Distilling Knowledge From Super-Resolution for Efficient Remote Sensing Salient Object Detection

Yanfeng Liu, *Student Member, IEEE*, Zhitong Xiong, *Member, IEEE*, Yuan Yuan, *Senior Member, IEEE*, and Qi Wang, *Senior Member, IEEE*

***Abstract*— Current state-of-the-art remote sensing salient object detectors always require high-resolution spatial context to ensure excellent performance, which incurs enormous computation costs and hinders real-time efficiency. In this work, we propose a universal super-resolution-assisted learning (SRAL) framework to boost performance and accelerate the inference efficiency of existing approaches. To this end, we propose to reduce the spatial resolution of the input remote sensing images (RSIs), which is model-agnostic and can be applied to existing algorithms without extra computation cost. Specifically, a transposed saliency detection decoder (TSDD) is designed to upsample interim features progressively. On top of it, an auxiliary SR decoder (ASRD) is proposed to build a multitask learning (MTL) framework to investigate an efficient complementary paradigm of saliency detection and SR. Furthermore, a novel task-fusion guidance module (TFGM) is proposed to effectively distill domain knowledge from the SR auxiliary task to the salient object detection task in optical RSIs. The presented ASRD and TFGM can be omitted in the inference phase without any extra computational budget. Extensive experiments on three datasets show that the presented SRAL with 224 × 224 input is superior to more than 20 algorithms. Moreover, it can be successfully generalized to existing typical networks with significant accuracy improvements in a parameter-free manner. Codes and models are available at *https://github.com/lyf0801/SRAL*.**

***Index Terms*— Auxiliary super-resolution (SR), cross-task knowledge transfer, multitask learning (MTL), optical remote sensing image (RSI), salient object detection (SOD).**

I. Introduction

SALIENT object detection (SOD), also named saliency detection, has recently attracted increasing research interest in optical remote sensing images (RSIs) [1], [2], [3], [4]. It seeks to identify objects/regions in aerial images, i.e., aircraft, ships, bridges, cars, buildings, and other objects, that most attract human attention. This task provides preprocessing for other visual tasks [5], [6], [7], [8] and facilitates numerous downstream topics, such as object detection [9], change detection [10], and super-resolution (SR) [11]. For this reason, inference speed is a crucial factor that needs to be considered for SOD models.

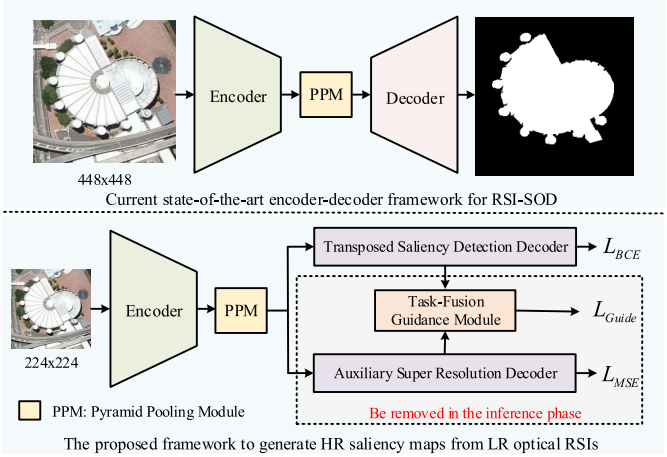


Fig. 1. Illustration of our SRAL and mainstream framework for RSI-SOD.

To achieve a fast inference speed, early SOD methods for natural scene imagery (NSI) introduce residual learning to refine saliency maps (SMs) with limited convolutional parameters [12], or deploy pooling layers to replace the usage of convolution [13]. However, these improvements have limitations in terms of efficiency gains. Recently, some approaches have been devoted to designing lightweight convolution modules or architectures to decrease the model size. For instance, Cheng et al. [14] present a flexible self-adaptive convolution and build an extremely lightweight holistic model based on it. Liu et al. [15] propose a stereoscopically attentive multiscale block based on depthwise separable convolution, achieving a considerable running efficiency for NSI-SOD. A hierarchical visual perception module is designed in [16] to facilitate the deployment of real-world SOD applications with a fast inference speed. These modules have significantly contributed to the research of lightweight models in the field of NSI-SOD. However, due to the characteristics of RSIs, it is nontrivial to directly apply these models to remote sensing scenarios.

In the remote sensing community, lightweight models have also been explored for RSI-SOD recently. Lin et al. [17] deploy MobileNetv2 [18] as the encoder and design a lightweight context block by using 1×3, 1×5, 3×1, and

Manuscript received 3 November 2022; revised 14 March 2023 and 11 April 2023; accepted 12 April 2023. Date of publication 14 April 2023; date of current version 19 May 2023. This work was supported by the National Natural Science Foundation of China under Grant U21B2041, Grant U1864204, and Grant 61825603. *(Corresponding author: Qi Wang.)*

Yanfeng Liu is with the School of Computer Science and the School of Artificial Intelligence, Optics and Electronics (iOPEN), Northwestern Polytechnical University, Xi'an 710072, China (e-mail: liuyanfeng99@gmail.com).

Zhitong Xiong is with the Chair of Data Science in Earth Observation, Technical University of Munich (TUM), 80333 Munich, Germany (e-mail: xiongzhitong@gmail.com).

Yuan Yuan and Qi Wang are with the School of Artificial Intelligence, Optics and Electronics (iOPEN), Northwestern Polytechnical University, Xi'an 710072, China (e-mail: y.yuan1.ieee@gmail.com; crabwq@gmail.com).

Digital Object Identifier 10.1109/TGRS.2023.3267271

5×1 convolutions to avoid heavy parameters. FSMINet [19] discards the common VGG backbone to reduce the parameters and operations and composes a multiscale network by stacking massive depthwise separable convolutions and dilation convolutions. Li et al. [20] also utilize depthwise separable operators to lighten the VGG backbone and decrease the cost in memory and computation to improve the inference speed.

However, these lightweight models [12], [13], [14], [15], [16], [17], [18], [19], [20] usually need sophisticated network designs and more training efforts due to lacking pretrained weights of backbone. Moreover, without using pretrained weights, the generalization ability and the performance are significantly limited due to the small size of the existing RSI-SOD datasets [2], [3], [4].

In addition, there are two other ways to boost the inference efficiency of deep models, i.e., pruning [21] and quantization [22]. The former prunes less important filters in deep networks, while the latter reduces parameter complexity by utilizing low-bit integers without changing the model structure. These methods require elaborate design and also depend on regularization techniques like weight decay to bring in sparsity filters [14], and thus, are rarely explored in the field of SOD.

Reducing the spatial resolution of input RSIs is a simple yet efficient approach to decrease the number of operations and improve the running efficiency without modifying the models. This strategy has been validated in some vision tasks, such as semantic segmentation [23], human pose estimation [24], and land cover classification [25]. However, by directly reducing the input resolution, the performance will significantly decline owing to the lack of efficient spatial context, especially clear boundary information. Typically, LR optical RSIs and small-scale datasets do not meet the learning demands of deep neural networks [23], while high-resolution (HR) images require expensive resources in acquisition cost and computational budget [26]. By reconstructing the edge and texture information of SR, we can effectively compensate for a certain degree of performance degradation due to low-resolution (LR) inputs for RSI-SOD. For instance, [23] and [25] investigate some vision applications in LR situations, introduce multitask learning (MTL) strategy for joint training, and explore domain knowledge to foster the main task. Based on the above-mentioned deficiency, it is imperative for the remote sensing community to develop universal algorithms that can be extended to real-world applications and resource-constrained devices with efficient running speed.

To address the above-mentioned issues, we present an SR-assisted learning (SRAL) paradigm for RSI-SOD tasks. Specifically, rather than designing lightweight networks to improve inference efficiency, we aim to investigate the spatial size of the input LR images. To achieve this, we propose a learning scheme that can distill SR domain knowledge to facilitate RSI-SOD and learn HR SMs. Meanwhile, the inference speed can also be boosted because of the reduced spatial resolution. Note that the proposed learning paradigm is different from knowledge distillation [27], i.e., we focus on cross-task knowledge transfer from SR into RSI-SOD.

As shown in Fig. 1, we present an end-to-end MTL architecture that feeds LR optical RSIs and produces HR SMs, for which we introduce an auxiliary SR subtask and construct an MTL-based model. With respect to the network structure, the proposed SRAL consists of a shared encoder, a transposed saliency detection decoder (TSDD), an auxiliary SR decoder (ASRD), and a novel task-fusion guidance module (TFGM). Among them, the shared encoder produces multiscale contextual features from LR optical RSIs, while the TSDD and ASRD predict the SOD and SR results, respectively. To enable the SR branch to explicitly guide the learning of the SOD branch, the TFGM first integrates predicted features and real labels and supervises them by a specialized objective function, which allows the SOD branch to acquire more HR texture details and focus more on regions of interest. Additionally, the proposed SRAL can also be easily and successfully extended to other SOD models and achieve significant performance benefits, as revealed in Fig. 2. In summary, the main contributions are presented as follows.

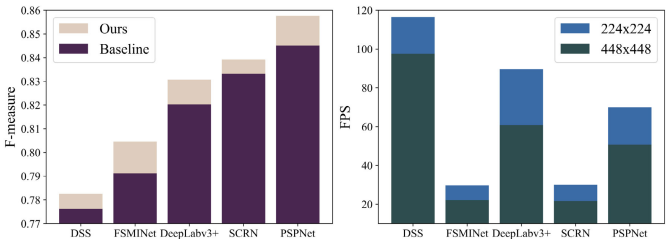


Fig. 2. Left describes the comparison of the five methods on the ORSI-4199 dataset trained at 224×224 input without and with the proposed SRAL. The right shows the inference speed at 224×224 and 448×448 inputs.

1) We first investigate a general and effective paradigm to decrease the computational cost for RSI-SOD, i.e., generating HR SMs by inputting LR optical RSIs to boost the inference speed for existing models.
2) We propose an MTL-based framework named SRAL, which exploits the task complementarity of SR and SOD to enhance the model representative capability and thus improve the final performance of RSI-SOD.
3) To guide SOD explicitly and effectively, we design TFGM, which distills the fine-grained structural knowledge from the SR branch into the SOD stream to enhance the representations without additional inference costs.
4) Sufficient experiments show that the presented SRAL is efficient, effective, and model-agnostic. Typically, it significantly boosts the performance and efficiency of five state-of-the-art models, which provides valuable insights for future research on efficient RSI-SOD models.

The other sections are organized as follows. We summarize the related studies in Section II and provide the methodology in Section III. Experimental results and model analysis are presented in Section IV. Section V illustrates the conclusion.

II. Related Work

This section presents related studies about SOD in optical RSIs, single-image SR (SISR), and MTL in the remote sensing community.

A. Salient Object Detection in Optical RSIs

Early works perform unsupervised strategies to learn low-level features of optical RSIs [28], [29], [30], such as brightness, color, texture, edges, and so on. For example, Huang et al. [30] propose dictionary learning of salient features and finding numerical solutions through constraint terms. In recent years, deep learning-based saliency detection for optical RSIs has become a research hotspot in the remote sensing community [2], [3], [4], [19], [31], [32], [33], [34], [35]. Many researchers propose several public datasets to promote the research of this topic, i.e., ORSSD [2], EORSSD [3], and ORSI-4199 [4]. Based on these datasets and state-of-the-art methods for NSI-SOD, numerous studies have been conducted to foster various problems of RSI-SOD, such as the complicated background of optical RSIs, irregular topology, and boundaries of salient objects. Most methods utilize the fully convolutional network (FCN) paradigms to tackle the scale variation of remote sensing scenes in a fully supervised manner. Researchers have introduced some novel approaches to investigate this topic from various perspectives. For instance, Huang et al. [32] propose a visual attention mechanism to obtain semantic refinement for RSI-SOD. EMFINet [31] combines SOD and boundary detection tasks to build a feature pyramid edge-aware detector. RRNet [33] first explores the combination of graph convolution and FCN to improve detection accuracy by reasoning channel and spatial relations. Currently, Wang et al. [35] first introduce a hybrid encoder consisting of convolutional neural network (CNN) and self-attention to capture local and global context adaptively and attain state-of-the-art performance on all RSI-SOD datasets. As revealed in [35], the above-mentioned studies can generate HR SMs from the HR optical RSIs. However, how to learn HR saliency results efficiently from LR spatial contexts is still an open and unstudied issue. In this article, we employ the MTL framework to build the SR-assisted network and exploit the intrinsic correlation between RSI-SOD and RSI-SR tasks while maintaining the detection capabilities as much as possible and reducing the computational cost.

B. Single Image Super-Resolution

SISR is an essential topic in computer vision, which draws on available image information to establish LR images and HR images, and researches on this topic have been widely investigated in natural and remotely sensed images. The earliest approach, SRCNN [36], only contains three convolutional layers, but its reconstruction performance outperforms various typical methods based on spatial interpolation. Motivated by residual learning, VDSR [37] and EDSR [38] can learn rich complex features by stacking multiple convolutional layers and improving the convergence speed. In the remote sensing community, SISR has also attracted the attention of many researchers. To address the multiscale problem of RSIs and the coupling relationship between surroundings and objects, Lei et al. [39] propose LGCNet to combine multiscale convolutional features of different layers and obtain local and global feature representations. Ma et al. [40] present an enhanced method with joint residual connection and wavelet transform to generate reconstructed RSIs with richer edge information compared with EDSR [38]. As illustrated in the above-mentioned literature, SR can benefit other downstream tasks of optical RSIs. In this study, our primary focus is not on designing new SR networks, but on how to foster RSI-SOD tasks with an auxiliary SR task. Furthermore, our work also paves a way for future research on MTL combined with SR.

C. MTL of Remote Sensing

MTL [23], [41], [42] is widely explored in remote sensing scenarios, which can be classified into two categories. The first kind of approach investigates the complementarity among different tasks to construct a unified framework that simultaneously implements multiple subtasks and obtains better performance across tasks as much as possible. For instance, Wang et al. [43] propose a unified model named boundary-aware multitask network to handle three tasks, i.e., height estimation, semantic segmentation, and boundary detection of RSIs. Recently, Mou and Zhu [44], Yuan et al. [45], and Heidler et al. [46] propose a series of algorithms to cope with joint learning among different tasks. A unified MTL framework is presented to learn vehicle region segmentation and semantic boundary detection based on residual networks [44]. An approach combining semantic segmentation and edge detection with a hierarchical attention mechanism is presented in [46] to monitor the Antarctic coastline effectively. The second type of approach is to introduce the MTL frameworks to facilitate the performance of the main tasks. Among such algorithms, the auxiliary tasks provide complementary supervision in the training phase and implicitly exploit domain knowledge to guide the primary task learning. For example, Liu et al. [47] explore how high-level visual tasks can promote image denoising and present an MTL solution. Xie et al. [25] propose an SR deep network for joint supervision of remote sensing land cover classification, which improves the pixelwise classification accuracy to some extent. Interestingly, Aakerberg et al. [48] present a scheme complementary to [25] that facilitates real-world SR by exploiting an auxiliary segmentation branch, which enables reconstructing sharp and noise-free HR images. However, only the MTL frameworks for edge detection with RSI-SOD have ever been investigated [31], [35], and joint learning of other tasks with RSI-SOD to foster saliency accuracy has never been exploited in the remote sensing community. In this work, we first investigate the MTL architecture of jointly SOD and SR from optical RSIs, which utilizes SR as auxiliary supervision and distills SR prior knowledge to guide RSI-SOD by the proposed TFGM explicitly.

III. Methodology

The limited resolution of optical RSIs restricts the performance of numerous state-of-the-art RSI-SOD approaches. As a fundamental image enhancement strategy, the SR-based methods can enhance the resolution of RSIs and facilitate other visual tasks to yield better results [49]. Based on the above-mentioned ideas, we design an MTL-based model that

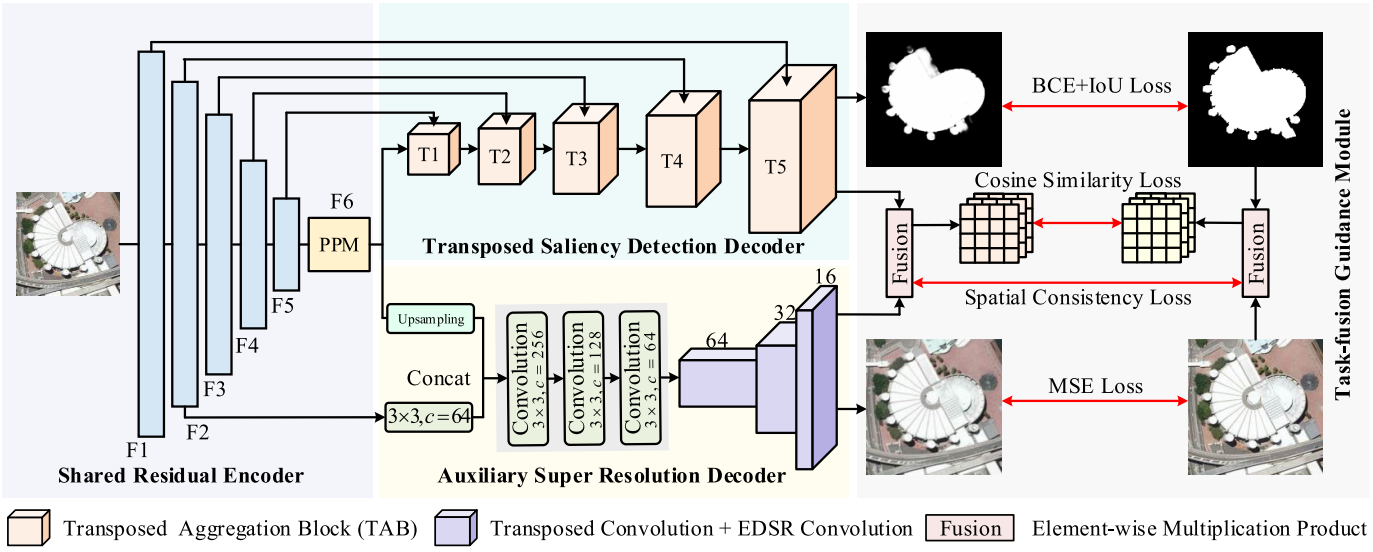


Fig. 3. Illustration of the proposed SRAL framework for optical RSI-SOD.

includes a shared residual encoder, TSDD, ASRD, and TFGM to accomplish the task of efficiently learning HR SMs from LR optical RSIs, as well as to reduce computational cost and boost inference speed from the input stream, rather than the model itself. That is, compared with the HR input-based algorithms, the proposed SRAL boosts the computational efficiency and can be extended to other SOD methods while exhibiting comparable detection results. In this section, we first describe an overview of the model in Section III-A and illustrate the framework of SRAL in Section III-B. Then, the methodology of the presented TFGM is introduced in Section III-C. Finally, the total loss function is provided in Section III-D.

A. Overview of SRAL

The presented SRAL employs an encoder–decoder framework to predict the HR SMs and super-resolved RSIs, which are twice the size of the input images. A shared residual encoder is equipped for feature encoding for SR and SOD simultaneously, and two heterogeneous decoders are designed to supervise the central SOD task and the auxiliary SR task, respectively. As presented in Fig. 3, the proposed learning paradigm includes a residual encoder for multiscale context extraction, a TSDD for the main SOD task, an ASRD for the auxiliary SR task, and cross-task knowledge distillation from the SR domain into the SOD pipeline to maintain HR saliency representation. Among them, ASRD is utilized to perform soft parameter sharing [50] between the SOD decoder and auxiliary SR decoder. To inject the knowledge of detail reconstruction captured from the SR task into the SOD branch, a TFGM is proposed, which considers the interactions between two branches explicitly by the specialized objective function.

B. Model Architecture

Fig. 3 illustrates the detailed framework of the presented model, which consists of a shared residual encoder, a TSDD, an ASRD, and a TFGM, while the detailed structures of the pyramid pooling module (PPM), EDSR convolution, and the transposed aggregation block (TAB) are presented in Fig. 4.

1) Shared Encoder: We utilize the ResNet50 backbone [51] with PPM [52] as the encoder to gradually exploit multiscale convolutional features and encode rich semantic information. Suppose the input RSI as $I \in \mathbb{R}^{3\times224\times224}$, and the five scales of features extracted by the shared encoder are

$$f_1, f_2, f_3, f_4, f_5 = \mathcal{F}_{\text{res}}(I) \tag{1}$$

where $\mathcal{F}_{\text{res}}(\cdot)$ denotes the process of the shared encoder, and $f_1 \in \mathbb{R}^{64\times112\times112}$, $f_2 \in \mathbb{R}^{256\times56\times56}$, $f_3 \in \mathbb{R}^{512\times28\times28}$, $f_4 \in \mathbb{R}^{1024\times14\times14}$, and $f_5 \in \mathbb{R}^{2048\times7\times7}$ are the produced out-stride multiscale contextual features, respectively.

To explore the multilevel global context of optical RSIs in a simple and effective manner, the introduced PPM [52] first reduces the channel dimension of f_5 with a 3×3 convolution to produce a reduced feature defined as f_5', then feeds it into global spatial average pooling layers of various scales, and obtains $f_6 \in \mathbb{R}^{256\times7\times7}$ by channelwise concatenation, that is,

$$f_5' = \mathcal{C}_{3\times3}(f_5) \in \mathbb{R}^{256\times7\times7} \tag{2}$$

$$f_6 = \mathcal{C}_{1\times1}\left(\left[\mathcal{P}_1\left(f_5'\right), \mathcal{P}_2\left(f_5'\right), \mathcal{P}_3\left(f_5'\right), \mathcal{P}_6\left(f_5'\right)\right]\right) \tag{3}$$

where $\mathcal{C}_{i\times i}(\cdot)$ indicates the operation of an $i \times i$ convolution with the BatchNorm and the PReLU function [53], $\mathcal{P}_i(\cdot)$ denotes the adaptive average pooling with $i \times i$ output, and $[\cdot,\cdot]$ represents the channelwise concatenation. The shared encoder network feeds optical RSIs to obtain multiscale contextual features, providing abundant deep semantic information for thereafter task-specific decoder learning.

2) Transposed Saliency Detection Decoder: To learn HR SMs from LR optical RSIs, the proposed SOD decoder should upsample SMs to $2\times$ higher resolution, which is challenging because of limited spatial contextual information. A simple solution is adding a spatial upsampling layer, such as bilinear interpolation or transposed convolution, at the end of the decoder part. However, directly learning a mapping from LR to HR is difficult owing to its ill-posed nature. To address this dilemma, we present a layer-by-layer TSDD based on the progressively upsampling fusion named TSDD.

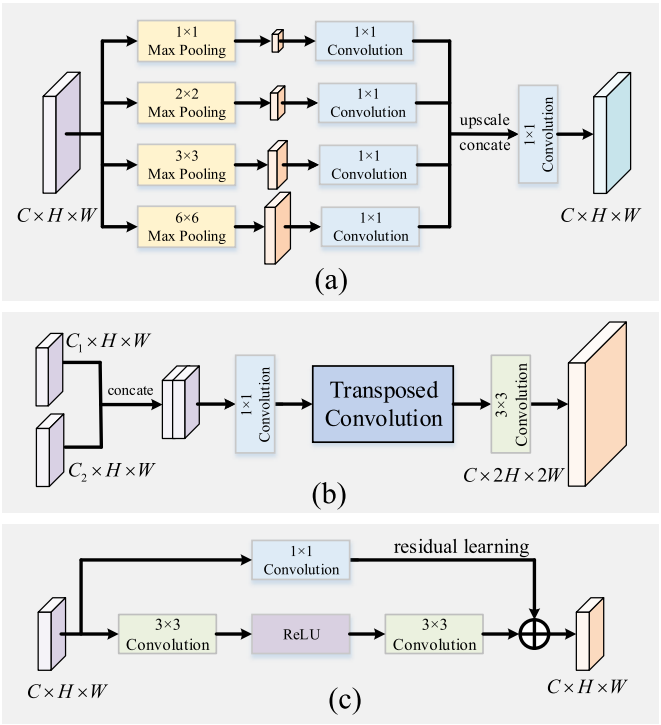


Fig. 4. (a) Illustration of PPM equipped in our model. (b) Detailed structure of the proposed TAB. (c) Framework of the EDSR convolution [54].

As shown in Fig. 3, TSDD adopts a similar structure to UNet [55] and FPN [56] for skip connection and layer-by-layer fusion with the encoder's features. With respect to how to integrate adjacent features and produce 2× HR output information, the TAB is designed. Specifically, there are five TABs to generate multiscale decoder features for SOD, which can be defined as $t_1 \sim t_5$. TAB first obtains the initial fused features by channelwise concatenation and a 1×1 convolution, introduces transposed convolution to learn the upsampled features adaptively, and finally refines them by using a 3×3 convolution. Regarding the input $x \in \mathbb{R}^{C_1 \times H \times W}$ and $y \in \mathbb{R}^{C_2 \times H \times W}$, TAB aggregates them and outputs features with the size of $64 \times 2H \times 2W$ as illustrated in Fig. 4(b). Mathematically

$$\mathcal{F}_{\mathrm{TAB}}(x, y) = \mathcal{C}_{3\times3}\big(\mathcal{F}_{\mathrm{TC}}\big(\mathcal{C}_{1\times1}\big([x, y]\big)\big)\big) \tag{4}$$

where $\mathcal{F}_{\mathrm{TC}}(\cdot)$ and $\mathcal{F}_{\mathrm{TAB}}(\cdot)$ denote the function of transposed convolution and TAB, respectively. As shown in Fig. 3, t_5 is completely mapped from the encoder, i.e., $t_5 = \mathcal{F}_{\mathrm{TAB}}(f_5, f_6)$, while the others are performed through the decoder's upper-level input and skip connection from the encoder as follows:

$$t_i = \mathcal{F}_{\mathrm{TAB}}(t_{i+1}, f_i), \quad i = 4, 3, 2, 1 \tag{5}$$

where $t_i \in \mathbb{R}^{64\times448/2^i\times448/2^i}$, the predicted HR SMs defined as $p_1 \sim p_5 \in \mathbb{R}^{1\times448\times448}$ are generated from these rich semantic features by a 3×3 convolution and $2^i\times$ bilinear interpolation for deep supervision [52], and p_5 is the final predicted SM. Inspired by FPN [56], TSDD also performs top-down feature fusion and continuously increases the spatial resolution of the out-stride features to recover HR SMs in an incremental manner, which alleviates the difficulty of direct upsampling by the factors of $2^{2\sim6}$.

3) Auxiliary Super-Resolution Decoder: Previous studies have proved that SR can help to facilitate other high-level tasks, e.g., semantic segmentation [57], object detection [58], and change detection [59]. Particularly, Wang et al. [24] propose an MTL-based model to achieve better performance for semantic segmentation and human pose estimation with the help of an auxiliary SR network. However, these learning paradigms have not been explored in the field of RSI-SOD, and we make the first attempt and some progress in this article. To promote the detection performance by MTL strategy, an effective and simple ASRD is presented motivated by well-known SR networks [51], [54] to reconstruct the HR optical RSIs from the multiscale decoder features. Specifically, we utilize the shallowest and deepest contextual information of the encoder to reconstruct HR RSIs and explore more generalized constraints. As shown in Fig. 3, ASRD employs several convolutional operations to reduce the channel dimension gradually, i.e., feeds f_2 and f_6 to obtain the integrated feature as

$$f_{\mathrm{mid}} = \mathcal{C}_{3\times3}\big(\mathcal{C}_{3\times3}\big(\mathcal{C}_{3\times3}\big(\big[\mathcal{C}_{3\times3}(f_2), \mathcal{F}_{\mathrm{up}}(f_6)\big]\big)\big)\big) \tag{6}$$

where the size of f_{mid} is $64 \times 56 \times 56$, and $\mathcal{F}_{\mathrm{up}}(\cdot)$ denotes the spatial bilinear interpolation function.

As for how to recover the HR representation of RSIs, we alternately deploy the deconvolution and EDSR convolution [54] to gradually recover the spatial resolution while continuously shrinking the feature channels (64→32→16→3 revealed in Fig. 3). Hence, the final predicted HR RSI is defined as

$$p_{\mathrm{sr}} = \mathcal{F}_{\mathrm{TE}}(\mathcal{F}_{\mathrm{TE}}(\mathcal{F}_{\mathrm{TE}}(f_{\mathrm{mid}}))) \oplus \mathcal{F}_{\uparrow2\times}(I) \tag{7}$$

where $\mathcal{F}_{\mathrm{TE}}(\cdot)$ represents the combined operations of deconvolution and EDSR convolution [54], $F_{\uparrow2\times}(\cdot)$ indicates the 2× upsampling interpolation, and $p_{\mathrm{sr}} \in \mathbb{R}^{3\times448\times448}$. Fig. 4(c) presents a detailed illustration of EDSR convolution, a modified residual convolution that performs better than the original one. Regarding the input $X \in \mathbb{R}^{C\times H\times W}$, the operation of EDSR convolution can be represented as follows:

$$\mathcal{F}_{\mathrm{EDSR}} = \mathcal{C}_{3\times3}(\mathcal{F}_{\mathrm{ReLU}}(\mathcal{C}_{3\times3}(X))) \oplus \mathcal{C}_{1\times1}(X) \tag{8}$$

where $\mathcal{F}_{\mathrm{ReLU}}(\cdot)$ denotes the ReLU function, $\mathcal{F}_{\mathrm{EDSR}}(\cdot)$ indicates the operation of EDSR convolution, and $\oplus$ denotes elementwise summation. By means of this simple scheme, the model can supervise both SR and SOD and implicitly perform knowledge transfer from the SR branch to the SOD stream.

C. Task-Fusion Guidance Module

How to further enhance the performance gain of the SR auxiliary task to the RSI-SOD main task? The most straightforward approach is to inject the representation learned from SR into the SOD branch by explicitly supervised learning [24], [49]. Previous works directly perform structural similarity [23] or spatial cosine similarity [25] between the output features of the dual streams. However, this strategy is shown not to work well for RSI-SOD through our experiments. Based on this deficiency, we present TFGM to distill the SR knowledge to the

saliency detection network via the specialized objective functions, which both consider the correlation between decoder features and ground truth (GT) products.

To obtain the knowledge representation from TSDD and ASRD, we deploy two 3×3 convolutions at the end of them to capture two feature expressions, i.e., $f_{\text{sod}}, f_{\text{sr}} \in \mathbb{R}^{3\times448\times448}$. As shown in Fig. 3, TFGM first integrates the decoder's output features and GT of two branches through elementwise multiplication. Thus, the subjects supervised by the loss function are $f^{\text{sod}} \cdot f^{\text{sr}}$ and $I^{\text{sod}} \cdot I^{\text{sr}}$, where I^{sod} and I^{sr} indicate the GT SMs and HR RSIs with the dimension of $3 \times 448 \times 448$. It can be seen that we optimize the TFGM by using two different streams from Fig. 3. With respect to the first branch, we present the pixel similarity loss to explore the spatial difference between the predicted elementwise product and the GT product. For $f^{\text{sod}} \cdot f^{\text{sr}}$, we first downsample it to a $3 \times 28 \times 28$ matrix ($16\times$ average pooling) because of the high memory overheads and reshape it as a 3×784 one. The affinity matrix can be represented as

$$S^f_{i,j} = \left(\frac{f^{\text{sod}}_i \cdot f^{\text{sr}}_i}{\left\| f^{\text{sod}}_i \cdot f^{\text{sr}}_i \right\|_1}\right)^{\text{T}} \left(\frac{f^{\text{sod}}_j \cdot f^{\text{sr}}_j}{\left\| f^{\text{sod}}_j \cdot f^{\text{sr}}_j \right\|_1}\right) \tag{9}$$

where $S^f_{i,j}$ refers to the self-correlation between ith and jth element of the fused $f^{\text{sod}}_i \cdot f^{\text{sr}}_i$, and we can calculate the cosine similarity matrix of GT product by the same way, i.e., S^f and S^I, the affinity matrix of $f^{\text{sod}} \cdot f^{\text{sr}}$ and $I^{\text{sod}} \cdot I^{\text{sr}}$, respectively. Then, the spatial similarity loss is captured as

$$\mathcal{L}_{\text{similar}} = \frac{1}{W^2H^2} \sum_{i=1}^{WH} \sum_{j=1}^{WH} \left\| S^f_{i,j} - S^I_{i,j} \right\|_2. \tag{10}$$

Here, W and H are both equal to 28. This supervision is comparable to a spatial attention mechanism and can guide the SOD branch to absorb the fine-grained structural information of the SR branch. As for the second one, we introduce a consistency-regularized loss to compute the L1 normalization between the fused features of both decoders and the GT product, which further aims to make the SR branch focus more on the salient regions' boundaries in RSIs and supervise the SOD branch to pay more attention to salient objects, that is,

$$\mathcal{L}_{\text{consist}} = \frac{1}{C'W'H'} \left\| f^{\text{sod}} \cdot f^{\text{sr}} - I^{\text{sod}} \cdot I^{\text{sr}} \right\|_1 \tag{11}$$

where the values of C', W', and H' are 3, 448, and 448, respectively. With the help of this function, our model will tend to detect the various salient objects precisely, as revealed in experiments. The total loss of TFGM is the combined summation as follows:

$$\mathcal{L}_{\text{TFGM}} = \mathcal{L}_{\text{similar}} + \mathcal{L}_{\text{consist}}. \tag{12}$$

D. Total Loss Function of the Model

To achieve the joint learning of the whole model in the training process, we integrate the loss item of TSDD, the loss item of ASRD, and the loss item of TFGM into a total loss function, and thus, the entire model can be optimized end-to-end. The total loss can be defined as

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{SOD}} + \lambda_2 \mathcal{L}_{\text{SR}} + \lambda_3 \mathcal{L}_{\text{TFGM}} \tag{13}$$

where $\mathcal{L}_{\text{SOD}}$ is the combination of the binary cross-entropy loss and IoU loss with deep supervision among $p_1 \sim p_5$, and $\mathcal{L}_{\text{SR}}$ is the mean square error loss between p^{sr} and I^{sr}, and λ_2 and λ_3 are utilized to balance the tasks while highlighting the main SOD task, and the above-mentioned loss items are represented as

$$\mathcal{L}^i_{\text{SOD}} = \mathcal{L}^i_{\text{BCE}} + \mathcal{L}^i_{\text{IoU}} \tag{14}$$

$$\mathcal{L}_{\text{SOD}} = \left(\mathcal{L}^1_{\text{BCE}} + \mathcal{L}^1_{\text{IoU}}\right) + \sum_{i=2}^{5} \frac{\mathcal{L}^i_{\text{BCE}} + \mathcal{L}^i_{\text{IoU}}}{2^{i-2}} \tag{15}$$

where $\mathcal{L}^i_{\text{BCE}}$ and $\mathcal{L}^i_{\text{IoU}}$ are imposed on the ith output of saliency decoder, namely, t_i. We perform deep supervision on each stage of TSDD and contribute to the final saliency results. Mathematically, the detailed definition of these loss functions is illustrated as

$$\mathcal{L}_{\text{BCE}}(x, y) = \frac{1}{N} \sum_{i=1}^{N} (-y_i \log(x_i) - (1 - y_i) \log(1 - x_i)) \tag{16}$$

$$\mathcal{L}_{\text{IoU}}(x, y) = 1 - \frac{\sum_1^N \left(x_j \otimes y_j\right) + 1}{\sum_1^N \left(x_j \oplus y_j - x_j \otimes y_j\right) + 1} \tag{17}$$

$$\mathcal{L}_{\text{SR}}\left(p^{\text{sr}}, I^{\text{sr}}\right) = \frac{1}{3N} \sum_{i=1}^{3N} \left(p^{\text{sr}}_i - I^{\text{sr}}_i\right)^2 \tag{18}$$

where N denotes the total number of pixels of x, y, and i reveals the index of each pixel. $x_j \oplus y_j$ and $x_j \otimes y_j$ indicate the summation product and multiplication product of the predicted SM and its label at pixel j, respectively. Since these loss terms have significantly different value ranges, even at an order of magnitude, it is very challenging to balance these parameters in a multitask manner. Empirically, we adopt manual hyperparameters tuning and observe that the values of λ_2 and λ_3 are equal to 100, and 1 work well in Section IV-C.

IV. Experiments

A. Experimental Protocol

1) Datasets: There are three public datasets released to the community, i.e., ORSSD [2], EORSSD [3], and ORSI-4199 [4], on which we organize experiments in this article.

ORSSD: It consists of 800 optical RSIs, including 600 images for training and 200 images for testing.

EORSSD: This dataset is an extension of ORSSD. Compared with ORSSD, covering more remote sensing scenes and a wider variety of salient objects is more challenging. It is divided into 1400 images for training and the rest for testing.

ORSI-4199: It is the most diverse dataset for RSI-SOD, containing 4199 optical RSIs, of which 2000 images are adopted for training and 2199 ones for testing. Furthermore, it defines nine scene attribute types to help us evaluate different algorithms more comprehensively.

2) Evaluation Metrics: To quantitatively measure various methods, we plot PR and F-measure curves to highlight the methods with better performance. Besides, three numerical metrics are employed for quantitative evaluation as follows.

MAE [60]: It defines the mean pixelwise error between the predicted SM with the GT, that is,

$$\mathrm{MAE} = \frac{1}{m \times n} \sum_{i=1}^{m} \sum_{j=1}^{n} |\mathrm{SM}(i, j) - \mathrm{GT}(i, j)| \tag{19}$$

where m and n denote the length and width of SMs, respectively.

F-Measure [61]: As a metric that balances the Precision and Recall of saliency detection, which is defined as

$$F_\beta = \frac{\left(1 + \beta^2\right) \times \mathrm{Precision} \times \mathrm{Recall}}{\beta^2 \times \mathrm{Precision} + \mathrm{Recall}} \tag{20}$$

where the value of β^2 is equal to 0.3 as suggested in [61].

S-Measure [62]: It estimates the structural information similarity between the predicted map and its label from region-aware (S_r) and object-aware (S_o) levels as follows:

$$S_m = \alpha \times S_o(\mathrm{SM}, \mathrm{GT}) + (1 - \alpha) \times S_r(\mathrm{SM}, \mathrm{GT}) \tag{21}$$

where α is 0.5 to balance S_o and S_r as recommended in [62].

Finally, we report the structure similarity (SSIM) scores [63] for attributes analysis on the ORSI-4199 dataset, defined as

$$\mathrm{SSIM}(x, y) = \frac{\left(2\mu_x\mu_y + C_1\right)\left(2\sigma_{xy} + C_2\right)}{\left(\mu_x^2 + \mu_y^2 + C_1\right)\left(\sigma_x^2 + \sigma_y^2 + C_2\right)}. \tag{22}$$

Here, σ_x and σ_y indicate the standard deviation of x and y, respectively. μ_x and μ_y are the means, and σ_{xy} is the covariance. We calculate SSIM by adjusting both C_1 and C_2 to 0, referring to [35].

3) Implementation Details: Following the previous study [35], we train and test the models on three separate datasets, respectively, and the previous data augmentation strategies are followed to enrich the training samples. We employ the source code of all compared algorithms for a fair reproduction, and all detection results are computed at 448 × 448. All CNN-based methods [3], [4], [12], [13], [14], [15], [16], [17], [19], [20], [24], [32], [33], [34], [35], [65], [66], [67], [68], [69], [70] are implemented on a single NVIDIA GeForce RTX 3090 GPU based on PyTorch 1.8 toolbox in the Linux system. With respect to the proposed method, we load the pretrained weights of ResNet50 [51] and utilize the SGD optimizer to train the model with a batch size of 8, an initial learning rate of 0.002, a momentum of 0.9, and a weight decay of 5e-4. The polynomial learning rate scheduler is employed in the iteration process with the learning rate updated by the formula: $lr = 0.002 \times (1 - (\mathrm{iter}/\mathrm{maxiter}))^{0.9}$. Referring to previous SR deep learning-based models, the SR branch is trained and supervised under YCbCr color space and then converted to RGB space during the inference phase. The proposed SRAL inputs 224 × 224 optical RSIs and outputs 448 × 448 saliency results. During the testing process, both the presented ASRD and TFGM are removed, so that no additional parameters or computational costs are introduced compared with the baseline. In contrast, the inference speed of the SRAL has more merit.

4) Baselines: As illustrated in Table I, to discover the pros and cons of the presented approach, we report the performance of 24 baselines on three datasets for a fair comparison. These algorithms include two traditional methods (i.e., LC [64] and FT [61]), six deep learning-based models for NSI-SOD (i.e., NLDF [65], DSS [66], PFAN [67], SCRN [68], GateNet [69], and F3Net [70]), seven deep learning-based methods for RSI-SOD (i.e., SARNet [32], DAFNet [3], MJRBM [4], RRNet [33], FSMINet [19], ACCoNet [34], and HFANet [35]), five lightweight models for NSI-SOD (i.e., RAS [12], PoolNet [13], CSNet [14], HVPNet [16], and SAMNet [15]), three lightweight LR input-based models for RSI-SOD (i.e., CorrNet [20], MSCNet [17], and FSMINet [19]), and an SR-guided model for semantic segmentation, i.e., DSRL [24].

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

In this section, we perform the quantitative analysis, qualitative analysis, attribute-based analysis, and running efficiency analysis of numerous state-of-the-art approaches.

1) Quantitative Comparison: As illustrated in Fig. 5, the PR and F-measure curves of our model in red outperform all competitors on three datasets, i.e., the PR curves of SRAL are closer to the upper right corner of Fig. 5 and its F-measure curves can cover the largest areas. Remarkably, our algorithm has the most significant superiority on the ORSI-4199 dataset.

To perform a more intuitive analysis, we report three metrics, i.e., F_β, MAE, and S_m of 24 approaches, in Table I. Four observations can be concluded as follows. First, any deep learning-based model outperforms the best-known traditional methods, LC [64] and FT [61], revealing the suitability of deep networks for RSIs. Second, since RSI-SOD methods are specialized for remote sensing scenes, most such methods are superior to NSI-SOD approaches, which illustrate the necessary design of specialized salient object detectors for RSIs. Third, SRAL outperforms various HR or LR input-based lightweight methods for NSI-SOD and RSI-SOD, i.e., CSNet [14], HVPNet [16], SAMNet [15], CorrNet [20], MSCNet [17], and FSMINet [19]. The above-mentioned phenomenon demonstrates that the lightweight algorithms we illustrated in Section I cannot achieve the desired performance due to the lack of pretraining weights and the small scale of the RSI-SOD datasets, despite the low number of parameters and operations. Finally and most importantly, the proposed SRAL can achieve comparable results to competitors by feeding fewer spatial contexts on all datasets. This provides a feasible solution for learning HR SMs from LR and low-quality RSIs. Table I provides the results of FSMINet [19] trained with 224 × 224 and 448 × 448 input, where the latter has better numerical results, which justifies the challenging issue that the performance of RSI-SOD decreases significantly when the input resolution of RSIs is reduced twice. This is precisely the motivation for this work. HFANet [35], F3Net [70], and GateNet [69] are the most competitive approaches among all competitors, while HFANet is our recent work that achieves the optimal results in four metrics. Compared with them, the

TABLE I

QUANTITATIVE PERFORMANCE ON THREE OPTICAL RSI-SOD TEST DATASETS WITH 24 STATE-OF-THE-ART APPROACHES. THE TOP THREE RESULTS ARE MARKED IN RED, GREEN AND BLUE, RESPECTIVELY

Methods	Publication	InputSize	OutputSize	ORSSD Dataset [2]			EORSSD Dataset [3]			ORSI-4199 Dataset [4]		
				$F_\beta \uparrow$	MAE $\downarrow$	$S_m \uparrow$	$F_\beta \uparrow$	MAE $\downarrow$	$S_m \uparrow$	$F_\beta \uparrow$	MAE $\downarrow$	$S_m \uparrow$
Traditional Methods												
LC [64]	MM'06	448×448	448×448	0.4275	0.1230	0.5941	0.4526	0.0864	0.5954	0.3573	0.1893	0.5270
FT [61]	CVPR'09	448×448	448×448	0.4526	0.1126	0.5916	0.4704	0.0715	0.6107	0.3680	0.1791	0.5256
SOD models for NSIs												
NLDF [65]	CVPR'17	448×448	448×448	0.8352	0.0267	0.8702	0.8060	0.0154	0.8706	0.7639	0.0584	0.8053
DSS [66]	CVPR'17	448×448	448×448	0.8469	0.0268	0.8688	0.7921	0.0167	0.8371	0.7672	0.0561	0.8115
PFAN [67]	CVPR'19	448×448	448×448	0.8755	0.0207	0.8853	0.8472	0.0127	0.8848	0.8024	0.0486	0.8373
SCRN [68]	ICCV'19	448×448	448×448	0.8687	0.0210	0.8799	0.8326	0.0158	0.8288	0.8232	0.0423	0.8524
GateNet [69]	ECCV'20	448×448	448×448	0.9083	0.0125	0.9103	0.8724	0.0091	0.9010	0.8443	0.0387	0.8660
F3Net [70]	AAAI'20	448×448	448×448	0.8927	0.0126	0.9245	0.8822	0.0077	0.9218	0.8175	0.0435	0.8520
SOD models for optical RSIs												
SARNet [32]	RS'21	448×448	448×448	0.8963	0.0185	0.8976	0.8865	0.0102	0.9097	0.8309	0.0448	0.8536
DAFNet [3]	TIP'21	448×448	448×448	0.8717	0.0161	0.8982	0.8378	0.0106	0.8824	0.8169	0.0473	0.8477
MJRBM-V [4]	TGRS'22	448×448	448×448	0.9028	0.0140	0.9156	0.8705	0.0099	0.9088	0.8352	0.0392	0.8601
MJRBM-R [4]	TGRS'22	448×448	448×448	0.9058	0.0129	0.9128	0.8685	0.0092	0.8980	0.8406	0.0379	0.8685
RRNet [33]	TGRS'22	448×448	448×448	0.8857	0.0142	0.9110	0.8511	0.0101	0.8964	0.8122	0.0448	0.8449
FSMINet [19]	GRSL'22	448×448	448×448	0.8623	0.0178	0.8949	0.8527	0.0100	0.8989	0.8046	0.0451	0.8344
ACCo-V [34]	TCYB'23	448×448	448×448	0.8514	0.0231	0.8717	0.8677	0.0137	0.9068	0.8182	0.0433	0.8490
HFANet [35]	TGRS'22	448×448	448×448	0.9224	0.0113	0.9324	0.9007	0.0082	0.9292	0.8419	0.0379	0.8659
Lightweight models or low-resolution input-based models												
RAS [12]	ECCV'18	448×448	448×448	0.8841	0.0185	0.8896	0.8636	0.0114	0.8800	0.7930	0.0595	0.8142
PoolNet [13]	CVPR'19	448×448	448×448	0.8291	0.0268	0.8610	0.8121	0.0207	0.8279	0.7777	0.0573	0.8184
CSNet [14]	PAMI'21	448×448	448×448	0.8258	0.0311	0.8403	0.7884	0.0179	0.8223	0.7858	0.0562	0.8203
HVPNet [16]	TCYB'21	448×448	448×448	0.8739	0.0151	0.8943	0.8554	0.0099	0.8878	0.8266	0.0408	0.8534
SAMNet [15]	TIP'21	448×448	448×448	0.8597	0.0178	0.8867	0.8487	0.0105	0.8806	0.8182	0.0434	0.8500
CorrNet [20]	TGRS'22	256×256	256×256	0.8813	0.0184	0.9026	0.8815	0.0095	0.9190	0.7906	0.0519	0.8211
MSCNet [17]	ICPR'22	224×224	224×224	0.8908	0.0144	0.9067	0.8629	0.0097	0.9034	0.8427	0.0384	0.8518
FSMINet [19]	GRSL'22	224×224	448×448	0.8308	0.0221	0.8769	0.8283	0.0124	0.8829	0.7912	0.0483	0.8230
DSRL [24]	CVPR'20	224×224	448×448	0.8765	0.0162	0.8895	0.8135	0.0120	0.8683	0.8229	0.0410	0.8546
SRAL (Ours)	–	224×224	448×448	0.9167	0.0105	0.9305	0.8964	0.0067	0.9234	0.8576	0.0321	0.8735

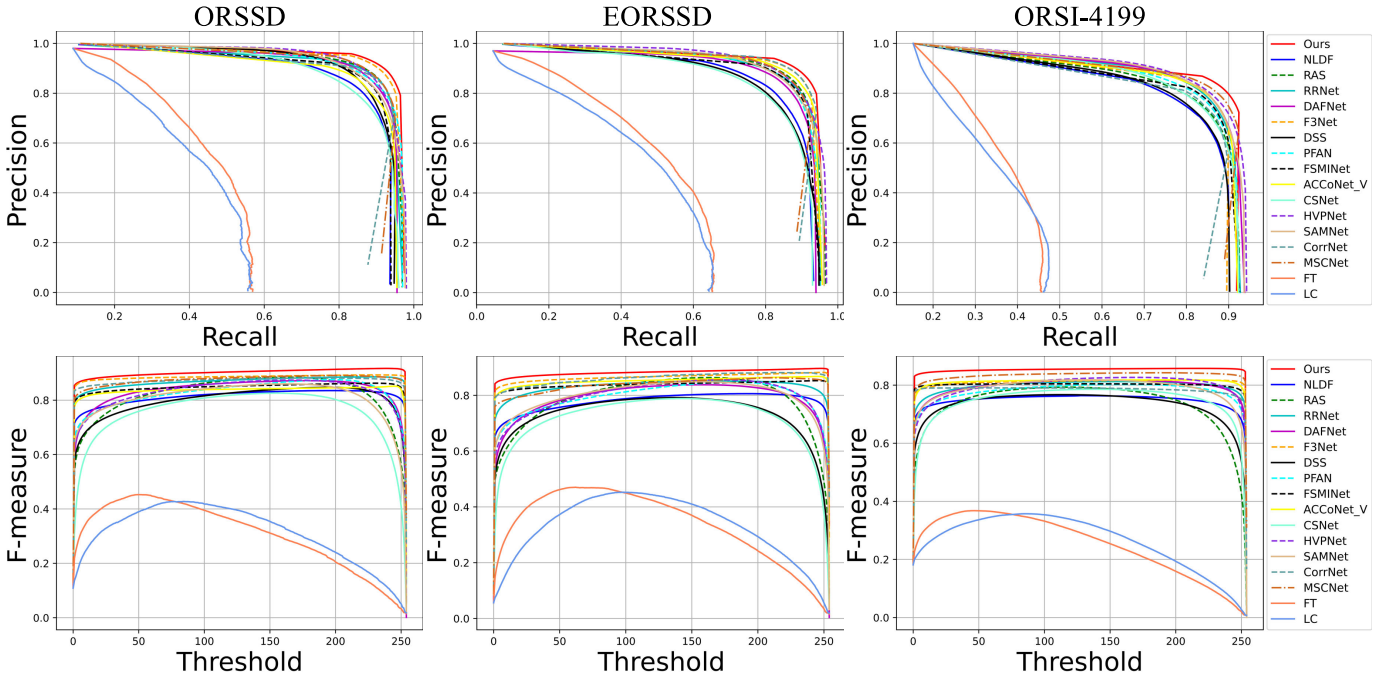


Fig. 5. Comparisons of PR and F-measure curves about 17 algorithms on three RSI-SOD datasets, and our method is marked in red.

presented SRAL reduces the input scale of RSIs but still maintains decent results, e.g., it offers the only solution with F_β greater than 0.85 and S_m greater than 0.87 on the most challenging ORSI-4199 dataset.

2) Qualitative Comparison: As revealed in Fig. 6, typical saliency results on the ORSI-4199 dataset are presented for qualitative visualization. Due to space limitations, we only show the detection results of two NSI-SOD algorithms, eight RSI-SOD methods, and two lightweight approaches with the most competitive performance. These typical scenes include large-scale aircraft, multiple airplanes, tiny vehicles, roads with irregular topology, man-made buildings with complicated edges, lakes with low contrast, and so on. We can observe that the proposed SRAL predicts the most complete and

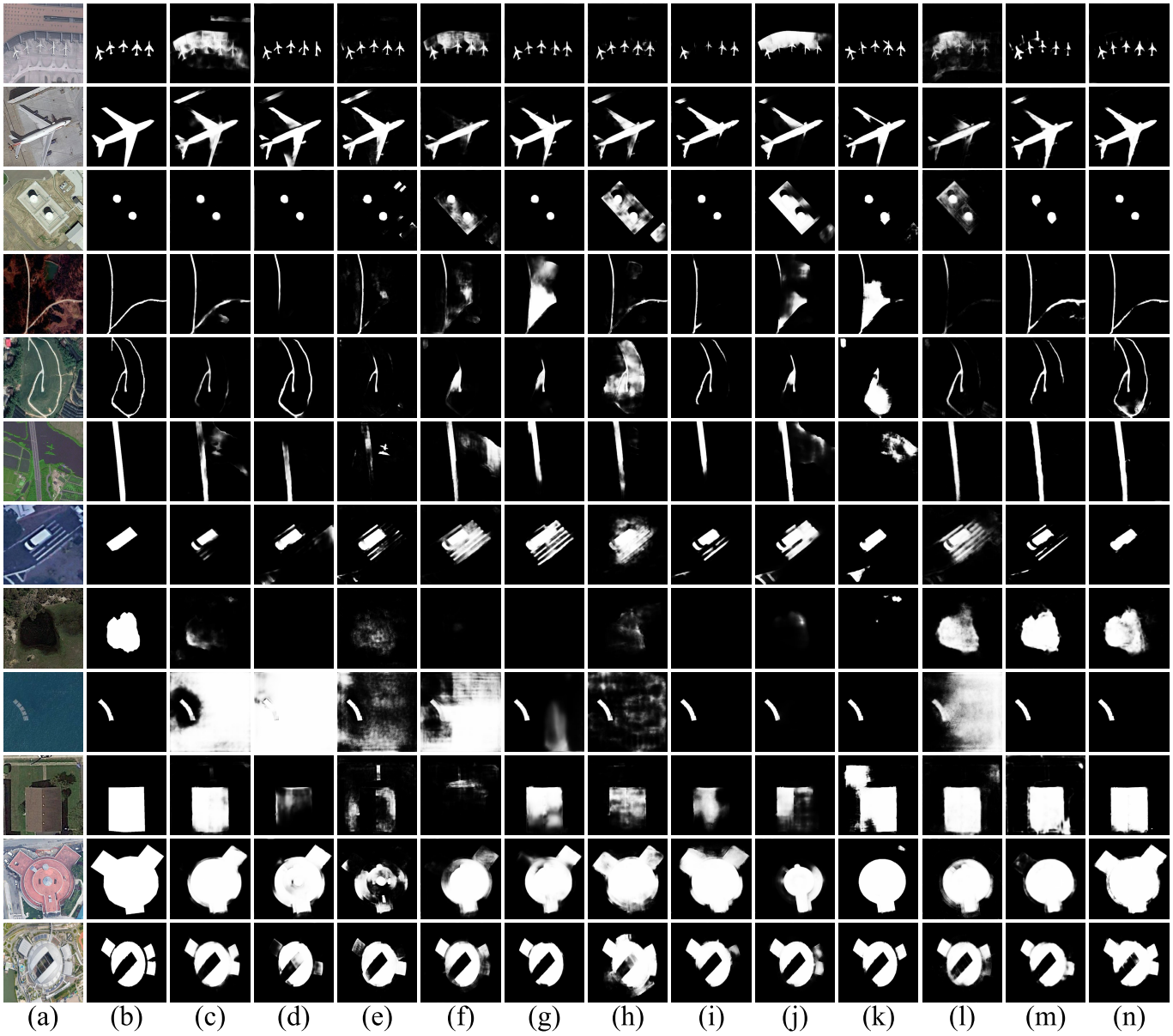


Fig. 6. Typical visualized maps with 12 state-of-the-art methods on the ORSI-4199 dataset, including two NSI-SOD approaches, eight RSI-SOD algorithms, and two lightweight methods on different patterns. (a) Optical RSIs. (b) GT. (c) GateNet [69]. (d) F3Net [70]. (e) SARNet [32]. (f) DAFNet [3]. (g) MJRBM-V [4]. (h) RRNet [33]. (i) FSMINet [19]. (j) ACCoNet-V [34]. (k) HFANet [35]. (l) HVPNet [16]. (m) MSCNet [17]. (n) SRAL (Ours).

precise SMs in these scenarios, while the competitors always suffer from deficiencies, such as incomplete salient objects, interference from the complex background, and unclear edges.

3) Attribute Analysis: The ORSI-4199 dataset provides nine scene patterns for a more intuitive comparison, i.e., big salient object (BSO), complex scene (CS), complex salient object (CSO), incomplete salient object (ISO), low contrast scene (LCS), multiple salient objects (MSO), narrow salient object (NSO), off center (OC), and small salient object (SSO). We present the SSIM scores of SRAL and 16 competitive approaches in Table II, which reveals some findings that are not shown in Table I. First, some methods that do not perform significantly in Table I, such as SARNet [32], HVPNet [16], and MSCNet [17], yet reach the top three performances in some scenarios. The above-mentioned phenomenon illustrates that the performance preferences of various models for different remote sensing scene attributes cannot be shown by the metrics MAE or F_β on the whole test set. Second, HFANet [35], which achieves competitive results in Table I, has mediocre performance in various scenarios, which reflects the inconsistency between the performance on individual scenarios and on the entire dataset. Besides, the proposed SRAL has the most advantages on CS and LCS attributes, yet has some deficiencies in both CSO and OC and is the most disadvantaged on MSO. We blame the difficulties exhibited by SRAL in both cases on the lack of contextual information for multiple objects owing to the LR of the input RSIs. Most intuitively, SRAL ranks first in the average score of all scenarios, indicating that the proposed method is most robust to all scenarios, while other algorithms perform very poorly in some attributes, e.g., the attribute scores of SARNet [32] in the first four scenarios.

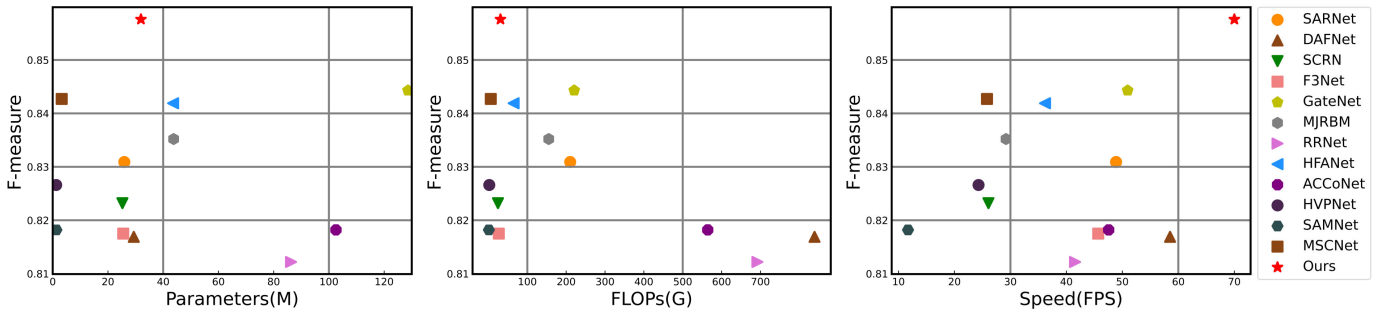


Fig. 7. Illustration of the tradeoffs between F-measure *versus* number of parameters, FLOPs, and inference speed for 13 state-of-the-art methods.

TABLE II

ATTRIBUTE-BASED PERFORMANCE ON THE ORSI-4199 DATASET [4]. THE AVERAGE SSIM SCORES FOR PARTICULAR ATTRIBUTES ARE PRESENTED. THE AVG. ROW REPORTS THE AVERAGE RESULTS FOR NINE ATTRIBUTES, AND TOP THREE SCORES IN EACH LINE ARE MARKED IN RED, GREEN, AND BLUE, RESPECTIVELY

Attributes	*NLDF [65]*	*DSS [66]*	*RAS [12]*	*PFAN [67]*	*SARNet [32]*	*DAFNet [3]*	*SCRN [68]*	*GateNet [69]*	*F3Net [70]*	*RRNet [33]*	*MJRBM [4]*	*FSMINet [19]*	*ACCo-V [34]*	*HVPNet [16]*	*HFANet [35]*	*MSCNet [17]*	*Ours*
BSO	.7365	.7604	.7472	.7984	.7984	.8064	.8295	.8546	.8256	.8226	.8344	.7866	.8186	.8448	.8222	.8081	.8530
CS	.7395	.7512	.7650	.8021	.8123	.8151	.8274	.8539	.8234	.8138	.8451	.7904	.8179	.8366	.8287	.8278	.8581
CSO	.7074	.7368	.7313	.7714	.7665	.7831	.8001	.8234	.7887	.7900	.8003	.7651	.7931	.8219	.7875	.7781	.8172
ISO	.6496	.7088	.6595	.7280	.7573	.7485	.7762	.8110	.7872	.7538	.7868	.7127	.7632	.7934	.7884	.7905	.8107
LCS	.6305	.6369	.6816	.6756	.7182	.7018	.7094	.7318	.7071	.6886	.7332	.6610	.7001	.7213	.7171	.7429	.7447
MSO	.7296	.7266	.7234	.7739	.8162	.7817	.7949	.8156	.7902	.7702	.8183	.7472	.7603	.7740	.7965	.7920	.8123
NSO	.6719	.6674	.7356	.7496	.7949	.8054	.7933	.8552	.7916	.7873	.8391	.7827	.7716	.8599	.8202	.8787	.8648
OC	.7085	.7131	.6915	.7457	.8071	.7780	.7474	.7730	.7481	.7301	.8055	.7128	.7553	.7365	.7967	.7731	.7983
SSO	.6632	.6585	.6703	.6984	.7596	.7219	.7092	.7311	.7179	.6897	.7456	.6721	.6972	.6929	.7308	.7209	.7490
Avg.	.6930	.7067	.7117	.7492	.7812	.7713	.7764	.8055	.7755	.7607	.8009	.7367	.7641	.7868	.7876	.7902	.8120

4) Efficiency Comparison: To show the compromise between accuracy and efficiency, we draw three scatter plots to demonstrate the relationship between F-measure *versus* the model parameters, floating-point operations (FLOPs), and inference speed of 13 approaches on the ORSI-4199 test dataset in Fig. 7. In the curves of F-measure *versus* parameters and F-measure *versus* FLOPs, SRAL is positioned at the top left, revealing that our method is less computationally expensive with high accuracy. Specifically, our model delivers excellent performance with 31.94M parameters, achieving the best tradeoff among all compared algorithms. It can be seen that although some lightweight algorithms, such as HVPNet [16], SAMNet [15], and MSCNet [17], significantly reduce the number of parameters by designing lightweight networks, and the FLOPs reach extremely low orders of magnitude. However, they do not achieve satisfactory results in terms of F-measure, which illustrates a common problem of lightweight methods, i.e., the limited parameters degrade the model's generalization ability, and thus, make it perform worse in complex remote sensing scenarios. In the curve of F-measure *versus* speed, the presented approach is located in the upper right corner, demonstrating an excellent tradeoff between detection accuracy and running speed. MSCNet [17] is a lightweight method designed explicitly for RSI-SOD, which has a competitive performance as revealed in Table I. However, we notice that its number of parameters and FLOPs are at the lowest level without an appreciable inference speed, which illustrates the inconsistency between the parameters and FLOPs *versus* inference speed. Based on this discovery, we determine that the approaches designed for NSI-SOD are not applicable to RSI-SOD due to the complex background of optical RSIs and the irregular topology of the salient objects, so we need to explore lightweight networks specifically designed for RSI-SOD in the future investigations. Notably, our algorithm dominates the inference speed because the FLOPs are only 1/4 of that in the case of 448 $\times$ 448 input, so the speed is considerably boosted. It shows that the way to improve the efficiency is not only to propose lightweight models but also to adjust the FLOPs by controlling the input resolution of RSIs effectively, e.g., the proposed SRAL. In summary, we can conclude that the presented model yields an excellent tradeoff between input resolution, accuracy, number of parameters, FLOPs, and speed.

C. Ablation Study

We first figure out the suitable settings of the combined objective function and reveal the effects of the presented components, i.e., TSDD, ASRD, and TFGM, respectively.

1) Baseline Setup: We adopt ResNet50 [51] with PPM [52] as encoder, and FPN [56] structure with deep supervision as a decoder to build the baseline model, which feeds 448 $\times$ 448 RSIs as input. The hyperparameters are set consistently in Tables III and IV for a fair comparison. As presented in Table IV, the baseline can reach an MAE of 0.0335 and F_β of 0.8540 with a speed of 50.64 ft/s on the ORSI-4199 dataset.

2) Hyperparameter Tuning: Table III presents hyperparameter tuning analysis on the ORSI-4199 dataset. First, when considering only $\mathcal{L}_{\text{SOD}} + \lambda_2 \mathcal{L}_{\text{SR}}$, we set the value of λ_2 from 10 to 500 based on the a priori knowledge that $\mathcal{L}_{\text{SR}}$ has a lower magnitude than $\mathcal{L}_{\text{SOD}}$ and SR performs

TABLE III
LOSS WEIGHTS TUNING ANALYSIS ON THE ORSI-4199 DATASET

$\mathcal{L}_{similar}$	$\mathcal{L}_{consist}$	$\lambda_2 \times 0.01$	λ_3	MAE ↓	F_β ↑	S_m ↑
×	×	–	–	0.0351	0.8451	0.8601
×	×	0.1	–	0.0343	0.8494	0.8670
×	×	1.0	–	0.0322	0.8505	0.8665
×	×	5.0	–	0.0336	0.8507	0.8672
✓	×	1.0	1.0	0.0330	0.8559	0.8722
×	✓	1.0	1.0	0.0327	0.8551	0.8706
✓	✓	1.0	1.0	**0.0321**	**0.8576**	**0.8735**

TABLE IV
ABLATION EXPERIMENTS ON THE ORSI-4199 [4] DATASET

TSDD	ASRD	TFGM	InputSize	MAE↓	F_β↑	S_m↑	FPS↑
			448×448	0.0335	0.8540	0.8726	50.64
			224×224	0.0349	0.8430	0.8541	**70.00**
✓			224×224	0.0351	0.8451	0.8601	
✓	✓		224×224	0.0322	0.8505	0.8665	
✓	✓	✓	224×224	**0.0321**	**0.8576**	**0.8735**	

secondary supervision as an auxiliary task. We observe that when $\lambda_2 = 100$, the model exceeds by a considerable margin compared with $\lambda_2 = 10$ or 500, and thus, we choose the value of 100 for λ_2 to train. On this basis, we define λ_3 as 1 to make these loss items ranges comparable for further experiments. Table III reveals that when a partial or complete TFGM module is introduced, there are certain boosts on all the F-measure metrics, justifying the significance of the proposed TFGM in facilitating the RSI-SOD task. We find that the performance boost is most noticeable when the $\mathcal{L}_{\text{SR}}$ and $\mathcal{L}_{\text{TFGM}}$ are introduced, i.e., both F-measure and MAE reach their peaks. In addition, two control experiments when only cosine similarity loss ($\mathcal{L}_{\text{similar}}$) or L1 constraint loss ($\mathcal{L}_{\text{consist}}$) is retained are also shown in Table III, and the strengths of both subloss functions are demonstrated separately according to the metrics.

3) Effects of TSDD, ASRD, and TFGM: In this section, ablation experiments and visualization are conducted to justify the effects of the presented components. Overall, we notice that the detection performance shows a continuous increasing trend in the quantitative results in Table IV. Observing the first and second rows about the numerical results of baseline under 224 × 224 and 448 × 448 training, respectively, the former performance is much worse than that of the latter, confirming that the LR optical RSIs significantly limit the prediction of HR representation. We propose TSDD to decode the HR SMs step by step in a spatial learning manner instead of fixed sampling calculations, thereby decreasing the difficulty of learning HR SMs from LR inputs. Comparing the second and third rows of Table IV reveals the incremental contribution of the proposed TSDD to RSI-SOD. To introduce an MTL framework for learning HR SMs from LR inputs, we design the ASRD to reconstruct the HR detailed representation and promote the results of SOD. In contrast to the third and fourth rows, it is shown that the MTL strategy contributes positively to the RSI-SOD task, i.e., implicit MSE supervision and joint MTL can boost the performance. Furthermore, we propose TFGM explicitly bootstrap the SOD task from the SR task and present two practical loss terms. The performance gains shown in Table IV demonstrate that it effectively transfers knowledge from SR. The reduction in input resolution also significantly shrinks the FLOPs of the model, which, in turn, enables the inference speed to increase from 50.64 to 70 ft/s.

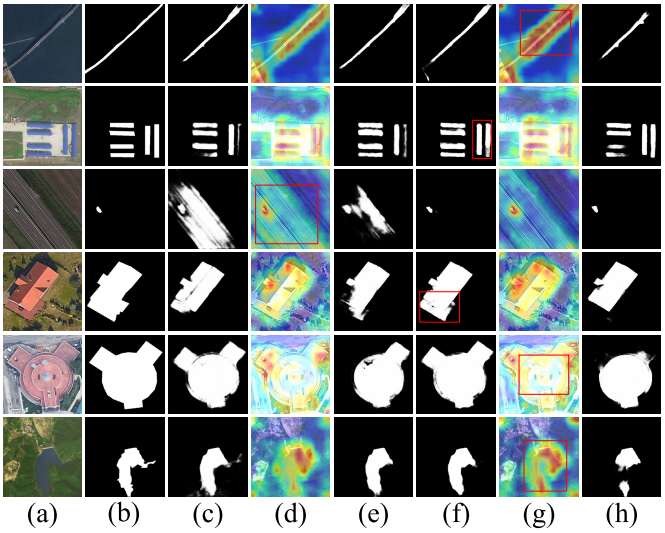


Fig. 8. SMs and feature visualizations of ablation study. (a) Optical RSIs. (b) GT. (c) and (d) Baseline + TSDD. (e) Baseline + TSDD + ASRD. (f) and (g) Baseline + TSDD + ASRD + TFGM. (h) Baseline with 448 × 448 input.

Fig. 8 displays some typical predictions and feature visualization results, which help us to draw the following findings: first, the proposed SRAL utilizes an MTL framework and TFGM to concentrate the model more on the saliency regions, e.g., rows 1, 2, 3, and 6; second, compared with baseline + TSDD, our entire model's antiinterference ability to complex background of optical RSIs is enhanced, e.g., row 3; third, the proposed model shows some benefits for learning boundaries and object completeness, e.g., rows 2, 4, and 5.

D. Model Analysis

We perform further experiments to show the convergence of the model and the effectiveness of the SR auxiliary network, compare the guidance efficiency of TFGM, and illustrate the model-agnostic ability and the failure analysis of SRAL, respectively.

1) Multiloss Convergence Analysis: To reveal the change of loss with training and convergence of our proposed model during training, we plot the curves of all types of losses *versus* epoch on the three datasets in Fig. 9. These curves about the ORSSD, EORSSD, and ORSI-4199 datasets show that all types of losses continue to decrease in value with training iterations on both the training and test datasets, leveling off and reaching convergence at the end of the training, and thus, the proposed model is stable, reliable, and convergent.

2) Visualization of Auxiliary Super-Resolution Task: To demonstrate the performance of SRAL for the SR subtask, we report the qualitative and quantitative results, as shown in Fig. 10 and Table V. The proposed model can reconstruct HR images with more complete structural details and smoother gradients compared with LR images in Fig. 10. We reproduce the DSRL [24], and modify the corresponding network structure to make it applicable to our task and conduct experiments. By observation, DSRL [24] is unable to achieve a quantitative performance comparable to our proposed SRAL for both the SR and SOD tasks. We summarize the mainly

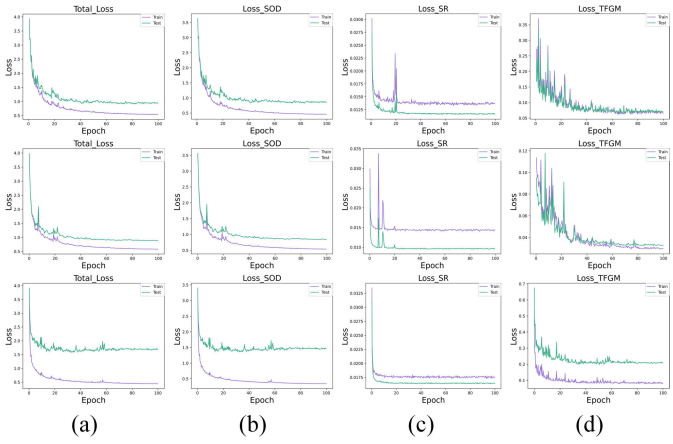


Fig. 9. Loss curves on three RSI-SOD datasets. Each row shows, in turn, the loss curves on the ORSSD, EORSSD, and ORSI-4199 datasets, respectively.

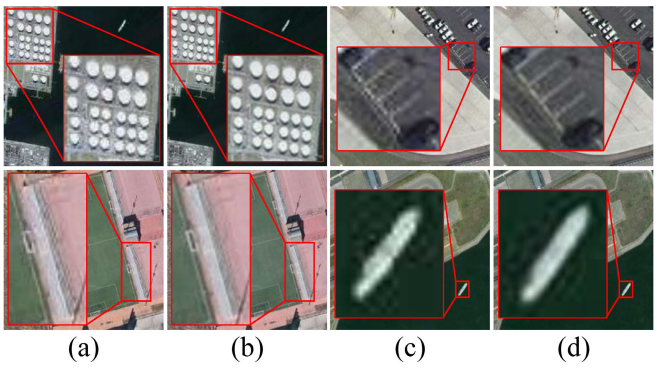


Fig. 10. Visualization of SR subnet of our SRAL. (a) and (c) LR images. (b) and (d) Predicted SR images of SRAL. Zoomed-in view for the best view.

possible reasons as follows: 1) the methodology designed by DSRL [24] aims at the semantic segmentation of natural images, which maybe result in the basic decoder network is not suitable for RSI-SOD. (2) There is no knowledge transfer module designed for the specific characteristics of the SOD task in [24], while we design a novel TFGM for efficient RSI-SOD, further increasing the performance gap between DSRL and our SRAL.

3) Guidance Comparison of TFGM: To reveal the superiority of TFGM for guiding RSI-SOD, we report several famous MTL training strategies (i.e., uncertainty [71], DWA [72], and GradVac [73]), and two multitask guidance methods for remote sensing (i.e., Zhang et al. [23] and Xie et al. [25]), as illustrated in Table VI. First, the method can obtain the results of 0.0322 and 0.8505 in terms of MAE and F_β, respectively, when only the SOD and SR branches are jointly supervised for training. Second, we discover that state-of-the-art MTL training strategies are not applicable to the RSI-SOD task, because the original intention of these methods is to exploit the union between multiple loss terms and improve the performance on all subtasks. This is contrary to our design philosophy, where our proposed SRAL treats SR as an auxiliary task and explores its possibility of facilitating RSI-SOD, and thus, these MTL-based methods yield poor results. Then, we introduce two recent learning approaches [23], [25] that use an auxiliary task to guide the main task in the remote

TABLE V

QUANTITATIVE SR PERFORMANCE COMPARISON STUDY

Datasets	Methods	Performance of SOD			Performance of SR	
		MAE ↓	F_β ↑	S_m ↑	SSIM ↑	PSNR ↑
ORSSD [2]	Nearest (SR)	–	–	–	0.9308	34.3265
	DSRL [24]	0.0162	0.8765	0.8895	0.9501	35.2629
	SRAL (Ours)	**0.0105**	**0.9167**	**0.9305**	**0.9513**	**35.2717**
EORSSD [3]	Nearest (SR)	–	–	–	0.9416	**36.6872**
	DSRL [24]	0.0120	0.8135	0.8683	0.9567	36.3348
	SRAL (Ours)	**0.0067**	**0.8964**	**0.9234**	**0.9571**	36.3354

TABLE VI

GUIDANCE COMPARISON OF TFGM ON THE ORSI-4199 DATASET

Guidance Strategies	Publications	MAE↓	F_β↑	S_m↑
None	—	0.0322	0.8505	0.8665
Uncertainty [71]	CVPR 2018	0.0351	0.8448	0.8570
DWA [72]	CVPR 2019	0.0337	0.8506	0.8641
GradVac [73]	ICLR 2021	0.0342	0.8485	0.8647
Zhang *et al.* [23]	TGRS 2022	0.0340	0.8520	0.8677
Xie *et al.* [25]	TGRS 2022	0.0338	0.8530	0.8700
Ours	—	**0.0321**	**0.8576**	**0.8735**

sensing community. Both methods have a gain, yet are limited in both S_m and F_β scores. The studies of Zhang et al. [23] and Xie et al. [25] are not specifically designed for the RSI-SOD task and do not consider feature integration of SR and SOD, as well as spatial guidance of SR for SOD and saliency enhancement of SOD for SR. Typically, they only introduce SSIM loss or cosine similarity loss on the features of two tasks for supervision in the training phase, while failing to achieve superior performance boosts for the specialized RSI-SOD task. Finally, the presented TFGM yields the optimal numerical results, emphasizing the effectiveness of the two loss terms designed explicitly for RSI-SOD.

4) Model-Agnostic Advantages of SRAL: To further prove the advantages of the SRAL, we apply our learning framework to five typical approaches, e.g., two famous networks for natural images, PSPNet [52], DeepLabv3+ [74], two mainstream NSI-SOD methods, DSS [66], SCRN [68], and an updated approach for RSI-SOD, FSMINet [19]. Among them, PSPNet, DeepLabv3+, and SCRN employ ResNet50 as the backbone, DSS utilizes VGG16 as the backbone, while FSMINet designs its particular backbone, and they propose various decoder structures. A series of compared experiments on three datasets about the above five algorithms trained at 224 × 224 input with and without the presented SRAL are reported in Table VII.

Overall, we find that the methods that integrate the proposed SRAL learning strategy show significant improvements in most metrics on all datasets. Among them, PSPNet [52] and DeeplabV3+ [74] adopt a plain encoder–decoder paradigm and utilize the efficient ResNet50 as the encoder without introducing any complex structured modules, which brings a significant performance boost to MTL. By contrast, SCRN [68] employs multiple stacked cross modules to refine the saliency results, and DSS [66] introduces unique and complicated shortcut connections with VGG16 as the encoder, which all limit the boosting effect brought by the proposed SRAL. Notably, the network structure of FSMINet [19] is completely customized without reference to any mainstream encoders, and our framework still has performance gain on this approach, justifying the generalization of SRAL to various network

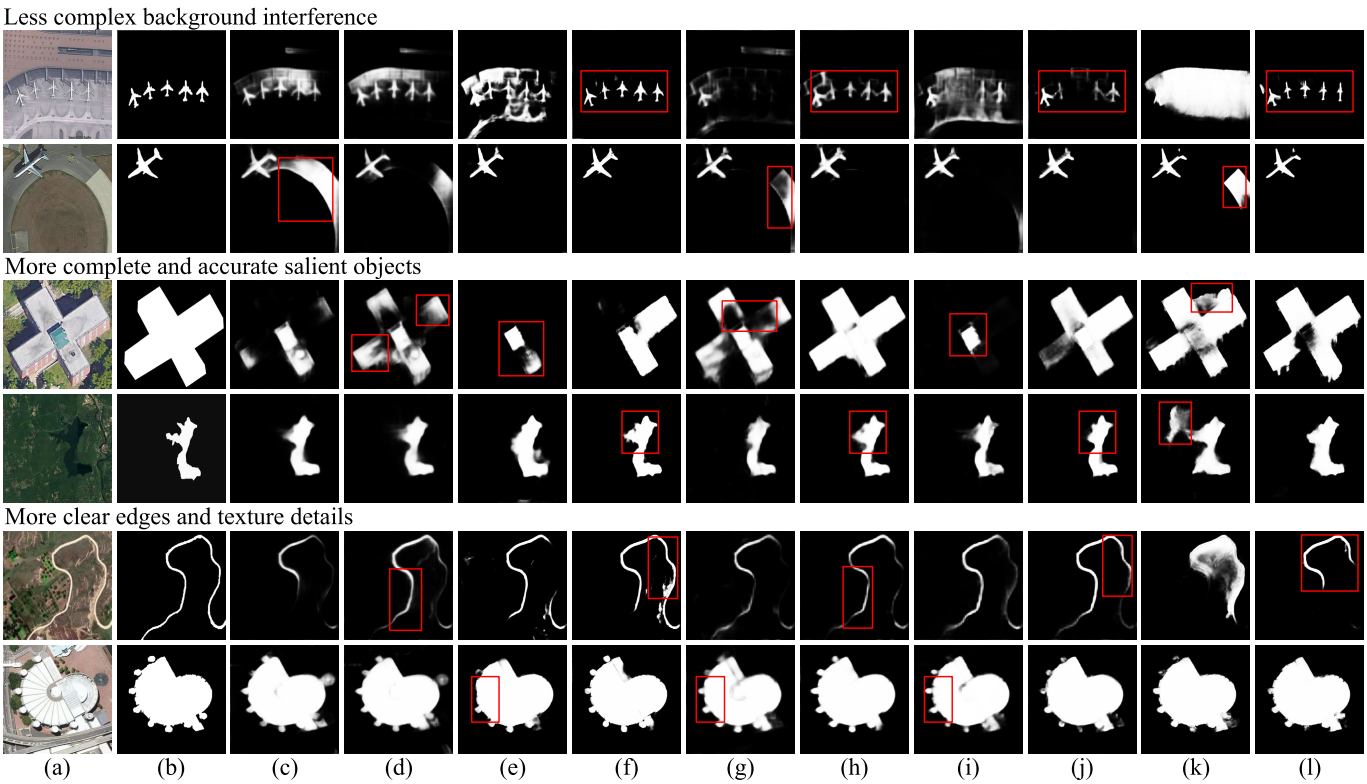


Fig. 11. Visualization of several typical SMs of five algorithms with and without the proposed SRAL, trained at 224 × 224 inputs on the ORSI-4199 test dataset. (a) Optical RSIs. (b) GT. (c) DSS [66]. (d) DSS (ours). (e) PSPNet [52]. (f) PSPNet (ours). (g) DeepLabv3+ [74]. (h) DeepLabv3+ (ours). (i) SCRN [68]. (j) SCRN (ours). (k) FSMINet [19]. (l) FSMINet (ours).

TABLE VII
QUANTITATIVE RESULTS OF FIVE TYPICAL METHODS ON THREE RSI-SOD DATASETS WITH AND WITHOUT THE SRAL FRAMEWORK

Methods	Publication	Input	ORSSD [2]			EORSSD [3]			ORSI-4199 [4]		
			$F_\beta\uparrow$	MAE↓	$S_m\uparrow$	$F_\beta\uparrow$	MAE↓	$S_m\uparrow$	$F_\beta\uparrow$	MAE↓	$S_m\uparrow$
DSS [66]	PAMI 2019	224×224	0.8422	0.0253	**0.8757**	**0.7761**	0.0167	0.8400	0.7762	0.0558	**0.8200**
DSS (Ours)		224×224	**0.8429**	**0.0229**	0.8723	0.7742	**0.0158**	**0.8424**	**0.7825**	**0.0545**	0.8191
PSPNet [52]	CVPR 2017	224×224	0.8993	0.0130	0.9154	0.8689	0.0089	0.9080	0.8430	0.0349	0.8541
PSPNet (Ours)		224×224	**0.9112**	**0.0107**	**0.9239**	**0.8886**	**0.0070**	**0.9210**	**0.8582**	**0.0322**	**0.8728**
DeepLabv3+ [74]	ECCV 2018	224×224	0.8759	0.0172	0.8883	0.8164	0.0123	0.8643	0.8203	0.0419	0.8524
DeepLabv3+ (Ours)		224×224	**0.8875**	**0.0157**	**0.8974**	**0.8325**	**0.0098**	**0.8834**	**0.8307**	**0.0391**	**0.8614**
SCRN [68]	ICCV 2019	224×224	0.8700	0.0175	0.8875	0.8046	0.0126	0.8524	0.8332	0.0385	0.8581
SCRN (Ours)		224×224	**0.8868**	**0.0162**	**0.8938**	**0.8282**	**0.0119**	**0.8711**	**0.8392**	**0.0372**	**0.8674**
FSMINet [19]	GRSL 2022	224×224	0.8308	0.0221	0.8769	0.8283	0.0124	0.8829	0.7912	0.0483	0.8230
FSMINet (Ours)		224×224	**0.8501**	**0.0202**	**0.8852**	**0.8414**	**0.0108**	**0.8920**	**0.8046**	**0.0464**	**0.8285**

structures of different algorithms. Therefore, the proposed SRAL is model-agnostic for the RSI-SOD task that is not specialized to a particular network structure but has a positive facilitation effect in many algorithms.

As illustrated in Fig. 2, our proposed SRAL can achieve considerable performance gains in the scenario of input LR RSIs under a condition that does not introduce additional computational costs. Compared with input HR RSIs, our strategy can boost the inference speed as much as possible, which creates favorable conditions for the algorithm to be deployed on resource-constrained and real-world devices.

To reveal how the proposed SRAL promotes the performance of the RSI-SOD task, we provide some typical prediction results of the above-mentioned five methods on the ORSI-4199 dataset, as shown in Fig. 11. The models trained with SRAL achieve better results than the original models in various scenarios, and we summarize the advantages as follows. First, SRAL enables the SOD branch to maintain rich HR representation and focus more on the foreground regions through the MTL framework and feature guidance learning, as in rows 1 and 2, which alleviates the interference issue of the complex background of optical RSIs. Second, observing rows 3 and 4, the proposed SRAL generates more accurate segmentation of local regions of saliency objects and enhanced learning ability for complete saliency objects by benefiting from the auxiliary supervision of HR and the consistency constraint of SOD and SR. Third, for saliency objects with irregular topology and complex edges, SRAL can enhance the boundary recognition capability, and the HR SMs predicted by SRAL have clearer image details instead of many blurs.

5) Failure Cases and Future Work: Despite the proposed SRAL achieving superior performance on all RSI-SOD datasets, there are still some hard samples. To analyze the

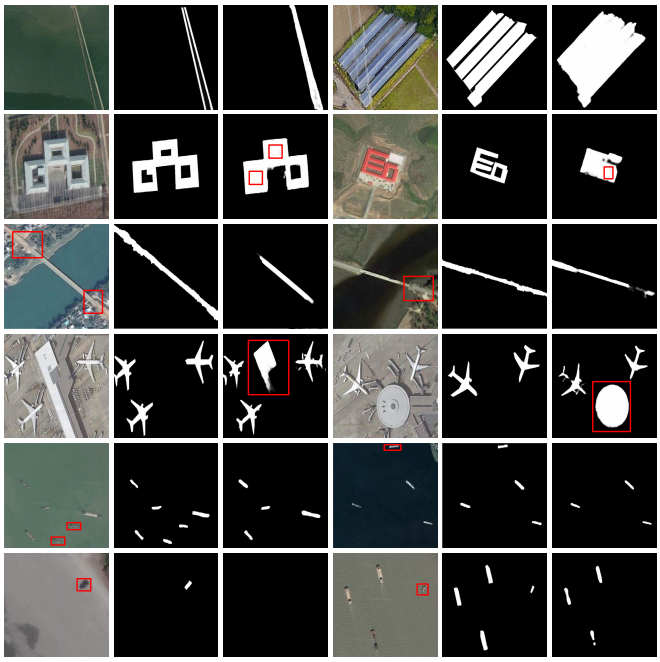


Fig. 12. Failure cases of our proposed approach. (a) and (d) Optical RSIs. (b) and (e) GT. (c) and (f) Predicted maps of our SRAL.

shortcomings of the SRAL and find directions for future investigations, some typically failed samples are illustrated in Fig. 12. First, the model always identifies the background regions in the middle of the neighboring objects or background regions within the salient regions as the foreground, and thus, results in false positive predictions, such as rows 1 and 2 in Fig. 12. Second, for the foreground object part that is confused with the background, our detector may cause a missed detection and determine it as a background region, as revealed in row 3. Third, subtle contrast differences between the salient and nonsalient objects cause our detector to potentially identify nonsalient targets as foreground. For instance, in row 4 of Fig. 12, the airplanes and airport buildings are very similar in color, with little difference other than the edges and target structure, and our method incorrectly recognizes the airport buildings as salient objects. In addition, the proposed SRAL fails to capture contextual information about the small objects close to the boundary part of the RSI, always causing missed detection. As presented in row 5, for individual vessels appearing at different locations in the optical RSIs, our detector can well segment the vessels located in the central region of the RSIs well while ignoring the vessels near the boundary part of the RSIs. Finally, SRAL performs poorly on some tiny salient objects in optical RSIs, such as the last row in Fig. 12.

Based on our analysis, the reasons for the above failures are as follows. First, since our model feeds LR RSIs and they cannot provide sufficient contextual information desired by the deep network, and thus, it has difficulty in recognizing tiny objects, especially those near the boundaries of RSIs, which always causes missed detection because of the absence of contextual information around themselves. Second, our model cannot extract the long-distance independence of salient objects because it is based entirely on the local inductive bias of convolution operations, which leads to partial incomplete or over-segmentation for some salient objects. Third, we propose TFGM to guide and supervise by feature fusion of SR and SOD branches for the main task, which makes the model focus more on the most salient objects and may reduce the attention to those foreground regions mixed with the background. Finally, to train the model on space-limited GPUs, the affinity matrices in TFGM are 16× downsampled, which makes the HR SMs absorb the SR representation, whereas some local details may be neglected.

In the future study, we plan to address the above-mentioned problems in the following directions: 1) further ameliorate the calculation strategy of the affinity matrices of TFGM instead of direct 16× spatial downsampling to capture more localized HR representations in optical RSIs. 2) Propose a lightweight network based on the long-distance context to obtain larger receptive fields, and thus, handle the challenge of incomplete or overcomplete detection. 3) Design a practical spatial attention module to cope with the complicated issue of identifying tiny objects, such as vehicles and vessels, in optical RSIs. 4) Propose an algorithm that adaptively adjusts the resolution of input RSIs according to the capacities of different models to boost the inference efficiency of various methods.

V. Conclusion

In this article, we present the first LR input-based framework with MTL strategy to generate accurate HR SMs for RSI-SOD, which can improve performance and inference speed in LR conditions. We first design a dual-branch network based on a residual encoder and two individual heterogeneous decoders, i.e., TSDD and ASRD, for SOD tasks and SR tasks, to achieve this goal. The former adopts a layer-by-layer upsampling learning and deep supervision strategy to gradually reconstruct the resolution of the predicted SMs, while the latter combines the shallowest and deepest features of the encoder to estimate the HR optical RSIs. To explicitly guide the learning process of RSI-SOD, we propose a novel TFGM to distill the domain knowledge with HR structural representation from the SR branch to the SOD branch. Extensive experiments reveal that SRAL outperforms more than 20 state-of-the-art algorithms under various input spatial resolutions and yields considerable efficiency. Both ASRD and TFGM can be omitted in the inference phase without any computational cost, and yield favorable conditions for deployment in resource-constrained and real-world applications. Besides, we extend the proposed SRAL to five typical methods to show its model-agnostic capability, and adequate failure analysis indicates our future research directions.

References

[1] Z. Xiong, F. Zhang, Y. Wang, Y. Shi, and X. X. Zhu, "EarthNets: Empowering AI in Earth observation," 2022, *arXiv:2210.04936*.

[2] 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.

[3] Q. Zhang et al., "Dense attention fluid network for salient object detection in optical remote sensing images," *IEEE Trans. Image Process.*, vol. 30, no. 10, pp. 1305–1317, Dec. 2021.

[4] Z. Tu, C. Wang, C. Li, M. Fan, H. Zhao, and B. Luo, "ORSI salient object detection via multiscale joint region and boundary model," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–13, 2022, Art. no. 5607913.

[5] Z. Xiong, Y. Yuan, and Q. Wang, "ASK: Adaptively selecting key local features for RGB-D scene recognition," *IEEE Trans. Image Process.*, vol. 30, pp. 2722–2733, 2021.

[6] Y. Yuan, Z. Xiong, and Q. Wang, "VSSA-NET: Vertical spatial sequence attention network for traffic sign detection," *IEEE Trans. Image Process.*, vol. 28, no. 7, pp. 3423–3434, Jul. 2019.

[7] Z. Xiong, Y. Yuan, N. Guo, and Q. Wang, "Variational context-deformable ConvNets for indoor scene parsing," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 3992–4002.

[8] Y. Liu, Q. Li, Y. Yuan, and Q. Wang, "Single-shot balanced detector for geospatial object detection," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2022, pp. 2529–2533.

[9] Y. Liu, Q. Li, Y. Yuan, Q. Du, and Q. Wang, "ABNet: Adaptive balanced network for multiscale object detection in remote sensing imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–14, 2022, Art. no. 5614914.

[10] M. Li, M. Li, P. Zhang, Y. Wu, W. Song, and L. An, "SAR image change detection using PCANet guided by saliency detection," *IEEE Geosci. Remote Sens. Lett.*, vol. 16, no. 3, pp. 402–406, Mar. 2019.

[11] H. Wu, L. Zhang, and J. Ma, "Remote sensing image super-resolution via saliency-guided feedback GANs," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–16, 2022, Art. no. 5600316.

[12] S. Chen, X. Tan, B. Wang, and X. Hu, "Reverse attention for salient object detection," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Jul. 2018, pp. 234–250.

[13] 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 Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 3912–3921.

[14] M.-M. Cheng, S.-H. Gao, A. Borji, Y.-Q. Tan, Z. Lin, and M. Wang, "A highly efficient model to study the semantics of salient object detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 11, pp. 8006–8021, Nov. 2022.

[15] 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.

[16] 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. 2021.

[17] Y. Lin, H. Sun, N. Liu, Y. Bian, J. Cen, and H. Zhou, "A lightweight multi-scale context network for salient object detection in optical remote sensing images," in *Proc. 26th Int. Conf. Pattern Recognit. (ICPR)*, Aug. 2022, pp. 238–244.

[18] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "MobileNetV2: Inverted residuals and linear bottlenecks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2018, pp. 4510–4520.

[19] K. Shen, X. Zhou, B. Wan, R. Shi, and J. Zhang, "Fully squeezed multiscale inference network for fast and accurate saliency detection in optical remote-sensing images," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, pp. 1–5, 2022.

[20] 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, pp. 1–12, 2022, Art. no. 5617712.

[21] P. Wimmer, J. Mehnert, and A. Condurache, "Interspace pruning: Using adaptive filter representations to improve training of sparse CNNs," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 12527–12537.

[22] Z. Liu, Y. Wang, K. Han, S. Ma, and W. Gao, "Instance-aware dynamic neural network quantization," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 12434–12443.

[23] Q. Zhang, G. Yang, and G. Zhang, "Collaborative network for super-resolution and semantic segmentation of remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–12, 2022, Art. no. 4404512.

[24] L. Wang, D. Li, Y. Zhu, L. Tian, and Y. Shan, "Dual super-resolution learning for semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 3773–3782.

[25] J. Xie, L. Fang, B. Zhang, J. Chanussot, and S. Li, "Super resolution guided deep network for land cover classification from remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–12, 2022, Art. no. 5611812.

[26] Y. Liu, W. Xie, Y. Li, Z. Li, and Q. Du, "Dual-frequency autoencoder for anomaly detection in transformed hyperspectral imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–13, 2022, Art. no. 5523613.

[27] W. Park, D. Kim, Y. Lu, and M. Cho, "Relational knowledge distillation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 3967–3976.

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

[29] L. Zhang, Y. Wang, and Y. Sun, "Salient target detection based on the combination of super-pixel and statistical saliency feature analysis for remote sensing images," in *Proc. 25th IEEE Int. Conf. Image Process. (ICIP)*, Oct. 2018, pp. 2336–2340.

[30] 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.

[31] X. Zhou, K. Shen, Z. Liu, C. Gong, J. Zhang, and C. Yan, "Edge-aware multiscale feature integration network for salient object detection in optical remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–15, 2022, Art. no. 5605315.

[32] 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.

[33] 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, 2022, Art no. 5613311.

[34] 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. 2023.

[35] Q. Wang, Y. Liu, Z. Xiong, and Y. Yuan, "Hybrid feature aligned network for salient object detection in optical remote sensing imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–15, 2022, Art. no. 5624915.

[36] C. Dong, C. C. Loy, K. He, and X. Tang, "Learning a deep convolutional network for image super-resolution," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Sep. 2014, pp. 184–199.

[37] Y. Tai, J. Yang, and X. Liu, "Image super-resolution via deep recursive residual network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 2790–2798.

[38] J. Kim, J. K. Lee, and K. M. Lee, "Accurate image super-resolution using very deep convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 1646–1654.

[39] S. Lei, Z. Shi, and Z. Zou, "Super-resolution for remote sensing images via local–global combined network," *IEEE Geosci. Remote Sens. Lett.*, vol. 14, no. 8, pp. 1243–1247, Aug. 2017.

[40] W. Ma, Z. Pan, J. Guo, and B. Lei, "Achieving super-resolution remote sensing images via the wavelet transform combined with the recursive res-net," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 6, pp. 3512–3527, Jun. 2019.

[41] C. Ma, Z. Jiang, Y. Rao, J. Lu, and J. Zhou, "Deep face super-resolution with iterative collaboration between attentive recovery and landmark estimation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 5568–5577.

[42] J. Feng et al., "A deep multitask convolutional neural network for remote sensing image super-resolution and colorization," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–15, 2022, Art. no. 5407915.

[43] Y. Wang, W. Ding, R. Zhang, and H. Li, "Boundary-aware multitask learning for remote sensing imagery," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 951–963, 2021.

[44] L. Mou and X. X. Zhu, "Vehicle instance segmentation from aerial image and video using a multitask learning residual fully convolutional network," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 11, pp. 6699–6711, Nov. 2018.

[45] Z. Yuan, L. Mou, Z. Xiong, and X. X. Zhu, "Change detection meets visual question answering," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–13, 2022, Art. no. 5630613.

[46] K. Heidler, L. Mou, C. Baumhoer, A. Dietz, and X. X. Zhu, "HED-UNet: Combined segmentation and edge detection for monitoring the Antarctic coastline," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–14, 2022, Art. no. 4300514.

[47] D. Liu, B. Wen, J. Jiao, X. Liu, Z. Wang, and T. S. Huang, "Connecting image denoising and high-level vision tasks via deep learning," *IEEE Trans. Image Process.*, vol. 29, pp. 3695–3706, 2020.

[48] A. Aakerberg, A. S. Johansen, K. Nasrollahi, and T. B. Moeslund, "Semantic segmentation guided real-world super-resolution," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. Workshops (WACVW)*, Jan. 2022, pp. 449–458.

[49] S. Lei, Z. Shi, X. Wu, B. Pan, X. Xu, and H. Hao, "Simultaneous super-resolution and segmentation for remote sensing images," in *Proc. IEEE Int. Geosci. Remote Sens. Symp. (IGARSS)*, Jul. 2019, pp. 3121–3124.

[50] T. Evgeniou, C. A. Micchelli, and M. Pontil, "Learning multiple tasks with kernel methods," *J. Mach. Learn. Res.*, vol. 6, no. 21, pp. 615–637, 2005.

[51] 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.

[52] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 6230–6239.

[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] B. Lim, S. Son, H. Kim, S. Nah, and K. Mu Lee, "Enhanced deep residual networks for single image super-resolution," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR) Workshops*, Jul. 2017, pp. 1132–1140.

[55] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent. (MICCAI)*, Oct. 2015, pp. 234–241.

[56] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Nov. 2017, pp. 2117–2125.

[57] Z. Guo et al., "Super-resolution integrated building semantic segmentation for multi-source remote sensing imagery," *IEEE Access*, vol. 7, pp. 99381–99397, 2019.

[58] J. Yang, K. Fu, Y. Wu, W. Diao, W. Dai, and X. Sun, "Mutual-feed learning for super-resolution and object detection in degraded aerial imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–16, 2022, Art. no. 5628016.

[59] M. Liu, Q. Shi, A. Marinoni, D. He, X. Liu, and L. Zhang, "Super-resolution-based change detection network with stacked attention module for images with different resolutions," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–18, 2022, Art. no. 4403718.

[60] 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. (CVPR)*, Jun. 2012, pp. 733–740.

[61] R. Achanta, S. Hemami, F. Estrada, and S. Susstrunk, "Frequency-tuned salient region detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2009, pp. 1597–1604.

[62] 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. 4558–4567.

[63] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Trans. Image Process.*, vol. 13, no. 4, pp. 600–612, Apr. 2004.

[64] Y. Zhai and M. Shah, "Visual attention detection in video sequences using spatiotemporal cues," in *Proc. 14th ACM Int. Conf. Multimedia*, Oct. 2006, pp. 815–824.

[65] Z. Luo, A. Mishra, A. Achkar, J. Eichel, S. Li, and P.-M. Jodoin, "Non-local deep features for salient object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Nov. 2017, pp. 6593–6601.

[66] Q. Hou, M.-M. Cheng, X. Hu, A. Borji, Z. Tu, and P. H. S. Torr, "Deeply supervised salient object detection with short connections," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 4, pp. 815–828, Apr. 2019.

[67] T. Zhao and X. Wu, "Pyramid feature attention network for saliency detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 3080–3089.

[68] Z. Wu, L. Su, and Q. Huang, "Stacked cross refinement network for edge-aware salient object detection," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 7263–7272.

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

[70] J. Wei, S. Wang, and Q. Huang, "F^3Net: Fusion, feedback and focus for salient object detection," in *Proc. AAAI Conf. Artif. Intell. (AAAI)*, vol. 34, 2020, pp. 12321–12328.

[71] A. Kendall, Y. Gal, and R. Cipolla, "Multi-task learning using uncertainty to weigh losses for scene geometry and semantics," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Dec. 2018, pp. 7482–7491.

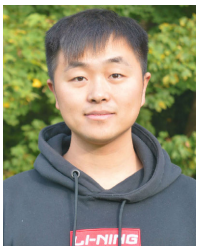
[72] S. Liu, E. Johns, and A. J. Davison, "End-to-end multi-task learning with attention," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 1871–1880.

[73] Z. Wang, Y. Tsvetkov, O. Firat, and Y. Cao, "Gradient vaccine: Investigating and improving multi-task optimization in massively multilingual models," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2021, pp. 1–12.

[74] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Sep. 2018, pp. 833–851.

Yanfeng Liu (Student Member, IEEE) received the B.E. degree in computer science and technology from Northeast Forestry University, Harbin, China, in 2021. He is currently pursuing the M.S. degree in computer science and technology with the School of Computer Science and School of Artificial Intelligence, Optics and Electronics (iOPEN), Northwestern Polytechnical University, Xi'an, China.

His research interests include computer vision, pattern recognition, and remote sensing.



Zhitong Xiong (Member, IEEE) received the Ph.D. degree in computer science and technology from Northwestern Polytechnical University, Xi'an, China, in 2021.

He is currently a Senior Scientist and the Leader of the ML4Earth Working Group, Chair of Data Science in Earth Observation, Technical University of Munich (TUM), Munich, Germany. His research interests include computer vision, machine learning, label-efficient learning, and Earth observation.

Yuan Yuan (Senior Member, IEEE) is currently a Full Professor with the School of Artificial Intelligence, Optics and Electronics (iOPEN), Northwestern Polytechnical University, Xi'an, China. She has authored or coauthored over 150 papers, including about 100 in reputable journals, such as the IEEE TRANSACTIONS AND PATTERN RECOGNITION, and conference papers in CVPR, BMVC, ICIP, and ICASSP. Her research interests include visual information processing and image/video content analysis.

Qi Wang (Senior Member, IEEE) received the B.E. degree in automation and the Ph.D. degree in pattern recognition and intelligent systems from the University of Science and Technology of China, Hefei, China, in 2005 and 2010, respectively.

He is currently a Professor with the School of Artificial Intelligence, Optics and Electronics (iOPEN), Northwestern Polytechnical University, Xi'an, China. His research interests include computer vision, pattern recognition, and remote sensing.