases and Analysis

As mentioned above, the proposed EMFI-Net can give good prediction for salient objects in optical RSIs. However, our model is still incapable of generating satisfactory results when dealing with some challenging scenes shown in Fig. 13. For instance, the two examples in the first and second rows of Fig. 13(a) present two salient objects, i.e., a road and a house, respectively. It can be seen that the road shares the similar color with the beachhead land, and the house roof shares similar appearance with background. As shown in Fig. 13(c), our model falsely highlights the background regions around the salient objects. For the bottom two examples of Fig. 13(a), curved slender river and three cars are surrounded by cluttered background, where the two objects also share similar appearance with the surroundings. As presented in Fig. 13(c), the salient objects are missed by our model. Therefore, we can conclude that the scene with low contrast (i.e., salient objects and background are quite similar) and cluttered background are still challenging for our network. To tackle this issue, more effort should be paid to design more effective integration method for multiscale deep features, so that the network can provide more discriminative representation for salient objects and background.

V. Conclusion

This article presents a novel EMFI-Net to detect salient objects in optical remote sensing images, where the two key components are the multiscale deep feature fusion and the edge cues exploitation. Specifically, the proposed EMFI-Net first generates effective multiscale deep features by using the three convolutional branches with different resolution inputs and the cascaded feature fusion module, so that a powerful representation of salient objects can be acquired. Then, the explicit and the implicit usage of the edge information not only further strengthens the multiscale deep features but also directly endows the saliency maps with clear boundaries.

Besides, the atrous dilated convolution-based bridge module is introduced to enhance the global context, which further promotes the ability of our method in depicting salient objects. Lastly, with the decoding network, the edge-enhanced multiscale deep features are progressively integrated to the final high-quality saliency maps, which show complete salient objects together with clear boundary details. Intensive experiments are conducted on two public optical RSI datasets, and both the quantitative and qualitative results clearly demonstrates the effectiveness and superiority of our model.

REFERENCES

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

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

[3] W. Wang *et al.*, "Learning unsupervised video object segmentation through visual attention," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 3064–3074.

[4] Z. Yu, Y. Zhuge, H. Lu, and L. Zhang, "Joint learning of saliency detection and weakly supervised semantic segmentation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 7223–7233.

[5] Y. Gu, L. Wang, Z. Wang, Y. Liu, M.-M. Cheng, and S.-P. Lu, "Pyramid constrained self-attention network for fast video salient object detection," in *Proc. AAAI Conf. Artif. Intell. (AAAI)*, vol. 34, no. 7, 2020, pp. 10869–10876.

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

[7] Y. Fang, W. Lin, Z. Chen, C.-M. Tsai, and C.-W. Lin, "A video saliency detection model in compressed domain," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 24, no. 1, pp. 27–38, Jan. 2014.

[8] Y.-C. Chen, K.-J. Chang, Y. C. F. Wang, Y.-H. Tsai, and W.-C. Chiu, "Guide your eyes: Learning image manipulation under saliency guidance," in *Proc. 30th Brit. Mach. Vis. Conf. (BMVC)*, 2019, pp. 1–12. [Online]. Available: https://openreview.net/pdf/c8a85d29b4b4c9d6d14cdd718cd5a2b459f62761.pdf

[9] X. Wang, L. Ma, S. Kwong, and Y. Zhou, "Quaternion representation based visual saliency for stereoscopic image quality assessment," *Signal Process.*, vol. 145, pp. 202–213, Apr. 2018.

[10] Y. Fang, J. Wang, Y. Yuan, J. Lei, W. Lin, and P. L. Callet, "Saliency-based stereoscopic image retargeting," *Inf. Sci.*, vol. 372, pp. 347–358, Dec. 2016.

[11] J. Han, X. Yao, G. Cheng, X. Feng, and D. Xu, "P-CNN: Part-based convolutional neural networks for fine-grained visual categorization," *IEEE Trans. Pattern Anal. Mach. Intell.*, early access, Aug. 6, 2020, doi: 10.1109/TPAMI.2019.2933510.

[12] M.-M. Cheng, G.-X. Zhang, N. J. Mitra, X. Huang, and S.-M. Hu, "Global contrast based salient region detection," in *Proc. CVPR*, Jun. 2011, pp. 409–416.

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

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

[15] N. Tong, H. Lu, X. Ruan, and M.-H. Yang, "Salient object detection via bootstrap learning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 1884–1892.

[16] G. Li and Y. Yu, "Visual saliency based on multiscale deep features," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 5455–5463.

[17] J.-J. Liu, Q. Hou, M.-M. Cheng, J. Feng, and J. Jiang, "A simple pooling-based design for real-time salient object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 3917–3926.

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

[19] C. Gong *et al.*, "Saliency propagation from simple to difficult," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 2531–2539.

[20] R. Cong, J. Lei, H. Fu, Q. Huang, X. Cao, and C. Hou, "Co-saliency detection for RGBD images based on multi-constraint feature matching and cross label propagation," *IEEE Trans. Image Process.*, vol. 27, no. 2, pp. 568–579, Feb. 2018.

[21] 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.

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

[23] 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. Eur. Conf. Comput. Vis. (ECCV)*. Cham, Switzerland: Springer, 2020, pp. 225–241.

[24] C. Li *et al.*, "ASIF-Net: Attention steered interweave fusion network for RGB-D salient object detection," *IEEE Trans. Cybern.*, vol. 51, no. 1, pp. 88–100, Jan. 2021.

[25] Z. Tu, T. Xia, C. Li, X. Wang, Y. Ma, and J. Tang, "RGB-T image saliency detection via collaborative graph learning," *IEEE Trans. Multimedia*, vol. 22, no. 1, pp. 160–173, Jan. 2020.

[26] Z. Liu, J. Li, L. Ye, G. Sun, and L. Shen, "Saliency detection for unconstrained videos using superpixel-level graph and spatiotemporal propagation," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 27, no. 12, pp. 2527–2542, Dec. 2017.

[27] X. Zhou, Z. Liu, C. Gong, and W. Liu, "Improving video saliency detection via localized estimation and spatiotemporal refinement," *IEEE Trans. Multimedia*, vol. 20, no. 11, pp. 2993–3007, Nov. 2018.

[28] Y. Piao, X. Li, M. Zhang, J. Yu, and H. Lu, "Saliency detection via depth-induced cellular automata on light field," *IEEE Trans. Image Process.*, vol. 29, pp. 1879–1889, 2020.

[29] 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.

[30] C. Li *et al.*, "A parallel down-up fusion network for salient object detection in optical remote sensing images," *Neurocomputing*, vol. 415, pp. 411–420, Nov. 2020.

[31] H. Lin, Z. Shi, and Z. Zou, "Fully convolutional network with task partitioning for inshore ship detection in optical remote sensing images," *IEEE Geosci. Remote Sens. Lett.*, vol. 14, no. 10, pp. 1665–1669, Oct. 2017.

[32] L. Mou, P. Ghamisi, and X. X. Zhu, "Deep recurrent neural networks for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 7, pp. 3639–3655, Jul. 2017.

[33] W. Song, S. Li, L. Fang, and T. Lu, "Hyperspectral image classification with deep feature fusion network," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 6, pp. 3173–3184, Jun. 2018.

[34] G. Cheng, Z. Li, J. Han, X. Yao, and L. Guo, "Exploring hierarchical convolutional features for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 11, pp. 6712–6722, Nov. 2018.

[35] S. Wan, C. Gong, P. Zhong, B. Du, L. Zhang, and J. Yang, "Multiscale dynamic graph convolutional network for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 5, pp. 3162–3177, May 2020.

[36] S. Wan, C. Gong, P. Zhong, S. Pan, G. Li, and J. Yang, "Hyperspectral image classification with context-aware dynamic graph convolutional network," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 1, pp. 597–612, May 2020.

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

[38] Z. Liu, D. Zhao, Z. Shi, and Z. Jiang, "Unsupervised saliency model with color Markov chain for oil tank detection," *Remote Sens.*, vol. 11, no. 9, p. 1089, May 2019.

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

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

[41] Q. Zhang, L. Zhang, W. Shi, and Y. Liu, "Airport extraction via complementary saliency analysis and saliency-oriented active contour model," *IEEE Geosci. Remote Sens. Lett.*, vol. 15, no. 7, pp. 1085–1089, Jul. 2018.

[42] N. Imamoglu, G. Ding, Y. Fang, A. Kanezaki, T. Kouyama, and R. Nakamura, "Salient object detection on hyperspectral images using features learned from unsupervised segmentation task," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2019, pp. 2192–2196.
[43] L. Zhang, Y. Liu, and J. Zhang, "Saliency detection based on self-adaptive multiple feature fusion for remote sensing images," *Int. J. Remote Sens.*, vol. 40, no. 22, pp. 8270–8297, Nov. 2019.
[44] 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.
[45] 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.
[46] W. Zhu, S. Liang, Y. Wei, and J. Sun, "Saliency optimization from robust background detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 2814–2821.
[47] Y. Yuan, C. Li, J. Kim, W. Cai, and D. D. Feng, "Reversion correction and regularized random walk ranking for saliency detection," *IEEE Trans. Image Process.*, vol. 27, no. 3, pp. 1311–1322, Mar. 2018.
[48] L. Zhou, Z. Yang, Z. Zhou, and D. Hu, "Salient region detection using diffusion process on a two-layer sparse graph," *IEEE Trans. Image Process.*, vol. 26, no. 12, pp. 5882–5894, Dec. 2017.
[49] T. Liu *et al.*, "Learning to detect a salient object," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 2, pp. 353–367, Feb. 2011.
[50] F. Huang, J. Qi, H. Lu, L. Zhang, and X. Ruan, "Salient object detection via multiple instance learning," *IEEE Trans. Image Process.*, vol. 26, no. 4, pp. 1911–1922, Apr. 2017.
[51] Q. Hou, M.-M. Cheng, X. Hu, A. Borji, Z. Tu, and P. Torr, "Deeply supervised salient object detection with short connections," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 3203–3212.
[52] X. Hu, L. Zhu, J. Qin, C.-W. Fu, and P.-A. Heng, "Recurrently aggregating deep features for salient object detection," in *Proc. 32nd AAAI Conf. Artif. Intell. (AAAI)*, 2018, pp. 1–8.
[53] L. Wang, L. Wang, H. Lu, P. Zhang, and X. Ruan, "Salient object detection with recurrent fully convolutional networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 7, pp. 1734–1746, Jul. 2019.
[54] Z. Wu, L. Su, and Q. Huang, "Cascaded partial decoder for fast and accurate salient object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 3907–3916.
[55] D. Zhang, J. Han, Y. Zhang, and D. Xu, "Synthesizing supervision for learning deep saliency network without human annotation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 7, pp. 1755–1769, Jul. 2020.
[56] J. Han, Y. Yang, D. Zhang, D. Huang, D. Xu, and F. De La Torre, "Weakly-supervised learning of category-specific 3D object shapes," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 4, pp. 1423–1437, Apr. 2021.
[57] X. Qin, Z. Zhang, C. Huang, M. Dehghan, O. R. Zaiane, and M. Jagersand, "U^2-Net: Going deeper with nested U-structure for salient object detection," *Pattern Recognit.*, vol. 106, Oct. 2020, Art. no. 107404.
[58] Z. Wu, L. Su, and Q. Huang, "Stacked cross refinement network for edge-aware salient object detection," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 7264–7273.
[59] C. Dong, J. Liu, F. Xu, and C. Liu, "Ship detection from optical remote sensing images using multi-scale analysis and Fourier HOG descriptor," *Remote Sens.*, vol. 11, no. 13, p. 1529, Jun. 2019.
[60] Q. Zhang *et al.*, "Dense attention fluid network for salient object detection in optical remote sensing images," *IEEE Trans. Image Process.*, vol. 30, pp. 1305–1317, 2021.
[61] X. Yao, X. Feng, J. Han, G. Cheng, and L. Guo, "Automatic weakly supervised object detection from high spatial resolution remote sensing images via dynamic curriculum learning," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 1, pp. 675–685, Jan. 2021.
[62] K. Li, G. Cheng, S. Bu, and X. You, "Rotation-insensitive and context-augmented object detection in remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 4, pp. 2337–2348, Apr. 2018.
[63] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.
[64] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 4, pp. 834–848, Apr. 2018.
[65] Z. Wang, E. P. Simoncelli, and A. C. Bovik, "Multiscale structural similarity for image quality assessment," in *Proc. 37th Asilomar Conf. Signals, Syst. Comput.*, Nov. 2003, pp. 1398–1402.
[66] P.-T. de Boer, D. P. Kroese, S. Mannor, and R. Y. Rubinstein, "A tutorial on the cross-entropy method," *Ann. Oper. Res.*, vol. 134, no. 1, pp. 19–67, Feb. 2005.
[67] G. Mattyus, W. Luo, and R. Urtasun, "DeepRoadMapper: Extracting road topology from aerial images," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 3438–3446.
[68] P. Zhang, D. Wang, H. Lu, H. Wang, and X. Ruan, "Amulet: Aggregating multi-level convolutional features for salient object detection," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 202–211.
[69] X. Glorot and Y. Bengio, "Understanding the difficulty of training deep feedforward neural networks," in *Proc. 13th Int. Conf. Artif. Intell. Statist.*, 2010, pp. 249–256.
[70] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*. [Online]. Available: http://arxiv.org/abs/1412.6980
[71] 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.
[72] 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.
[73] A. Borji, M.-M. Cheng, H. Jiang, and J. Li, "Salient object detection: A benchmark," *IEEE Trans. Image Process.*, vol. 24, no. 12, pp. 5706–5722, Dec. 2015.
[74] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2015, pp. 1–14.

Xiaofei Zhou received the Ph.D. degree from Shanghai University, Shanghai, China, in 2018.

He is a Lecturer with the School of Automation, Hangzhou Dianzi University, Hangzhou, China. His research interests include saliency detection, video segmentation, and video quality assessment.

Kunye Shen is pursuing the B.E. degree with the School of Automation, Hangzhou Dianzi University, Hangzhou, China.

His research interests include computer vision and visual saliency analysis.

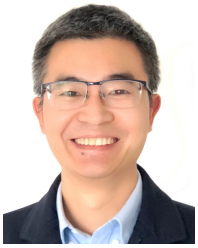
Zhi Liu (Senior Member, IEEE) received the B.E. and M.E. degrees from Tianjin University, Tianjin, China, in 1999 and 2002, and the Ph.D. degree from the Institute of Image Processing and Pattern Recognition, Shanghai Jiaotong University, Shanghai, China, in 2005.

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

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

Chen Gong (Member, IEEE) received the B.E. degree from the East China University of Science and Technology (ECUST), Shanghai, China, in 2010, and dual Ph.D. degrees from Shanghai Jiao Tong University (SJTU), Shanghai, in 2016 and University of Technology Sydney (UTS), Sydney, NSW, Australia, in 2017.

He is a Full Professor with the School of Computer Science and Engineering, Nanjing University of Science and Technology, Nanjing. His research interests mainly include machine learning, data mining, and learning-based vision problems. He has authored or coauthored more than 100 technical articles at prominent journals and conferences such as *Journal of Machine Learning Research (JMLR)*, the IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE (TPAMI), the IEEE TRANSACTIONS ON NEURAL NETWORKS AND LEARNING SYSTEMS (TNNLS), the IEEE TRANSACTIONS ON IMAGE PROCESSING (TIP), the IEEE TRANSACTIONS ON CYBERNETICS (TCYB), the IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY (TCSVT), the IEEE TRANSACTIONS ON MULTIMEDIA (TMM), the IEEE TRANSACTIONS ON INTELLIGENT TRANSPORTATION SYSTEMS (TITS), *ACM Transactions on Intelligent Systems and Technology* (ACM TIST), International Conference on Machine Learning (ICML), Neural Information Processing Systems (NeurIPS), International Conference on Learning Representations (ICLR), IEEE Conference on Computer Vision and Pattern Recognition (CVPR), the Association for the Advance of Artificial Intelligence (AAAI), International Joint Conference on Artificial Intelligence (IJCAI), International Conference on Data Mining (ICDM), etc. He also serves as the Reviewer for more than 30 international journals such as *Artificial Intelligence* (AIJ), *International Journal of Computer Vision* (IJCV), JMLR, the IEEE TPAMI, the IEEE TNNLS, the IEEE TIP, and also the SPC/PC Member of several top-tier conferences such as ICML, NeurIPS, ICLR, CVPR, IEEE International Conference on Computer Vision (ICCV), AAAI, IJCAI, ICDM, International Conference on Artificial Intelligence and Statistics (AISTATS), etc. He received the Excellent Doctoral Dissertation awarded by Shanghai Jiao Tong University (SJTU) and Chinese Association for Artificial Intelligence (CAAI). He was enrolled by the Young Elite Scientists Sponsorship Program of Jiangsu Province and China Association for Science and Technology. He was also the recipient of Wu Wen-Jun AI Excellent Youth Scholar Award.'



Jiyong Zhang received the B.S. and M.S. degrees in computer science from Tsinghua University, Beijing, China, in 1999 and 2001, respectively, and the Ph.D. degree in computer science from the Swiss Federal Institute of Technology at Lausanne (EPFL), Lausanne, Switzerland, in 2008.

He is a Distinguished Professor with Hangzhou Dianzi University, Hangzhou, China. His research interests include *Artificial Intelligence* (AIJ), machine learning, data mining, and image processing.

Chenggang Yan received the B.S. degree in control science and engineering from Shandong University, Shandong, China, in 2008, and the Ph.D. degree in computer science from the Chinese Academy of Sciences University, Beijing, China, in 2013.

He is a Professor with the Department of Automation, Hangzhou Dianzi University, Hangzhou, China. His research interests include computational photography and pattern recognition and intelligent system.