Adaptive Dual-Stream Sparse Transformer Network for Salient Object Detection in Optical Remote Sensing Images

Jie Zhao, Yun Jia, Lin Ma, *Senior Member, IEEE*, and Lidan Yu

***Abstract*—Excellent performance has been demonstrated by convolutional neural network (CNN) in salient object detection for optical remote sensing images (ORSI-SOD). However, the limitations of CNN's feature extraction using sliding window approach hinder the capture of global representations. Therefore, an end-to-end detection model, known as adaptive dual-stream sparse transformer network (ADSTNet), has been proposed for ORSI-SOD and is assisted by the vision transformer. It effectively addresses the compensation issue of global and local information in ORSI-SOD. In particular, an adaptive interaction encoder has been devised, amalgamating the multiscale sparse transformer and the pyramid atrous attention to constitute the adaptive dual-stream sparse encoder. This encoder collaborates with the CNN to enhance long-range dependency modeling and preserve global information more effectively base on local features. In addition, a directional feature reconfiguration is constructed to extract texture details from multiple directional dimensions. Finally, we propose the adaptive feature cascade decoder that synthesizes content information from the foreground, edges, and background to enhance the representational capacity of the image. Furthermore, a structural loss function, known as the weight compensation mechanism, is introduced to balance the performance of boundary and salmap segmentation losses. The proposed model has been demonstrated to outperform 26 state-of-the-art ORSI-SOD methods across eight evaluation metrics on two standard datasets, as evidenced by extensive experiments. Furthermore, to verify its robustness, the generalization performance of the model on the latest challenging ORSI-4199 dataset is reported.**

***Index Terms*—Adaptive, boundary detection operator, optical remote sensing images (ORSIs), salient object detection (SOD), sparse transformer.**

I. Introduction

SALIENT object detection (SOD) [1], [2], inspired by human visual attention mechanisms, aims to identify the most distinctive objects and determine their locations in complex images. Widely applied in image and video processing, SOD finds diverse applications in areas such as camouflaged detection [3], [4], human–computer interaction [5], [6], [7], [8], video surveillance [9], [10], [11], and medical imaging [12], [13], [14], [15]. Our focus is on salient object detection for optical remote sensing images (ORSI-SOD) [16], [17], [18], [19], distinguishing it from classical salient object detection in natural scene images (NSI-SOD) [20], [21]. ORSI-SOD, a rapidly growing subfield, has proven successful in ship detection [22], [23], airport detection [24], [25], [26], instance, and semantic segmentation [27], [28]. Compared with prominent semantic segmentation [29], [30], [31], [32], [33], [34], [35] in computer vision, ORSI-SOD shares some resemblances but distinguishes itself with unique characteristics. Concerning task objectives, both contribute to segmentation; however, semantic segmentation entails pixel-level delineation of the entire image, while optical remote sensing image saliency models concentrate on isolating a specific object or region within remote sensing images. The objective is to accentuate the most salient object against the surrounding backdrop, rather than segmenting the entire image comprehensively. In recent years, numerous saliency detection methods have emerged, garnering attention and falling into two categories: 1) traditional methods; and 2) convolutional neural network-based (CNN-based) methods.

The emergence of traditional methods and CNN has stimulated the development of ORSI-SOD [36], [37], [38]. Traditional methods for SOD [26], [39], [40] often rely on low-level attributes such as color information content [41] and saliency feature analysis [42]. However, they fail to generate accurate information representations for some deep and low-level features. In contrast, CNN can automatically learn features through large-scale data, and exhibiting stronger adaptability to complex scenes and noise. MCCNet [43] utilizes multiple content feature information for complementarity, and ACCoNet [44] employs multiscale information interaction. CNN is more adept at extracting local region features. As shown in Fig. 1, the local attention of a standard CNN structure tends to focus on neighboring features around a key point, making it difficult to capture global representations. In response to this issue, a number of CNN-based methods [16], [45], [46], [47], [49] have been proposed to capture a wide receptive field by utilizing deeper network architectures. They also explore global cues through different techniques such as global pooling or nonlocal modules. However, the adoption of deeper network layers

Manuscript received 14 September 2023; revised 5 December 2023; accepted 9 February 2024. Date of publication 16 February 2024; date of current version 28 February 2024. This work was supported by the Yantai Science and Technology Innovation Development Plan under Grant 2022XDRH031. *(Corresponding author: Yun Jia.)*

Jie Zhao, Yun Jia, and Lidan Yu are with the School of Information and Electronic Engineering, Shandong Technology and Business University, Yantai 264005, China (e-mail: jiezhao_99@163.com; jiayun0223@sdtbu.edu.cn; lidan_yu@163.com).

Lin Ma is with the School of Electronics and Information Engineering, Harbin Institute of Technology, Harbin 150006, China (e-mail: malin@hit.edu.cn).

The code and results for this work can be found at https://github.com/JieZzzoo/ADSTNet.

Digital Object Identifier 10.1109/JSTARS.2024.3365729

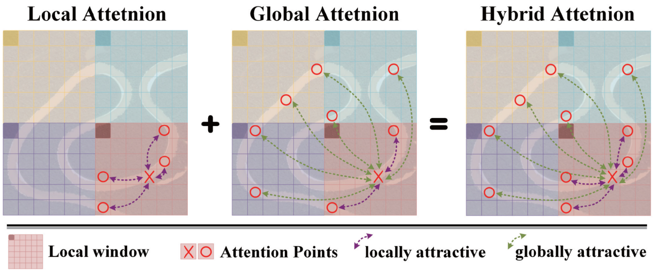


Fig. 1. Visualization of the local attention of CNN, global attention of Transformer and hybrid attention of the proposed ADSTNet in the feature space. It could be get that the hybrid could capture the more comprehensive and accurate information by compare above attention. Best viewed in color.

unavoidably incurs considerable computational overhead, while maintaining the standard structure of deep neural networks can pose challenges in achieving long-range dependencies.

Therefore, we believe that a framework capable of global information stimulated the development of ORSI-SOD is also possible. VST [50] was a pioneer in introducing the transformer to the SOD, replacing conventional CNN models with self-attention mechanisms to explore global information. Subsequently, ASTT [51] designed the adaptive spatial tokenization module to mitigate the impact of optical image features on SOD and also employed the transformer to explore global information. These works have demonstrated the necessity of replacing CNN with transformer architectures to explore global information in the ORSI-SOD. Moreover, various transformer variants have been developed by researchers for other domains of SOD, resulting in substantial advances in RGB SOD [52], RGB-D/T SOD [53], [54], and Video SOD (VSOD) [55]. However, transformers cannot extract local information as effectively as CNN in Fig. 1. This is because the lack of CNN's inductive biases results in less effective extraction of local information compared to CNN, leading to a degradation in performance.

Inspired by CNN- and Transformer-based approaches in SOD, it is worthwhile to explore the fusion of these two methods to achieve the maximum representation of SOD techniques. However, optical images captured from high altitudes are typically characterized by small and varying scales, presenting significant limitations when directly applying NSI-SOD methods to ORSI-SOD, resulting in unsatisfactory performance. In addition, incorporating boundary information as supervision can compel the network to learn more accurate pixel-level edge information, which is crucial in ORSI-SOD. Currently, there is no comprehensive CNN with transformer fusion architecture suitable for ORSI-SOD while incorporating boundary supervision.

In this regard, we propose the adaptive dual-stream sparse transformer network (ADSTNet), which effectively combines features at different levels from both local and global perspectives and achieves precise detection and localization of ORSI-SOD through boundary-guided assistance. The encoding stage employs a sparse framework that balancing the encoding of local region information and global object relationships. Specifically, one branch of the encode extracts spatial features using CNN are combined with global dependencies established by the adaptive dual-stream sparse encode (ADSE), interacting to alleviate discrepancies. Features at different levels are adaptively captured by this encoder, thereby enhancing the interpretability of the model results. To acquire more accurate representations of salient object boundary features in ORSIs, we introduce a directional feature reconfiguration (DFR) as a plug-and-play component, enhancing boundary information. It is noteworthy that, to the best of our knowledge, this is the initial attempt to apply dedicated boundary detection operators to the ORSI-SOD task.

Furthermore, we propose adaptive feature cascade decoder (AFCD) to guide the decoder learning process using boundary masks as explicit supervision. We construct a comprehensive loss function that balances boundary loss and saliency map loss through a weight compensation mechanism, further improving the accuracy and robustness of ORSI-SOD. In this manner, the ADSTNet network achieves the best results compared to 26 state-of-the-art (SOTA) models, achieving optimal performance in terms of S-measure, adaptive F-measure, adaptive E-measure, and other evaluation metrics.

Our contributions can be summarized as follows.

1) We propose an encoder–decoder structure, namely ADSTNet, which combines the strengths of CNN and Transformer to efficiently complement local and global information. With the support of boundary information, it achieves better feature extraction at different levels and enhances representation learning. In addition, we guide the model's loss through a weight compensation mechanism for implicit supervised feature learning, thus improving the model's robustness.
2) We introduce the ADSE, which enhances the global perception of local features and local details of global representations in the adaptive interaction encoding. We also propose the plug-and-play DFR, which strengthens the representation capability of boundary information using a dedicated boundary detection operator.
3) We design an AFCD that explicitly enhances the encoding feature representation through complementary learning from multiple contents. Intraclass and interclass consistency within the feature space are effectively captured by it.

The rest of this article is organized as follows. Section II offers an overview of related research, while Section III provides a detailed description of the proposed model. Section IV reports on the comprehensive evaluation of our model, including ablation analyses and an analysis of failure cases. Finally, Section VI concludes this article.

II. Related Work

In this section, we begin by summarizing the work on ORSI-SOD. Subsequently, a concise overview of the advancements in vision transformer and the utilization of transformers in SOD is provided. Lastly, we elucidate the boundary detection operators employed in image processing.

A. *Salient Object Detection for ORSI*

Recently, extensive research has been conducted by scholars to address the various challenges faced by the emerging task of

ORSI-SOD within the SOD community. Among these efforts, the CNN-based encoder–decoder structures have gained significant popularity [48], [57], [58], [59]. Hou et al. [60] introduced deep supervision into SOD, implicitly enhancing the multiscale feature representation of salient objects. The implementation of this approach significantly enhances detection accuracy and has had a profound impact on subsequent CNN-based methods. Li et al. [61] extracted feature information from images at three different resolutions to optimize the detection drawbacks caused by varying object scales. Bai et al. [62] proposed a global–local–global context-aware network to obtain the final comprehensive representation of salient objects in a spatially and semantically global manner. Furthermore, cross-scale interaction is achieved through an enlarged receptive field in the network proposed by Zheng et al. in [37], which utilizes dilated convolutions and attention mechanisms to capture potential fine-grained information.

In addition, some works have drawn inspiration from NSI-SOD and proposed strategies to incorporate contextual and boundary information to adapt to ORSI-SOD features, alleviating challenges in ORSI. For instance, the fusion of contextual features enhances the encoded representations [44], [63], exploring the contributions of boundaries, foreground, and background to global information through the complementary integration of multiple content information [43], and utilizing additional edge labels to improve the model's boundary perception capability [47], [64], [65]. Despite the foundation laid by CNN-based ORSI-SOD models in improving performance, their performance is limited due to the constrained long-range semantic contextual relationships of CNN, as convolutions have a limited receptive field. To address this limitation, we propose a sparse transformer-assisted dual-stream encoder that enhances the global perception of local features and captures local details of global representation, thus compensating for the deficiency in capturing global information by CNN.

B. Vision Transformer

While CNN [66] have demonstrated excellent performance in visual tasks [67], [68], [69], [70], they are still constrained by the limitation of employing a strategy involving the gradual expansion of the receptive field through the use of local window movements, hindering effective modeling of long-distance relationships [72], [73], [74]. In a parallel field such as natural language processing, another popular technique called transformer has emerged. Transformer leverages its self-attention mechanism to capture extensive global relationships and has achieved notable success. Recognizing the significance of global information in visual tasks, researchers have introduced transformers into image processing to overcome the limitations of CNNs, thereby mitigating the risks associated with compromising feature resolution and representational capacity. The effective integration of both aspects can significantly address the challenges associated with a singular focus. Some works [75], [76], [77] proposed to linearly combine CNN and Transformer to achieve the combination of local mechanism and dynamic attention. And [78], [79] proposed a dual-stream network based on CNN and Transformer to fully explore the representation ability of local and global pattern features in image classification. In addition, the authors in [80] constructed a dual-transformer with two parallel pathways, integrating pixel pathways and semantic pathways to enhance self-attention. In contrast, the authors in [72] built an interactive structure to achieve information exchange and joint feature learning between CNN and Transformer, fully learning the relationships between different positions.

The accomplishments of vision transformers in NSI-SOD have also been showcased for ORSI-SOD. For instance, a pioneering study by Liu et al. [50] presented a unified RGB and RGB-D SOD model based on a vision transformer achieving saliency and boundary detection by introducing task-specific labels. Wang et al. [52] proposed a transformer architecture consisting of an FCN decoder and three additional modules to capture salient local and global information in RGB images. The interplay of information from different modalities facilitates the learning of deeper information representations by the network model. To fully exploit the essence of different modalities, Liu et al. [53] introduced a dual-stream Swin Transformer equipped with spatial alignment and channel calibration modules, effectively integrating multimodal information and aggregating intralayer features. A transformer-based model was proposed by Zhang et al. [81] to capture implicit details and create the challenging RGBD COSAL1K dataset. This model incorporates two class labels to extract intrasaliency and intersaliency information, respectively. Moreover, a cross-reference transformer model that integrates appearance and motion cues from VSOD was presented by Huang et al. [55]. Furthermore, Zhang et al. [82] presented a transformer-guided dual-stream structure that enhances information features through cascading. However, these methods are limited by a large number of parameters, making optimization challenging. In this article, we suggest utilizing a sparse transformer-assisted CNN to acquire global information while reducing noise caused by irrelevant information, thereby enhancing foreground-background discrimination.

C. Operators in Image Processing

In the field of digital image processing, operators play a crucial role as fundamental components. Among them, boundary detection operators, as one of the most central elements, have garnered widespread attention and research. Two types of commonly used edge detection operators exist, namely: 1) first-order derivative operators; and 2) second-order derivative operators. The first-order derivative operators include Roberts, Prewitt, and Sobel, while the second-order derivative operators include Laplacian [83]. In recent years, boundary detection operators have regained importance in pixel-level computer vision tasks, such as camouflage object detection [3], [84], manipulation detection [85], and MISEG [12], gaining wide applications and research interests. Within this article, we utilize edge detection operators to construct the AFCD as an explicit mask extractor. The purpose of this is to guide the implicit feature learning process in ORSI-SOD. Our research is the first to apply boundary detection operators in ORSI-SOD, synthesizing high-quality information predictions from the feature maps transmitted by the

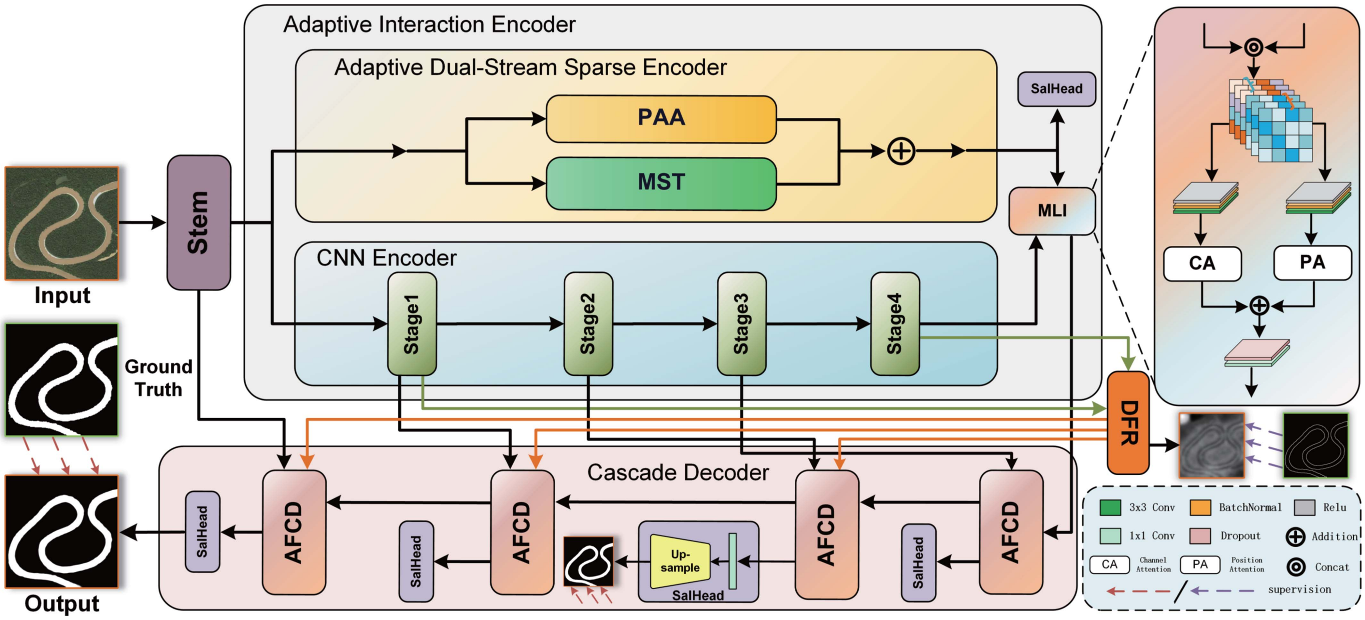


Fig. 2. Overall architecture of the proposed ADSTNet, including the ADSE, DFR, AFCD. Meanwhile, we guide the model's loss through a weight compensation mechanism for implicit supervised feature learning, thus improving the model's robustness.

backbone encoder, resulting in satisfactory results. This work not only expands the application scope of edge detection operators, but also provides new insights and approaches for research in the field of ORSI-SOD.

III. Proposed Method

In the present section, the proposed ADSTNet network is introduced. Subsequently, each component is described in order, as presented in Sections III-B to III-D. Finally, the loss function, which balances the weighted compensation mechanism, is elucidated.

A. Overview of the Proposed Architecture

The proposed ADSTNet follows an encoder–decoder structure, and its main framework is illustrated in Fig. 2, consisting of three parts: adaptive interaction encoder, DFR, and cascade decoder. Firstly, the input image $I \in \mathbb{R}^{3\times 256\times 256}$ is fed into the stem to obtain initial fine features $f_{\text{stem}} \in \mathbb{R}^{64\times 128\times 128}$. Then, adaptive interaction encoder is employed to capture more extensive local and global information. Here, adaptive interaction encoder comprises two components: the well-known Res2Net serves as the CNN encoder, extracting multiscale and multilevel local features f_{le}, while the ADSE composed of pyramid atrous attention (PAA) and multiscale sparse transformer (MST) is designed to extract more comprehensive global information f_{ge}. Specifically, Res2Net is divided into four stages, sequentially extracting information denoted as $f_{ce}^{t} \in R^{h_t \times w_t \times c_t}$, where h_t being $^{256}/_{2^{t+1}}$, w_t being $^{256}/_{2^{t+1}}$, $c_t = \{256, 512, 1024, 2048\}$, t is the stage index and belongs to $\{1, 2, 3, 4\}$. Meanwhile, the features f_{ADSE} generated by ADSE are also outputted for the primary saliency map ADSE with reverse supervision, forcing the learning of more accurate information and laying the foundation for subsequent feature refinement. The multilevel integration (MLI) eliminates semantic discrepancies between f_{le} and f_{ge} interactively, greatly enhancing the global perception ability of local features and the local details of global representations, delivering rich high-level semantic information f_e to the AFCD for feature parsing. Due to the multiscale information captured by encoder, each detection head includes a 1×1 convolutional layer and upsampling to restore the resolution and obtain saliency maps. In addition, to enhance the auxiliary function of boundaries, we employ a mixed loss function assisted by a weighted compensation mechanism to implicitly assign clear boundaries to the saliency maps, ensuring plausible acceptance. We also show the inference details based on the proposed ADSTNet in Algorithm 1. Next, we will provide detailed explanations for each component.

B. Adaptive Dual-Stream Sparse Encoder

To address the dilemma of CNN getting trapped in global information extraction, we propose an ADSE that complements global and local information, as shown in Fig. 3. Specifically, the features from the stem are concurrently fed into two branches, each compensating for the other's deficiencies and progressively enhancing the missing information, thereby achieving the purification of high-quality and effective features.

To achieve a balance between computational efficiency and global information, we construct a new mechanism, MST, for capturing global information in ADSE. Similar to ViT, the encoding layer consists of a multihead attention layer and a feed-forward network (FFN). However, the multihead attention enriches the object information by transforming the received Q, K, and V information through dimensionality changes. The transformer information is then fed into the attention module to

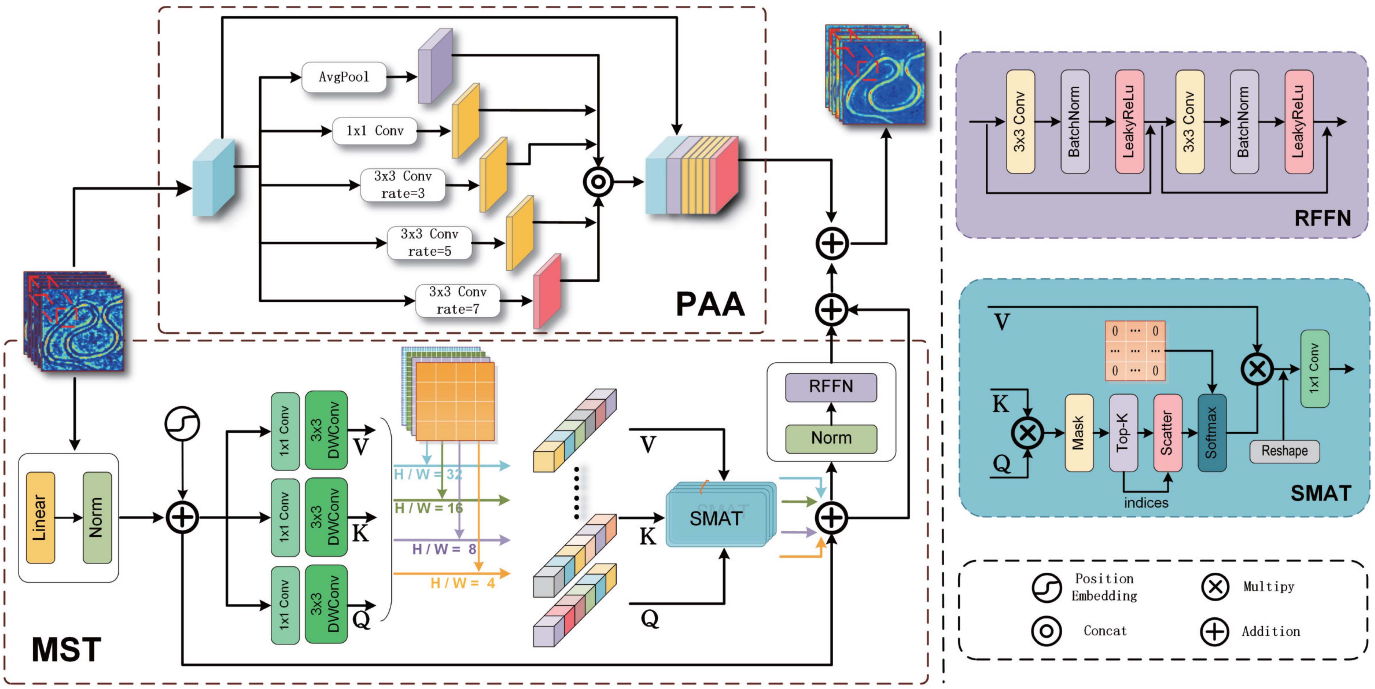


Fig. 3. Structure of ADSE, including the PAA and MST.

compute correlation scores. Finally, all the results are summed and enhanced using convolution, batch normalization, and ReLU operations to improve the feature extraction capability of remote sensing images and enhance information accuracy. Specifically, to maximize pixel information, we partition the image f_{stem} into $s \times s$ patches, where s takes values of 4, 8, 16, and 32. As a result, the original image is resized to a size of $\frac{HW}{s^2}$ with $s \times s$ and then flattened into a vector $v_i \in \mathbb{R}^{s^2 \times c}$. Subsequently, a linear projection is utilized to transform each patch vector into the embedding $e_i \in \mathbb{R}^c$ that encodes the patch representation. Subsequently, the patches and positional encodings are fed into the encoder to obtain the output.

Furthermore, we introduce a sparse attention mechanism to reduce noise and additional computational overhead caused by irrelevant information. The sparse attention also aims to improve foreground-background discrimination and alleviate blurriness in the foreground edge regions, addressing the issue of blurriness in the foreground edge regions caused by the naive ViT's attention computation for all pixels. Inspired by [86], we conduct a sparse multihead attention (SMAT) and apply self-attention across channels instead of spatial dimensions to reduce time and memory complexity. We compute the similarity between pairs of reshaped queries and keys, considering only the k most similar pixel values, which leads to a more concentrated foreground and more discriminative foreground edge regions. This can also find an approximate match for a particular region or object in the image. We then normalize the k largest pixels in each row of the similarity matrix using softmax, setting other elements to zero, as derived below

$$SparseAtt(Q, K, V) = Softmax\left(T_k\left(\frac{QK^T}{\sqrt{d_k}}\right)\right)V \quad (1)$$

where $T_k(\cdot)$ is the learnable top-k selection operator. Finally, the similarity matrix is multiplied with the values matrix to obtain the final result. Here, k is an adjustable parameter that dynamically controls the level of sparsity. It is obtained through a weighted average of certain fractions, and we set it to [1/2, 3/4]. This dynamic selection enables attention to transition from dense to sparse, as derived below

$$[T_k(P)]_{i,j} = \begin{cases} P_{i,j}, & \mathrm{P}_{i,j} \in top - k(rowj) \\ 0, & \text{other points.} \end{cases} \quad (2)$$

The prepared information is fed into residual FFN (RFFN) to complete the composition, achieving feature enhancement and extraction through linear projection. This reduces information loss in the sequence and effectively enhances semantic modeling capability. The aforementioned procedure can be delineated as follows:

$$f_{\text{MST}} = \text{RFFN}\left(\text{SMAT}\left(Norm\left(Linear\left(f_{\text{stem}}\right)\right)\right)\right) \quad (3)$$

where f_{MST} is the output of the MST, SMAT$(\cdot)$ is the SMAT, and RFFN$(\cdot)$ is the RFFN.

To assist in the preliminary extraction of multilevel information by the Transformer, we design a PAA. While transformers may discount local information extraction, we believe that incorporating additional information supervision can guide the learning process. On the other hand, although conventional convolutions extract more detailed information, they come at the cost of enormous computational complexity, contradicting our initial goal of achieving a balance between accuracy and real-time performance. Therefore, in this study, we leverage the power of pyramid hollow convolutions with four different receptive field sizes ($r = 1, 3, 5, 7$). To further enhance performance, we introduce a 1×1 convolutional layer for feature smoothing.

Algorithm 1: Inference Details of Proposed ADSTNet.

```
Input: Optical RSI $I \in \mathbb{R}^{3\times256\times256}$, Mutil-Scale Patches
  $mp$ (32, 16, 8, 4) for MST
Output: Salmap $S \in \mathbb{R}^{1\times256\times256}$
  $//Step1$: Information coding in Adaptive Interaction
  Encoder
  $f_{stem} \Leftarrow Conv_{\varsigma}$
  $f_{le} = f_{ce}^{4}, f_{ce}^{t} \Leftarrow \mathcal{F}_{CE}(f_{stem})$, $t = 1$ to 4
  $f_{PAA} \Leftarrow \mathcal{F}_{PAA}(f_{stem})$
  while H=W in mp do
    while $k$ in (1/2, 2/3, 3/4) do
      $f_{MST} \Leftarrow \mathcal{F}_{RFFN}(\mathcal{F}_{SMST}(f_{stem}, attn(k)))$
    end while
    $f_{MST} = \sum_{\mathrm{i}=1}^{4} f_{MST}$
  end while
  $f_{ge} = f_{ADSE} \Leftarrow f_{PAA} + f_{MST}$
  $f_{e} \Leftarrow \mathcal{F}_{MLI}(f_{le}, f_{ge})$
  $//Step2$: The boundary Extraction base-on stage 1 and
  stage 4 in CNN Encoder
  $f_{DFR} = \mathcal{F}_{DFR}(f_{ce}^{1}, f_{ce}^{4})$
  $//Step3$: Complementary fusion of multiple contents
  in AFCD
  if t=4 then
    $//f_{AFCD}^{4} \Leftarrow \mathcal{F}_{AFCD}(f_{ce}^{4}, f_{DFR}, f_{e})$
  else
    $//f_{AFCD}^{t} \Leftarrow \mathcal{F}_{AFCD}(f_{ce}^{t}, f_{DFR}, f_{AFCD}^{t+1})$
  end if
  Output: $S \Leftarrow f_{AFCD}^{1}$
  Return $S$
```

Then, we refine the features by summing up all the results and eliminating noise in the output. Finally, we combine the obtained features with the original information to facilitate information propagation. This process can be expressed as follows:

$$f_{\text{PAA}} = conv_{1\times1}(concat(f_{\text{stem}}, conv_{r=1}, conv_{r=3}, conv_{r=5}, conv_{r=7})) + f_{\text{stem}} \tag{4}$$

where f_{PAA} is the output of the PAA, $conv_{1\times1}$ is a 1×1 convolutional layer, and $conv_{r=1}$ is a convolution with the receptive field of 1.

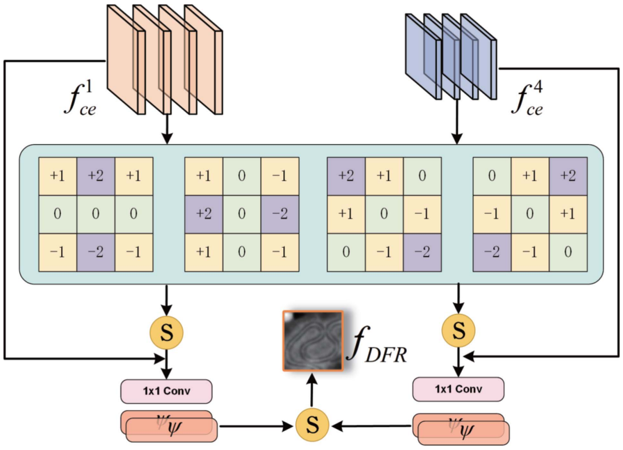


Fig. 4. Illustration of the DFR.

C. Directional Feature Reconfiguration

To enhance the representation of boundary information, we propose a DFR, as illustrated in Fig. 4. It has been observed that salient images in remote sensing exhibit topological structures where objects such as buildings, ships, and airplanes are often arranged in a cluttered manner, deviating from horizontal or vertical orientations. Therefore, in addition to conventional horizontal and vertical boundary detection computations, we consider inclined boundaries to have significant influence on salient objects. Drawing inspiration from the utilization of the Sobel operator in traditional image processing, we design a dedicated gradient-based boundary detection operator to extract boundary information in four directions: 0°, 45°, 90°, and 135°. Specifically, we construct four convolution kernels of size 3 × 3 with fixed parameters and apply convolution operations with a stride of 1. These aforementioned four convolutions are defined as follows:

$$K_x = \begin{bmatrix} 1 & 0 & -1 \\ 2 & 0 & -2 \\ 1 & 0 & -1 \end{bmatrix}, K_m = \begin{bmatrix} 0 & 1 & 2 \\ -1 & 0 & 1 \\ -2 & -1 & 0 \end{bmatrix}$$

$$K_y = \begin{bmatrix} 1 & 2 & 1 \\ 0 & 0 & 0 \\ -1 & -2 & -1 \end{bmatrix}, K_n = \begin{bmatrix} 2 & 1 & 0 \\ 1 & 0 & -1 \\ 0 & -1 & -2 \end{bmatrix} \tag{5}$$

where, K_x, K_m, K_y, and K_n represent specialized operators for feature extraction along the horizontal, inclined at 45°, vertical, and inclined at 135° directions, respectively. The DFR is employed to obtain the boundary gradient maps by applying it to the output features of stage 1 and stage 4. The output of stage 1 exhibits finer features compared to the stem's output, devoid of rough information interference, while the features from stage 4 encompass rich semantic information of the overall image [64]. Subsequently, we apply four basic convolution groups to the input features to obtain the gradient map G_{xymn}^{t} $(t = 1, 4)$. Then, the features are smoothed using a 1 × 1 convolution and normalized through the sigmoid function to attenuate noise. Finally, the boundary-enhanced feature map is obtained by integrating the normalized features with the input features. The aforementioned procedure can be delineated as follows:

$$G_{xymn}^{t} = \sum_{\mu} f_{ce}^{t} \odot K_{\mu}, \mu = x, y, m, n; t = 1, 4 \tag{6}$$

$$f_{ce}^{t\,\prime} = f_{ce}^{t} \odot \sigma\left(conv_{1\times1}(G_{xymn}^{t})\right) \tag{7}$$

where $\odot$ denotes element-wise multiplication, σ represents the sigmoid operation. The G_{xymn}^{t} is obtained by applying a specialized boundary detection operator on f_{ce}^{t}, where G_{x}, G_{y}, G_{m}, and G_{n} are concatenated along the channel dimension. The boundary information obtained from f_{ce}^{t} is represented by the variable $f_{ce}^{t\,\prime}$. In particular, our initial step involves the application of a 1 × 1 convolution coupled with bilinear upsampling to the product of stage 4, thereby facilitating feature alignment commensurate

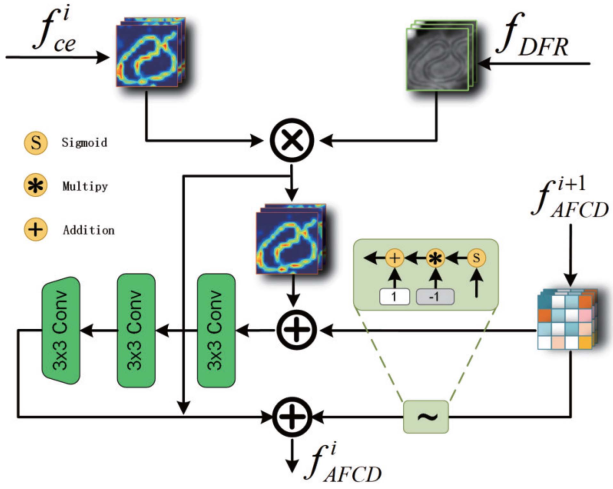


Fig. 5. Architecture of AFCD.

with the dimensions of stage 1. Subsequently, we utilize distinct 1×1 convolutions to ascertain uniformity in the channel dimensions of the two feature maps under consideration. This is succeeded by the implementation of a pair of convolution layers to derive the ultimate feature map. The aforementioned procedure can be delineated as follows:

$$f_{\mathrm{DFR}} = \sigma\left[\sum_{t}^{t=1,4} \psi\left(\psi\left(\mathrm{Conv}_{1\times 1}\left(f_{ce}^{t}{}'\right)\right)\right)\right] \tag{8}$$

where the variable ψ represents a convolution group consisting of 3 × 3 Conv, BatchNorm, and ReLU. f_{DFR} denotes the output of DFR. To mitigate the influence of internal edge noise, we utilize the boundary information generated from the ground truth (GT) saliency map as a supervisory signal, disregarding the interference from internal boundary information. In addition, we employ a weight compensation mechanism to enhance the supervision on boundary information, better serving the compensation of information in the decoder.

D. Adaptive Feature Cascade Decoder

The boundary features obtained from DFR are utilized as a valuable source of prior knowledge to boost the image representation capability of the encoder. We propose the AFCD as illustrated in Fig. 5. The integration of boundary features enables AFCD to employ a cascaded structure that enhances the representation of both foreground and background features. This, in turn, facilitates the complementary fusion of multiple contents within the image. Specifically, AFCD consists of three inputs: 1) the prior boundary knowledge extracted by DFR; 2) the multiscale features from the encoder; and 3) the features from the upper-level AFCD. Within AFCD, three separate pathways are implemented, each dedicated to strengthening the feature representation in edges, foreground, and background. With regard to the boundary information, a fusion of the prior knowledge and encoder features is performed to acquire boundary-enhanced information. This process can be succinctly expressed as follows:

$$f_{\mathrm{edge}} = f_{\mathrm{DFR}} \times f_{ce}^{t} \tag{9}$$

where f_{edge} represents the output from the fusion boundary and CNN coded information. For foreground information, f_{edge} is aligned and fused with features f_{AFCD} from the previous AFCD to strengthen the representation. Specifically, f_{AFCD} is adjusted in scale using bilinear interpolation, followed by SA [87] and CA [70] attention mechanisms and three convolutional layers for cascaded fusion, enabling complementary fusion of multiple contents and enhancing information representation. Simultaneously, the output f_{AFCD} from the previous decoder is reshaped and passed through a sigmoid function to obtain background information. Subsequently, three convolutional layers with batch normalization and ReLU activation are applied to obtain optimized features. This process can be described as follows:

$$f_{\mathrm{fg}} = \vartheta(\vartheta(\vartheta(\mathrm{CA}(\mathrm{SA}(f_{\mathrm{AFCD}}))))) \tag{10}$$

$$f_{\mathrm{bg}} = -1 * sigmoid(x) + 1 \tag{11}$$

where the terms spatial attention and channel attention are represented by the acronyms SA and CA, respectively. ϑ denoted the 3 × 3 Conv, and f_{fg} and f_{bg} denote the feature of foreground and background, respectively.

Finally, the aforementioned results are summed together to obtain the final output f_{AFCD}, which includes foreground, background, edges, and features from the previous decoder.

E. Loss Function

Given that ADSTNet is a multitask model, addressing both interior and boundary segmentation, we introduce a comprehensive loss function to simultaneously optimize these two tasks. Moreover, a weight compensation mechanism is incorporated to facilitate effective feature learning. The definition of interior segmentation loss involves a weighting of both the cross-entropy loss ($\mathcal{L}_{\mathrm{CE}}$) and the mean intersection-over-union loss ($\mathcal{L}_{\mathrm{mIoU}}$). This combination is expressed mathematically as follows:

$$\mathcal{L}_{\mathrm{CE}} = -\frac{1}{N}\sum_{i=1}^{N}\left(\tilde{G}_i \log(\tilde{S}_i) + (1-\tilde{G}_i)\log(1-\tilde{S}_i)\right) \tag{12}$$

$$\mathcal{L}_{\mathrm{mIoU}} = 1 - \frac{\sum_{i=1}^{N}(\tilde{G}_i * \tilde{S}_i)}{\sum_{i=1}^{N}(\tilde{G}_i + \tilde{S}_i - \tilde{G}_i * \tilde{S}_i)} \tag{13}$$

where $\tilde{G}_i$ and $\tilde{S}_i$ denote the GT and the predicted label for the ith pixel in image, respectively, and the total number of pixels in the image is denoted by N. Due to class imbalance between foreground and background pixels in boundary detection, the training effectiveness of our model on highly imbalanced datasets is enhanced by employing the Dice Loss. The Dice Loss ($\mathcal{L}_{\mathrm{Dice}}$) is expressed as follows:

$$\mathcal{L}_{\mathrm{Dice}} = 1 - \frac{2\sum_{i=1}^{N}(\tilde{G}_i * \tilde{S}_i)}{\sum_{i=1}^{N}(\tilde{G}_i + \tilde{S}_i)}. \tag{14}$$

In summary, our designed overall loss consists of the salmap loss ($\mathcal{L}_{\mathrm{sal}}$) and the boundary loss ($\mathcal{L}_{\mathrm{bnd}}$). It is crucial to note that, with respect to the boundary detection loss, only the predictions generated by the DFR, which reconstructs directional features, are taken into account. On the other hand, for the primary image salmap loss, a deep supervision strategy is adopted to obtain predictions from decoder features at different levels. As a result, the total loss ($\mathcal{L}$) is expressed as follows:

$$\mathcal{L} = \mathcal{L}_{\text{sal}} + \mathcal{L}_{\text{bnd}} = \sum_{i}^{D} \left(\mathcal{L}_{\text{CE}} + \mathcal{L}_{\text{mIoU}}\right) + \gamma \mathcal{L}_{\text{Dice}} \tag{15}$$

where the weight factor γ is introduced, with γ value of three chosen to enhance the auxiliary role of the boundary and achieve dynamic balance, and the D is the number of AFCDs. A series of ablation experiments were carried out to investigate the optimal value of parameter γ.

IV. Experiments

In this section, extensive experiments were conducted to evaluate the proposed ADSTNet. The datasets, evaluation criteria, and experimental settings are described in Section IV-A. A comprehensive comparison between our proposed model and all competing methods is presented in Section IV-B. Section IV-C provides ablation studies and related discussions. Lastly, an analysis of failure cases is performed.

A. Datasets and Implementation Details

1) Datasets: Our model was comprehensively evaluated on three datasets, namely ORSSD [16], EORSSD [17], and ORSI-4199 [65], to demonstrate its superiority. These datasets were annotated and provided convenience during model training. ORSSD is the first publicly available dataset designed for investigating saliency detection performance, consisting of 600 training and 200 testing images. The dataset encompasses a wide variety of object scales, types, and backgrounds. EORSSD serves as a supplement to ORSSD, enhancing the diversity and complexity of the dataset with 1400 training and 600 testing images. ORSI-4199 is a more challenging dataset compared to ORSSD and EORSSD, comprising 2000 training and 2199 testing images. To better train the model, we applied data augmentation techniques inspired by methods such as [44], [88]. Specifically, we employed methods like mirror flipping and rotations at 90°, 180°, 270°, resulting in an augmented dataset of 4200, 9800, and 14000 images for ORSSD, EORSSD, and ORSI-4199, respectively.

2) Evaluation Criteria: To objectively evaluate the performance of all models, we employed eight quantitative analysis metrics, namely S-measure (S_α, $\alpha = 0.5$) [89], max F-measure (F_β^{max}, β^2 to 0.3) [90], mean F-measure (F_β^{mean}), adaptive F-measure (F_β^{adp}), max E-measure ($E_\xi^{\max}$) [91], mean E-measure (E_ξ^{mean}), adaptive E-measure (E_ξ^{adp}), and MAE ($\mathcal{M}$). Among these metrics, a smaller $\mathcal{M}$ value is preferred, while larger values are desirable for the other seven metrics. In addition, we utilized two qualitative indicators, namely the F-measure curve and the precision-recall (PR) curve, to visually illustrate the variations among the models through the tool [92]. A model with a curve approaching 1 in the F-measure curve indicates superior performance, and likewise, a model's curve approaching (1, 1) in the PR curve represents optimal performance.

3) Implementation Details: To maximize the performance of our model, we utilized Res2Net [56] as the initial weight for the backbone and resized each image to 256×256 as input. The evaluation of our model was conducted on a NVIDIA TITAN RTX GPU. We employed the Adam [93] optimizer with a learning rate of 1e-4 and a batch size of 8 on PyTorch [94]. The model was trained for 50 epochs, and the learning rate was reduced to 0.1 every 30 epochs. To prevent exploding gradients during training, we implemented gradient clipping with a maximum norm of 0.5 using the clip gradient function of the optimizer. The performance of the resultant models in terms of saliency was subsequently assessed using test sets in the ORSSD [17], EORSSD [16], and ORSI-4199 [65].

B. Comparison With SOTA Methods

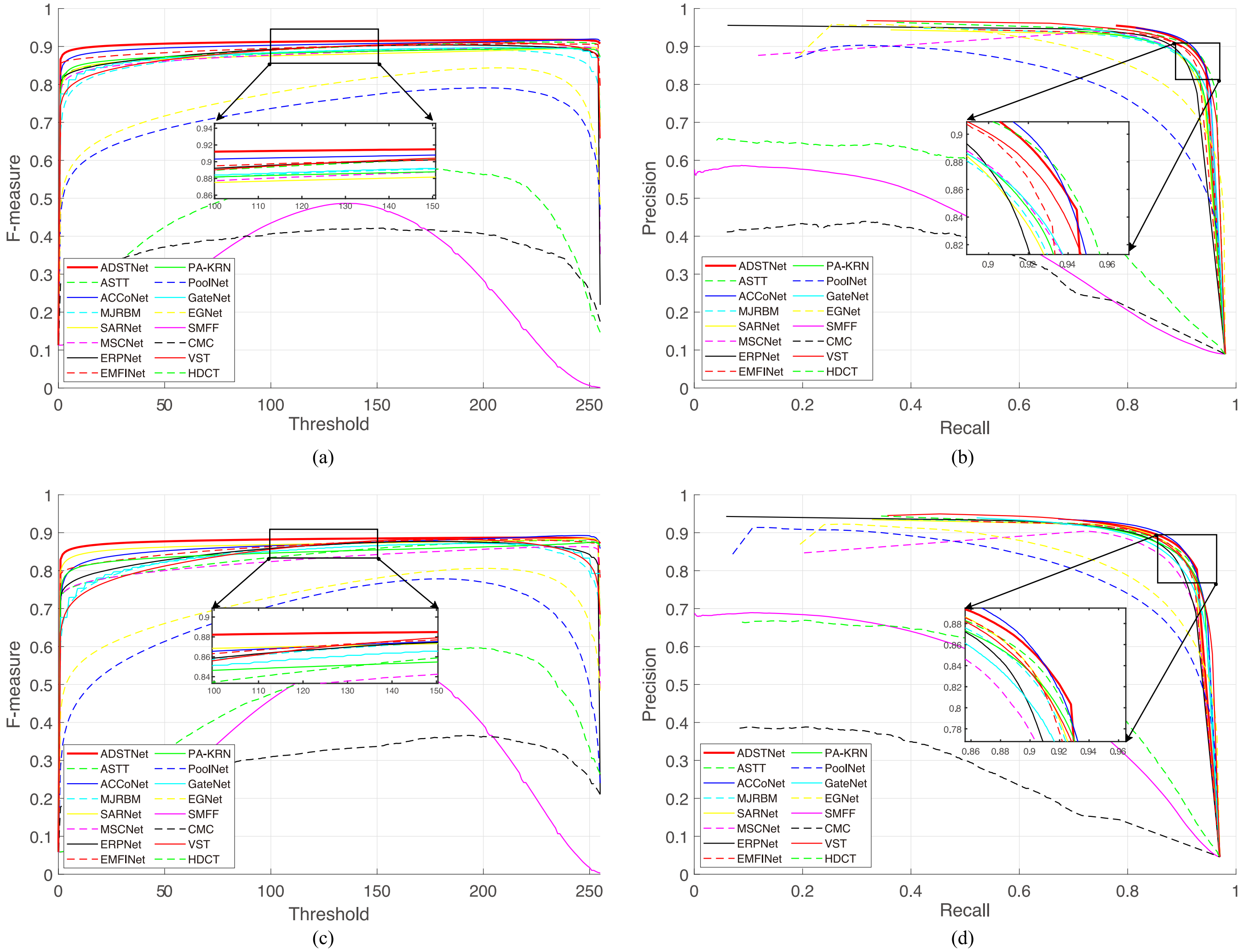
In this study, we propose ADSTNet and compare it with 26 SOTA networks on the EORSSD and ORSSD datasets to demonstrate its superiority. The evaluation involves both qualitative and quantitative analyses. These methods can be classified into five traditional NSI-SOD (RRWR [95], HDCT [96], DSG [97], RCRR [98], VST [50]), seven CNN-based NSI-SOD (DSS [60], RADF [99], EGNet [100], PoolNet [101], GateNet [102], SUCA [103], PA-KRN [104]), three traditional ORSI-SOD (VOS [26], CMC [40], SMFF [39]), and eleven CNN-based ORSI-SOD (LVNet [16], DAFNet [17], EMFINet [64], ERPNet [47], ACCoNet [44], MSCNet [105], SARNet [106], CorrNet [48], FSMINet [107], MJRBM [65], ASTT [51]). Furthermore, among the aforementioned comparative models, we select eleven representative models for further evaluation on the ORSI-4199 dataset to assess their robust generalization. To ensure fair competition, all methods adopt the same training and testing sample settings. In addition, consistent parameter settings and environments are maintained during the evaluation of predicted maps. The predicted maps of all comparative models are generated using the authors' provided code.

1) Quantitative Comparison: The quantitative comparison results between ADSTNet and 26 other models on the EORSSD and ORSSD can be found in Table I. It is evident that ADSTNet demonstrates superior or competitive performance compared to other SOTA methods across all benchmark datasets. On the EORSSD, although our model falls behind ACCoNet in terms of $F_\beta^{\max}$, it exhibits significant advantages in other metrics. Particularly noteworthy is the 0.0563 lead of ADSTNet over ACCoNet in F_β^{adp}. Similarly, when compared to DAFNet, which excels in $E_\xi^{\max}$, our model shows room for improvement in $\mathcal{M}$ but achieves a comprehensive victory in other metrics, such as a 0.2105 improvement in F_β^{adp} and a 0.1235 improvement in E_ξ^{adp}. Furthermore, when comparing ADSTNet to VST and ASTT, both utilizing the transformer framework, our model slightly lags behind ASTT in $\mathcal{M}$ but emerges as the leader in other aspects, which is forgivable. Likewise, on the ORSSD, our model consistently maintains a top-three position in all comparative results. Specifically, compared to the second-best performing ACCoNet, our proposed model exhibits a marginal difference of 0.0058 in S_α, but compensates significantly by leading with 0.0173 in F_β^{adp}. In addition, we present the comparative results of all methods on the PR curve and F-measure curve in Fig. 6. It is evident that our proposed model achieves

TABLE I
QUANTITATIVE RESULTS ON EORSSD AND ORSSD

Methods	Publication	Type	Speed (fps) ↑	#Params (M) ↓	FLOPs (G) ↓	EORSSD								ORSSD							
						S_α ↑	F_β^{max} ↑	F_β^{mean} ↑	F_β^{adp} ↑	E_ξ^{max} ↑	E_ξ^{mean} ↑	E_ξ^{adp} ↑	$\mathcal{M}$ ↓	S_α ↑	F_β^{max} ↑	F_β^{mean} ↑	F_β^{adp} ↑	E_ξ^{max} ↑	E_ξ^{mean} ↑	E_ξ^{adp} ↑	$\mathcal{M}$ ↓
RRWR [97]	2015 CVPR	TN	0.3	–	–	.5992	.3993	.3686	.3344	.6894	.5943	.5639	.1677	.6835	.5590	.5125	.4874	.7649	.7017	.6949	.1324
HDCT [98]	2015 TIP	TN	7	–	–	.5971	.5407	.4018	.2658	.7861	.6376	.5192	.1088	.6197	.5257	.4235	.3722	.7719	.6495	.6291	.1309
DSG [99]	2017 TIP	TN	0.6	–	–	.6420	.5232	.4597	.4012	.7260	.6954	.6188	.1246	.7195	.6238	.5747	.5657	.7912	.7337	.7532	.1041
RCRR [100]	2017 TIP	TN	0.3	–	–	.6007	.3995	.3685	.3347	.6882	.5946	.5636	.1644	.6849	.5591	.5126	.4876	.7651	.7021	.6950	.1277
VST [50]	2021 ICCV	TN	23	44.03	23.2	.9203	.8724	.8257	.7015	.9736	.9417	.8872	.0067	.9318	.8999	.8710	.8126	.9755	.9596	.9417	.0100
DSS [60]	2017 CVPR	CN	8	62.23	114.6	.7868	.6849	.5801	.4597	.9186	.7631	.6933	.0186	.8262	.7467	.6962	.6206	.8860	.8362	.8085	.0363
RADF [101]	2018 AAAI	CN	7	62.54	214.2	.8179	.7446	.6582	.4933	.9140	.8567	.7162	.0168	.8259	.7619	.6856	.5730	.9130	.8298	.7678	.0382
EGNet [102]	2019 ICCV	CN	9	108.07	291.9	.8601	.7880	.6967	.5379	.9570	.8775	.7566	.0110	.8721	.8332	.7500	.6452	.9731	.9013	.8226	.0216
PoolNet [103]	2019 CVPR	CN	25	53.63	123.4	.8207	.7545	.6406	.4611	.9292	.8193	.6836	.0210	.8403	.7706	.6999	.6166	.9343	.8650	.8124	.0358
GateNet [104]	2020 ECCV	CN	25	100.02	108.3	.9114	.8566	.8228	.7109	.9610	.9385	.8909	.0095	.9186	.8871	.8679	.8229	.9664	.9538	.9428	.0137
SUCA [105]	2020 TMM	CN	24	117.71	56.4	.8998	.8229	.7949	.7260	.9520	.9277	.9082	.0097	.8989	.8484	.8237	.7748	.9584	.9400	.9194	.0145
PA-KRN [106]	2021 AAAI	CN	16	141.06	617.7	.9192	.8639	.8358	.7993	.9619	.9536	.9416	.0104	.9239	.8890	.8727	.8548	.9680	.9620	.9579	.0139
VOS [26]	2018 GRSL	TR	–	–	–	.5082	.2765	.2107	.1836	.5982	.4886	.4767	.2096	.5366	.3471	.2717	.2633	.6514	.5352	.5826	.2151
CMC [40]	2019 RS	TR	–	–	–	.5798	.3268	.2692	.2007	.6803	.5894	.4890	.1057	.6033	.3913	.3454	.3108	.7064	.6417	.5996	.1267
SMFF [39]	2019 IJRS	TR	–	–	–	.5401	.5176	.2992	.2083	.7744	.5197	.5014	.1434	.5312	.4417	.2684	.2496	.7402	.4920	.5676	.1854
LVNet [16]	2019 TGRS	CR	–	–	–	.8630	.7794	.7328	.6284	.9254	.8801	.8445	.0146	.8815	.8263	.7995	.7506	.9456	.9259	.9195	.0207
SARNet [108]	2021 RS	CR	47	25.91	129.7	.9240	.8719	.8541	.8304	.9620	.9555	.9536	.0099	.9134	.8850	.8619	.8512	.9557	.9477	.9464	.0187
DAFNet [17]	2022 TIP	CR	26	29.35	68.5	.9166	.8614	.7845	.6427	.9861	.9291	.8446	.0060	.9191	.8928	.8511	.7876	.9771	.9539	.9360	.0113
EMFINet [64]	2022 TGRS	CR	25	107.26	480.9	.9290	.8720	.8486	.7984	.9711	.9604	.9501	.0084	.9366	.9002	.8856	.8617	.9737	.9671	.9663	.0109
ERPNet [47]	2022 TCYB	CR	50	56.48	87.2	.9210	.8632	.8304	.7554	.9603	.9401	.9228	.0089	.9254	.8974	.8745	.8356	.9710	.9566	.9520	.0135
MSCNet [107]	2022 ICPR	CR	55	3.29	23.38	.9071	.8539	.8151	.7553	.9689	.9551	.9329	.0090	.9227	.8927	.8676	.8350	.9754	.9653	.9584	.0129
FSMINet [110]	2022 GRSL	CR	28	3.6	75.9	.9255	.8678	.8436	.8015	.9666	.9567	.9490	.0079	.9361	.9041	.8878	.8710	.9759	.9672	.9693	.0101
MJRBM [65]	2022 TGRS	CR	32	43.54	95.7	.9197	.8656	.8239	.7066	.9646	.9350	.8897	.0099	.9204	.8842	.8556	.8022	.9623	.9415	.9328	.0163
CorrNet [109]	2023 TGRS	CR	100	4.09	21.1	.9289	.8778	.8620	.8311	.9696	.9646	.9593	.0083	.9380	.9129	.9002	.8875	.9790	.9746	.9721	.0098
ACCoNet [44]	2023 TCYB	CR	81	102.55	368.6	.9290	.8837	.8552	.7969	.9727	.9653	.9450	.0074	.9437	.9149	.8971	.8806	.9796	.9754	.9721	.0088
ASTT [51]	2023 TGRS	CR	13	43.12	23.4	.9253	.8741	.8297	.7534	.9757	.9584	.9247	.0059	.9348	.9061	.8811	.8456	.9795	.9699	.9587	.0094
Ours	–	CR	39.5	62.09	27.72	.9311	.8804	.8716	.8532	.9769	.9709	.9681	.0065	.9379	.9124	.9042	.8979	.9807	.9740	.9785	.0086

The 26 SOTA methods can be classified into five traditional NSI-SOD (TN), seven CNN-based NSI-SOD (CN), three traditional ORSI-SOD (TR), and eleven CNN-based ORSI-SOD (CR). ↑/↓ means a larger/smaller score is better. "–" symbol indicates that the results are not available. The top three results are highlighted in red, blue, and green, respectively.



(a)

(b)

(c)

(d)

Fig. 6. F-measure and PR curves of our method compared with other representative methods: (a) F-measure curves on ORSSD, (b) PR curves on ORSSD, (c) F-measure curves on EORSSD, and (d) PR curves on EORSSD.

TABLE II
ELEVEN REPRESENTATIVE MODELS FOR FURTHER EVALUATION ON THE ORSI-4199 DATASET

Methods	ORSI-4199							
	$S_\alpha \uparrow$	$F_\beta^{\max} \uparrow$	$F_\beta^{\text{mean}} \uparrow$	$F_\beta^{\text{adp}} \uparrow$	$E_\xi^{\max} \uparrow$	$E_\xi^{\text{mean}} \uparrow$	$E_\xi^{\text{adp}} \uparrow$	$\mathcal{M} \downarrow$
EGNet	.8464	.8267	.8041	.7650	.9161	.8947	.8620	.0440
PoolNet	.8271	.8010	.7779	.7382	.8964	.8676	.8531	.0541
GateNet	.8680	.8626	.8414	.7946	.9369	.9199	.8816	.0357
SUCA	**.8794**	**.8692**	**.8590**	.8415	**.9438**	**.9356**	**.9186**	**.0304**
PA-KRN	.8491	.8415	.8324	.8200	.9280	.9168	.9063	.0382
DAFNet	.8552	.8458	.8261	.7819	.9220	.9007	.8905	.0396
EMFINet	.8675	.8584	.8479	.8186	.9340	**.9257**	.9136	.0330
ERPNet	.8670	.8553	.8374	.8024	.9290	.9149	.9024	.0357
MJRBM	.8593	.8493	.8309	.7995	.9311	.9102	.8891	.0374
CorrNet	.8623	.8560	.8513	**.8534**	.9330	.9206	.9142	.0366
ACCoNet	**.8761**	**.8657**	**.8591**	**.8562**	**.9398**	**.9297**	**.9163**	**.0314**
Ours	**.8710**	**.8698**	**.8653**	**.8655**	**.9433**	**.9356**	**.9212**	**.0318**

These methods can be classified into five CNN-based NSI-SOD (CN) and six CNN-based ORSI-SOD (CR). ↑/↓ means a larger/smaller score is better. The top three results are highlighted in red, blue, and green, respectively.

satisfactory performance on the F-measure curve, both on the EORSSD and ORSSD, with curves closer to 1 compared to other models. On the PR curve, ADSTNet demonstrates a competitive performance with other comparative methods, but exhibits a surpassing trend, approaching the (1, 1) coordinate point in the later stages. Particularly on the EORSSD, ADSTNet stands out among its peers, and the gap between ADSTNet and the first-ranked ASTT on the ORSSD is also minimal.

ADSTNet consistently demonstrates excellent performance on the ORSI-4199, mirroring its impressive results on the EORSSD and ORSSD, as shown in Table II. Our model achieves top rankings in five out of eight metrics, secures the second position in one metric, and attains the third position in two metrics. Specifically, compared to the highly competitive SUCA method, our model achieves parity in E_ξ^{mean} and outperforms it by 0.0240 in F_β^{adp}, albeit with a slight lag of 0.0084 in S_α. Moreover, our method remains competitive across all metrics. For instance, in terms of $F_\beta^{\max}$, we achieve 0.8698 (ours) compared to 0.8560 (CorrNet), for $E_\xi^{\max}$, we attain 0.9433 (ours) compared to 0.9369 (GateNet), and for $\mathcal{M}$, we obtain 0.0318 (Ours) compared to 0.0357 (ERPNet). It is noteworthy that our model is the sole approach to surpass the threshold of 0.92 in E_ξ^{adp}.

2) Computational Complexity Comparison: The computational complexity of our method was evaluated based on three perspectives, which encompassed inference speed (I/O time excluded), network parameters, and FLOPs. Table I reports on the data acquired from the publicly available ORSI-SOD benchmark [43], [62], as well as our own retraining efforts. Upon evaluation, we discovered that the majority of CNN-based techniques could perform in real-time (at a rate of 25–30 *fps*). In contrast, our method excels with an impressive inference speed of 39.5 *fps* on the ORSSD, EORSSD, and ORSI-4199. The parameters of network are also at a middle level. However, compared with the second-ranked ACCoNet, our network has much better parameters, such as Params: 62.09 M (ours) versus 102.55 M (ACCoNet). Meanwhile, significant progress was made in the FLOPs competition, with only 6.62 G separating it from CorrNet, which ranks first on the FLOPs leaderboard. This achievement is noteworthy in the overall comparison. Compared with the third-ranked SUCA in ORSI-4199, our model still has a significant advantage in speed, parameters, and FLOPs, with an improvement of 15.5, 55.62, and 29.68, respectively. Based on the quantitative and computational complexity comparisons above, it can be inferred that our method is both highly competitive in the field.

3) Qualitative Comparison: As illustrated in Fig. 7, we present representative ORSI-SOD methods for each category, along with their corresponding timelines. Five different scenarios are compared, including morphologically regular buildings, elongated rivers with topographic structures, low-light conditions on complex shorelines, multiple tiny objects in complex backgrounds, and more challenging scenes.

Our model demonstrates satisfactory results in all five showcased scenarios. For the first scenario, our model excels in capturing detailed information in local regions, effectively highlighting salient objects compared to other models. However, ASTT and MSCNet show deficiencies in accurately localizing the overall contours of small buildings and suffer from misidentifications in certain instances (1st and 2nd instance). Our model, on the other hand, distinguishes itself by accurately discriminating salient regions of the building and providing well-defined boundaries, which is a key advantage over other models (3rd instance).

In the second scenario, which involves elongated rivers with irregular topographic structures and varying background colors, most compared models exhibit detection incompleteness and fail to capture global information or accurately represent the true width of the rivers. This deficiency can hinder subsequent processes in practical applications, as seen in GateNet and DSG. In contrast, our model overcomes these challenges, producing saliency maps that closely resemble the GT, with clear and distinct boundary information.

In the third scenario, which features rich boundary information, some methods suffer from blurred boundaries and detection omissions due to the interference of objects and surrounding scenes. Examples include ERPNet, CMC, PoolNet, PA-KRN, and HDCT. In addition, in the third example, where islands have elongated extensions, a challenging aspect, some methods such as ASTT overlook the importance of this easily neglected information. In comparison, our proposed method excels by achieving highly accurate detection, surpassing these obstacles.

The fourth scenario involves the detection of multiple tiny objects, a well-known challenge in remote sensing saliency detection where objects may be missed due to their small size. While some methods successfully detect all tiny objects, they inevitably make errors in boundary details. ASTT and MSCNet exhibit imperfect boundary information and produce blurry detections. In addition, ERPNet and PoolNet have omissions in detecting small vehicles. In contrast, our model showcases notable regional information, accurately capturing small objects in local regions of remote sensing images.

Lastly, the fifth scenario combines challenges from the previous four scenarios, incorporating various shapes of tiny objects and complex background noise. In traditional natural scene and remote sensing methods, such as CMC, HDCT, and DSG, there are instances where color-salient background regions are erroneously identified as salient objects, deviating significantly from

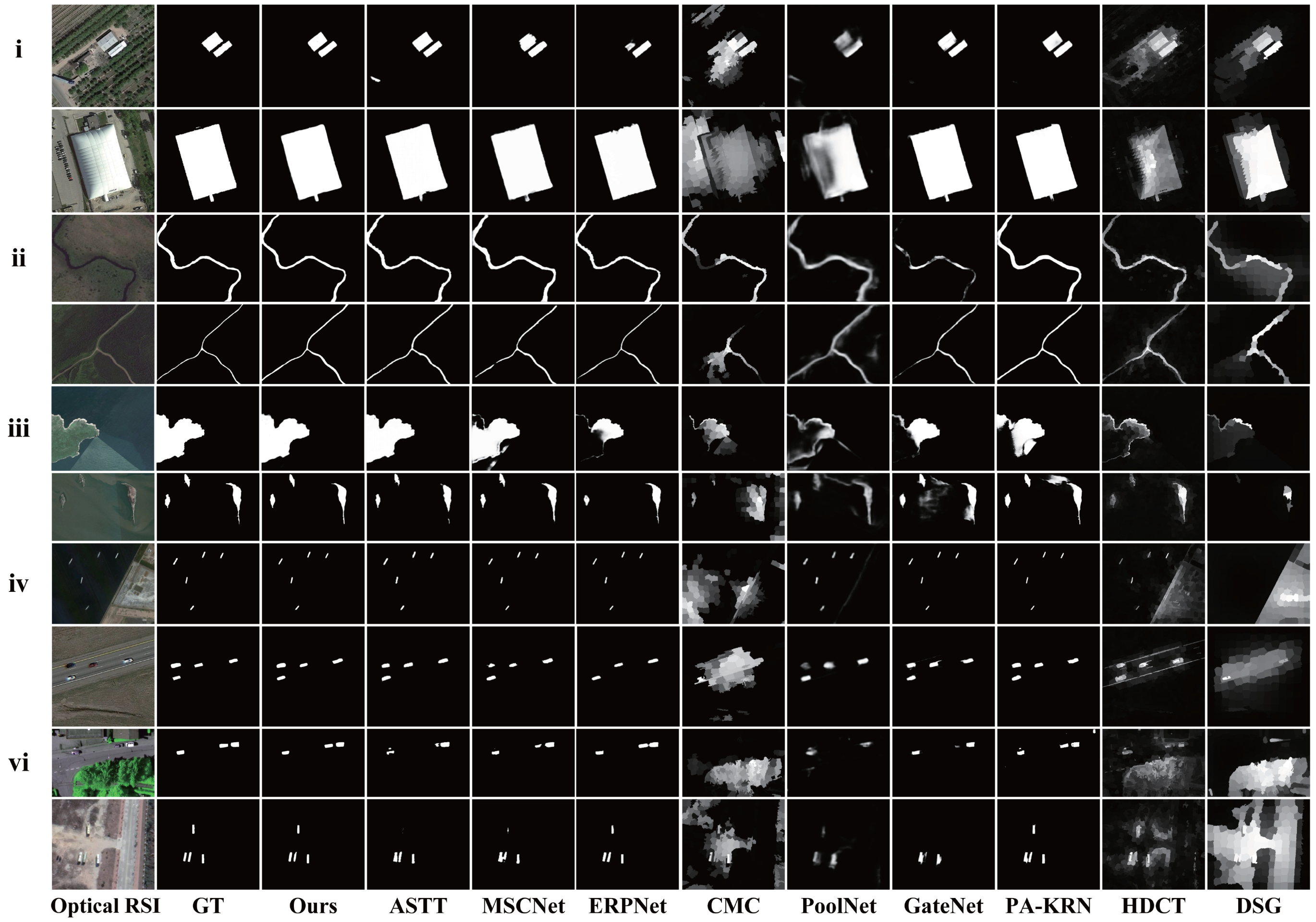


Fig. 7. Qualitative comparison with nine representative SOTA methods, including three CNN-based ORSI-SOD methods (ASTT, MSCNet, ERPNet), three CNN-based NSI-SOD methods (PoolNet, GateNet, PA-KRN), one traditional RSI-SOD method (CMC), and two traditional NSI-SOD methods (HDCT, DSG) on various scenes.

our expectations. Moreover, CNN-based methods are prone to mistakenly classifying object parts as background due to color similarities with the surrounding environment. These methods also encounter difficulties in detecting adjacent objects, leading to adhesive detections, as observed in MSCNet, PoolNet, and PA-KRN. In contrast, our model minimizes the influence of complex interference factors and achieves satisfactory results.

Overall, the proposed model exhibits high detection accuracy for multiple tiny objects, low-light interference, and topographic structures. It excels in boundary delineation and outperforms other methods in capturing global information.

TABLE III
COMPARE WITH SOTA METHODS OF SEMANTIC SEGMENTATION, INCLUDING DEEPLABV3+, HRNET, POINTREND, SEGMENTER, SEGNEXT, AND SAN

Method	Publication	EORSSD				ORSSD			
		$S_\alpha \uparrow$	$F_\beta^{\max} \uparrow$	$E_\xi^{\max} \uparrow$	$\mathcal{M} \downarrow$	$S_\alpha \uparrow$	$F_\beta^{\max} \uparrow$	$E_\xi^{\max} \uparrow$	$\mathcal{M} \downarrow$
DeepLabV3+	2018 CVPR	.8667	.7725	.9158	.0124	.8981	.8532	.9522	.0163
HRNet	2019 CVPR	.8818	.7971	.9340	.0094	.8809	.8410	.9274	.0173
PointRend	2020 CVPR	.8710	.7722	.9290	.0107	.8961	.8526	.9453	.0171
Segmenter	2021 ICCV	.8405	.7391	.8767	.0152	.8770	.8324	.9159	.0191
SegNeXt	2022 NeurIPS	.8769	.7858	.9505	.0078	.9142	.8670	.9730	.0102
SAN	2023 CVPR	.9019	.8342	.9470	.0069	.9176	.8877	.9624	.0096
ADSTNet	-	**.9311**	**.8804**	**.9769**	**.0065**	**.9379**	**.9124**	**.9807**	**.0086**

The best result in each column is given in bold.

4) Compare With SOTA Methods of Semantic Segmentation: To better illustrate the difference of ORSI-SOD compared to semantic segmentation models, we retrained six SOTA semantic segmentation models on the EORSSD and ORSSD using the authors' specified parameters. These models include DeepLabV3+ [30], HRNet [31], PointRend [32], Segmenter [33], SegNeXt [34], and SAN [35]. As shown in Table III, although these six models demonstrated commendable performance on remote sensing datasets, a notable performance gap still exists when compared to our proposed ADSTNet. Specifically, among the semantic segmentation models, SAN achieved the best results, with S_α scores of 0.9019 and 0.9176 on the two datasets, respectively. However, when compared to ADSTNet, a significant disparity remains, with differences of 0.0292 and 0.0203 on the EORSSD and ORSSD, respectively. Notably, on the ORSSD, ADSTNet surpassed the 0.009 bottleneck on the $\mathcal{M}$, achieving a 55% gain over the Segmenter. In conclusion, the dissimilarities in object localization between the two tasks hinder the effective alignment of semantic segmentation models with salient object detection tasks. Future efforts will focus on synergizing these tasks to design a universal model that can unlock greater value.

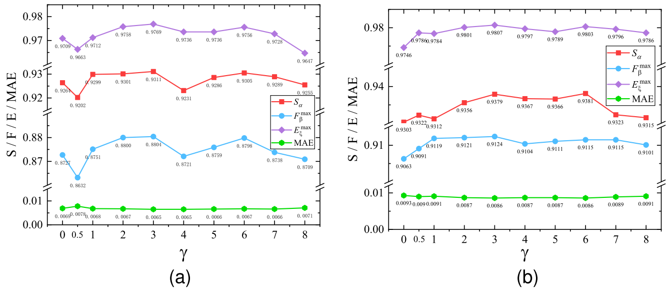


Fig. 8. Performance trends produced with different values of the hyperparameter γ. The curves of S_α, $F_\beta^{\max}$, $E_\xi^{\max}$, and $\mathcal{M}$ on (a) EORSSD and (b) ORSSD.

C. Ablation Studies and Related Discussions

To evaluate the effectiveness and indispensability of our proposed global–local–boundary information compensation scheme and key modules, we conducted extensive ablation experiments on the EORSSD and ORSSD. To ensure fairness in the experiments, each variant was retrained under the consistent experimental settings described in Section IV-A.

1) Loss Ablation: The hyperparameter γ is incorporated into the coarse prediction branch, specifically subitem $\mathcal{L}_{\text{Dice}}$, of the overall loss function ($\mathcal{L}$). In contrast to the conventional loss weight parameters in [65], [51], and [64], we set γ to be greater than 1, aligning with our initial intention of constructing boundary detection. While mimicking the DFR for edge extraction, we inadvertently detected internal boundaries within objects, which deviated from our expectations. Therefore, we introduced a weight compensation mechanism to amplify the importance of boundary loss, forcing the DFR to primarily learn external boundary information and disregard internal fine details. In our investigation of γ ranging from 0 to 8, we observed that a value of 3 yielded the optimal performance, as shown in Fig. 8. In addition, we conducted an additional experiment with γ set to 0.5, and it was evident that its performance on ORSSD followed the overall upward trend, while exhibiting a dramatic anomaly on EORSSD. This indirectly confirms that γ does not belong to this range. Thus, setting γ to a value greater than 1 is deemed necessary.

2) Verification Process of the Individual Modules: In order to evaluate the effectiveness of the individual modules proposed, we defined a simple U-shaped baseline network with Res2Net [56] as the encoder and a decoder consisting of three consecutive convolutional layers for conducting ablation experiments. The verification process of ADSE, DFR, and AFCD was performed by employing a controlled variable method, where only one component was modified at a time, ensuring strict experimental operations. The numerical results for different combinations of these modules are presented in Table IV. In addition, visualizations are provided in Fig. 9.

TABLE IV
ABLATION STUDY ON EVALUATING THE INDIVIDUAL CONTRIBUTION OF EACH CONTENT IN ADSTNET

No.	Baseline	DFR	ADSE	AFCD	EORSSD				ORSSD			
					S_α ↑	$F_\beta^{\max}$ ↑	$E_\xi^{\max}$ ↑	$\mathcal{M}$ ↓	S_α ↑	$F_\beta^{\max}$ ↑	$E_\xi^{\max}$ ↑	$\mathcal{M}$ ↓
1	✓				.9218	.8646	.9667	.0074	.9277	.8981	.9724	.0110
2	✓	✓			.9271	.8691	.9676	.0072	.9353	.9009	.9713	.0108
3	✓		✓		.9286	.8781	*.9735*	.0066	*.9374*	.9077	.9735	.0096
4	✓			✓	.9240	.8754	.9684	*.0067*	.9350	.9103	.9753	*.0095*
5	✓	✓	✓		.9280	.8721	.9708	.0068	**.9380**	*.9113*	.9810	**.0086**
6	✓	✓		✓	**.9315**	*.8783*	.9759	.0068	.9368	.9099	.9762	*.0095*
7	✓		✓	✓	*.9295*	**.8819**	.9759	.0070	.9379	.9118	**.9817**	.0088
8	✓	✓	✓	✓	.9311	.8804	**.9769**	**.0065**	.9379	**.9124**	*.9807*	**.0086**

Baseline is simple U-shaped baseline network. The DFR, ADSE, AFCD mean directional feature reconfiguration, adaptive dual-stream sparse encoder, and adaptive feature cascade decoder. Best results are given in bold, second results are indicated by using underscores, and third results are indicated by slanting.

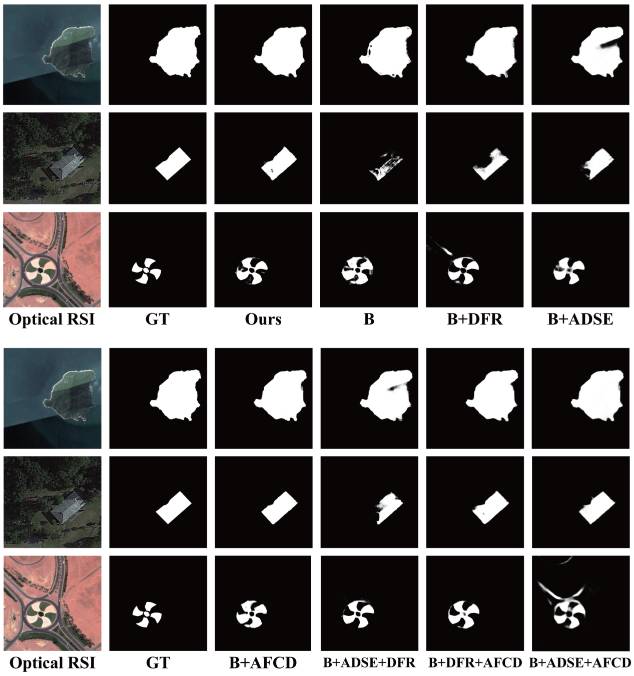


Fig. 9. Visual comparison results of individual contribution of Baseline (B), DFR, ADSE and AFCD.

Our proposed model demonstrates comprehensive optimization from both quantitative and qualitative perspectives. In quantitative comparisons, as shown in Table IV, the baseline network achieved only 0.9277 in S_α, 0.8981 in $F_\beta^{\max}$, 0.9724 in $E_\xi^{\max}$, and 0.0110 in $\mathcal{M}$ on ORSSD. It is evident that the addition of ADSE and DFR significantly improved performance. Specifically, S_α increased to 0.9353 and 0.9374, showing an improvement of 0.0076 and 0.0097 compared to the baseline, respectively. While the improvements within individual components may appear marginal, the synergistic effects among these components yield a discernible enhancement. Specifically, on the EORSSD, the incorporation of ADSE alone results in a modest increase of only 0.003 in the $F_\beta^{\max}$ compared to scenarios without it (No.2 and No.5). However, the additional inclusion of both ADSE and AFCD surpasses the sole presence of DFR, exhibiting a substantial improvement of 0.0113 in the $F_\beta^{\max}$. Similarly, on the ORSSD, although the introduction of DFR alone leads to a marginal improvement of only 0.0002 in $\mathcal{M}$, the collaborative action of all components boosts $\mathcal{M}$ by 0.0024, a commendable achievement. We also conducted an experiment wherein we eliminated the ADSE while simultaneously introducing DFR and AFCD (i.e., No.6) in Table IV. This ablation of components translates to using only CNN as the encoding part. It is evident that, with the removal of ADSE's encoding information, and thus utilizing only locally acquired information, there is a notable

TABLE V
EFFECTIVE CONTRIBUTION OF COMPONENTS IN ADSE, INCLUDING MST, PAA, AND SUPERVISION (SV)

No.	Varient			EORSSD				ORSSD			
	MST	PAA	SV	$S_\alpha \uparrow$	$F_\beta^{\max} \uparrow$	$E_\xi^{\max} \uparrow$	$\mathcal{M} \downarrow$	$S_\alpha \uparrow$	$F_\beta^{\max} \uparrow$	$E_\xi^{\max} \uparrow$	$\mathcal{M} \downarrow$
1	✓			.9300	.8782	$\underline{.9758}$	*.0067*	.9324	*.9116*	.9796	**.0086**
2		✓		$\underline{.9309}$	.8638	.9650	.0069	.9314	.9083	.9793	.0104
3	✓	✓		.9261	.8676	.9698	**.0064**	.9339	.9108	.9779	$\underline{.0088}$
4	✓		✓	.9301	*.8799*	.9751	.0073	$\underline{.9365}$	**.9163**	$\underline{.9804}$	*.0097*
5		✓	✓	*.9304*	$\underline{.8801}$	*.9752*	.0071	*.9362*	.9113	*.9798*	.0103
6	✓	✓	✓	**.9311**	**.8804**	**.9769**	$\underline{.0065}$	**.9379**	$\underline{.9124}$	**.9807**	**.0086**

Best results are given in bold, second results are indicated by using underscores, and third results are indicated by slanting.

reduction in $\mathcal{M}$. For instance, on EORSSD, it decreased from 0.0068 (No.5) to 0.0065 (No.6), and on ORSSD, it decreased from 0.0095 (No.5) to 0.0086 (No.6). Simultaneously, under the AFCD-based condition, combining the global information obtained by ADSE with CNN encoding information resulted in a notable enhancement of 0.0065 on EORSSD for $F_\beta^{\max}$. Furthermore, ADSTNet, which incorporates all components, achieved comprehensive superiority. When comparing ADSTNet to individual components for DFR, ADSE, and AFCD, this amalgamation also achieved superior results on $F_\beta^{\max}$ in ORSSD: 0.9713 (No.2) versus 0.9735 (No.3) versus 0.9753 (No.4) versus 0.9807 (No.8). These findings underscore the pronounced synergy among the proposed components, demonstrating that their collective impact exerts a more substantial influence on network performance.

From a qualitative standpoint, the model incorporating all components demonstrated superiority over individual elements in object saliency, as illustrated in Fig. 9. This aspect better complements the limitations of achieving limited performance. In the first example, the component with DFR achieved more accurate boundary delineation compared to the one without it. With the assistance of ADSE, the model captured more comprehensive global information, as observed in the second example. For irregular objects, especially those with sharp edges, such as the third example, the combined effect of boundary and global information resulted in a clearer overall object contour. In particular, the synergistic pairing of DFR with AFCD produces saliency maps characterized by enhanced boundary clarity, surpassing the clarity achieved with individual components. When solely relying on ADSE and AFCD without the incorporation of DFR, inaccuracies arise in outlining object boundaries. This nuanced observation underscores the formidable contributions of each component and highlights the substantial potential for performance enhancement through the collaborative interplay among these components.

3) Effectiveness of Each Component in ADSE: To validate the role of the transformer in guided learning, ablation studies were conducted on each component of ADSE. As depicted in Table V, although PAA can contribute to local information acquisition to some extent, its impact is limited due to the information loss during the dilation convolution process. The addition of PAA resulted in only marginal performance improvements in $\mathcal{M}$ for EORSSD (0.0003) and S_α for ORSSD (0.0015), while exhibiting weaker performance in other metrics. However, the deficiency of PAA was effectively compensated under the supervision of saliency maps. Notably, this compensation in the significant improvements of $F_\beta^{\max}$ by 0.0011 on EORSSD and S_α by 0.0055 on ORSSD. In the absence of the comprehensive global information provided by MST, the network's performance experiences varying degrees of decline as shown in Table V. For instance, on EORSSD, $F_\beta^{\max}$ exhibits a decrease from 0.8638 (No. 2) to 0.8676 (No. 3), and on ORSSD, $\mathcal{M}$ shows a reduction from 0.0104 (No. 2) to 0.0088 (No. 3). Applying implicit supervision to the global information extracted by MST leads to performance enhancement, notably exemplified by a 0.0099 improvement in $E_\xi^{\max}$ on EORSSD. Conversely, when MST is omitted alone (i.e., No.5), in contrast to the comprehensive ADSTNet with all components (i.e., No.6), there is a decrement of 0.0017 in $\mathcal{M}$ on ORSSD. These research findings underscore the collaborative contribution of each component within ADSE. Furthermore, introducing salient map supervision in the early stages proves to be a judicious and effective strategy, encouraging ADSE to assimilate more valuable information. In addition, the role of each component was further demonstrated through visualization, which can be found in Fig. 10. The constructive impact of MST on global information capture is evident; nevertheless, it falters in accurately localizing crucial objects. Simultaneously, PAA excels in delineating local object features. The introduction of MST and PAA enhances the model's capacity to capture comprehensive information. Specifically, the collaborative action of MST and PAA efficiently redirects attention from broad global contexts to essential objects, facilitating precise localization across all positions. The collective contribution of these components surpasses that of individual elements. However, the sparse nature

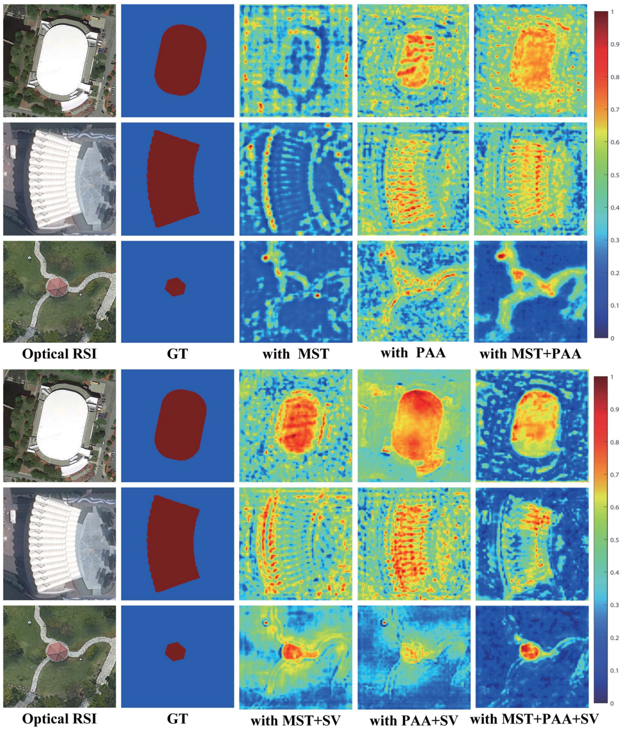
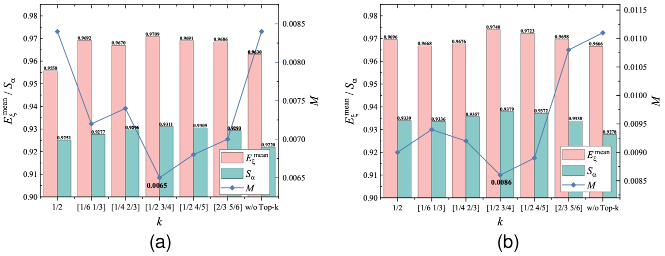


Fig. 10. Effective visualization of components in ADSE. Best viewed in color.



(a)

(b)

Fig. 11. Ablation analysis for the implication of the number k in SMAT on (a) EORSSD and (b) ORSSD.

of components imposes constraints on detailed information. Implicit supervision is introduced to address this limitation. Under this guidance, MST and PAA are compelled to refine and acquire more precise information features, yielding results more satisfactory than in previous scenarios. Various visual cues affirm that the proposed components significantly enhance the overall network performance, aligning with prior data analyses and reinforcing the critical importance of each component in cooperation.

4) Implication of the Number K*:* In the SMAT architecture we proposed, the choice of parameter k significantly influences overall performance, as illustrated in Fig. 11. Our findings underscore the critical significance of optimally choosing k to regulate boundary sparsity. Specifically, when k assumes a particular value, such as $k = 1/2$, or when not considering k (w/o Top-K), the overall network demonstrates heightened sensitivity to k, resulting in unsatisfactory performance. Consequently, we introduced a dynamic range to ascertain the optimal value of k, leading to performance surpassing that of using a singular fixed value. Throughout our experiments, we observed that a too-small k prevented the network from capturing comprehensive information, causing a notable performance decline. Conversely, when k was excessively large, the network incorporated irrelevant information and noise, placing a burden on performance. Through meticulous adjustments, we determined that optimal overall performance is achieved when k falls within the range [1/2, 3/4]. Within this range, the network demonstrated performance for 0.0086 and 0.0065 on $\mathcal{M}$ and superior results of 0.9709 and 0.9740 on E_ξ^{mean}. These experiments validate the efficacy of dynamically selecting k, empowering the network to adapt more flexibly to various scenarios and datasets, thereby achieving a more robust and efficient optical remote sensing salient object detection performance.

5) Different Directions Role in the DFR: The DFR ablation study was conducted to further demonstrate the necessity of capturing boundary information in different directions. We split the directions into conventional horizontal and vertical orientations, as well as oblique orientations of 45° and 135° while ensuring the completeness of information pairs. As shown in Table VI, without the assistance of boundary information in the $K_x K_y$ and $K_m K_n$ directions (i.e., No.1), the overall detection performance suffered varying degrees of degradation. Specifically, compared to the complete DFR configuration, the absence of $K_x K_y$ boundary information led to a reduction of 0.0037 in $F_\beta^{\max}$ for EORSSD and 0.0094 for ORSSD. Similarly, $E_\xi^{\max}$ decreased by 0.0026 and 0.0084, respectively. However, there was still room for improvement compared to the variant that had $K_x K_y$ but lacked $K_m K_n$. Specifically, there was minimal difference in performance on EORSSD, but a notable gap of 0.0020 in S_α and 0.0034 in $F_\beta^{\max}$ on ORSSD. This indirectly confirms our observation that objects in remote sensing images are not strictly oriented horizontally. We further substantiated the criticality of our proposed components through qualitative comparisons. In Fig. 12, we present three instances and annotate the conventional indications of different detection directions. It is evident that information from any of these directions contributes to the model's performance enhancement. In the first instance, the extraction of boundary information at 90° allows for a complete depiction of the object's overall contour in horizontal orientations. In addition, the boundary and contour information obtained under the joint effect of multidirectional boundary detection is superior to that obtained under the aforementioned single condition, with particular emphasis on the third scenario. It has been validated that proposed DFR can significantly contribute to the detection performance of the overall model.

TABLE VI
NECESSITY OF CAPTURING BOUNDARY INFORMATION IN DIFFERENT DIRECTIONS

No.	Variant	EORSSD				ORSSD			
		$S_\alpha \uparrow$	$F_\beta^{\max} \uparrow$	$E_\xi^{\max} \uparrow$	$\mathcal{M} \downarrow$	$S_\alpha \uparrow$	$F_\beta^{\max} \uparrow$	$E_\xi^{\max} \uparrow$	$\mathcal{M} \downarrow$
1	*None*	.9214	.8706	.9664	.0068	.9334	.9029	.9704	.0102
2	w/o $K_x K_y$	.9309	.8767	.9743	**.0064**	.9359	.9030	.9723	.0098
3	w/o $K_m K_n$	.9295	.8741	.9742	**.0064**	**.9399**	.9090	.9765	.0087
4	Ours	**.9311**	**.8804**	**.9769**	.0065	.9379	**.9124**	**.9807**	**.0086**

"None" means without $K_x K_y$ and $K_m K_n$. The best result in each column is given in bold.

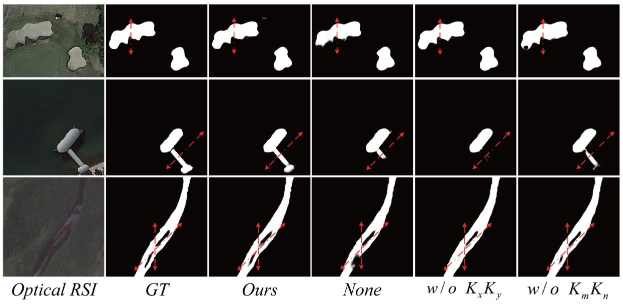


Fig. 12. Visual comparison results of the capturing boundary information in different directions by DFR. Best viewed in color.

6) Impact of DFR Information Sources: In further validating the impact of DFR information sources on the aggregation of boundary information, a series of experiments was conducted to explore optimal boundary information sources. As shown in Table VII, information from stages 1 to 4 was combined using various methods. The results indicate that under the condition of only two information sources, performance is superior when either f_{ce}^1 or f_{ce}^4 is essential, compared to scenarios where this information is lacking, such as No1, No2, No3, and No4. Specifically, when based on f_{ce}^3, the performance with f_{ce}^1 is that S_α was improved by 0.0045 on EORSSD and ORSSD. Similarly, when based on f_{ce}^4, F_β^{mean} is increased by 0.0210 and 0.0093 on the two datasets, respectively. However, when three information sources

TABLE VII
ABLATION ANALYSIS FOR MULTIPLE COMBINATIONS OF INFORMATION FROM ENCODERS

No.	f_{ce}^1	f_{ce}^2	f_{ce}^3	f_{ce}^4	EORSSD				ORSSD			
					$S_\alpha \uparrow$	$F_\beta^{mean} \uparrow$	$E_\xi^{mean} \uparrow$	$\mathcal{M} \downarrow$	$S_\alpha \uparrow$	$F_\beta^{mean} \uparrow$	$E_\xi^{mean} \uparrow$	$\mathcal{M} \downarrow$
1	✓		✓		$\underline{.9297}$	*.8624*	$\underline{.9665}$	*.0077*	**.9445**	*.9027*	.9746	.0098
2	✓			✓	**.9311**	**.8716**	**.9709**	**.0065**	.9379	**.9042**	.9740	**.0086**
3		✓	✓		.9252	.8516	.9635	.0080	.9400	.8896	.9737	.0102
4		✓		✓	*.9259*	.8506	.9645	.0079	.9216	.8949	.9715	.0111
5	✓	✓	✓		.9057	.8476	.9645	.0078	.9320	.8945	.9713	.0096
6	✓	✓		✓	.9225	.8404	*.9652*	$\underline{.0074}$	$\underline{.9444}$	.9009	*.9756*	*.0095*
7		✓	✓	✓	.9245	$\underline{.8591}$	.9584	.0099	*.9434*	$\underline{.9033}$	**.9761**	$\underline{.0088}$
8	✓	✓	✓	✓	.9194	.8337	.9623	.0082	.9407	.8806	$\underline{.9759}$	.0097

The best results are given in bold, second results are indicated by using underscores, and third results are indicated by slanting.

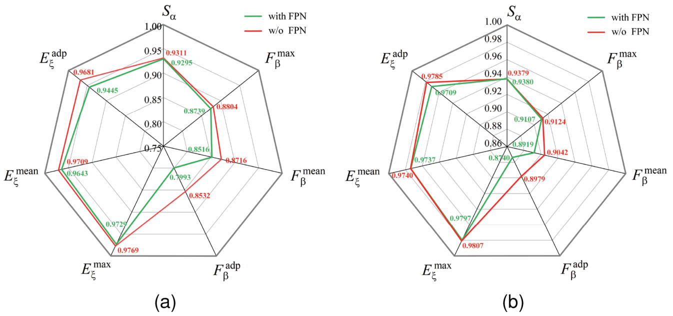


Fig. 13. Ablation analysis for the information fusion of f_{ce}^1 and f_{ce}^4 by the neck like FPN. (a) EORSSD. (b) ORSSD.

TABLE VIII
ABLATION ANALYSIS FOR THE TRANSPLANTABILITY PROPERTIES OF THE DFR

Method	EORSSD				ORSSD				#Params(M) $\downarrow$
	$S_\alpha \uparrow$	$F_\beta^{mean} \uparrow$	$E_\xi^{mean} \uparrow$	$\mathcal{M} \downarrow$	$S_\alpha \uparrow$	$F_\beta^{mean} \uparrow$	$E_\xi^{mean} \uparrow$	$\mathcal{M} \downarrow$	
MJRBM	.9197	.8239	.9350	.0099	.9204	.8566	.9415	.0163	**63.52**
+DFR	**.9266**	**.8328**	**.9468**	**.0089**	**.9237**	**.8629**	**.9552**	**.0148**	63.67(*+0.15*)
ERPNet	.9210	.8304	.9401	.0089	.9254	.8745	.9566	.0135	**77.19**
+DFR	**.9255**	**.8383**	**.9485**	**.0086**	**.9281**	**.8792**	**.9590**	**.0129**	77.84(*+0.64*)
EMFINet	.9290	.8486	.9604	.0084	.9366	.8856	.9671	.0109	**95.09**
+DFR	**.9326**	**.8546**	**.9651**	**.0080**	**.9401**	**.8928**	**.9769**	**.0099**	95.81(*+0.72*)

The best result in each column is given in bold.

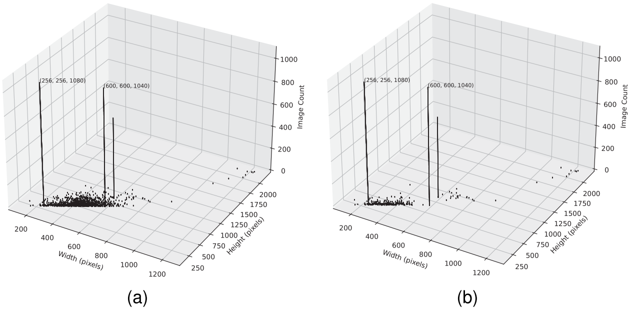


Fig. 14. Spatial analysis of image pixels chart on (a) EORSSD and (b) ORSSD. Please zoom in for better view.

TABLE IX
ABLATION ANALYSIS FOR IMAGE SIZE

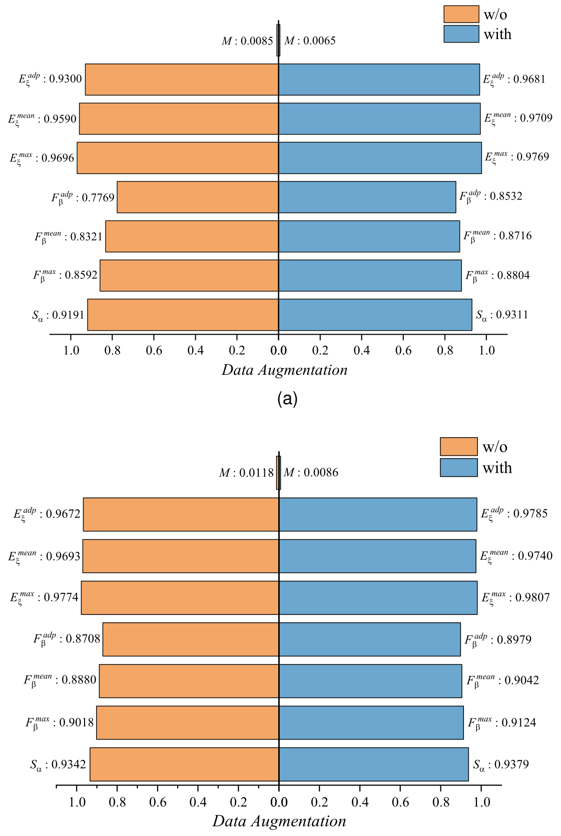
No.	Image_Size	EORSSD							
		$S_\alpha \uparrow$	$F_\beta^{max} \uparrow$	$F_\beta^{mean} \uparrow$	$F_\beta^{adp} \uparrow$	$E_\xi^{max} \uparrow$	$E_\xi^{mean} \uparrow$	$E_\xi^{adp} \uparrow$	$\mathcal{M} \downarrow$
1	224×224	*.9280*	.8751	*.8628*	*.8281*	*.9751*	.9640	*.9588*	.0080
2	256×256	$\underline{.9311}$	*.8804*	**.8716**	**.8532**	$\underline{.9769}$	$\underline{.9709}$	**.9681**	**.0065**
3	288×288	**.9328**	$\underline{.8825}$	$\underline{.8675}$	$\underline{.8376}$	**.9772**	**.9713**	$\underline{.9633}$	$\underline{.0069}$
4	600×600	$\underline{.9311}$	**.8846**	.8625	.8271	.9733	*.9670*	.9540	*.0074*

Best results are given in bold, second results are indicated by using underscores, and third results are indicated by slanting.

are used as DFR inputs, performance did not meet expectations, showing varying degrees of decline compared to the input with two information sources. For instance, on EORSSD, 0.9635 (No.3) versus 0.9584 (No.7) for E_ξ^{mean}; on ORSSD, 0.0086 (No.2) versus 0.0095 (No.6) for $\mathcal{M}$. Aggregating all information sources (f_{ce}^1, f_{ce}^2, f_{ce}^3, f_{ce}^4) for boundary information did not yield satisfactory performance in all ablation experiments. This further affirms that not all boundary information in the encoder is suitable for aggregating texture features, and the effective combination of boundary information from lower levels and region information from higher levels significantly enhances network performance. In addition, exploration of combining f_{ce}^1 and f_{ce}^4 in a manner similar to FPN [71] to determine overall network efficiency was conducted, as shown in Fig. 13. It is evident that model performs better on benchmark datasets without using FPN, achieving improvements of 0.02 and 0.0123 on F_β^{mean} based on EORSSD and ORSSD, respectively. In the next steps of our research, we will continue to explore the optimal boundary information interaction mode suitable for this network to ensure smooth information transfer.

7) Influence of DFR With Plug-and-Play: To further validate the portability of the proposed DFR module, we selected three network architectures from the ORSI-SOD method that simultaneously incorporate boundary information and region features, owing to the accessibility of their source code. The results are presented in Table VIII, demonstrating that DFR exhibits favorable plug-and-play characteristics. Transferring DFR to MJRBM, ERPNet, and EMFINet led to performance improvements. On the EORSSD, the F_β^{mean} increased by 0.0089, 0.0079, and 0.0060, respectively. On the ORSSD, the E_ξ^{mean} showed enhancements of 0.0137, 0.0024, and 0.0058. Notably, with the support of DFR, EMFINet surpassed the 0.01 threshold for $\mathcal{M}$ on the ORSSD, establishing itself as a frontrunner. In addition, we presented the computational overhead introduced by transplanting the DFR module. While maintaining the original architecture, the parameter count increased by less than 1 M for all architectures. MJRBM experienced a modest increase of 0.15 M, attributed to the architectural similarity between ADSTNet and MJRBM, facilitating the seamless transplantation of the module. In conclusion, DFR exhibits plug-and-play characteristics and can contribute to performance enhancements in other SOTA networks.

8) Effect of Image Size: To further investigate the impact of remote sensing image dimensions on saliency detection models, we conducted extensive experiments to explore the optimal training dimensions, aiming for enhanced model robustness. Initially, we analyzed the distribution of image sizes in the EORSSD and ORSSD, as illustrated in Fig. 14. Given that EORSSD is an expansion of the ORSSD, increasing from the original 800 images to 2000, the overall distribution is similar, with 256×256 being the most common size (1080 sheets), followed by 600×600 (1040 sheets). Taking this as a starting point, following [48], [51], we conducted experiments on sizes neighboring 256×256, namely 224×224 and 288×288 as shown in Table IX. It is evident that the performance of the 224×224 size is the poorest, while 256×256 and 288×288 show similar performance, with 256×256 exhibiting slightly better results. Achieving four top positions and three second



(a)

(b)

Fig. 15. Bar chart depicting comparative experiments with and without data augmentation on (a) EORSSD and (b) ORSSD.

positions in the experiments. Subsequently, experiments were conducted on the 600×600, revealing that the increase in image size did not lead to a significant performance improvement. This is because the upsampling of the majority of images to 600×600 introduces noninherent information, which can be considered as noise and interferes with image information representation. Larger images transformed into smaller ones might lose information but retain intrinsic details, resulting in less interference with the information representation during training compared to the noise introduced by upsampling. This also explains why the performance of 288×288 is slightly worse than 255×255. In summary, redefining the image size as 256×256 optimally unleashes the network's potential, demonstrating satisfactory robustness.

9) Effect of Date Augmentation: To meticulously unravel the ramifications of data augmentation on ADSTNet's performance, a meticulously designed series of experiments was executed, aiming to rigorously validate the efficacy of this augmentation strategy. In our experimental framework, ADSTNet underwent its initial training phase on the pristine dataset, with performance metrics meticulously documented and visually depicted by the distinctive yellow bar in Fig. 15. Subsequently, an advanced phase ensued where data augmentation techniques were adroitly applied to amplify the breadth of the training set, facilitating a comprehensive model evolution, vividly represented by the discernible blue bar. A meticulous comparative analysis between the two model iterations brought to light a discernible augmentation-induced surge in ADSTNet's prowess for salient object detection tasks. The augmented ADSTNet showcased notable performance strides, manifesting as increments of 0.0763 and 0.0271 for F_β^{adp} on the EORSSD and ORSSD, respectively. Furthermore, commendable enhancements of 0.0119 and 0.0047 surfaced in E_ξ^{mean} across these datasets, adding layers of robustness and reliability to the model's repertoire. In summary, the judicious application of data augmentation emerged as a pivotal catalyst, facilitating the assimilation of a more nuanced understanding of diverse object features into the model.

TABLE X
Performance Evaluation Across Various Encoder Backbones in ADSTNet

Method	EORSSD				ORSSD			
	$S_\alpha \uparrow$	$F_\beta^{\text{mean}} \uparrow$	$E_\xi^{\text{mean}} \uparrow$	$\mathcal{M} \downarrow$	$S_\alpha \uparrow$	$F_\beta^{\text{mean}} \uparrow$	$E_\xi^{\text{mean}} \uparrow$	$\mathcal{M} \downarrow$
ADSTNet-VGG	.9283	.8582	.9647	.0075	.9330	.8879	.9685	.0102
ADSTNet-ResNet	**.9317**	.8630	.9678	.0075	**.9419**	.9013	**.9761**	.0087
ADSTNet-Res2Net	.9311	**.8716**	**.9709**	**.0065**	.9379	**.9042**	.9740	**.0086**

The best result in each column is given in bold.

10) Flexibility of Our Approach: To substantiate the efficacy of our proposed method across diverse backbone networks, two variants, namely ADSTNet-VGG and ADSTNet-ResNet, were introduced. These variants utilized VGG [67] and ResNet [68] as encoding backbone networks, respectively, and underwent validation on the EORSSD and ORSSD. As presented in Table X, when compared to our initial method employing the Res2Net backbone (ADSTNet-Res2Net), these two variants exhibited slightly inferior performance, indirectly indicating the superior feature encoding capabilities of Res2Net over ResNet and VGG. Particularly noteworthy is that, while ADSTNet-ResNet demonstrated striking similarities in performance representation to ADSTNet-Res2Net, the comprehensive evaluation favored ADSTNet-Res2Net, securing the top position in all five evaluated metrics. This underscores the advantages of Res2Net. In summary, ADSTNet showcases robust adaptability to different backbone networks, adept at leveraging information obtained from various encoders and manifesting its intrinsic capabilities.

D. Analysis of Failure Samples

As stated previously, the aim of this article is twofold: first, to propose a novel framework for ORSI-SOD that effectively combines global and local information extraction, and second, to enhance the contribution of features at different levels throughout the entire image and improve the delineation of salient object contours through multidirectional boundary assistance so as to achieve accurate localization. However, ADSTNet still faces certain limitations when confronted with challenging scenarios as shown in Fig. 16. For instance, due to inherent biases, our model struggles to differentiate highly camouflaged noise information. In the first instance, the target object bears a striking resemblance to the surrounding distractors, and the connecting bridge seamlessly blends with the background color, resulting in minimal discernible changes. Although our model successfully detects the objects, it also exhibits instances of both false positives and

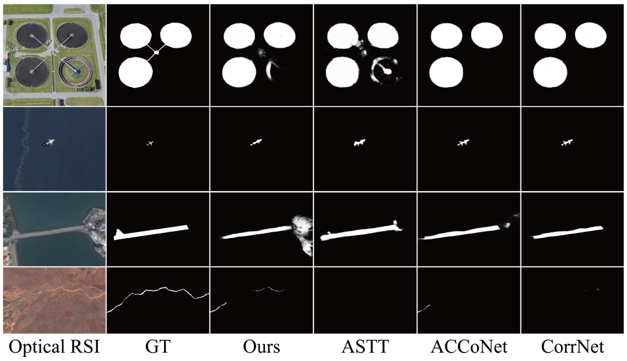


Fig. 16. Visualization results depict the failure cases encountered on the EORSSD, particularly with challenging scenes.

false negatives, indicating the potential for further advancements in camouflage detection. In addition, effectively addressing complex shadows still poses a challenge for ADSTNet. Specifically, while the shadows on the fuselage can be effectively suppressed, those at the rear of the aircraft are inaccurately preserved in the second row of Fig. 16. Similarly, achieving comprehensive detection of salient objects still presents challenges. For instance, the bridge and winding river fail to be completely detected in the third and fourth rows of Fig. 16. This is primarily due to their close resemblance to the surrounding environment, making their complete detection a formidable task. These observations reflect the inherent challenges of the scene, which persist even in the latest network architectures combining transformer and CNN components, for instance, ASTT, ACCoNet, and CorrNet.

V. Discussion

A. Effectiveness

The proposed ADSTNet in this article aims to acquire comprehensive image information and enhance information representation through progressive multilevel information interaction and boundary feature assistance, as confirmed by a series of experiments in Section IV. Such conceptualizations hold promise for bringing a more integrated detection framework to ORSI-SOD and are potentially transferable to other computer vision domains. However, there is still room for improvement in terms of real-time performance, particularly for deployment on airborne or spaceborne satellite equipment to realize greater value.

B. Hierarchical Information Adaptive Interaction

Acquiring more comprehensive information in the encoder is crucial as it significantly impacts the decoder's capacity to perform. This article introduces an adaptive iterative encoder, seeking to integrate local and global information complementarily. However, the current approach confines information matching solely to the terminal stage. Despite the supervision of matched information, certain limitations endure. Future work will concentrate on optimizing the fusion of both pieces of information, aiming to effectively diminish noise interference and irrelevant features during the fusion process.

C. Boundary-Assisted Optimization

The introduced DFR demonstrates effective adaptability to the complex features of remote sensing information, establishing a robust foundation for the accurate localization of regional information. While DFR can distill multidirectional boundary information, this process is currently manually defined. In the future, inspiration from domain-adaptive principles could be considered to develop an automated boundary information extraction module, tailored for a broader range of complex scenarios, thereby assisting various visual downstream tasks.

VI. Conclusion

In this work, we address the issue of compensating global and local information in the encoder-decoder framework and propose a novel end-to-end SOD model called ADSTNet. In the encoder, we introduce an adaptive interaction encoder that combines CNN with ADSE to capture multiscale feature processing enables a more comprehensive understanding of image details and contextual information. Moreover, the sparse attention mechanism selectively focuses on critical features, enhancing and the model's decision-making process can be interpreted with high accuracy. For the decoder, we propose an AFCD that dynamically adjusts and adapts the decoding process based on multipath input data. Furthermore, we introduce a plug-and-play DFR, which analyzes image information across multiple directional dimensions to extract distinctive features. This module assists the AFCD in compensating for diverse content information. In order to strengthen the learning ability of the model, we incorporate a structural loss function with a weight compensation mechanism. This loss function enhances the model's capacity to capture salient objects accurately. Comprehensive experiments on three benchmark datasets exhibit the advantages of our proposed model compared to 26 SOTA methods. Our model effectively combines global and local information, showcasing the effectiveness of each component. Despite the advantages of our approach, we plan to further refine our research by developing a lightweight transformer-based ORSI-SOD model. This model aims to enable practical deployment and achieve precise saliency detection, particularly in challenging scenarios such as camouflage situations.

References

[1] J. Zheng, Y. Gu, Y. Feng, J. Xu, and M. Zhang, "Boosting feature-aware network for salient object detection," in *Proc. Int. Conf. Artif. Neural Netw.*, 2022, pp. 14–26.

[2] Y. Gu, H. Xu, Y. Quan, W. Chen, and J. Zheng, "ORSI salient object detection via bidimensional attention and full-stage semantic guidance," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5603213.

[3] D.-P. Fan, G.-P. Ji, G. Sun, M.-M. Cheng, J. Shen, and L. Shao, "Camouflaged object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 2777–2787.

[4] F. Yang et al., "Uncertainty-guided transformer reasoning for camouflaged object detection," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 4146–4155.

[5] Y. Sugano, Y. Matsushita, and Y. Sato, "Calibration-free gaze sensing using saliency maps," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, 2010, pp. 2667–2674.

[6] A. Borji and L. Itti, "Defending yarbus: Eye movements reveal observers' task," *J. Vis.*, vol. 14, no. 3, pp. 29–29, 2014.

[7] A. Karpathy, S. Miller, and L. Fei-Fei, “Object discovery in 3D scenes via shape analysis,” in *Proc. IEEE Int. Conf. Robot. Automat.*, 2013, pp. 2088–2095.

[8] S. Frintrop, G. M. García, and A. B. Cremers, “A cognitive approach for object discovery,” in *Proc. 22nd Int. Conf. Pattern Recognit.*, 2014, pp. 2329–2334.

[9] R. Cong, W. Song, J. Lei, G. Yue, Y. Zhao, and S. Kwong, “PSNet: Parallel symmetric network for video salient object detection,” *IEEE Trans. Emerg. Topics Comput. Intell.*, vol. 7, no. 2, pp. 402–414, Apr. 2023.

[10] R. Zhao, W. Ouyang, and X. Wang, “Unsupervised salience learning for person re-identification,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2013, pp. 3586–3593.

[11] S. Bi, G. Li, and Y. Yu, “Person re-identification using multiple experts with random subspaces,” *J. Image Graph.*, vol. 2, no. 2, pp. 151–157, 2014.

[12] Y. Lin, D. Zhang, X. Fang, Y. Chen, K.-T. Cheng, and H. Chen, “Rethinking boundary detection in deep learning models for medical image segmentation,” in *Proc. Int. Conf. Inf. Process. Med. Imag.*, 2023, pp. 730–742.

[13] N. T. Duc, N. T. Oanh, N. T. Thuy, T. M. Triet, and V. S. Dinh, “Colonformer: An efficient transformer based method for colon polyp segmentation,” *IEEE Access*, vol. 10, pp. 80575–80586, 2022.

[14] D.-P. Fan et al., “Inf-Net: Automatic COVID-19 lung infection segmentation from CT images,” *IEEE Trans. Med. Imag.*, vol. 39, no. 8, pp. 2626–2637, Aug. 2020.

[15] Y.-H. Wu et al., “JCS: An explainable COVID-19 diagnosis system by joint classification and segmentation,” *IEEE Trans. Image Process.*, vol. 30, pp. 3113–3126, 2021.

[16] 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.

[17] 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.

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

[19] M. Yang, Z. Liu, W. Dong, and Y. Wu, “An important pick-and-pass gated refinement network for salient object detection in optical remote sensing images,” *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 16, pp. 6505–6516, 2023.

[20] A. Borji, M.-M. Cheng, H. Jiang, and J. Li, “Salient object detection: A benchmark,” *IEEE Trans. Image Process.*, vol. 24, no. 12, pp. 5706–5722, Dec. 2015.

[21] R. Cong, J. Lei, H. Fu, M.-M. Cheng, W. Lin, and Q. Huang, “Review of visual saliency detection with comprehensive information,” *IEEE Trans. Circuits Syst. Video Technol.*, vol. 29, no. 10, pp. 2941–2959, Oct. 2019.

[22] C. Dong, J. Liu, and F. Xu, “Ship detection in optical remote sensing images based on saliency and a rotation-invariant descriptor,” *Remote Sens.*, vol. 10, no. 3, 2018, Art. no. 400.

[23] Z. Ren, Y. Tang, Z. He, L. Tian, Y. Yang, and W. Zhang, “Ship detection in high-resolution optical remote sensing images aided by saliency information,” *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5623616.

[24] D. Zhu, B. Wang, and L. Zhang, “Airport target detection in remote sensing images: A new method based on two-way saliency,” *IEEE Geosci. Remote Sens. Lett.*, vol. 12, no. 5, pp. 1096–1100, May 2015.

[25] T. Zhu, Y. Li, Q. Ye, H. Huo, and T. Fang, “Integrating saliency and resnet for airport detection in large-size remote sensing images,” in *Proc. 2nd Int. Conf. Image, Vis. Comput.*, 2017, pp. 20–25.

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

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

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

[29] H. Huang et al., “UNet 3+: A full-scale connected unet for medical image segmentation,” in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2020, pp. 1055–1059.

[30] 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.*, 2018, pp. 801–818.

[31] K. Sun, B. Xiao, D. Liu, and J. Wang, “Deep high-resolution representation learning for human pose estimation,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 5693–5703.

[32] A. Kirillov, Y. Wu, K. He, and R. Girshick, “Pointrend: Image segmentation as rendering,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 9799–9808.

[33] R. Strudel, R. Garcia, I. Laptev, and C. Schmid, “Segmenter: Transformer for semantic segmentation,” in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 7262–7272.

[34] M.-H. Guo, C.-Z. Lu, Q. Hou, Z. Liu, M.-M. Cheng, and S.-M. Hu, “Segnext: Rethinking convolutional attention design for semantic segmentation,” in *Proc. Int. Conf. Adv. Neural Inf. Process. Syst.*, 2022, vol. 35, pp. 1140–1156.

[35] M. Xu, Z. Zhang, F. Wei, H. Hu, and X. Bai, “Side adapter network for open-vocabulary semantic segmentation,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 2945–2954.

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

[37] J. Zheng, Y. Quan, H. Zheng, Y. Wang, and X. Pan, “Orsi salient object detection via cross-scale interaction and enlarged receptive field,” *IEEE Geosci. Remote Sens. Lett.*, vol. 20, 2023, Art. no. 6003205.

[38] 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.

[39] L. Zhang, Y. Liu, and J. Zhang, “Saliency detection based on self-adaptive multiple feature fusion for remote sensing images,” *Int. J. Remote Sens.*, vol. 40, no. 22, pp. 8270–8297, 2019.

[40] Z. Liu, D. Zhao, Z. Shi, and Z. Jiang, “Unsupervised saliency model with color Markov chain for oil tank detection,” *Remote Sens.*, vol. 11, no. 9, 2019, Art. no. 1089.

[41] 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.*, 2015, pp. 1877–1880.

[42] 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 25th Int. Conf. Image Process.*, 2018, pp. 2336–2340.

[43] 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.

[44] 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.

[45] Y. Jia, J. Zhao, and L. Yu, “Aadh-Yolov5: Improved Yolov5 based on adaptive activate decoupled head for garbage detection,” *J. Electron. Imag.*, vol. 32, no. 4, 2023, Art. no. 043017.

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

[47] X. Zhou et al., “Edge-guided recurrent positioning network for salient object detection in optical remote sensing images,” *IEEE Trans. Cybern.*, vol. 53, no. 1, pp. 539–552, Jan. 2023.

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

[49] Y. Liu, S. Zhang, Z. Wang, B. Zhao, and L. Zou, “Global perception network for salient object detection in remote sensing images,” *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5617212.

[50] N. Liu, N. Zhang, K. Wan, L. Shao, and J. Han, “Visual saliency transformer,” in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 4722–4732.

[51] L. Gao, B. Liu, P. Fu, and M. Xu, “Adaptive spatial tokenization transformer for salient object detection in optical remote sensing images,” *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5602915.

[52] Z. Wang, Y. Zhang, Y. Liu, Z. Wang, S. Coleman, and D. Kerr, “TF-SOD: A novel transformer framework for salient object detection,” *Neural Comput. Appl.*, vol. 34, no. 14, pp. 11789–11806, 2022.

[53] Z. Liu, Y. Tan, Q. He, and Y. Xiao, “SwinNet: Swin transformer drives edge-aware RGB-D and RGB-T salient object detection,” *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 7, pp. 4486–4497, Jul. 2022.

[54] Z. Liu, Y. Wang, Z. Tu, Y. Xiao, and B. Tang, “TritransNet: RGB-D salient object detection with a triplet transformer embedding network,” in *Proc. 29th ACM Int. Conf. Multimedia*, 2021, pp. 4481–4490.

[55] K. Huang, C. Tian, J. Su, and J. C.-W. Lin, "Transformer-based cross reference network for video salient object detection," *Pattern Recognit. Lett.*, vol. 160, pp. 122–127, 2022.

[56] S.-H. Gao, M.-M. Cheng, K. Zhao, X.-Y. Zhang, M.-H. Yang, and P. Torr, "Res2Net: A new multi-scale backbone architecture," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 2, pp. 652–662, Feb. 2022.

[57] R. Cong et al., "Global-and-local collaborative learning for co-salient object detection," *IEEE Trans. Cybern.*, vol. 53, no. 3, pp. 1920–1931, Mar. 2023.

[58] X. Qin, Z. Zhang, C. Huang, M. Dehghan, O. R. Zaiane, and M. Jagersand, "U2-Net: Going deeper with nested u-structure for salient object detection," *Pattern Recognit.*, vol. 106, 2020, Art. no. 107404.

[59] R. Cong et al., "Does thermal really always matter for RGB-T salient object detection?," *IEEE Trans. Multimedia*, vol. 25, pp. 6971–6982, 2023.

[60] Q. Hou, M.-M. Cheng, X. Hu, A. Borji, Z. Tu, and P. H. Torr, "Deeply supervised salient object detection with short connections," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 3203–3212.

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

[62] Z. Bai, G. Li, and Z. Liu, "Global-local-global context-aware network for salient object detection in optical remote sensing images," *ISPRS J. Photogrammetry Remote Sens.*, vol. 198, pp. 184–196, 2023.

[63] D. Song, Y. Dong, and X. Li, "Adjacent complementary network for salient object detection in optical remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5403111.

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

[65] 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, 2021, Art. no. 5607913.

[66] A. Dosovitskiy et al., "An image is worth 16x16 words: Transformers for image recognition at scale," 2020, *arXiv:2010.11929*.

[67] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*.

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

[69] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," *Commun. ACM*, ACM New York, NY, USA, vol. 60, no. 6, pp. 84–90, 2017.

[70] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 7132–7141.

[71] 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.*, 2017, pp. 2117–2125.

[72] Z. Peng et al., "Conformer: Local features coupling global representations for recognition and detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 8, pp. 9454–9468, Aug. 2023.

[73] Y.-H. Wu, Y. Liu, X. Zhan, and M.-M. Cheng, "P2T: Pyramid pooling transformer for scene understanding," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 11, pp. 12760–12771, Nov. 2023.

[74] C.-M. Fan, T.-J. Liu, and K.-H. Liu, "SUNet: Swin transformer UNet for image denoising," *IEEE Int. Symp. Circuits Syst.*, pp. 2333–2337, 2022.

[75] H. Wu et al., "CVT: Introducing convolutions to vision transformers," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 22–31.

[76] Y. Li, K. Zhang, J. Cao, R. Timofte, and L. Van Gool, "Localvit: Bringing locality to vision transformers," 2021, *arXiv:2104.05707*.

[77] J. Guo et al., "CMT: Convolutional neural networks meet vision transformers," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 12175–12185.

[78] M. Mao et al., "Dual-stream network for visual recognition," in *Proc. Int. Conf. Adv. Neural Inf. Process. Syst.*, 2021, vol. 34, pp. 25346–25358.

[79] Y. Xu, Q. Zhang, J. Zhang, and D. Tao, "Vitae: Vision transformer advanced by exploring intrinsic inductive bias," in *Proc. Int. Conf. Adv. Neural Inf. Process. Syst.*, 2021, vol. 34, pp. 28522–28535.

[80] T. Yao, Y. Li, Y. Pan, Y. Wang, X.-P. Zhang, and T. Mei, "Dual vision transformer," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 9, pp. 10870–10882, Sep. 2023.

[81] N. Zhang, J. Han, and N. Liu, "Learning implicit class knowledge for RGB-D co-salient object detection with transformers," *IEEE Trans. Image Process.*, vol. 31, pp. 4556–4570, 2022.

[82] Y. Zhang, J. Guo, H. Yue, X. Yin, and S. Zheng, "Transformer guidance dual-stream network for salient object detection in optical remote sensing images," *Neural Comput. Appl.*, vol. 35, pp. 17733–17747, 2023.

[83] N. Kanopoulos, N. Vasanthavada, and R. L. Baker, "Design of an image edge detection filter using the Sobel operator," *IEEE J. Solid-State Circuits*, vol. 23, no. 2, pp. 358–367, Apr. 1988.

[84] X. Zhang, B. Yin, Z. Lin, Q. Hou, D.-P. Fan, and M.-M. Cheng, "Referring camouflaged object detection," 2023, *arXiv:2306.07532*.

[85] X. Chen, C. Dong, J. Ji, J. Cao, and X. Li, "Image manipulation detection by multi-view multi-scale supervision," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 14185–14193.

[86] S. W. Zamir, A. Arora, S. Khan, M. Hayat, F. S. Khan, and M.-H. Yang, "Restormer: Efficient transformer for high-resolution image restoration," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 5728–5739.

[87] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "CBAM: Convolutional block attention module," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 3–19.

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

[89] 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.*, 2017, pp. 4548–4557.

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

[91] D.-P. Fan, C. Gong, Y. Cao, B. Ren, M.-M. Cheng, and A. Borji, "Enhanced-alignment measure for binary foreground map evaluation," 2018, *arXiv:1805.10421*.

[92] D.-P. Fan, Z. Lin, Z. Zhang, M. Zhu, and M.-M. Cheng, "Rethinking RGB-D salient object detection: Models, data sets, and large-scale benchmarks," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 32, no. 5, pp. 2075–2089, May 2021.

[93] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*.

[94] A. Paszke et al., "Pytorch: An imperative style, high-performance deep learning library," in *Proc. Int. Conf. Adv. Neural Inf. Process. Syst.*, 2019, vol. 32, pp. 8026–8037.

[95] C. Li, Y. Yuan, W. Cai, Y. Xia, and D. Dagan Feng, "Robust saliency detection via regularized random walks ranking," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 2710–2717.

[96] J. Kim, D. Han, Y.-W. Tai, and J. Kim, "Salient region detection via high-dimensional color transform and local spatial support," *IEEE Trans. Image Process.*, vol. 25, no. 1, pp. 9–23, Jan. 2016.

[97] L. Zhou, Z. Yang, Z. Zhou, and D. Hu, "Salient region detection using diffusion process on a two-layer sparse graph," *IEEE Trans. Image Process.*, vol. 26, no. 12, pp. 5882–5894, Dec. 2017.

[98] Y. Yuan, C. Li, J. Kim, W. Cai, and D. D. Feng, "Reversion correction and regularized random walk ranking for saliency detection," *IEEE Trans. Image Process.*, vol. 27, no. 3, pp. 1311–1322, Mar. 2018.

[99] X. Hu, L. Zhu, J. Qin, C.-W. Fu, and P.-A. Heng, "Recurrently aggregating deep features for salient object detection," *Proc. AAAI Conf. Artif. Intell.*, vol. 32, no. 1, pp. 6943–6950, 2018.

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

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

[102] X. Zhao, Y. Pang, L. Zhang, H. Lu, and L. Zhang, "Suppress and balance: A simple gated network for salient object detection," in *Proc. 16th Eur. Conf. Comput. Vis.*, 2020, pp. 35–51.

[103] J. Li, Z. Pan, Q. Liu, and Z. Wang, "Stacked U-Shape network with channel-wise attention for salient object detection," *IEEE Trans. Multimedia*, vol. 23, pp. 1397–1409, 2020.

[104] B. Xu, H. Liang, R. Liang, and P. Chen, "Locate globally, segment locally: A progressive architecture with knowledge review network for salient object detection," in *Proc. AAAI Conf. Artif. Intell.*, 2021, vol. 35, pp. 3004–3012.

[105] 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.*, 2022, pp. 238–244.

[106] 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, 2021, Art. no. 2163.

[107] 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.

Jie Zhao received the B.E. degree in computer science and technology from Shandong Youth University of Political Science, Jinan, China in 2021. He is currently working toward the M.E. degree in remote sensing image interpretation with the School of Information and Electronic Engineering, Shandong Technology and Business University, Yantai, China.

His research interests include remote sensing image interpretation, saliency detection and segmentation, and deep learning.

Yun Jia received the B.E., M.E., and Ph.D. degrees from the College of Information and Communication Engineering, Harbin Institute of Technology, Heilongjiang, China, in 2006, 2009, and 2014, respectively, all in information and communication engineering.

From 2011 to 2012, he was a Visiting Scholar with the School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore. He is currently with Shandong Technology and Business University, Yantai, China. His research interests include topics related to cognitive radio/software radio, deep learning, and intelligent signal processing.

Lin Ma (Senior Member, IEEE) received the B.E., M.E., and Ph.D. degrees in communication engineering from the Harbin Institute of Technology, Heilongjiang, China, in 2003, 2005 and 2009, respectively, all in communication engineering.

From 2013 to 2014, he was a Visiting Scholar with the Edward S. Rogers, Sr. Department of Electrical and Computer Engineering, University of Toronto, Toronto, ON, Canada. He is currently a Professor with the School of Electronics and Information Engineering, Harbin Institute of Technology. His research interests include image processing, artificial intelligence-based cognition, and recognition technology for complex environments and deep learning.

Lidan Yu received the B.E. degree in computer science and technology from Qingdao University of Science and Technology, Qingdao, China, in 2021. She is currently working toward the M.E. degree in image denoising with the School of Information and Electronic Engineering, Shandong Technology and Business University, Yantai, China.

Her research interests include image processing and computer vision.