Transcending Pixels: Boosting Saliency Detection via Scene Understanding From Aerial Imagery

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

***Abstract*— Existing remote sensing image salient object detection (RSI-SOD) methods widely perform object-level semantic understanding with pixel-level supervision, but ignore the image-level scene information. As a fundamental attribute of remote sensing images (RSIs), the scene has a complex intrinsic correlation with salient objects, which may bring hints to improve saliency detection performance. However, existing RSI-SOD datasets lack both pixel- and image-level labels, and it is non-trivial to effectively transfer the scene domain knowledge for more accurate saliency localization. To address these challenges, we first annotate the image-level scene labels of three RSI-SOD datasets inspired by remote sensing scene classification. On top of it, we present a novel scene-guided dual-branch network (SDNet), which can perform cross-task knowledge distillation from the scene classification to facilitate accurate saliency detection. Specifically, a scene knowledge transfer module (SKTM) and a conditional dynamic guidance module (CDGM) are designed for extracting saliency key area as spatial attention from the scene subnet and guiding the saliency subnet to generate scene-enhanced saliency features, respectively. Finally, an object contour awareness module (OCAM) is introduced to enable the model to focus more on irregular spatial details of salient objects from the complicated background. Extensive experiments reveal that our SDNet outperforms over 20 state-of-the-art algorithms on three datasets. Moreover, we prove that the proposed framework is model-agnostic, and its extension to six baselines can bring significant performance benefits. Code is available at https://github.com/lyf0801/SDNet.**

***Index Terms*— Conditional guidance learning, dynamic class activation map (CAM), optical remote sensing image (RSI), salient object detection (SOD), scene knowledge distillation.**

I. Introduction

RECENTLY, salient object detection (SOD) in optical remote sensing images (RSIs) [1] has attracted much research interest in remote sensing community. It aims to identify and locate various visually salient objects in complex RSIs and compute clear saliency maps (SMs). As revealed in [2], SOD has extensive applications as a preprocessing technique in numerous areas [3], [4], [5], [6]. However, the irregular topology and scale diversity of man-made salient objects, as well as the low contrast and complicated background of aerial imagery, have presented many challenges to RSI-SOD.

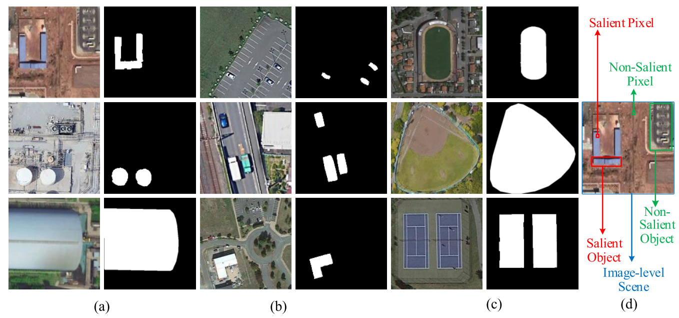


Fig. 1. Illustrations of different saliency objects in several various scenes. (a) Industrial facilities. (b) Transportation facilities. (c) Sports facilities. (d) Scene-, object-, and pixel-level elements in an RSI.

Benefiting from the development of convolutional neural networks (CNNs) and SOD in natural scene images (NSIs) [7], [8], [9], RSI-SOD has witnessed substantial progress in recent years. There are several public datasets [10], [11], [12], [13] released as benchmarks for evaluating different approaches. To tackle the above challenges, many sophisticated methods have been proposed from various perspectives and delivered excellent performance [14], [15], [16]. For instance, SARNet [14] constructs semantic attention mechanisms at channel-wise and spatial levels to refine SMs. HFANet [15] combines CNN and Transformer to model local and global contexts simultaneously, and utilizes adjacent feature alignment modules to aggregate multiscale salient features. SRAL [16] designs a novel multitask network of RSI-SOD with super-resolution and efficiently transfers the image reconstructed knowledge to the learning procedure of saliency detection. Nevertheless, the existing algorithms all employ the encoder-decoder frameworks for training in a fully pixel-wise supervised manner. They only perform semantic understanding at object and pixel levels, but ignore an essential property of RSIs, i.e., the scene. That is, the correlation between scenes and salient objects, and the impact of intrinsic dependencies on RSI-SOD, are still unexplored.

As shown in Fig. 1(d), scene, as an image-level attribute, is higher than pixel- and object-level elements of RSIs, and

Manuscript received 9 May 2023; revised 28 June 2023; accepted 21 July 2023. Date of publication 25 July 2023; date of current version 9 August 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.2023.3298661

can be exploited as a fundamental research subject [17]. The scenes of RSIs are closely connected with objects, and this correlation is complex and intrinsic, with great interclass similarity and intraclass variability [18]. In fact, salient objects are highly correlated with the contextual information of their remote sensing scenes, i.e., instances of a particular class of objects appear with a very high probability in some specific scenes [19]. As in Fig. 1(a)–(c), in an industrial scene, the salient objects in RSIs are most likely to be storage tanks, factories, etc. In a transportation scenario, the salient objects might be cars, buildings, roads, etc. In a sports facility scenario, the salient objects will most likely be ground track fields, tennis courts, baseball fields, etc. To foster other tasks, learning scene information from aerial images is a potentially effective way [20]. For instance, Zhang et al. [21] introduce a scene preprocessing step before performing super-resolution and design multiple scene-independent super-resolution networks for fine-grained learning, while a similar idea is adopted by Tao et al. [22] for vehicle detection tasks.

Thus, it would be very insightful to model the relation between scenes and salient objects as well as to facilitate the final RSI-SOD performance. Based on this observation, our goal is to introduce more universal scene information of RSIs, and to distill the knowledge of precise spatial localization in scenes into the RSI-SOD task for more effective saliency detection from complicated aerial images.

However, it is non-trivial to introduce scene information into the existing RSI-SOD models and contribute to the final detection results. We consider that the specific challenges can be summarized in the following three points.

1) The existing RSI-SOD datasets only provide pixel-level ground truths (GTs) of salient objects, and lack sound definition and classification of specific scene categories.
2) How to effectively model and distill the accurate location and edge information of category-agnostic salient objects in aerial images from scene classification?
3) How to deploy scene knowledge to enhance the spatial features of salient objects and perform dynamically and conditionally guided learning for RSI-SOD?

To cope with the above challenges, we first consider the specific knowledge of remote sensing scene classification [18], observe the primary scene distribution of the three existing RSI-SOD datasets, and then define 12 unified scenes. Fig. 2 illustrates the distribution of 12 types of scenes in these three datasets. Specifically, we describe these scenes as airplane facilities, industrial facilities, bridges, ships, rural buildings, transportation facilities, highways, rivers, lakes, islands, sports facilities, and others. It can be seen that these scenes show various percentage distributions in several datasets, which also reflects that the variability of scenes among datasets is a great challenge for the generalizability of the model.

Then, we propose a multitask learning (MTL)-based scene-guided dual-branch network (SDNet), which distills precise localization knowledge of salient objects from scene classification and then performs dynamic guided learning to further compensate for saliency detection. As shown in Fig. 3, the current mainstream RSI-SOD methods all utilize the encoder-decoder architectures, feeding RSI and producing SMs. In contrast, our proposed SDNet employs an MTL framework to simultaneously learn image-level scene and pixel-level saliency, and strengthen the decoding for saliency parsing by scene knowledge transfer. To achieve this goal, we fully exploit the scene subnet to generate multiscale scene features and dynamic class activation maps (CAMs) [23]. With these ingredients, we design a novel scene knowledge transfer module (SKTM) to obtain the scene-saliency representational knowledge. On top of that, a conditional dynamic guidance module (CDGM) is proposed to further guide the learning of scene-enhanced saliency features for more accurate salient object localization in RSIs. Furthermore, motivated by Laplacian pyramid, we introduce an object contour awareness module (OCAM) to enable the model to focus more on irregular spatial details of salient objects from the complicated background. Extensive experiments demonstrate that the presented SDNet can reach excellent performance among the three datasets. In addition, it can be successfully extended to other SOD models and achieve noticeable performance benefits over six baselines. The main contributions of this article are listed as follows.

1) We first manually annotate scene labels for three existing RSI-SOD datasets, providing supervised signals and research insights to enable scene-aware multitask modeling and image-level weakly supervised learning.
2) We present an MTL-based SDNet that combines scene classification and SOD, and design a scene knowledge distillation strategy to facilitate saliency detection.

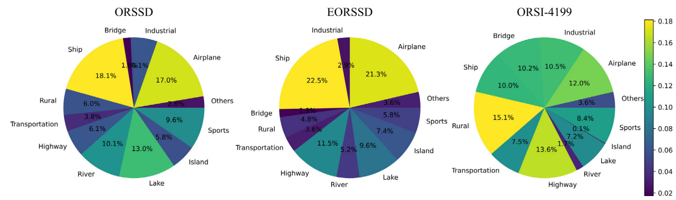


Fig. 2. Illustration of the percentage of 12 scenes in three RSI-SOD datasets.

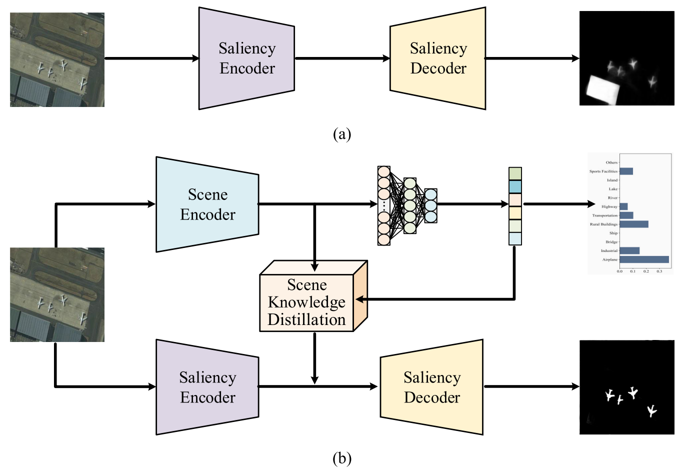


Fig. 3. Illustration of the proposed approach versus mainstream RSI-SOD models. (a) Current state-of-the-art encoder-decoder frameworks for RSI–SOD. (b) Our proposed scene knowledge distillation model for RSI-SOD.

3) Two plug-and-play modules, SKTM and CDGM, are proposed to extract scene-saliency knowledge and generate scene-enhanced saliency representation, respectively.
4) Extensive experiments show the superiority and model-agnostic capability of SDNet. Typically, it exhibits significant performance boosts in six models.

The remainder is presented as follows. Section II draws the work related to scene classification and RSI-SOD. We describe the methodology in Section III, and conduct adequate experiments in Section IV. Section V concludes this article.

II. Related Work

In this section, we present related studies of RSI-SOD, scene classification, and scene-related visual models, respectively.

A. SOD in Optical RSIs

In the remote sensing community, researches on RSI-SOD emerge early, when researchers apply traditional algorithms to extract and model low-level features of RSIs [24], [25], [26], such as luminance, texture, gradient, color, edges, or design handcrafted features for saliency recognition. However, these methods are not universally applicable due to their fixed feature production and limited saliency accuracy.

With the development of CNN and the release of large-scale datasets [10], [11], [12], [13], RSI-SOD has made unprecedented progress. There are numerous algorithms [14], [15], [16], [27], [28], [29], [30], [31] proposed to cope with the difficulties of complex background and low contrast of RSIs, scale diversity of man-made objects, irregular topology, and complex edges. For instance, Wang et al. [15] propose an interactive guidance loss function with joint edge prediction to extract irregular boundaries of salient objects effectively. Cong et al. [29] introduce graph convolutions for reasoning the spatial-wise and channel-wise relations of RSI-SOD. Several edge-aware models have been proposed, such as EMFINet [30] and MJRBM [12], which introduce boundary supervision and embeds edge attention modules, respectively. Additionally, many researchers investigate attention mechanisms for RSI-SOD, i.e., utilizing channel and spatial attention blocks for semantic feature refinement [14] or salient information aggregation [32]. Recently, Liu et al. [16] present a cross-task distillation-based model, which brings many insights to transfer knowledge from super-resolution into RSI-SOD efficiently, and significantly reduce the computation costs as well as accelerate the inference of model.

Nonetheless, the existing methods all adopt the encoder-decoder frameworks to perform pixel-wise saliency and non-saliency prediction, considering only pixel-level features for RSI-SOD, but never taking into account the intrinsic correlation between the scenes of RSIs and salient objects. To address this issue, we first propose the assisted learning of remote sensing scene classification, and design an MTL framework to combine both subtasks collaboratively in this article.

B. Scene Classification in Optical RSIs

The scene, as a basic attribute of optical RSIs, its classification is one of the most fundamental tasks in the remote sensing community, and has been widely studied since the era of deep learning [17]. Numerous scene classification datasets have been proposed in the past six years, e.g., AID [17], OPTIMAL31 [33], and NWPU-RESISC45 [18], which cover 30, 31, and 45 scene categories, respectively. They encompass most scenarios in RSIs, such as bridge, industrial, river, beach, airport, ship, highway, buildings, overpass, islands, etc.

On the basis of these datasets, the researchers have conducted extensive, thorough investigations. Early studies focus on fully supervised manners, utilizing multiscale convolutional features for scene recognition [18], introducing spatial attention mechanisms, channel attention mechanisms [34], and recurrent attention mechanisms [33] to enhance features on key regions, joint learning of local features of objects and global features of images [35], etc. After that, unsupervised learning, semi-supervised learning, and few-shot classification have been greatly explored. For example, Lu et al. [36] present the first unsupervised representation learning framework for remote sensing scene classification. Huang et al. [37] explore semi-supervised representation learning among different datasets, and investigate a bidirectional alignment strategy for cross-domain adaption. Li et al. [38] present an adaptive match network with a few-shot attention mechanism in a parallel manner, to focus on discriminative regions.

Recently, some novel network architectures and technologies have been introduced to build deep models for scene parsing, such as vision transformer [39] and neural architecture search [40]. For instance, a spatial-channel feature preserving vision transformer [41] is proposed for remote sensing scene classification. Furthermore, some fine-grained scene tasks have been proposed to contribute to the community, such as multilabel classification [34] and multiscene classification [42]. In this article, inspired by the task of scene classification, we introduce this task for auxiliary supervision and knowledge transfer to build a novel RSI-SOD framework.

C. Scene-Related Visual Models for Image Analysis

The scene is a widespread knowledge that has been applied in many visual tasks, including NSIs and RSIs, and researchers have proposed a variety of scene-related models. To our best knowledge, these models can be divided into three categories.

1) Without scene labels, this type of models simply learns scene-aware contextual information [43], [44], [45], [46]. For instance, Liu and Han [45] present a deep spatial recurrent convolutional network, and combine scene context modulation with local object context for saliency parsing in NSIs.
2) According to individual scene categories, design scene-independent models for various tasks, e.g., super-resolution [21], radar object detection [47], building extraction [48], vehicle detection [22]. For example, Zhang et al. [21] employ a two-stage approach, in which a scene classifier is designed to classify low-resolution RSIs in the first stage, and then propose separate super-resolution networks for each scene in the second stage for specific learning among scenes.

3) With scene labels, design MTL-based visual models that consist of supervised signals for the main task and auxiliary scene classification. These main tasks contain object detection [49], point cloud segmentation [20], oriented object detection [19], cloud segmentation [50], etc. For instance, Xu et al. [20] illustrate binary vector-based scene descriptors for different scenes supervision and utilize global information of point clouds as prior knowledge for semantic segmentation [51].

Motivated by these studies, we first annotate the scene labels for RSI-SOD, and present a scene-guided saliency model with multitask supervision. Different from them, the proposed framework not only learns both SOD and scene classification, but strives to learn knowledge from scene classification that facilitates saliency localization, and performs conditional dynamic guided learning by cross-task knowledge distillation.

III. Methodology

In this section, we provide the methodology of the proposed framework point by point. First, we present the overview of the proposed SDNet in Section III-A, and describe the methodology of SKTM, CDGM, and OCAM in Sections III-B–III-D, respectively. Finally, the hybrid multitask loss function is illustrated in Section III-E.

A. Overview of the Model Architecture

As shown in Fig. 4, the proposed framework consists of four components. First, we utilize ResNet18 [52] as the scene subnet for scene classification and scene-specific feature extraction. For an input RSI defined as $I \in \mathbb{R}^{3\times 448\times 448}$, we downsample its spatial size to 224 × 224, then feed it into the scene subnet, and obtain the multilevel features S_1–S_5 as well as the scene probability vector p_{scene} as follows:

$$S_1, S_2, S_3, S_4, S_5, p_{\text{scene}} = \mathcal{F}_{\text{res18}}\left(\mathcal{F}_{\downarrow 2\times}(I)\right) \tag{1}$$

where $\mathcal{F}_{\text{res18}}(\cdot)$ and $\mathcal{F}_{\downarrow 2\times}(\cdot)$ refer to the function of ResNet18 and 2× spatial downsampling operation, respectively.

To obtain object localization information, we also define a 3 × 3 convolution and fully connected weights to generate a dynamic optimal CAM [23], an input for scene knowledge transfer in the training phase, defined as

$$C = \text{argmax}(\ \mathcal{F}_{\text{CAM}}(\mathcal{C}_{3\times 3}(S_5) \otimes \mathcal{W}_{\text{fc}})) \in \mathbb{R}^{1\times 448\times 448} \tag{2}$$

where $\mathcal{F}_{\text{CAM}}(\cdot)$ denotes CAM generation, $\otimes$ is element-wise multiplication, and argmax$(\cdot)$ indicates to select the optimal CAM among scene categories. $\mathcal{C}_{i\times i}(\cdot)$ is the $i\times i$ convolution, and $\mathcal{W}_{\text{fc}}$ illustrates the weights of the fully-connected layer.

Then, an encoder-decoder baseline named PSPNet [53] is equipped as the saliency subnet to generate multiscale saliency features F_1–F_5, which also serves as an input for scene knowledge distillation in the subsequent process. After that, the model enters a scene knowledge-induced learning process with multilevel SKTM and CDGM for scene-saliency knowledge and condition-guided saliency feature generation, respectively. For an SKTM, it captures saliency feature F_i, scene feature S_i, and dynamic CAM map C as inputs, and obtains the fine scene-saliency context vector K_i for accurate localization of salient objects, which can be defined as

$$K_i = \text{SKTM}_i(F_i, S_i, C) \in \mathbb{R}^{1\times H\times W} \tag{3}$$

where $\text{SKTM}_i(\cdot)$ is the function of SKTM. Then, CDGM employs the fine scene-saliency context vector K_i to conditionally guide the original saliency feature F_i to project the scene-enhanced saliency feature for final RSI-SOD supervision, i.e.,

$$G_i = \text{CDGM}_i(F_i, K_i) \in \mathbb{R}^{256\times H\times W} \tag{4}$$

where $\text{CDGM}_i(\cdot)$ denotes the process of CDGM, and G_i indicates the scene-enhanced saliency feature. Here, G_1–G_5 are fed into a five-layer top-down, layer-by-layer summation decoder to produce the predicted SMs as p_1–p_5, where the decoder consists of a 3 × 3 and a 1 × 1 convolutions

$$p_i = \begin{cases} \mathcal{F}_{\text{decoder}}\left(G_i \oplus \mathcal{F}_{\uparrow 2\times}(G_{i+1})\right), & i = 4, 3, 2, 1 \\ \mathcal{F}_{\text{decoder}}(G_i), & i = 5 \end{cases} \tag{5}$$

where $\oplus$ is element-wise summation, $\mathcal{F}_{\uparrow 2\times}(\cdot)$ indicates 2× bilinear interpolation, and p_1 is the final saliency prediction.

Finally, Laplacian pyramid-based OCAM is introduced to compensate for the object contours as well as foster RSI-SOD in a supervised manner, as shown in Fig. 4(c).

B. Scene Knowledge Transfer Module

If we intend to facilitate RSI-SOD through scene classification, the model must capture auxiliary and accurate spatial localization of salient objects to compensate for the SOD subnet. However, scene classification eventually delivers a relatively weak category probability vector, and it is also non-trivial to produce a considerable boost by merely concatenating scene features into the main decoder [50]. Fortunately, existing deep learning techniques through the classification probability, such as CAM [23], [54], can find the spatial region with the optimal activation [55], which we believe is inextricably associated with salient objects. To this end, we present SKTM, which combines scene features and CAM to distill knowledge from the scene branch, and project fine scene-saliency context vector with accurate saliency spatial information.

As illustrated in Fig. 5, the proposed SKTM first considers the correlation between multiscale scene features and saliency features to filter the negative spatial activation, and explore Hadamard product as the scene-saliency representation, i.e.,

$$w_i^{j,k} = \frac{\exp\left(\kappa\left(F_i^j, S_i^k\right)\right)}{\sum_{t=1}^{H\times W}\exp\left(\kappa\left(F_i^j, S_i^t\right)\right)}. \tag{6}$$

Here, $\kappa(x, y) = \varphi(x)\phi(y)$ is Hadamard relation production, $\varphi(\cdot)$ and $\phi(\cdot)$ are transform functions with 1 × 1 convolutions, and w_i refers to the ith scene-saliency representation.

To exploit scene information more adequately, we introduce a dynamic CAM map as a priori spatial attention mechanism to enhance the coarse scene-saliency representation, as follows:

$$m_i = w_i \otimes \text{Repeat}(C) \in \mathbb{R}^{C\times H\times W} \tag{7}$$

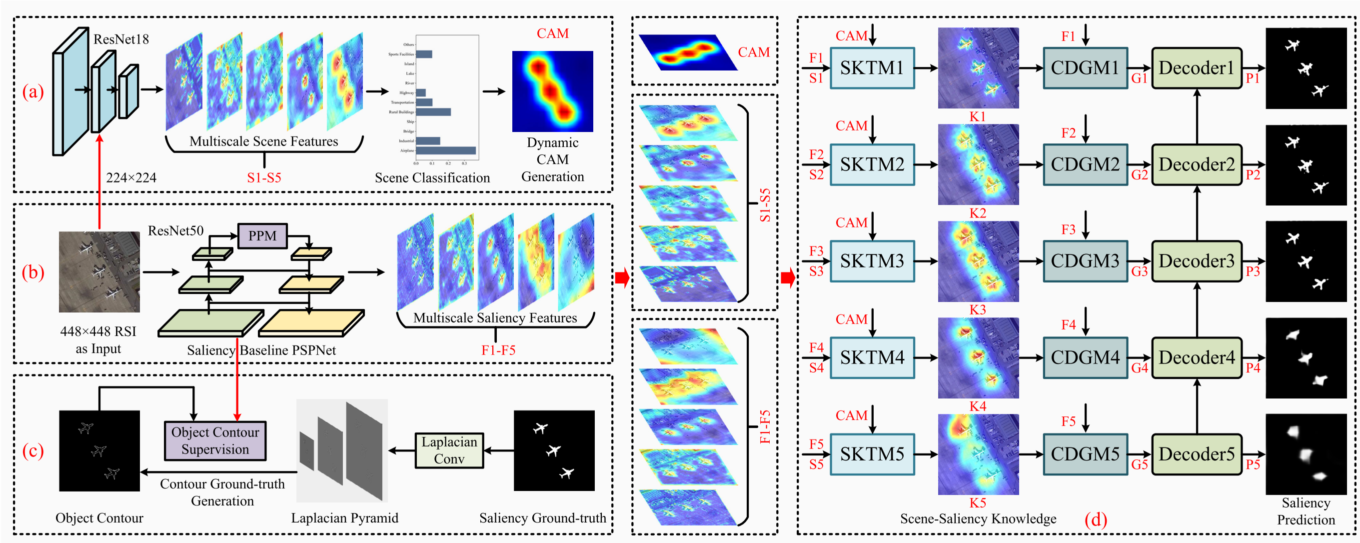


Fig. 4. Illustration of the proposed framework. (a) Scene subnet for multiscale scene features and dynamic CAM generation. (b) Saliency baseline (PSPNet) for multiscale saliency feature generation. (c) Illustration of OCAM. (d) Scene-saliency knowledge distillation for conditionally dynamic saliency prediction.

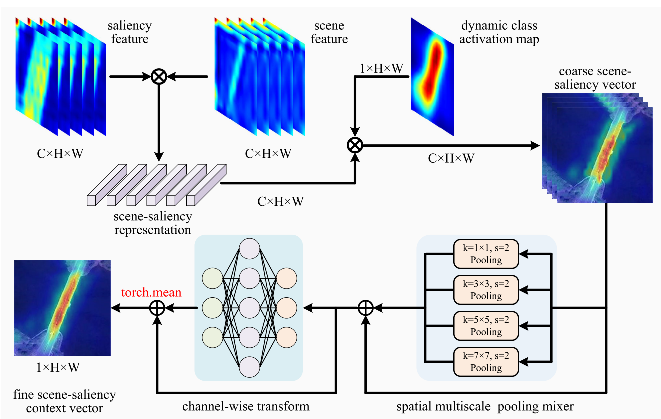


Fig. 5. Illustration of the proposed SKTM with the multiscale pooling mixer.

where $\otimes$ denotes element-wise production, Repeat$(\cdot)$ indicates channel-wise repetition, and m_i is the ith coarse scene-saliency knowledge as shown in Fig. 5.

Motivated by various kinds of token mixers [56], we then deploy an effective strategy to refine the coarse knowledge across spaces and channels, respectively, and generate the fine scene-saliency context vector as accurately as possible. Specifically, we first employ multiscale spatial pooling layers with strides of 2, kernels of 1×1, 3×3, 5×5, 7×7, to model the intricate correlation among different spatial locations with a very low number of parameters, which can be defined as

$$r_i = \mathcal{P}_1(m_i) \oplus \mathcal{P}_3(m_i) \oplus \mathcal{P}_5(m_i) \oplus \mathcal{P}_7(m_i) \oplus m_i \quad (8)$$

where r_i is the intermediate vector of stage i, and $\mathcal{P}_j(\cdot)$ refers to $j \times j$ spatial pooling layer to conduct basic token mixing.

Finally, a channel-wise multilayer perceptron (MLP) is introduced as the transform function to explore the relation among different channels. In this phase, we also utilize a residual learning strategy to avoid useful information degradation and produce a single-channel distilled knowledge as K_i. The above-mentioned function can be mathematically defined as

$$K_i = \text{Mean}(\mathcal{M}_c(r_i) \oplus r_i) \in \mathbb{R}^{1\times H\times W} \quad (9)$$

where $\mathcal{M}_c(\cdot)$ represents MLP operating among channels, and Mean$(\cdot)$ indicates the average function of channels to project a single-channel K_i, i.e., fine scene-saliency knowledge with accurate spatial activation. Compared to simple channel concatenation or element-wise multiplication strategies, the presented SKTM can simultaneously perform a mutually supervised fusion of scene features, saliency features, and dynamic CAMs, filtering and eliminating spurious regions of activation in an explicit manner. It could generate high-quality spatial attention knowledge, which can serve as excellent compensation and guidance information for RSI-SOD.

C. Conditional Dynamic Guidance Module

The scene subnet always has some samples with incorrect classification predictions, and thus the model will generate the wrong spatial activation in dynamic CAM, which cannot deliver precisely localized spatial attention to salient objects. Hence, if we completely trust the designed scene-saliency knowledge, such as simply multiplying it with the original saliency features to yield predicted saliency results, it will result in an output that severely degrades the detection performance and reduces the generalizability of the model. Based on this deficiency, we propose CDGM to exploit the designed scene-saliency knowledge for a conditionally dynamic guided learning process of RSI-SOD, as revealed in Fig. 6.

To achieve this goal, we propose to adopt the fine scene-saliency knowledge as the input to compute independent attention for each RSI, and introduce dynamic convolutional kernels [57], [58] and biases that can be optimized and differentiated in the training phase to enhance the conditional capability. As for a fine scene-saliency knowledge K_i, we first utilize the global average pooling function to produce a

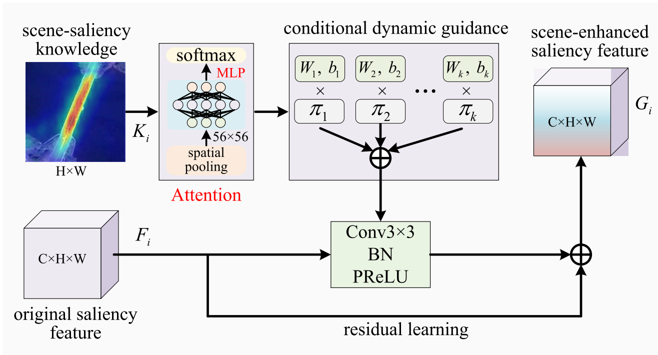


Fig. 6. Illustration of the proposed CDGM with residual learning strategy.

hardware-friendly matrix $\mathcal{P}_{\text{Avg}}(K_i) \in \mathbb{R}^{56\times 56}$, then flatten it, and deploy an MLP with softmax operation and a large temperature to project k-dimensional attention weights, which can be defined as below

$$\text{Att}_i = \mathcal{M}\left(\mathcal{F}_{\text{flatten}}\left(\mathcal{P}_{\text{Avg}}(K_i)\right)\right) \in \mathbb{R}^k \tag{10}$$

$$\pi_k = \frac{\exp\left(\text{Att}_{i,k}/\tau\right)}{\sum_j \exp\left(\text{Att}_{i,j}/\tau\right)} \tag{11}$$

where Att_i is the k-dimensional output vector of MLP ($\mathcal{M}$), $\mathcal{F}_{\text{flatten}}(\cdot)$ indicates the flattening operation, $\mathcal{P}_{\text{Avg}}(\cdot)$ denotes 56×56 global average pooling function, and $\tau = 30$ refers to the temperature to control the sparsity of the output attention weights $\{\pi_k\}$. Obviously, the value of $\{\pi_k\}$ depends on the scene-saliency context vector K_i of each optical RSI and thus is not fixed. It varies from the individual inputs and serves as the optimal integrated signal, which significantly increases the dynamics of the conditional guidance capability of CDGM compared with the static convolutional aggregation modules. After gathering dynamic attention weights $\pi(x)$, a series of parallel convolutional kernels $\{W_k, b_k\}$ are deployed, which are the differentiable parameters and can be optimized during the training of the model. We integrate them with these weights in a nonlinear way for every individual input x (e.g., RSI, K_i) dynamically, to yield more powerful convolutional weights and biases as follows:

$$W(x) = \sum_{i=1}^{k} \pi_i(x) W_i, \quad b(x) = \sum_{i=1}^{k} \pi_i(x) b_i$$
$$\text{s.t.} \quad 0 \le \pi_i \le 1, \quad \sum_{i=1}^{k} \pi_i(x) = 1. \tag{12}$$

Here, to simplify the learning process of attention weights $\pi(x)$ and compress the kernel space, we utilize the sum-to-one strategy and constrain $\sum_{i=1}^{k}(\pi_i(x)) = 1$, where k is equal to 16. Note that these weights and biases are assembled variously for different fine-grained scene-saliency knowledge and share the same attention as the input of conditional dynamic guidance part in Fig. 6. Besides, all parallel convolutional kernels share the output 256-D channels by combination.

Finally, we introduce a residual learning-based skip connection to accelerate the convergence of the module and facilitate enhanced saliency features, and the total conditional guidance process of CDGM can be represented as

$$G_i = \sigma\left(W^{\text{T}}(K_i) \cdot F_i + b(K_i)\right) \oplus F_i \tag{13}$$

where $\sigma(\cdot)$ indicates PReLU activation function, W^{T} is the weight matrix of 3×3 convolution, and b is the bias vector. In contrast to simple feature integration modules, the proposed CDGM guides, compensates for, and improves the learning efficiency of enhanced saliency features in a dynamic manner. It can distill the scene-specific knowledge into the saliency decoder in an effective way, and further facilitate the localization results of salient objects from complicated RSIs.

D. Object Contour Awareness Module

It is crucial for RSI-SOD to recognize the edges and corners of salient objects accurately. However, whether it is salient feature F_i, scene feature S_i, or CAM map C, they all focus on the contextual features of the regions of interest and cannot capture the contours of salient objects completely. Existing adaptive saliency losses, e.g., CT [59], ACT [60], are only able to enhance the supervised weights of edge information in the decoding stage and particularly rely on the model to ensure successful edge detection. This means that these losses must be combined with excellent models to meet a considerable supplement. To cope with the above issue, we propose to guide the learning of object spatial details in a supervised manner at the shallow level of the saliency baseline, instead of the decoder's output. Hence, a simple yet effective module, termed OCAM, is introduced to compensate for the contours of salient objects from complicated RSIs, as illustrated in Fig. 4(c).

We simplify contour detection for salient objects in RSIs as a pixel-level binary classification problem. First, we introduce multiscale Laplacian convolutional operators to generate contour GT of salient objects from saliency GT map g_s. With the kernel

$$\begin{bmatrix} -1 & -1 & -1 \\ -1 & 8 & -1 \\ -1 & -1 & -1 \end{bmatrix},$$

Laplacian convolution with various strides (i.e., 1, 2, 4) could extract abundant detailed information about edges and corners of salient objects, and delivers Laplacian pyramid [61] as shown in Fig. 4(c). Then, we deploy a 1×1 convolution to aggregate these pyramidal contour maps for the trainable reweighting and produce the dynamic object contour GT map g_c as follows:

$$g_c = \mathcal{C}_{1\times 1}\left(\left[\mathcal{L}_1(g_s), \mathcal{F}_{\uparrow 2\times}(\mathcal{L}_2(g_s)), \mathcal{F}_{\uparrow 4\times}(\mathcal{L}_4(g_s))\right]\right) \tag{14}$$

where $g_c \in \mathbb{R}^{1\times 448\times 448}$, $[\cdot, \cdot]$ indicates the channel-wise concatenation, $\mathcal{L}_i(\cdot)$ denotes Laplacian convolution with stride of i, and $\mathcal{F}_{\uparrow 4\times}(\cdot)$ represents $4\times$ bilinear upsampling.

Then, we perform the supervised learning of contour for the shallow F_1–F_3 of the saliency baseline, and equip several convolutional layers as contour heads to produce the predicted contour maps p_c^1, p_c^2, p_c^3. Since contour detection is a task of extremely imbalanced categories, effective supervision cannot be performed by relying on binary cross-entropy (BCE) solely. Following [62] and [63], we introduce dice loss as an auxiliary part for BCE to jointly optimize the contour supervision, and thus the combined contour loss can be defined as

$$\mathcal{L}_{\text{contour}} = \sum_{j=1}^{3}\left(\mathcal{L}_{\text{bce}}\left(g_c, p_c^j\right) + \mathcal{L}_{\text{dice}}\left(g_c, p_c^j\right)\right) \tag{15}$$

where $\mathcal{L}_{\text{contour}}$ is the contour loss, $\mathcal{L}_{\text{bce}}$ and $\mathcal{L}_{\text{dice}}$ denote BCE and dice loss, respectively. They are clearly defined as

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

$$\mathcal{L}_{\text{dice}}\left(p_c^i, g_c^i\right) = 1 - \frac{2\sum_i^{H\times W} p_c^i g_c^i + \varepsilon}{\sum_i^{H\times W}\left(p_c^i\right)^2 + \sum_i^{H\times W}\left(g_c^i\right)^2 + \varepsilon} \tag{17}$$

where N indicates the total number of pixels, and ε is the smoothing factor to avoid zero division which we set to 1.

E. Total Loss Function

As shown in Fig. 4(d), we perform multiscale supervision of saliency detection among p_1–p_5, on which we impose the joint function of BCE loss and WIoU loss as follows:

$$\mathcal{L}_{\text{saliency}} = \sum_{i=1}^{5}(L_{\text{bce}}(p_i, g_s) + L_{\text{wiou}}(p_i, g_s))/2^{i-1} \tag{18}$$

where $\mathcal{L}_{\text{wiou}}$ refers to WIoU loss function as follows:

$$\mathcal{L}_{\text{wiou}}(x, y) = 1 - \frac{\sum_{j=1}^{N}\left(x_j \otimes y_j\right) + \varepsilon}{\sum_{j=1}^{N}\left(x_j \oplus y_j - x_j \otimes y_j\right) + \varepsilon}. \tag{19}$$

As our proposed framework has explicitly supervised signals for saliency detection, scene classification, and contour detection. Therefore, considering the empirically similar orders of magnitude of these loss terms, and to train the model end-to-end, the total loss $\mathcal{L}_{\text{total}}$ could be simply defined as their weighted summation as follows:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{saliency}} + \mathcal{L}_{\text{scene}} + \mathcal{L}_{\text{contour}} \tag{20}$$

where $\mathcal{L}_{\text{scene}}$ is a 12-category multiple cross-entropy loss, i.e.,

$$\mathcal{L}_{\text{scene}} = \sum_{c=1}^{12} -\log\left(p_{\text{scene}}^c\right) g_{\text{scene}}^c \tag{21}$$

where g_{scene} denotes the scene label vector we annotated.

IV. Experiments

In this section, we first present experimental protocol, then conduct the quantitative and qualitative comparison experiments, ablation study, and model analysis, respectively.

A. Experimental Protocol

1) Datasets: We perform the following experiments on three public RSI-SOD datasets in this article, i.e., as follows.

ORSSD [10] contains 800 optical RSIs, 600 ones of which are used for training and 200 images for testing.

EORSSD [11] is an updated version of ORSSD that includes 2000 RSIs with more complex scenarios. It divides 1400 images as the training set and 600 as the test set.

ORSI-4199 [12] is currently the most complicated RSI-SOD dataset containing 4199 RSIs with complex background and diverse salient object types, of which 2000 are used for testing and 2199 for training. In addition, it also divides nine attribute patterns for a more comprehensive evaluation.

2) Evaluation Metrics: In this article, we report four common quantitative indicators in the field of SOD as follows.

MAE [64] measures the pixel-level difference between the predicted SM and the GT, i.e.,

$$\text{MAE} = \frac{1}{m \times n}\sum_{i=1}^{m}\sum_{j=1}^{n}|\text{SM}(i, j) - \text{GT}(i, j)| \tag{22}$$

where m and n are the height and width of RSI, respectively.

F-measure [65] is a weighted and combined metric to define precision and recall between SM and GT as follows:

$$F_\beta = \frac{\left(1 + \beta^2\right) \times \text{Precision} \times \text{Recall}}{\beta^2 \times \text{Precision} + \text{Recall}} \tag{23}$$

where β^2 is utilized to balance the precision over recall and is equal to 0.3 according to [65]. In this article, we calculate the max F-measure under different thresholds in [0, 255].

S-measure [66] employs the balanced structural information of object-aware (S_o) and region-aware (S_r) levels to measure the structural similarity between SM and GT, i.e.,

$$S_m = \alpha \times S_o(\text{SM}, \text{GT}) + (1 - \alpha) \times S_r(\text{SM}, \text{GT}) \tag{24}$$

where α is an equilibrium factor of value 0.5 referring to [66].

E-measure [67] is an indicator close to 1, which quantifies both pixel-level matching and image-level statistics, defined as

$$E_m = \frac{1}{m \times n}\sum_{x=1}^{m}\sum_{y=1}^{n}\xi_{\text{SM}}(x, y) \tag{25}$$

where ξ_{SM} indicates the enhanced alignment matrix of SM.

3) Implementation Details: For the saliency subnet, we apply PSPNet [53] as saliency baseline. Following [15] and [16], we perform the learning of the model on three datasets separately and unify all images as 448 × 448 for training and testing to calculate various quantitative metrics. Additionally, the data augmentation techniques used in the training process are consistent with the above work. For a fair comparison, all deep learning-based comparative algorithms [7], [8], [9], [11], [12], [14], [15], [16], [27], [28], [29], [30], [31], [68], [69], [70], [71], [72], [73], [74], with the same input and output settings and uniform data augmentation scheme, are reproduced by their publicly available source code. All of these deep learning-based algorithms are deployed on the PyTorch1.8 toolbox, running on a single NVIDIA GeForce RTX 3090 GPU, and equipped with the Ubuntu18.04 system.

As for the proposed model, we load the pretraining weights of ResNet18 and ResNet50 for the scene branch and saliency baseline, respectively, and utilize *kaiming-normal* initialization for the other parametric layers uniformly. The model is trained iteratively via the stochastic gradient descent (SGD) algorithm with the polynomial learning rate scheduler, where the number

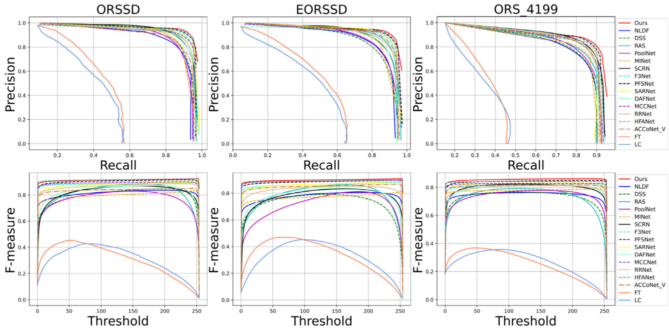


Fig. 7. PR and F-measure curves on the three RSI-SOD datasets of 17 state-of-the-art methods, where our proposed approach is marked in red.

of epochs is 100, the batch size is 8, the initial learning rate is 0.002, the momentum is 0.9, the weight decay is $5e^{-4}$, and the learning rate update formula is $0.002 \times (1-(\text{iter}/\text{maxiter}))^{0.9}$. As for the proposed framework, it contains 55.98M parameters with a considerable inference speed 38.89 ft/s.

4) Baselines: As revealed in Table I, we report 24 baselines among three RSI-SOD datasets for a fair comparison. These approaches consist of two conventional methods (i.e., LC [75] and FT [65]), ten deep learning-based algorithms for NSI-SOD (i.e., NLDF [7], DSS [8], RAS [9], PoolNet [68], PFAN [69], MINet [70], SCRN [71], GateNet [72], F3Net [73], and PFSNet [74]), and 12 deep learning-based methods for RSI-SOD (i.e., SARNet [14], DAFNet [11], FSMINet [27], MCCNet [28], RRNet [29], MJRBM-V [12], MJRBM-R [12], EMFINet-R [30], HFANet [15], ACCoNet-V [31], ACCoNet-R [31], SRAL [16]).

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

This section presents quantitative results, qualitative evaluation, and attribute-based analysis of numerous state-of-the-art methods and the proposed model, respectively.

1) Quantitative Comparison: For a comprehensive comparison, we select the most competitive 16 baselines and the proposed method to plot the PR and F-measure curves, as shown in Fig. 7. By observation, our method has the most superior performance on both EORSSD and ORSI-4199 datasets, i.e., the areas under the curves in red are the largest. However, the curves of our approach overlap with several methods on the ORSSD dataset and do not maintain a significant superiority. We blame it on the small scale of the ORSSD dataset, and thus the presented model cannot explore adequate scene knowledge from this dataset.

Table I shows the quantitative results of four metrics, i.e., F-measure (F_β), MAE, S-measure (S_m), and E-measure (E_m), on three datasets of 25 algorithms. Among all competitors, traditional methods provide no advantage in all metrics. This is because they utilize low-level features or manual operators to compute SM directly, thus, are not generalizable in complex RSIs. These methods designed for natural images all yield considerable performance. However, except for PFSNet [74], which reaches the top three in very few metrics, all other metrics are inferior to the algorithms for RSI-SOD. The above fully justifies the previous work [15], [16] that SOD methods designed for natural images cannot be adapted to the characteristics of optical RSIs, such as complex background, low contrast, object scale diversity, complicated edges, and irregular topology. Most competitive results are in those methods designed for RSI-SOD, including EMFINet-R [30], HFANet [15], ACCoNet-R [31], SRAL [16], and our proposed framework. Of these, SRAL is our most recent work, notably achieving the second-best results on the ORSI-4199 dataset.

As we can see, the presented model achieves the most favorable results on both EORSSD and ORSI-4199 datasets, with state-of-the-art performance on all metrics. In particular, on the most challenging ORSI-4199 dataset, our SDNet is the only one that reaches an F_β over 0.86 and an MAE less than 0.032. The proposed model employs the fundamental PSPNet as the saliency baseline, introduces a scene classification task, and performs cross-task scene knowledge transfer via SKTM and CDGM. It achieves such remarkable performance that is attributed to the effectiveness of the proposed scene knowledge distillation framework, especially SKTM and CDGM. With respect to the differences in effectiveness among individual modules, we will reveal them in detail in Section IV-C.

The proposed model, however, lacks excellent performance on the ORSSD dataset, i.e., F_β is lower than ACCoNet-R [31] and S_m is lower than EMFINet-R [30] and HFANet [15]. This is a critical phenomenon that deserves our attention, why our model performs well on EORSSD and ORSI-4199 datasets but shows mediocre results on ORSSD? The proposed model strives to transfer scene knowledge into an existing saliency baseline, PSPNet [53], rather than designing sophisticated structures to extract multiscale contexts or global attention to facilitate RSI-SOD. Therefore, combining the above findings, we suggest that the ORSSD dataset, with only 200 test images, is too small to adequately benefit from an MTL framework and distill valid scene knowledge by the proposed SKTM and CDGM for saliency performance boosting.

2) Qualitative Comparison: Fig. 8 shows the visualized prediction results of 13 state-of-the-art algorithms in 13 typical scenarios, including five ones designed for NSIs and seven RSI-SOD algorithms. Overall, the presented SDNet predicts the most complete and accurate SMs among various samples over the competitors in Fig. 8(o). For instance, in the first and fifth rows, our predictions successfully overcome background interference (i.e., road, crosswalk). In the third and fourth rows, when competitors are in ambiguous recognition or incomplete detection, our method still achieves effective localization, and the predicted SMs are closest to GTs.

By vertical comparison, we find that all these approaches designed for natural images in Fig. 8(c)–(g) do not perform as well as those RSI-SOD ones, which shows that the algorithms for natural scenes are not applicable to RSI scenes, and the design of specialized SOD algorithms for RSI is essential.

3) Attribute-Based Analysis: The ORSI-4199 dataset describes nine attribute scenarios, namely big salient object (BSO), complex scene (CS), complex salient object (CSO), incomplete salient object (ISO), low contrast scene (LCS), multiple salient objects (MSOs), narrow salient object (NSO), off-center (OC), and small salient object (SSO). Comparison of the results under each attribute pattern can reveal more in-depth performance drawbacks of different methods, providing insights for further algorithm investigations.

TABLE I

RESULTS OF FOUR METRICS ON THREE OPTICAL RSI-SOD DATASETS. TOP THREE ARE MARKED IN RED, GREEN, AND BLUE, RESPECTIVELY

Methods	Published	ORSSD Dataset [10]				EORSSD Dataset [11]				ORSI-4199 Dataset [12]			
		$F_\beta\uparrow$	MAE↓	$S_m\uparrow$	$E_m\uparrow$	$F_\beta\uparrow$	MAE↓	$S_m\uparrow$	$E_m\uparrow$	$F_\beta\uparrow$	MAE↓	$S_m\uparrow$	$E_m\uparrow$
LC [75]	MM'06	0.4275	0.1230	0.5941	0.5353	0.4526	0.0864	0.5954	0.5412	0.3573	0.1893	0.5270	0.5052
FT [65]	CVPR'09	0.4526	0.1126	0.5916	0.5186	0.4704	0.0715	0.6107	0.5304	0.3680	0.1791	0.5256	0.4883
NLDF [7]	CVPR'17	0.8352	0.0267	0.8702	0.8967	0.8060	0.0154	0.8706	0.8739	0.7639	0.0584	0.8053	0.8496
DSS [8]	CVPR'17	0.8469	0.0268	0.8688	0.8842	0.7921	0.0167	0.8371	0.8191	0.7672	0.0561	0.8115	0.8301
RAS [9]	ECCV'18	0.8841	0.0185	0.8896	0.8842	0.8636	0.0114	0.8800	0.8593	0.7930	0.0595	0.8142	0.8268
PoolNet [68]	CVPR'19	0.8291	0.0268	0.8610	0.8527	0.8121	0.0207	0.8279	0.8155	0.7777	0.0573	0.8184	0.8468
PFAN [69]	CVPR'19	0.8755	0.0207	0.8853	0.8975	0.8472	0.0127	0.8848	0.8801	0.8024	0.0486	0.8373	0.8787
MINet [70]	CVPR'20	0.8380	0.0227	0.8640	0.8975	0.8178	0.0129	0.8569	0.8846	0.7891	0.0473	0.8232	0.8862
SCRN [71]	ICCV'19	0.8687	0.0210	0.8799	0.8924	0.8326	0.0158	0.8288	0.8637	0.8232	0.0423	0.8524	0.8867
GateNet [72]	ECCV'20	0.9083	0.0125	0.9103	0.9267	0.8724	0.0091	0.9010	0.8961	0.8443	0.0387	0.8660	0.9101
F3Net [73]	AAAI'20	0.8927	0.0126	0.9245	0.9309	0.8822	0.0077	0.9218	0.9241	0.8175	0.0435	0.8520	0.9013
PFSNet [74]	AAAI'21	0.9153	0.0101	0.9303	0.9469	0.8979	0.0077	0.9287	0.9350	0.8496	0.0374	0.8686	0.9245
SARNet [14]	RS'21	0.8963	0.0185	0.8976	0.9307	0.8865	0.0102	0.9097	0.9277	0.8309	0.0448	0.8536	0.9112
DAFNet [11]	TIP'21	0.8717	0.0161	0.8982	0.9047	0.8378	0.0106	0.8824	0.8838	0.8169	0.0473	0.8477	0.8861
FSMINet [27]	GRSL'22	0.8623	0.0178	0.8949	0.9266	0.8527	0.0100	0.8989	0.9132	0.8046	0.0451	0.8344	0.8906
MCCNet [28]	TGRS'22	0.9005	0.0212	0.9040	0.9393	0.8976	0.0083	0.9306	0.9392	0.8284	0.0439	0.8506	0.9093
RRNet [29]	TGRS'22	0.8857	0.0142	0.9110	0.9232	0.8511	0.0101	0.8964	0.8979	0.8122	0.0448	0.8449	0.8905
MJRBM-V [12]	TGRS'22	0.9028	0.0140	0.9156	0.9265	0.8705	0.0099	0.9088	0.9040	0.8352	0.0392	0.8601	0.9103
MJRBM-R [12]	TGRS'22	0.9058	0.0129	0.9128	0.9252	0.8685	0.0092	0.8980	0.8956	0.8406	0.0379	0.8685	0.9075
EMFINet-R [30]	TGRS'22	0.9132	0.0107	0.9350	0.9458	0.8972	0.0075	0.9286	0.9337	0.8469	0.0352	0.8678	0.9215
HFANet [15]	TGRS'22	0.9224	0.0113	0.9324	0.9517	0.9007	0.0082	0.9292	0.9410	0.8419	0.0379	0.8659	0.9181
ACCoNet-V [31]	TCYB'23	0.8514	0.0231	0.8717	0.9008	0.8677	0.0137	0.9068	0.9164	0.8182	0.0433	0.8490	0.8992
ACCoNet-R [31]	TCYB'23	0.9249	0.0102	0.9187	0.9498	0.9009	0.0084	0.9290	0.9360	0.8486	0.0354	0.8702	0.9238
SRAL [16]	TGRS'23	0.9167	0.0105	0.9305	0.9532	0.8964	0.0067	0.9234	0.9406	0.8576	0.0321	0.8735	0.9338
SDNet (Ours)	–	0.9243	0.0099	0.9315	0.9565	0.9089	0.0063	0.9338	0.9429	0.8618	0.0313	0.8771	0.9358

TABLE II

ATTRIBUTE-BASED EVALUATION ON THE ORSI-4199 DATASET [12]. THE AVERAGE SSIM METRICS FOR NINE ATTRIBUTES ARE REPORTED. THE AVG. ROW SHOWS THE AVERAGE RESULTS FOR ALL ATTRIBUTES, AND TOP THREE IN EACH LINES ARE MARKED IN RED, GREEN, AND BLUE, RESPECTIVELY

Attributes	*PFAN [69]*	*MINet [70]*	*SCRN [71]*	*GateNet [72]*	*F3Net [73]*	*PFSNet [74]*	*SARNet [14]*	*DAFNet [11]*	*MCCN [28]*	*RRNet [29]*	*MJRBM [12]*	*EMFI-R [30]*	*HFANet [15]*	*ACCo-V [31]*	*ACCo-R [31]*	*SDNet*
BSO	0.7984	0.7929	0.8295	0.8546	0.8256	0.8531	0.7984	0.8064	0.8200	0.8226	0.8344	0.8547	0.8222	0.8186	0.8589	0.8626
CS	0.8021	0.7858	0.8274	0.8539	0.8234	0.8519	0.8123	0.8151	0.8210	0.8138	0.8451	0.8509	0.8287	0.8179	0.8521	0.8664
CSO	0.7714	0.7630	0.8001	0.8234	0.7887	0.8196	0.7665	0.7831	0.7935	0.7900	0.8003	0.8250	0.7875	0.7931	0.8245	0.8325
ISO	0.7280	0.7184	0.7762	0.8110	0.7872	0.8166	0.7573	0.7485	0.7778	0.7538	0.7868	0.8102	0.7884	0.7632	0.8222	0.8326
LCS	0.6756	0.6649	0.7094	0.7318	0.7071	0.7342	0.7182	0.7018	0.7119	0.6886	0.7332	0.7225	0.7171	0.7001	0.7319	0.7482
MSO	0.7739	0.7429	0.7949	0.8156	0.7902	0.8218	0.8162	0.7817	0.7720	0.7702	0.8183	0.7905	0.7965	0.7603	0.8076	0.8181
NSO	0.7496	0.7304	0.7933	0.8552	0.7916	0.8318	0.7949	0.8054	0.8028	0.7873	0.8391	0.8412	0.8202	0.7716	0.8224	0.8743
OC	0.7457	0.6707	0.7474	0.7730	0.7481	0.7861	0.8071	0.7780	0.7721	0.7301	0.8055	0.7662	0.7967	0.7553	0.7887	0.7889
SSO	0.6984	0.6447	0.7092	0.7311	0.7179	0.7493	0.7596	0.7219	0.7094	0.6897	0.7456	0.7130	0.7308	0.6972	0.7359	0.7444
Avg.	0.7492	0.7237	0.7764	0.8055	0.7755	0.8072	0.7812	0.7713	0.7756	0.7607	0.8009	0.7971	0.7876	0.7641	0.8049	0.8187

As illustrated in Table II, we select 15 competitive approaches, and report their SSIM scores with our method on nine scene patterns. Overall, the proposed framework achieves the highest average score and reaches the best results among six attributes, and is competitive in MSO and SSO attributes, but is weak in OC scenarios. We find that SARNet [14] and MJRBM [12] perform best in OC attributes. Their common feature is that they both introduce spatial attention mechanisms, resulting in better localization of OC salient objects, which provides us with ideas for further investigations.

C. Ablation Study

In this section, we conduct abundant experiments on the ORSI-4199 dataset to reveal the effects of each module in the proposed framework both quantitatively and qualitatively.

TABLE III

ABLATION EXPERIMENTS ON THE ORSI-4199 DATASET. THE BEST RESULTS ARE MARKED IN BOLD

No.	SKTM	CDGM	OCAM	MAE↓	$F_\beta\uparrow$	$S_m\uparrow$	$E_m\uparrow$
1	baseline (PSPNet [53])			0.0377	0.8466	0.8657	0.9226
2	baseline + scene subnet			0.0346	0.8468	0.8673	0.9244
3			✓	0.0364	0.8473	0.8663	0.9220
4	✓			0.0338	0.8576	0.8686	0.9297
5	✓	✓		0.0322	0.8598	0.8752	0.9339
6	✓	✓	✓	**0.0313**	**0.8618**	**0.8771**	**0.9358**

1) Baseline Setup: We adopt the classical fully convolutional network, PSPNet [53], as the baseline, which utilizes ResNet50 as a backbone and pyramid pooling module for multiscale context modeling. As shown in Fig. 9(c), it performs poorly in these scenarios, producing too much missed and false detection. Also, as revealed in Table III, it scores 0.0377 and

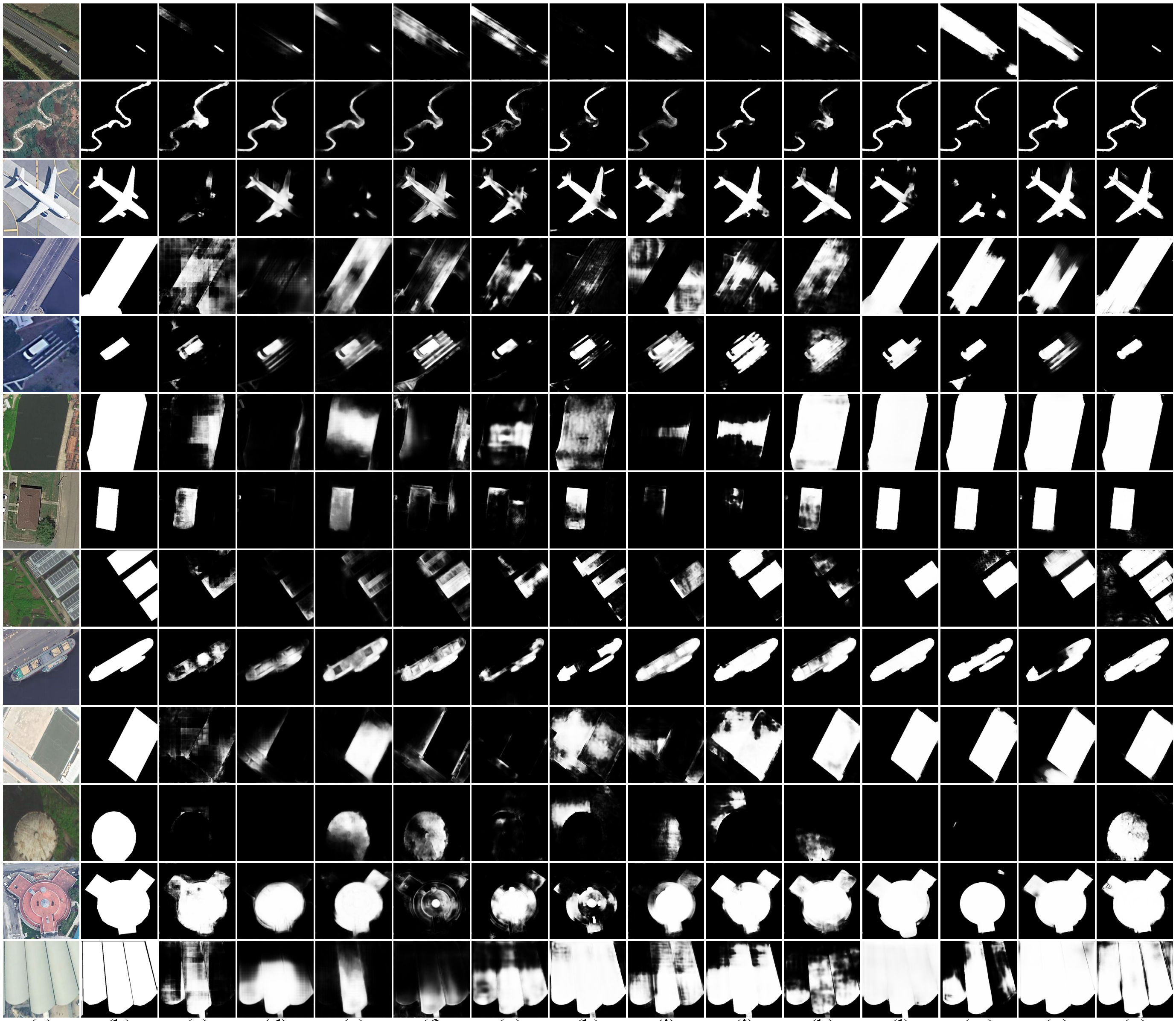


Fig. 8. Typical visualization of 13 state-of-the-art methods. (a) RSIs. (b) GTs. (c) NLDF [7]. (d) DSS [8]. (e) RAS [9]. (f) PoolNet [68]. (g) PFAN [69]. (h) SARNet [14]. (i) DAFNet [11]. (j) MCCNet [28]. (k) RRNet [29]. (l) EMFINet-R [30]. (m) HFANet [15]. (n) ACCoNet-R [31]. (o) SDNet (Ours).

0.8466 in terms of MAE and F_β, respectively, which are far inferior to state-of-the-art competitors in Table I.

2) Effects of Only Scene Supervision: A worthy concern is the impact of simple dual-branch supervised learning by saliency and scene labels (i.e., without scene knowledge transfer) on RSI-SOD performance. To verify this, we construct the model variant named "baseline+scene subnet" and yield its performance as shown in Table III. Compared with the baseline, it brings some MAE and S_m benefits, but contributes nothing to F_β. The above proves that only joining the supervised signal by weak image-level labels does not bring a sufficient boost to spatial saliency understanding, i.e., it cannot improve the model's detection ability of Precision and Recall.

3) Effects of SKTM: The intention of proposing SKTM is to combine scene features with CAM spatial activation information to filter and exploit scene-saliency context vector that facilitates salient object localization. As shown in No. 4 of Table III, compared to baseline, "baseline+SKTM" shows considerable performance gains in MAE, F_β, and E_m. In particular, it increased by 1.1% on the F_β metric, which fully illustrates the beneficial effect of the scene-saliency vectors extracted by SKTM on the RSI-SOD task, i.e., the spatial awareness of salient regions is greatly improved. Compared with Fig. 9(c) and (d), the implementation of SKTM can reduce the missed and false detection rates of the model in these typical samples, such as buildings, bridge, and vehicles.

4) Effects of CDGM: After SKTM captures the fine-grained scene knowledge vector, how to utilize this cross-task knowledge to guide the optimal saliency interpretation dynamically is a remaining issue. To this end, we consider the scene knowledge as conditional guided attention to aggregate dynamic convolutional kernels and build a novel module named CDGM.

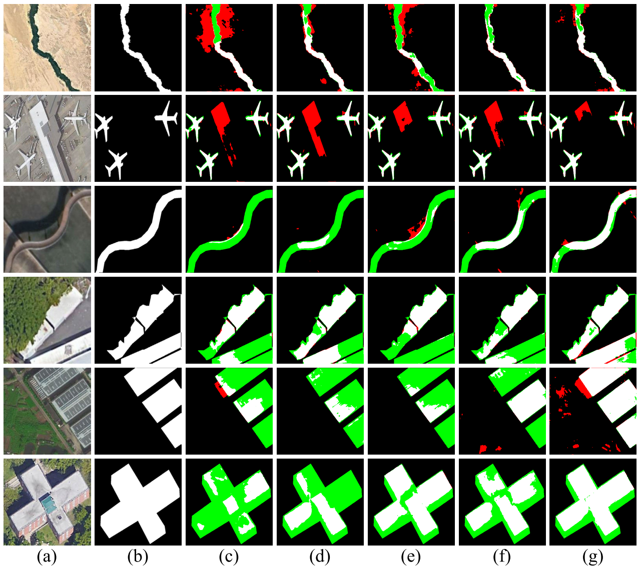


Fig. 9. Typical saliency error maps on the ORSI-4199 dataset, where red and green pixels mean false negatives and missed detection, respectively. (a) RSIs. (b) GTs. (c) baseline. (d) baseline+SKTM. (e) baseline+OCAM. (f) baseline+SKTM+CDGM. (g) baseline+SKTM+CDGM+OCAM.

To reveal its effectiveness, we organize the ablation experiments of No. 4 and No. 5 in Table III. By observation, noticeable performance improvements are shown on all four metrics, e.g., a 0.66% increase in terms of S_m. As shown in Fig. 9(d) and (f), with the integration of CDGM into the model, more salient regions can be distinguished, and in particular, the rate of missed detection is greatly reduced. Overall, it is the dynamic and conditional guidance mechanism of CDGM that ensures the adaptability and generalization of the model, yielding both quantitative and qualitative gains in saliency understanding for various scenarios.

5) Effects of OCAM: The essential purpose of SKTM and CDGM is to explore potential regions of salient objects from the scene subnet, yet the edges and corners of such regions are mismatched with those of salient objects. To enhance the model's perception of the object contours in the shallow spatial feature layers, and inspired by Laplacian operators and pyramids, we propose a simple contour awareness module, i.e., OCAM. How about its validity? We perform a comparative analysis by No. 3 and No. 6 in Table III, and Fig. 9(f) and (g). In contrast to No. 1 and No. 3 or No. 5 and No. 6 of Table III, the inclusion of OCAM in the model witnesses a certain gain in quantitative indicators, e.g., No. 1 and No. 3 exhibit MAE gains greater than 0.1%, No. 5 and No. 6 show F_β gains of 0.2%. Furthermore, we also demonstrate that OCAM indeed qualitatively improves object contour awareness and further contributes to the detection performance of the proposed MTL model for RSI-SOD. As shown in Fig. 9, with the assistance of OCAM, the predicted SMs for bridges, vehicles, and buildings in rows 2–6 are closest to the GTs in terms of edges, corners, and completeness, and also reach the most superior results among all ablation competitors.

D. Model Analysis

Here, we conduct further experiments and visual analysis to reveal the scene classification performance, visualization analysis of scene knowledge distillation, comparison of scene guidance capability of SKTM and CDGM, comparison of OCAM with other methods for object contour perception, and design rationale of the dual-branch framework.

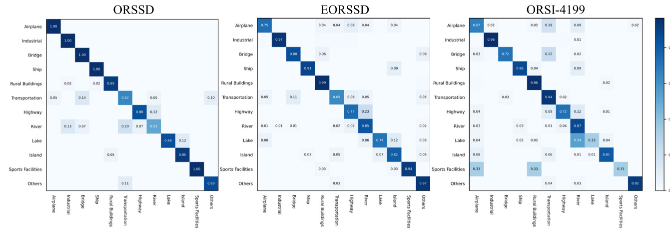


Fig. 10. Confusion matrix of scene classification on three RSI-SOD datasets.

1) Performance Analysis of Scene Classification Task: The most common metric to reveal the performance of scene classification is the confusion matrix [35] and overall accuracy (OA), and thus we report confusion matrices of scene branch on three RSI-SOD datasets as shown in Fig. 10. By observation, the majority of probabilities on the diagonals dominate these confusion matrices, reflecting that most test samples have been correctly predicted with their scene labels. Notably, in the confusion matrix of ORSI-4199, the model shows some recognition failures in lakes and rivers. We blame this problem on the tremendous interclass similarity between these two scenes. Additionally, we calculate the OA of scene classification on three datasets, i.e., 89.00%, 88.50%, and 85.72%. This further illustrates the validity and rationality of our annotated scene labels. In conclusion, these qualitative metrics fit the distribution of remote sensing scene classification and serve as solid support to verify the soundness of our model.

2) Visual Illustration of Scene Knowledge Guidance: To reveal how our model performs accurate saliency understanding and achieves state-of-the-art results both qualitatively and quantitatively, we show an extensive feature map visualization as shown in Fig. 11. Overall, the scene-saliency knowledge (K_1–K_5) explored by SKTM is the most accurate for locating salient objects, and CAM can provide rough location information of salient objects, while saliency features F_1–F_5 and scene features S_1–S_5 always suffer from some failures.

As illustrated in Fig. 11(a) and (b), for multiobject scenes, the saliency branch struggles to cope with the interference of complex background, and fails to focus on salient objects such as airplanes, vehicles at deep layers, e.g., F_4. For Fig. 11(c) and (e) of the single object scenes, the scene features also extract some object spatial information, such as the edges of salient objects or the central region where they are located, which shows that the scene features also might be helpful for RSI-SOD. With respect to Fig. 11(d) and (f), the F_4 and F_5 explore partial regions of salient objects, but lack attention to cover the complete objects. In contrast, the scene features suffer from low contrast problems and mainly produce false activation. In this case, CAM plays a critical role by combining scene features and saliency features to exploit more accurate and complete feature activation regions through SKTM.

To summarize, the scene features and CAMs output from the scene subnet are helpful for localizing salient objects in different scenes, **and they can complement and compensate**

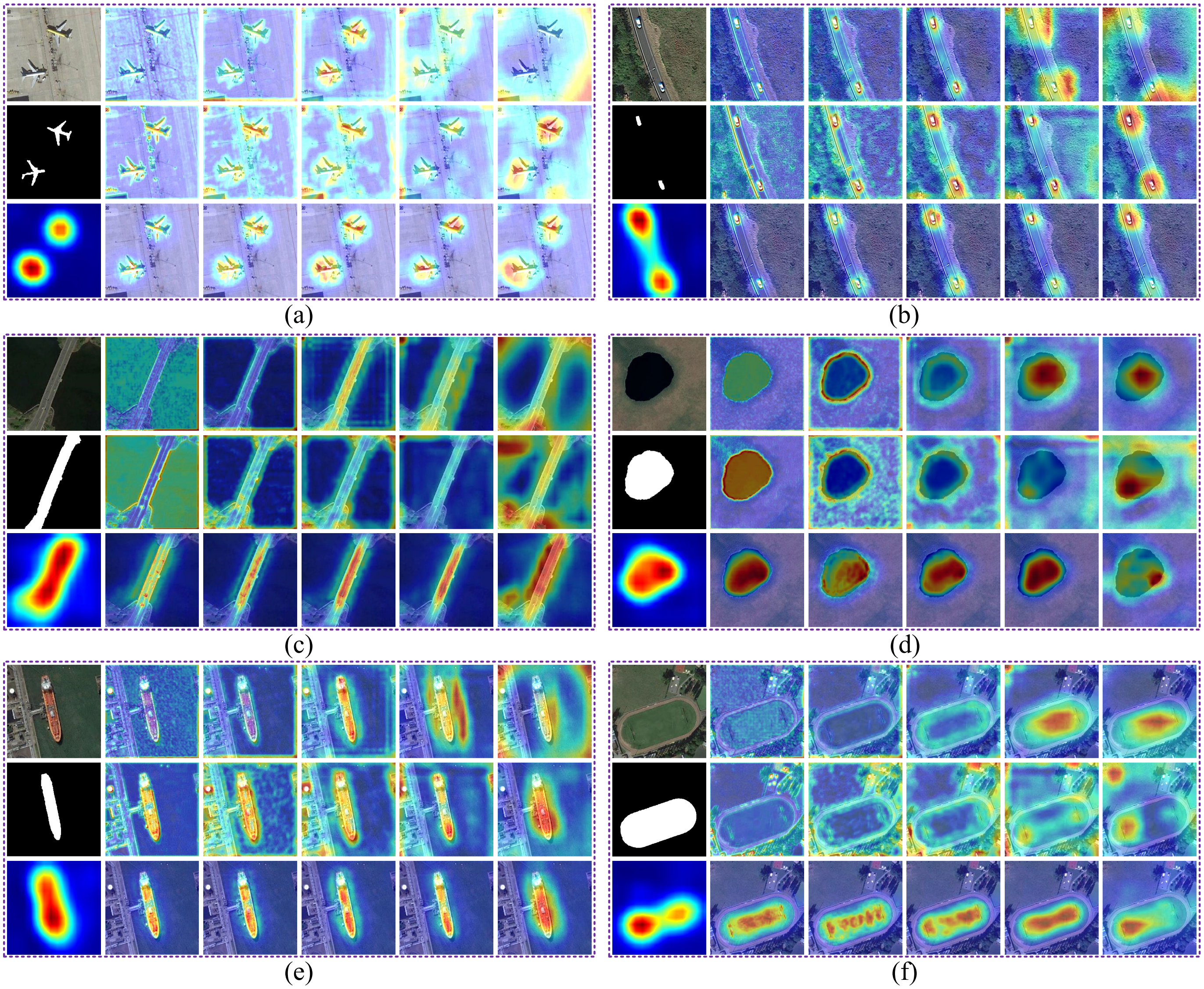


Fig. 11. Typical visualization of saliency features (F_1–F_5), scene features (S_1–S_5), dynamic CAM (C) and scene-saliency knowledge (K_1–K_5) delivered by SKTM. For each sample, the first row is RSI and F_1–F_5, the second row is GT and S_1–S_5, and the third row is C and K_1–K_5, respectively.

TABLE IV

Scene Guidance Comparison of SKTM and CDGM on ORSI-4199

Scene Aggregation Strategies	CAM	Feat.	MAE↓	F_β↑	S_m↑	E_m↑
baseline (PSPNet [53])			0.0377	0.8466	0.8657	0.9226
+ channel-wise concatenation	✓		0.0358	0.8503	0.8698	0.9267
		✓	0.0356	0.8481	0.8684	0.9251
+ element-wise summation	✓		0.0350	0.8508	0.8685	0.9267
		✓	0.0355	0.8499	0.8686	0.9266
+ element-wise multiplication	✓		0.0368	0.8533	0.8645	0.9260
		✓	0.0354	0.8511	0.8678	0.9264
+ proposed SKTM	✓	✓	0.0338	0.8576	0.8686	0.9297
+ proposed SKTM + CDGM	✓	✓	**0.0322**	**0.8598**	**0.8752**	**0.9339**

for the inadequacy of the saliency baseline for object localization and detection. Furthermore, Fig. 11 also demonstrates the effectiveness of the proposed SKTM, which can positively combine saliency features, scene features, and dynamic CAM to yield more complete and accurate activation maps for salient regions by the filtering and distillation mechanisms of multiscale spatial mixers and channel transforms efficiently.

3) Scene Guidance Comparison of SKTM and CDGM: Is it possible to integrate scene knowledge in a simple way? Obviously, we can integrate salient features with scene features or CAM by channel-wise concatenation, element-wise multiplication or summation. To compare the differences between the above schemes and the proposed SKTM and CDGM, we organize experiments on the ORSI-4199 dataset, as shown in Table IV. Our findings are as follows. First, no matter what plain approach is deployed, the scene features and dynamic CAM both have a certain promotive effect on RSI-SOD, with various metrics exceeding the baseline. Second, when only CAM or scene features are introduced, the single-channel dynamic CAM indeed delivers a better guidance capability for RSI-SOD, although the latter has multiple scales

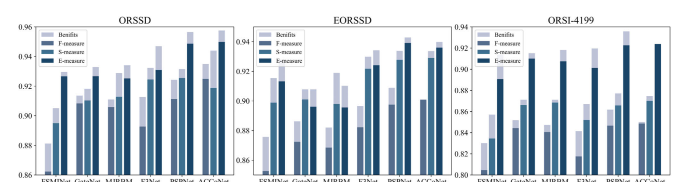


Fig. 12. Performance improvements histogram for three quantitative metrics.

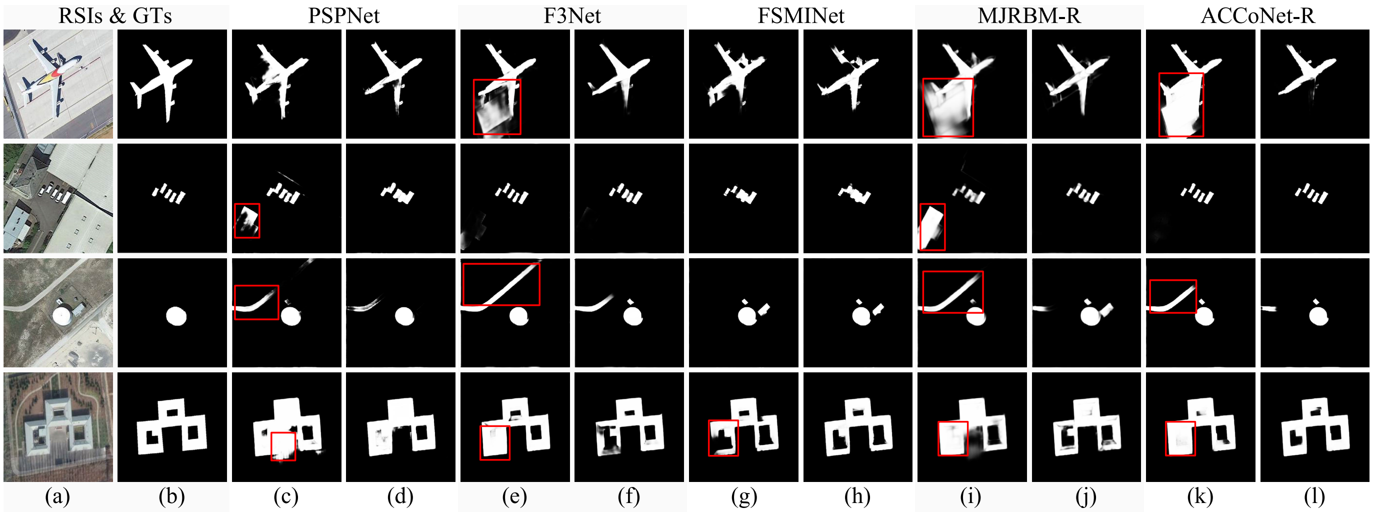


Fig. 13. Typical visualized prediction results of six algorithms with and without the proposed framework. (a) RSIs. (b) GTs. (c) PSPNet [53]. (d) PSPNet+Ours. (e) F3Net [73]. (f) F3Net+Ours. (g) FSMINet [27]. (h) FSMINet+Ours. (i) MJRBM-R [12]. (j) MJRBM-R+Ours. (k) ACCoNet-R [31]. (l) ACCoNet-R+Ours.

TABLE V
OBJECT AWARENESS COMPARISON OF OCAM ON ORSI-4199

Object Auxiliary Strategies	MAE↓	F_β↑	S_m↑	E_m↑
baseline + SKTM + CDGM	0.0322	0.8598	0.8752	0.9339
+ CT Loss [59]	0.0341	0.8592	0.8753	0.9332
+ ACT Loss [60]	0.0328	0.8592	0.8744	0.9322
+ proposed OCAM	**0.0313**	**0.8618**	**0.8771**	**0.9358**

TABLE VI
BACKBONE COMPARISON OF SALIENCY AND SCENE BRANCHES

SOD Branch	Scene Branch	ORSSD [10]		EORSSD [11]		ORSI-4199 [12]	
		F_β↑	MAE↓	F_β↑	MAE↓	F_β↑	MAE↓
Shared ResNet50 [52]		0.9151	0.0110	0.9011	0.0072	0.8561	0.0329
ResNet50 [52]	ResNet18 [52]	**0.9243**	**0.0099**	**0.9089**	**0.0063**	**0.8618**	**0.0313**
ResNet50 [52]	PVT b1 [76]	0.9142	0.0111	0.9045	0.0067	0.8611	0.0316

and channels. Third, the proposed SKTM outperforms these three vanilla aggregation approaches, and it combined with CDGM can achieve a more desirable performance among these competitors. 1) No matter what plain approach is deployed, the scene features and dynamic CAM both have a certain promotive effect on RSI-SOD, with various metrics exceeding the baseline. 2) When only CAM or scene features are introduced, the single-channel dynamic CAM indeed delivers a better guidance capability for RSI-SOD, although the latter has multiple scales and channels. 3) The proposed SKTM outperforms these three vanilla aggregation approaches, and it combined with CDGM can achieve a more desirable performance among these competitors.

4) Object Awareness Comparison of OCAM: As presented in Table V, we show the comparison of OCAM with two improved cross-entropy losses, i.e., CT [59] and ACT [60]. We find that these two modified losses do not offer gains in quantitative metrics and are even marginally weaker than the ordinary BCE loss. We infer that with the support of SKTM and CDGM, the above improved losses cannot make efforts on this basis anymore. Instead, the explicit and complete supervision of the object contours and boosting the awareness ability of shallow features for object edges by OCAM is a feasible solution to further promote RSI-SOD.

5) Design Rationale of Dual-Branch Framework: If we do not introduce a scene branch (like ResNet18) to extract scene features, and just equip a fully connected layer on the end of saliency baseline for scene classification, would such a framework also have comparable performance? To verify this idea, we conduct the experiments shown in Table VI. Unfortunately, the shared ResNet50 [52] as backbone struggles to achieve the performance of the dual-branch architecture. We believe that the two tasks, scene classification and RSI-SOD, are characteristically independent and mutually exclusive. A shared encoder would only confuse the spatial distribution of these two features and fail to achieve the goal of scene knowledge distillation. Besides, we consider PVT [76] as a scene branch to encode self-attention-based scene features, but such a model is inferior to ResNet18. A possible reason is that the inductive bias of convolution is very favorable for image-level classification of small-scale datasets. In summary, although the introduction of ResNet18 to design the dual-branch model brings a certain number of additional parameters, it achieves our goals and significantly contributes to RSI-SOD.

E. Scene Knowledge Transfer Is Model-Agnostic

In this section, we show that the proposed model is agnostic to the structure of saliency baselines. Specifically, we extend the proposed scene-aware framework to six state-of-the-art SOD algorithms, i.e., three NSI-SOD methods, PSPNet [53], GateNet [72], F3Net [73], and three RSI-SOD approaches, FSMINet [27], MJRBM [12], and ACCoNet [31]. These six algorithms employ various encoders and design distinctive network structures with widely varying model capacities and parameter numbers. To demonstrate the model-agnostic capability of our framework, we perform a full comparison from both quantitative and qualitative aspects.

Table VII reports the results of a series of experiments under the same input settings on three RSI-SOD datasets. Overall, the model integrated with the proposed MTL framework exhibits significant improvements in most metrics on all datasets. As for the parameter cost of six baselines, the total number of parameters of proposed three contributive modules, SKTMs, CDGMs, and OCAMs, is 5.99M, 7.03M,

TABLE VII
QUANTITATIVE RESULTS OF SIX STATE-OF-THE-ART METHODS ON THREE RSI-SOD DATASETS WITH AND WITHOUT THE PROPOSED FRAMEWORK

Methods	Published	Parameters↑	ORSSD Dataset [10]				EORSSD Dataset [11]				ORSI-4199 Dataset [12]			
			F_β↑	MAE↓	S_m↑	E_m↑	F_β↑	MAE↓	S_m↑	E_m↑	F_β↑	MAE↓	S_m↑	E_m↑
PSPNet [53]	CVPR'17	+5.99M	0.9113	0.0117	0.9255	0.9487	0.8975	0.0078	0.9278	0.9391	0.8466	0.0377	0.8657	0.9226
PSPNet + Ours			**0.9243**	**0.0099**	**0.9315**	**0.9565**	**0.9089**	**0.0063**	**0.9338**	**0.9429**	**0.8618**	**0.0313**	**0.8771**	**0.9358**
GateNet [72]	ECCV'20	+7.03M	0.9083	0.0125	0.9103	0.9267	0.8724	0.0091	0.9010	0.8961	0.8443	0.0387	0.8660	0.9101
GateNet + Ours			**0.9136**	**0.0115**	**0.9182**	**0.9329**	**0.8862**	**0.0077**	**0.9078**	**0.9078**	**0.8517**	**0.0362**	**0.8712**	**0.9151**
F3Net [73]	AAAI'20	+6.01M	0.8927	0.0126	0.9245	0.9309	0.8822	0.0077	0.9218	0.9241	0.8175	0.0435	0.8520	0.9013
F3Net + Ours			**0.9126**	**0.0103**	**0.9324**	**0.9469**	**0.8965**	**0.0068**	**0.9299**	**0.9342**	**0.8414**	**0.0360**	**0.8670**	**0.9196**
FSMINet [27]	GRSL'22	+6.27M	0.8623	0.0178	0.8949	0.9266	0.8527	0.0100	0.8989	0.9132	0.8046	0.0451	0.8344	0.8906
FSMINet + Ours			**0.8812**	**0.0158**	**0.9051**	**0.9296**	**0.8757**	**0.0083**	**0.9154**	**0.9280**	**0.8301**	**0.0397**	**0.8570**	**0.9138**
MJRBM-R [12]	TGRS'22	+12.84M	0.9058	0.0129	0.9128	0.9252	0.8685	0.0092	0.8980	0.8956	0.8406	0.0379	0.8685	0.9075
MJRBM-R + Ours			**0.9110**	**0.0102**	**0.9288**	**0.9341**	**0.8820**	**0.0073**	**0.9190**	**0.9103**	**0.8473**	**0.0350**	**0.8712**	**0.9181**
ACCoNet-R [31]	TCYB'23	+11.49M	0.9249	0.0102	0.9187	0.9498	**0.9009**	0.0084	0.9290	0.9360	0.8486	0.0354	0.8702	**0.9238**
ACCoNet-R + Ours			**0.9349**	**0.0089**	**0.9440**	**0.9576**	0.9003	**0.0066**	**0.9336**	**0.9398**	**0.8501**	**0.0335**	**0.8746**	**0.9238**

6.01M, 6.27M, 12.84M, and 11.49M, respectively. That is, the number of parameters introduced by the proposed modules is relatively light when added to the existing models. As shown in Fig. 12, the F_β metric still exceeds the baseline by a large margin when our scene-guided SDNet is implemented on this baseline. The above illustrates that the proposed framework is a universal model that does not depend on any specific model structure, but instead performs scene knowledge distillation that facilitates saliency localization.

To reveal how the proposed framework contributes to the performance of RSI-SOD, we provide some comparative prediction results of the five models on the ORSI-4199 dataset in Fig. 13. By comparison, the upgraded models combined with scene-saliency knowledge show better detection results in various scenarios. Specifically, the updated models reduce the false detection rate of background and non-salient objects, and improve the saliency accuracy in rows 1–3, including airplanes, vehicles, and storage tanks. **This is precisely because of the introduction of scene knowledge, which avoids the inadequacy of pixel-level supervision and overcomes the erroneous detection of background and non-salient objects.** We believe that it is the most critical factor for which scene knowledge distillation can potentially boost RSI-SOD.

V. CONCLUSION

In this work, we design a universal, effective, and model-agnostic scene knowledge distillation framework for RSI-SOD. To achieve this goal, we first define 12 types of scene categories and annotate three RSI-SOD datasets with image-level scene labels. Then, we couple the saliency baseline with a parallel scene subnet to extract both saliency features and scene features. Considering the weakness of image-level supervision, we introduce dynamic CAM in the scene subnet to explore the localization of salient objects. Then, we propose a novel SKTM to integrate scene features, dynamic CAM, and saliency features to obtain the saliency region activation as accurately as possible. To achieve conditionally dynamic guidance strategies, an adaptive module named CDGM is presented to deliver guidance from scene knowledge to enhanced salience features. Furthermore, a simple yet effective module named OCAM is proposed to boost the learning of spatial details and contours of salient objects in a supervised manner at shallow levels. Extensive experiments demonstrate that the proposed algorithm exceeds more than 20 state-of-the-art methods both quantitatively and qualitatively.

In the experiments, we employ ablation studies to analyze the effectiveness of the proposed SKTM, CDGM, and OCAM. We also show the soundness of our annotated scene classification labels by means of confusion matrices. In addition, we justify the superiority of the proposed modules over various existing methods or simple feature fusion strategies. Finally, we also explain why the presented model can facilitate saliency understanding, and reveal the model-agnostic ability of scene knowledge distillation through feature visualization.

In the future, we will exploit the scene labels presented in this article as weakly supervised signals for RSI-SOD to investigate image-level weakly supervised learning algorithms.

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] R. Xu, C. Wang, J. Zhang, S. Xu, W. Meng, and X. Zhang, "RSSFormer: Foreground saliency enhancement for remote sensing land-cover segmentation," *IEEE Trans. Image Process.*, vol. 32, pp. 1052–1064, 2023.

[3] 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, 2022, Art. no. 5614914.

[4] 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.

[5] C. Zhang, K.-M. Lam, and Q. Wang, "CoF-Net: A progressive coarse-to-fine framework for object detection in remote-sensing imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5600617.

[6] C. Zhang, J. Su, Y. Ju, K.-M. Lam, and Q. Wang, "Efficient inductive vision transformer for oriented object detection in remote sensing imagery," *IEEE Trans. Geosci. Remote Sens.*, early access, Jul. 5, 2023, doi: 10.1109/TGRS.2023.3292418.

[7] 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)*, Jul. 2017, pp. 6593–6601.

[8] 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.

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

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

[11] Q. Zhang et al., "Dense attention fluid network for salient object detection in optical remote sensing images," *IEEE Trans. Image Process.*, vol. 30, pp. 1305–1317, 2021.

[12] 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, 2022, Art. no. 5607913.

[13] Z. Xiong, Y. Liu, Q. Wang, and X. X. Zhu, "RSSOD-bench: A large-scale benchmark dataset for salient object detection in optical remote sensing imagery," in *Proc. IEEE Int. Geosci. Remote Sens. Symp. (IGARSS)*, Jul. 2023, pp. 1–4.

[14] 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.

[15] 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, 2022, Art. no. 5624915.

[16] Y. Liu, Z. Xiong, Y. Yuan, and Q. Wang, "Distilling knowledge from super-resolution for efficient remote sensing salient object detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5609116.

[17] G.-S. Xia et al., "AID: A benchmark data set for performance evaluation of aerial scene classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 7, pp. 3965–3981, Jul. 2017.

[18] G. Cheng, J. Han, and X. Lu, "Remote sensing image scene classification: Benchmark and state of the art," *Proc. IEEE*, vol. 105, no. 10, pp. 1865–1883, Oct. 2017.

[19] J. Liu, S. Li, C. Zhou, X. Cao, Y. Gao, and B. Wang, "SRAF-Net: A scene-relevant anchor-free object detection network in remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5405914.

[20] J. Xu, J. Gong, J. Zhou, X. Tan, Y. Xie, and L. Ma, "SceneEncoder: Scene-aware semantic segmentation of point clouds with a learnable scene descriptor," in *Proc. Int. Joint Conf. Artif. Intell. (IJCAI)*, Jul. 2020, pp. 601–607.

[21] S. Zhang, Q. Yuan, J. Li, J. Sun, and X. Zhang, "Scene-adaptive remote sensing image super-resolution using a multiscale attention network," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 7, pp. 4764–4779, Jul. 2020.

[22] C. Tao, L. Mi, Y. Li, J. Qi, Y. Xiao, and J. Zhang, "Scene context-driven vehicle detection in high-resolution aerial images," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 10, pp. 7339–7351, Oct. 2019.

[23] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, "Learning deep features for discriminative localization," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 2921–2929.

[24] D. Zhao, J. Wang, J. Shi, and Z. Jiang, "Sparsity-guided saliency detection for remote sensing images," *J. Appl. Remote Sens.*, vol. 9, pp. 1–14, Sep. 2015.

[25] 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.

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

[27] 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, 2022, Art. no. 6507705.

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

[29] 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, 2022, Art. no. 5613311.

[30] 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, 2022, Art. no. 5605315.

[31] 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.

[32] Y. Liu, Y. Yuan, and Q. Wang, "Uncertainty-aware graph reasoning with global collaborative learning for remote sensing salient object detection," *IEEE Geosci. Remote Sens. Lett.*, vol. 20, pp. 1–5, 2023.

[33] Q. Wang, S. Liu, J. Chanussot, and X. Li, "Scene classification with recurrent attention of VHR remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 2, pp. 1155–1167, Feb. 2019.

[34] P. Li, P. Chen, and D. Zhang, "Cross-modal feature representation learning and label graph mining in a residual multi-attentional CNN-LSTM network for multi-label aerial scene classification," *Remote Sens.*, vol. 14, no. 10, p. 2424, May 2022.

[35] Q. Wang, W. Huang, Z. Xiong, and X. Li, "Looking closer at the scene: Multiscale representation learning for remote sensing image scene classification," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 33, no. 4, pp. 1414–1428, Apr. 2022.

[36] X. Lu, X. Zheng, and Y. Yuan, "Remote sensing scene classification by unsupervised representation learning," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 9, pp. 5148–5157, Sep. 2017.

[37] W. Huang, Y. Shi, Z. Xiong, Q. Wang, and X. X. Zhu, "Semi-supervised bidirectional alignment for remote sensing cross-domain scene classification," *ISPRS J. Photogramm. Remote Sens.*, vol. 195, pp. 192–203, Jan. 2023.

[38] L. Li, J. Han, X. Yao, G. Cheng, and L. Guo, "DLA-MatchNet for few-shot remote sensing image scene classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 9, pp. 7844–7853, Sep. 2021.

[39] J. Zhang, H. Zhao, and J. Li, "TRS: Transformers for remote sensing scene classification," *Remote Sens.*, vol. 13, no. 20, p. 4143, Oct. 2021.

[40] J. Shen, B. Cao, C. Zhang, R. Wang, and Q. Wang, "Remote sensing scene classification based on attention-enabled progressively searching," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 4707513.

[41] P. Lv, W. Wu, Y. Zhong, F. Du, and L. Zhang, "SCViT: A spatial-channel feature preserving vision transformer for remote sensing image scene classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 4409512.

[42] Y. Hua, L. Mou, P. Jin, and X. X. Zhu, "MultiScene: A large-scale dataset and benchmark for multiscene recognition in single aerial images," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5610213.

[43] G. Zhai, G. Liu, X. He, Z. Wang, C. Ren, and Z. Chen, "Adaptive scene-aware deep attention network for remote sensing image compression," *J. Electron. Imag.*, vol. 30, no. 5, Sep. 2021, Art. no. 053008.

[44] C. Chen, W. Gong, Y. Chen, and W. Li, "Object detection in remote sensing images based on a scene-contextual feature pyramid network," *Remote Sens.*, vol. 11, no. 3, p. 339, Feb. 2019.

[45] N. Liu and J. Han, "A deep spatial contextual long-term recurrent convolutional network for saliency detection," *IEEE Trans. Image Process.*, vol. 27, no. 7, pp. 3264–3274, Jul. 2018.

[46] W. Li, W. Wei, and L. Zhang, "GSDet: Object detection in aerial images based on scale reasoning," *IEEE Trans. Image Process.*, vol. 30, pp. 4599–4609, 2021.

[47] Z. Zheng, X. Yue, K. Keutzer, and A. S. Vincentelli, "Scene-aware learning network for radar object detection," in *Proc. Int. Conf. Multimedia Retr. (ICMR)*, Aug. 2021, pp. 573–579.

[48] H. Guo, Q. Shi, B. Du, L. Zhang, D. Wang, and H. Ding, "Scene-driven multitask parallel attention network for building extraction in high-resolution remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 5, pp. 4287–4306, May 2021.

[49] Y. Li, Y. Zhang, X. Huang, and A. L. Yuille, "Deep networks under scene-level supervision for multi-class geospatial object detection from remote sensing images," *ISPRS J. Photogramm. Remote Sens.*, vol. 146, pp. 182–196, Dec. 2018.

[50] X. Wu and Z. Shi, "Scene aggregation network for cloud detection on remote sensing imagery," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, pp. 1–5, 2022.

[51] 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. 3991–4001.

[52] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Mar. 2016, pp. 770–778.

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

[54] Z. Chen, T. Wang, X. Wu, X.-S. Hua, H. Zhang, and Q. Sun, "Class re-activation maps for weakly-supervised semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 959–968.

[55] J. Wang, A. Bhalerao, T. Yin, S. See, and Y. He, "CAMANet: Class activation map guided attention network for radiology report generation," 2022, *arXiv:2211.01412*.

[56] W. Yu et al., "MetaFormer is actually what you need for vision," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 10809–10819.

[57] Y. Chen, X. Dai, M. Liu, D. Chen, L. Yuan, and Z. Liu, "Dynamic convolution: Attention over convolution kernels," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 11027–11036.

[58] Y. Li et al., "Revisiting dynamic convolution via matrix decomposition," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, May 2021, pp. 1–11.

[59] Z. Chen, H. Zhou, J. Lai, L. Yang, and X. Xie, "Contour-aware loss: Boundary-aware learning for salient object segmentation," *IEEE Trans. Image Process.*, vol. 30, pp. 431–443, 2021.

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

[61] P. Burt and E. Adelson, "The Laplacian pyramid as a compact image code," *IEEE Trans. Commun.*, vol. 31, no. 4, pp. 532–540, Apr. 1983.

[62] R. Deng, C. Shen, S. Liu, H. Wang, and X. Liu, "Learning to predict crisp boundaries," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Sep. 2018, pp. 562–578.

[63] M. Fan et al., "Rethinking BiSeNet for real-time semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 9711–9720.

[64] 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.

[65] 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.

[66] 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.

[67] D.-P. Fan, C. Gong, Y. Cao, B. Ren, M.-M. Cheng, and A. Borji, "Enhanced-alignment measure for binary foreground map evaluation," in *Proc. Int. Joint Conf. Artif. Intell. (IJCAI)*, Jul. 2018, pp. 698–704.

[68] 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.

[69] 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.

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

[71] 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.

[72] 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.

[73] J. Wei, S. Wang, and Q. Huang, "F^3Net: Fusion, feedback and focus for salient object detection," in *Proc. AAAI Conf. Artif. Intell. (AAAI)*, Apr. 2020, vol. 34, no. 7, pp. 12321–12328.

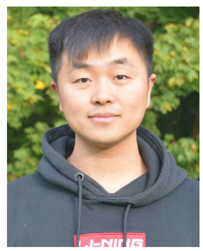
[74] M. Ma, C. Xia, and J. Li, "Pyramidal feature shrinking for salient object detection," in *Proc. AAAI Conf. Artif. Intell. (AAAI)*, May 2021, vol. 35, no. 3, pp. 2311–2318.

[75] 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.

[76] W. Wang et al., "Pyramid vision transformer: A versatile backbone for dense prediction without convolutions," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 548–558.

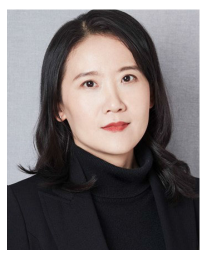
Yanfeng Liu (Student Member, IEEE) received the B.E. degree in computer science and technology from Northeast Forestry University, Harbin, China, in 2021. He is currently pursuing the M.S. degree in computer science and technology with the School of Computer Science and the School of Artificial Intelligence,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

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



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 the 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 more than 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.