IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

IEEE GEOSCIENCE AND REMOTE SENSING SOCIETY SECTION

Received 20 November 2025, accepted 22 December 2025, date of publication 24 December 2025,
date of current version 31 December 2025.

Digital Object Identifier 10.1109/ACCESS.2025.3648292

RESEARCH ARTICLE

Edge-Guided Shallow and Contextual Deep Feature Learning via Bidirectional Attention for Salient Object Detection in Optical Remote Sensing Images

SAMRA KANWAL[1], NAZAR WAHEED[2], BUSHRA RASHID[3], NAYEF ALQAHTANI[4], (Member, IEEE), AND ALI ALQAHTANI[5]

[1]Department of Computer Software Engineering, National University of Sciences and Technology, Islamabad 44000, Pakistan
[2]Faculty of Computer Information Systems, Higher Colleges of Technology, Abu Dhabi, United Arab Emirates
[3]Department of Biomedical Engineering, College of Engineering, King Faisal University, Al-Ahsa 31982, Saudi Arabia
[4]Department of Electrical Engineering, College of Engineering, King Faisal University, Al-Ahsa 31982, Saudi Arabia
[5]Department of Networks and Communications Engineering, College of Computer Science and Information Systems, Najran University, Najran 61441, Saudi Arabia

Corresponding author: Samra Kanwal (samrakanwal@mcs.edu.pk)

This work was supported in part by the Deanship of Scientific Research, Vice Presidency for Graduate Studies and Scientific Research, King Faisal University, Saudi Arabia, under Grant KFU254724; and in part by the Deanship of Graduate Studies and Scientific Research at Najran University through the Growth Funding Program under Grant NU/GP/SERC/13/358-7.

ABSTRACT Salient object detection from optical remote sensing images seeks to detect and segment out objects that stand out visually by mimicking human interpretation. Accurately detecting salient objects in remote sensing imagery is hindered by the broad variability in object appearances and background complexity. Despite recent advancements, there exist two major challenges. Firstly, it remains unclear how to fuse shallow features with deep features optimally. Secondly, defining an optimal strategy for multi-scale feature processing is crucial, as patterns can vary significantly across different levels. We introduce a novel bidirectional attention mechanism to tackle these challenges. Our framework features: 1) a Parallel Convolution-Channel Attention (PCCA) module that boosts edge representation and highlights salient feature channels, and 2) a Holistic Reverse Attention (HRA) module that preserves crucial details in high-level features. Coupled with a multiscale progressive decoder, our method enables precise feature integration. Extensive evaluation on benchmark datasets demonstrates the superiority of our proposed model compared with 17 state-of-the-art (SOTA) models, achieving a significant 11.57% reduction in Mean Absolute Error and setting a new state-of-the-art on S-measure with a 0.4% improvement.

INDEX TERMS Bidirectional attention mechanism, feature alignment, shallow and deep features fusion, optical remote sensing images, salient object detection.

I. INTRODUCTION

Salient object detection (SOD) deals with identifying the visually most prominent object in an image. Most studies of saliency detection task have only been carried out on natural scene images (NSI-SOD) [1], [2], [3]. Salient object detection is an important downstream computer vision task, with applications in field of medical image segmentation [4], object tracking [5], image retrieval [6], and action recognition [7] etc. In contrast to NSI-SOD, salient object detection in optical remote sensing images (ORSI-SOD) generally faces more challenging scenarios, featuring small objects, irregular structures, and highly complex backgrounds. Optical remote sensing images are acquired by remote sensing satellites, thereby, exhibiting high-altitude overhead view. Due to significant difference between NSI and ORSI, the

The associate editor coordinating the review of this manuscript and approving it for publication was Monidipa Das.

NSI-SOD models cannot detect objects with high accuracy from remote sensing images [10], [22]. Furthermore, ORSI-SOD is needed in various applications such as environmental monitoring, agriculture and crop monitoring, urban planning, disaster management, and reconnaissance and surveillance etc [33], [34]. Therefore, ORSI-SOD has emerged as an increasingly important task in computer vision field. Unlike, hyperspectral remote sensing images that include large spectral band information, saliency detection in optical remote sensing images deals with identifying visually prominent man-made targets or river system [24]. In certain real-world outdoor scenes — for example, deserts, forests, or oceans — salient regions may be completely absent. Considering these particular imaging characteristics, the first annotated ORSI dataset was published by Li et al. [24]. Afterwards, several salient object detection models for remote sensing images were evolved, each formulated from different perspective. For example, Yan et al. [16] propose ASNet model having a dual branch encoder, which includes a Convolution Neural Network (CNN) stream and a Transformer stream. The intuition behind is that CNN is good in capturing local details while Transformer captures long range dependencies. The ORSI-SOD models need local fine-grained information because of intense topology of remote sensing images. However, prediction inaccuracies arise if global contextual representation is ignored. Furthermore, when features are hierarchically combined, the distinction between global and local contexts can be lost. As a result, ASNet model using this pattern achieves lower accuracy. Liang and Luo [22] propose a model integrated with multiscale edge-embedded attention and multilevel semantic guidance modules. Although this model provides a lightweight solution, however, supervised edge-embedded attention may not add value because deeper features lose fine details. While AESI [40] combines high-level features via pixel-wise multiplication, it does not model the cross-layer affinities that link shallow and deep representations. The absence of multi-scale relational information limits its ability to integrate spatial detail with semantic context, yielding less effective fusion.

A. PURPOSE OF THE PROPOSED MODEL

The purpose of the proposed model is to address the unresolved challenge of effectively integrating shallow spatial details with deep semantic representations while preserving multi-scale feature affinity across adjacent layers. The challenges and responses of existing models on optical remote sensing images are featured in Fig. 1. Existing methods fail to establish an optimal strategy for fusing shallow and deep features, as well as for effectively processing multiscale features. Thereby, these methods often struggle with both contextual saliency detection and structure-preserving object segmentation [15], [16], represented as bounding box in 3^{rd} and 1^{st} row of Fig. 1, respectively. These limitations highlight the necessity of designing more sophisticated ORSI-SOD models.

After carefully examining the existing strategies, we derived three main insights, which serve as the foundation

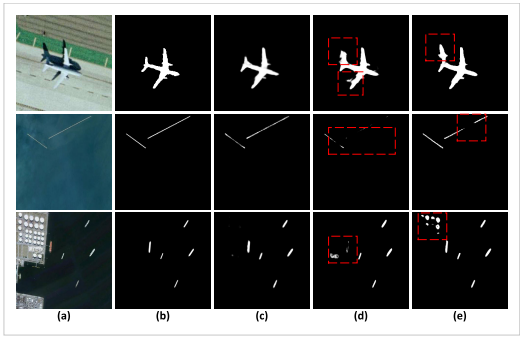

FIGURE 1. Some examples of missed/incorrect detection, area highlighted by bounding boxes. (a)ORSI Input Image (b)Ground Truth (c)Ours (d)ASNet [16] (e)BCARNet [15].

for our proposed model. Firstly, shallow and deep features exhibit distinct characteristics, therefore, it is necessary to develop dedicated methods to leverage shallow features for local detail refinement and deep features for global context enhancement. Secondly, in remote sensing imagery, the salient object frequently resembles surrounding non-salient areas in terms of low-level details like edges, texture, and color. This leads to redundant or missed segmented area in prediction map. Thirdly, progressive refinement of multilevel features often leads to inconsistent results due to non-homogeneous patterns in hierarchical representation. Unlike hybrid Transformer-CNN encoder in ASNet model [16], where global semantic guidance from Transformer stream is added with CNN stream, we utilize Conformer [35] as backbone in which dual Transformer-CNN streams have bidirectional feature coupling units (FCU). CNN features (edges, textures) help the Transformer to attend meaningful fine-grained details while Transformer's global context helps the CNN suppress noise or irrelevant local patterns. This mutual reinforcement will generate enhanced multi-level representation. Shallow features are refined to produce a supervised edge map obtained by applying a Sobel filter to the saliency ground truth. Deeper features are processed to generate contextual saliency. We then build a feature pyramid network to combine low and high level features and to generate the final saliency map. In summary the main contributions of our proposed work is as follows:

- We propose a novel bidirectional attention strategy, to optimally extract shallow and deep features from Conformer [35] backbone network.
- To preserve structural details, we propose Parallel Convolution-Channel Attention (PCCA) module which takes shallow features as input and generates supervised edge map.
- We propose a Holistic Reverse Attention (HRA) module to generate global contextual saliency map from deep features.
- We aggregate the manifestations obtained from PCCA and HRA in feature pyramid network to generate the predicted saliency map.

The rest of paper is organized as follows. In Section II, we comprehensively review the related work on ORSI-SOD. The proposed methodology is discussed in Section III. Evaluation results along with ablation study is provided in Section IV. Finally we conclude the article in Section VI.

II. RELATED WORK

In this section, we will provide a review of salient object detection on optimal remote sensing images (ORSI-SOD). We have structured literature review into two sections including Traditional ORSI-SOD methods and Deep ORSI-SOD methods.

A. TRADITIONAL ORSI-SOD METHODS

Traditional ORSI-SOD methods use handcrafted features for remote sensing images. In [36], RSI image is first converted to image superpixels, which are used to extract contrast distribution and their spatial relation. In second step, line density distribution is formed. The fusion of two-way saliency is specifically beneficial for airport extraction. Zhang et al. [37] proposed self-adaptively multiple feature fusion model for remote sensing saliency detection. In [38] remote sensing imagery is used for oil tank detection by applying Color Markov Chain and generate a bottom-up latent saliency map. Although, these conventional methods have laid the foundation for salient object detection using ORSI, however, paucity of global semantic guidance results in limited performance.

B. DEEP ORSI-SOD METHODS

With the widespread adoption of deep learning in various fields, the ORSI-SOD research to date has tended to focus on deep methods also. LVNet proposed by Li et al. [24] was the first systematic CNN model for ORSI, comprises of a L-shaped pyramid module and V-shaped encoder-decoder module and was evaluated on the first publicly available ORSI dataset. In [25], Global Context-aware Attention is realized in end-to-end network. This leads to low detection accuracy for small objects or where objects have higher resemblance with the surroundings. For resource-constrained circumstances various lightweight saliency detection models are proposed [17], [22]. They offer reduced computational burden with compromised accuracy due to fine-grained features missclassification.

Li et al. [39] highlight the importance of attention-based multimodal fusion, inspiring more robust feature integration but it is limited to image captioning. Recently, Dong et al. [43] introduced a Transformer-based hybrid edge-fusion perception network for the ORSI-SOD task. However, the model struggles in challenging scenarios, particularly those involving elongated objects or low-contrast scenes. These limitations may stem from inadequate preservation of fine image details during the edge-enhancement stage. In [47], scribble-based weakly supervised multi-modality salient object detection model is proposed. This model is training on general saliency images and lacks the comprehensive evaluation on optical remote sensing images.

Yan et al. [16] presented hybrid Transformer-CNN encoder to overcome the deficiencies of using standalone CNN or Transformer. The top-down progressive refinement bring-out poor boundary localization. Gu et al. [41] used non-local attention to capture global context, but this lead to attention sparsity, where the model overlooks certain regions. To overcome this issue Gu et al. in another recent work BCARNet [15] proposed a novel Bidirectional Cross-Attention and Attention Restoration Neural Network. However, BCARNet relies on full-parameter fine-tuning for each dataset, underutilizing the knowledge from pretrained backbones. Zeng et al. [40] fused high-level features via pixel-wise multiplication but ignored shallow–deep layer affinities, limiting multi-scale spatial–semantic integration. To tackle this issue in [42], CMNFNet uses two heterogeneous encoders, however CMNFNet may still struggle to effectively differentiate multiple salient objects of varying sizes or complex structures, especially when shallow and deep cues are not explicitly coordinated.

Liu et al. [8] propose an Efficient Global Perception Module (EGPM) to capture long range dependencies and utilize graph convolution to further refine it. Along with that they highlight the object boundaries using supervised approach. This method provided a computationally effective solution but failed to perform reliably under low-contrast and occlusion conditions. In [10], spit-and-concatenate operation is opted to capture salient region in top-down manner. In [11], a dual-stream CNN Transformer encoder is used to extract multi-level correlation between features via cross-attention complementary mining module (CCMM). Furthermore, they proposed cross-layer feature interaction module (CFIM) to combine features in top-down approach. This model struggles with irregular shapes and fine edges, leading to misclassification in complex structures. Li et al. [12], proposed a novel dynamically enhanced aggregation module (DEAM) to deconstruct–interact–recombine channel features manifestations and dynamically represent spatial details. In [13], a novel progressive self-prompting Segment Anything Model (SAM) is proposed. Although, it performed well in various challenging scenarios but has limited contextual perception.

In [14], edge and object regions are simultaneously learned in a joint multi-scale framework. This model suffers from spatial inconsistency due to the aggregation of branches with multiple receptive fields, leading to missing regions in the predicted map. In [9], a novel furcate skip connection module with large receptive fields is used to capture salient objects of various sizes. The focus of this approach was handling global context. Huang et al. [18], introduced a memory-based context propagation network (MCP-Net) designed to exploit dataset-level contextual information. The model fails to cater low-contrast and occluded images, due to saliency supervision at all levels irrespective of level intrinsic characteristics.

In [19], a novel heterogeneous feature collaboration network (HFCNet), which employs a CNN-Transformer encoder to adaptively fuse the complementary strengths of global and local features.

Yao and Gao [20] proposed a bidirectional encoder-decoder framework, where, in the feedforward path, a saliency aggregation mechanism integrates multi-level saliency cues to generate complementary information and in the feedback path, a saliency assignment strategy injects region and edge based prompts into the encoder, guiding hierarchical feature updates and strengthening salient representations. Li et al. [21] utilized coarse prediction to refine saliency information from global features via cross-layer correlation operation and local features interaction. However, the dilated convolution involved resulted missed or noisy prediction. Another work from Di et al. [23] proposed a multiscale and multidimensional weighted network to overcome the noisy prediction due to excessive features fusion. However, this model lacks generalization ability. Despite these prominent works, ORSI-SOD still face challenges to accurately predict the salient objects because of highly complex structures, nonuniform illumination, object reflections and occlusions.

Existing studies overlook the inherent properties of CNN and Transformer architectures, as well as the discrepancies among multi-level feature characteristics. Therefore, we propose a novel strategy to extract complementarity of various architectures and integration of multi-level feature. We aim to reduce the inconsistent predictions from previous works by optimally fusing the non-homogeneous features manifestation from multiple layers.

III. METHODOLOGY

The overall framework of proposed method is illustrated in Fig. 2, which consists of following four modules.

- Pre-trained backbone network
- Parallel Convolution-Channel Attention (PCCA)
- Holistic Reverse Attention (HRA)
- Feature Pyramid Network

The details of each module is provided below. Furthermore, we have provided the training pipeline of proposed framework in Algorithm 1.

A. PRE-TRAINED BACKBONE NETWORK

For multi-level features extraction from the optical remote sensing images, the Conformer Network [35] proposed for image classification is used as backbone network. The Conformer architecture adopts a dual-branch design, integrating both a CNN branch and a Transformer branch. Conformer leverages both convolutional operations from CNN branch and self-attention mechanisms from Transformer branch to achieve stronger representation learning. The architecture of Conformer Network is presented in Table 1. The multi-level features are hierarchically represented in five stages (c1,c2,c3,c4,c5) with dimensions $F_{c\_i} \in \mathbb{R}^{C\times H\times W}$ (C, H, W are channels, height and width respectively), for CNN feature maps and $F_{t\_i} \in \mathbb{R}^{(K+1)\times E}$ (K, 1, and E represent the number of image patches, class token and embedding dimensions respectively), for Transformer feature maps. Here, i ∈ {initial,s,d1,d2,d3}, where the index s refers to shallow features, while d1,d2,d3 indicate progressively deeper feature levels. At its core lies the Feature Coupling Unit (FCU), which integrates local features with global representations across multiple resolutions in an interactive manner.

In the proposed methodology, the c2-stage features obtained from the CNN and Transformer branches after the FCU are extracted and further refined to capture fine-grained details in Parallel Convolution-Channel Attention (PCCA). The CNN branch captures shallow features rich in edges, textures, and colors, whereas the Transformer branch complements them by highlighting structural details. The notation

Algorithm 1 Training Pipeline of Proposed ORSI-SOD Model

Input:

1) Training dataset $\mathcal{D} = \{(ORSI\_i, S_{sal}\_i, S_{edge}\_i)\}_{i=1}^{N}$, learning rate η, number of epochs E, batch size B, model parameters θ

Output:

1) Final saliency map S_f, Saliency map from CNN branch last layer F_{c_sal}, Saliency map from transformer branch last layer F_{t_sal}, Edge saliency map S_{edge_pred}

Initialize model parameters θ
for epoch = 1 **to** E **do**
  Shuffle training dataset $\mathcal{D}$
  for each batch $\{(ORSI\_j, S_{sal}\_j, S_{edge}\_j)\}_{j=1}^{B}$ in $\mathcal{D}$ **do**
    1. Extract intermediate feature maps from conformer backbone network: F_{c_i}, F_{t_i}.
    2. Sample shallow features from F_{c_i}, F_{t_i}.
      i. Sample f_{ce_i} using eq. 1 to 3 from shallow CNN branch features.
      ii. Sample f_{tg_s} using eq. 4 to 6, to align shallow Transformer branch features with CNN features.
      iii. Sample Out_{CA_g} by applying shuffle channel attention using eq. 7 to 12.
      iv. Compute S_{edge_pred} using eq. 13.
    3. Sample deep features from F_{c_i}, F_{t_i}.
      i. Compute F_{c_sal} and F_{t_sal} from eq. 14 and 15.
      ii. Sample HRA_out_i using eq. 16 to 20.
    4. Compute S_f using eq. 21
    5. Compute loss: $\mathcal{L}_{total}$ using eq. 22
    6. Backward pass: compute gradients $\nabla_\theta \mathcal{L}_{total}$
    7. Update parameters: $\theta \leftarrow \theta - \eta \nabla_\theta \mathcal{L}_{total}$
  end for
  Adjust learning rate
end for
Return trained model parameters θ

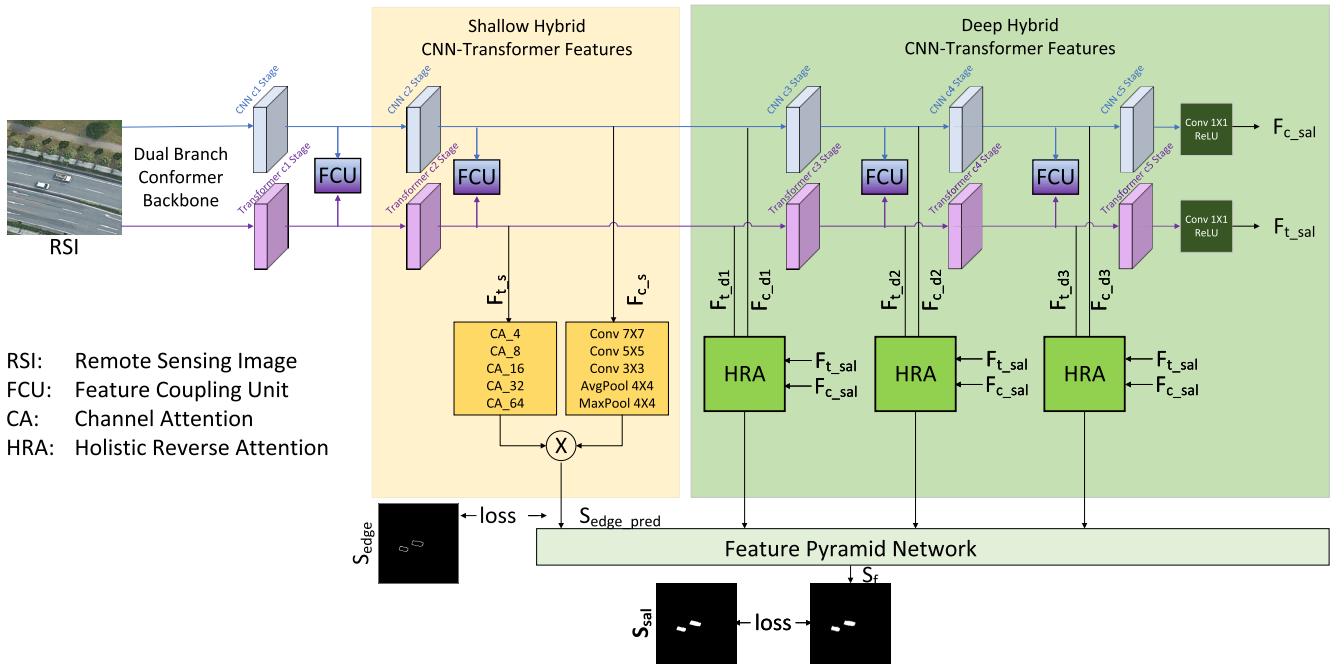

FIGURE 2. Overview of proposed model. We use conformer network as backbone. Stage c2 features are used for edge refinement and Stage c3, c4 and c5 are used for global contexual details.

TABLE 1. Dual branch conformer [35] backbone network.

Stage	CNN Branch	Transformer Branch
Initial Feature Maps		
c1	$F_{c\_initial} \in \mathbb{R}^{64\times160\times160}$	$F_{t\_initial} \in \mathbb{R}^{64\times160\times160}$
Shallow Feature Maps		
c2	$F_{c\_s} \in \mathbb{R}^{384\times80\times80}$	$F_{t\_s} \in \mathbb{R}^{(400+1)\times576}$
Deep Feature Maps		
c3	$F_{c\_d1} \in \mathbb{R}^{768\times40\times40}$	$F_{t\_d1} \in \mathbb{R}^{(400+1)\times576}$
c4	$F_{c\_d2} \in \mathbb{R}^{1536\times20\times20}$	$F_{t\_d2} \in \mathbb{R}^{(400+1)\times576}$
c5	$F_{c\_d3} \in \mathbb{R}^{1536\times10\times10}$	$F_{t\_d3} \in \mathbb{R}^{(400+1)\times576}$

used for these feature manifestations are presented in Fig. 2 and Table 1 as $F_{c\_s} \in \mathbb{R}^{384\times80\times80}$ and $F_{t\_s} \in \mathbb{R}^{(400+1)\times576}$. Additionally, deeper features extracted from dual branch backbone are represented as $F_{c\_d1} \in \mathbb{R}^{768\times40\times40}$, $F_{c\_d2} \in \mathbb{R}^{1536\times20\times20}$ and $F_{c\_d3} \in \mathbb{R}^{1536\times10\times10}$ CNN representations and $F_{t\_d1} \in \mathbb{R}^{(400+1)\times576}$, $F_{t\_d2} \in \mathbb{R}^{(400+1)\times576}$ and $F_{t\_d3} \in \mathbb{R}^{(400+1)\times576}$ Transformer representations. These features capture long range dependencies via Holistic Reverse Attention (HRA). Feature Pyramid Network decodes the encoded low and high level features to obtained final saliency map.

B. PARALLEL CONVOLUTION-CHANNEL ATTENTION (PCCA)

This module is designed to handle the hybrid shallow features. The architecture of PCCA is shown in Fig. 3. It is widely recognized that CNN is good in extracting local details while Transformer network provides long range dependencies. The aim of Parallel Convolution-Channel Attention (PCCA) is to output edge saliency map in supervised manner. For edge ground truth, edge map from saliency ground truth is obtained from Sobel operator. The rational behind this training strategy is to focus on boundary details of salient object only. As shallow features extracted from CNN branch of Conformer network are rich in color, texture and edge details, therefore, for multi-scale object's boundary detection a number of convolution and pooling operations are performed with different kernel sizes as presented in Fig. 2 and 3 and given by equation 1. The various receptive fields obtained by different kernel sizes ($3 \times 3, 5 \times 5$ and 7×7) helps to identify multiple salient objects with different sizes. Moreover, pooling operations helps to capture salient semantics globally. These operations are performed on $F_{c\_s} \in \mathbb{R}^{C\times H\times W}$ (C, H, W are channels, height and width respectively) in parallel manner.

Optical remote sensing images exhibit significant variations in object scales and highly diverse contextual backgrounds, making it challenging to accurately capture the edge details of salient objects. Inspired by [45], multi-scale objects are extracted using multiple convolution kernel sizes. Furthermore, the shuffled channel attention mechanism enhances the diversity of local contextual information. To model multi-scale channel dependencies, we deploy five parallel shuffle channel attention modules with different group sizes: smaller groups capture fine-grained features of

TABLE 2. Parallel convolution-channel attention (PCCA).

CNN Branch				Transformer Branch			
Input Feature Map	Output Feature Map	Operations	Remarks	Input Feature Map	Output Feature Map	Channel Attention(CA)	Remarks
F_{c_s}	f_{ce_1}	Convolution, kernel size (7×7)	Broader details	F_{t_s}	Out_{CA_4}	Shuffle channel attention with group 4	Limited channel mixing
F_{c_s}	f_{ce_2}	Convolution, kernel size (5×5)	Mid-level details	F_{t_s}	Out_{CA_8}	Shuffle channel attention with group 8	Less channel mixing
F_{c_s}	f_{ce_3}	Convolution, kernel size (3×3)	Fine-grained details	F_{t_s}	Out_{CA_16}	Shuffle channel attention with group 16	Moderate channel mixing
F_{c_s}	f_{ce_4}	Average Pooling, kernel size (4×4)	Smoother representation	F_{t_s}	Out_{CA_32}	Shuffle channel attention with group 32	Large channel mixing
F_{c_s}	f_{ce_5}	Max Pooling, kernel size (4×4)	Strongest features	F_{t_s}	Out_{CA_64}	Shuffle channel attention with group 64	Very large channel mixing

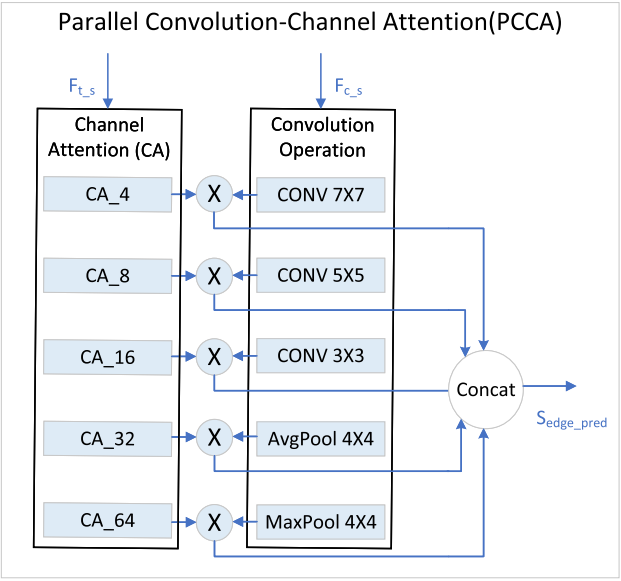

FIGURE 3. Architecture of parallel convolution-channel attention (PCCA).

small objects, while larger groups encode broader contextual information. The channel shuffle operation ensures cross-group interactions, improving feature diversity and saliency representation. This design is inspired by [46] and multi-scale attention mechanisms [44]. In Table 2, we have further summarized these hyperparameters.

$$f_{ce\_i} = Conv_{k\times k}(\mathrm{F_{c\_s}}), \tag{1}$$

where, k and i represents kernel size and index respectively. We have used three kernel sizes $k \in \{3, 5, 7\}$ to get three feature maps at index $i \in \{1, 2, 3\}$.

$$f_{ce\_4} = AvgPool_{4\times4}(\mathrm{F_{c\_s}}). \tag{2}$$

The resultant feature map from the fourth operation is represented by f_{ce_4}. Average Pooling is used to preserve the overall distribution given by equation 2.

$$f_{ce\_5} = MaxPool_{4\times4}(\mathrm{F_{c\_s}}). \tag{3}$$

Max Pooling operation keeps the strongest activation presented as f_{ce_5} in equation 3. The complementarity of these operations will be fused with channel attended features as given below.

The architecture used to process shallow features from Transformer branch is presented in Fig. 4. To differentiate the low-level details of salient and non-salient objects, global context from shallow features of Transformer branch is utilized. Furthermore, shuffled channel attention with different groups sizes are also applied in parallel manner. The intuition behind this is to emphasize stronger channels form global semantics representations. To do so, the shallow features $\mathrm{F_{t\_s}}$ from transformer branch having dimension $(K + 1) \times E$ are converted to dimension $C \times W \times H$. First we split the input $\mathrm{F_{t\_s}}$ into class token and patch embedding as in equation 4.

$$\mathrm{F}_{t\_s,cls} = \mathrm{F_{t\_s,0}} \in \mathbb{R}^{1\times E}, \quad \mathrm{F}_{t\_s,patch} = \mathrm{F_{t\_s,1:}}_K \in \mathbb{R}^{K\times E} \tag{4}$$

To obtain grid representation from token embeddings, we ignore the class token. Assuming $K = H \times W$, we reshape the $\mathrm{F}_{t\_s,patch}$ using equation 5

$$\mathrm{F}^{reshape}_{t\_s,patch} = \mathrm{reshape}(\mathrm{F}_{t\_s,patch}, (H, W, E)). \tag{5}$$

After applying linear projection as in equation 6, we obtain $f_{tg\_s} \in \mathbb{R}^{C\times H\times W}$ feature map.

$$f_{tg\_s} = \sum_{e=1}^{E} \mathrm{F}^{reshape}_{t\_s,patch}(h, w, e) \cdot W(e, c), \tag{6}$$

where $W \in \mathbb{R}^{E\times C}$. Global average pooling (GAP) is applied on f_{tg_s} which suppress the spatial details and emphasize the strong activation along channels presented in equation 7.

$$z(c) = \frac{1}{HW} \sum_{h=1}^{H} \sum_{w=1}^{W} f_{tg\_s}(c, h, w), \quad c = 1, \ldots, C \tag{7}$$

Reshaping $z \in \mathbb{R}^{C\times1\times1}$ into groups g using equation 8 and then apply channel shuffling using equation 9. These equations implement channel shuffling. Equation 8 reshapes the feature tensor z into g groups, each containing $\frac{C}{g}$ channels, and equation 9 transposes this tensor to mix channels across groups, producing $z_{shuffled}$ of shape $\frac{C}{g} \times g$ for better inter-group information flow. Table 2, further clarifies the choice of g: smaller values of g allow limited channel mixing,

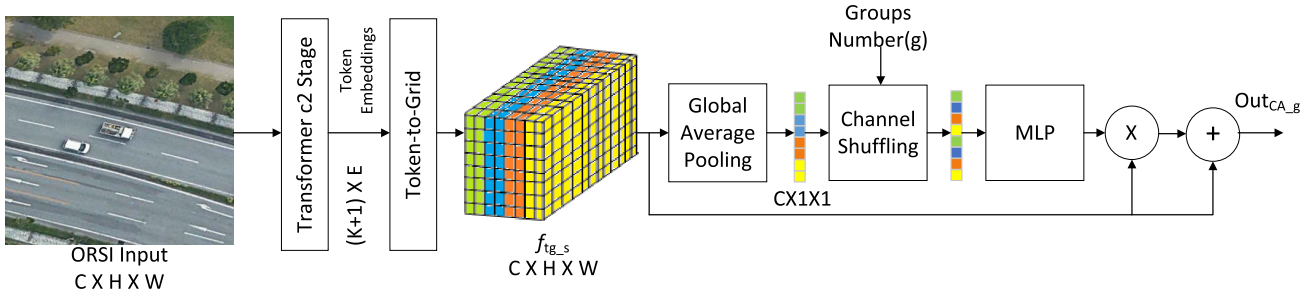

FIGURE 4. Transformer shallow features processing using shuffled channel attention. Transformed features are represented by F_{t_s}, g represent group number and CA_g represent channel attention with g number of groups.

while larger values enable more extensive mixing, which results in enhanced feature representation through shuffle channel attention. Using multiple values of g in parallel (e.g., 4, 8, 16, 32, 64) allows the network to capture both fine-grained local interactions (with small g) and broader global interactions (with large g) simultaneously, providing a richer and more diverse feature embedding that improves overall model performance.

$$z = \text{reshape}(z, (g, \frac{C}{g})) \tag{8}$$

$$z_{shuffled} = \text{transpose}(z) \in \mathbb{R}^{\frac{C}{g} \times g} \tag{9}$$

Again $z_{shuffled}$ is reshaped to dimension $C \times 1 \times 1$

$$z_{shuffled} = \text{reshape}(z_{shuffled}, (C,)) \tag{10}$$

$z_{shuffled}$ is forwarded to MLP block comprises of two fully connected layers with non-linearity is given by equation 11.

$$s = \sigma\left(W_2\, \delta\left(W_1 \mathbf{z}_{shuffled}\right)\right). \tag{11}$$

Here, $W_1 \in \mathbb{R}^{\frac{C}{r} \times C}$ and $W_2 \in \mathbb{R}^{C \times \frac{C}{r}}$. And δ, σ and r represents ReLU, Sigmoid and reduction ratio respectively. Then we apply these attention weights channel wise using equation 12.

$$\text{Out}_{\text{CA\_g}} = s \odot f_{tg\_s} + f_{tg\_s}. \tag{12}$$

The number of groups g in proposed model are {4,8,16,32,64}, which enables the cross-group information exchange and learns diverse and specialized features within each group, improving representation without excessively increasing parameters.

After extracting five spatial representation f_{ce_i} and five channel attended representations $\text{Out}_{\text{CA\_g}}$, the next step is to select complementarity between these features using equation 13. This process identifies mutually reinforcing information across spatial and channel features, while filtering out redundancy and suppressing the noise typically present in shallow layers. To integrate complementary spatial cues from CNN features with the global contextual representations of the Transformer, we perform an element-wise (pointwise) product. This pointwise multiplication functions as a context-aware gating mechanism, allowing global semantic cues to modulate the local feature responses so that salient objects of varying scales can be more effectively captured.

$$\begin{aligned} S_{\text{edge\_pred}} = &Concat((\text{Out}_{\text{CA\_4}} \odot f_{ce\_1}), \\ &(\text{Out}_{\text{CA\_8}} \odot f_{ce\_2}), (\text{Out}_{\text{CA\_16}} \odot f_{ce\_3}), \\ &(\text{Out}_{\text{CA\_32}} \odot f_{ce\_4}), (\text{Out}_{\text{CA\_64}} \odot f_{ce\_5})), \end{aligned} \tag{13}$$

here, Concat represents concatenation operation. This stacks all 5 fused feature maps along the channel dimension, producing a single large enriched feature tensor. The resultant feature map $S_{\text{edge\_pred}}$ provides edge prediction map and is supervised using smooth L1 loss.

C. HOLISTIC REVERSE ATTENTION (HRA)

The high-level features representations extracted from CNN and Transformer branches are rich in global information and can play a vital role in locating the salient regions. The last layer from both streams of Conformer Network generate coarse saliency map using binary cross-entropy loss. The coarse prediction from CNN and Transformer streams are obtained using equation 14 and 15, respectively.

$$\text{F}_{\text{c\_sal}} = ReLU(Conv_{1\times1}(\text{F}_{\text{c\_d4}})). \tag{14}$$

$$\text{F}_{\text{t\_sal}} = ReLU(Conv_{1\times1}(\text{F}_{\text{t\_d4}})). \tag{15}$$

The coarse prediction provides the contextual reference for applying reverse attention. This holistic reverse attention is applied on three deep features $\text{F}_{\text{c\_di}}$ and $\text{F}_{\text{t\_di}}$ from dual branch backbone network as represented in Fig. 2 and the detailed architecture of HRA module is illustrated in Fig. 5. The deepest backbone features are at index i=3, extracted from c4-Stage of Conformer carrying rich global semantics details. Additionally, we exploit c3- and c2-stage features (indexed as i=2 and i=1), which preserve medium- and fine-level details.

A transformation operation is performed to obtain single channel feature map from $\text{F}_{\text{c\_di}}$ and $\text{F}_{\text{t\_di}}$ using equations 16 and 17.

$$f_{\text{c\_di}} = Conv_{1\times1}(ReLU(Conv_{1\times1}(\text{F}_{\text{c\_di}}))). \tag{16}$$

$$f_{\text{t\_di}} = Conv_{1\times1}(ReLU(Conv_{1\times1}(h(\text{F}_{\text{t\_di}})))), \tag{17}$$

where, h is similar reshaping function as provided in equation 4 and 5. Subsequently, an upsampling operation

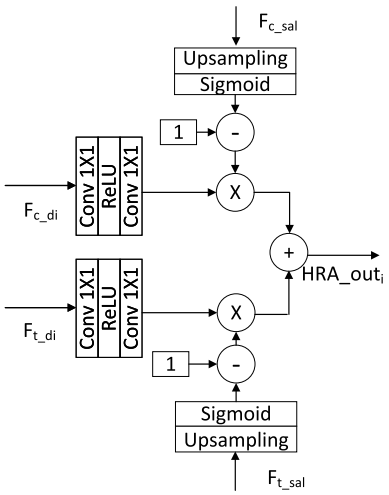

FIGURE 5. Deep hybrid CNN-transformer features processed using holistic reverse attention.

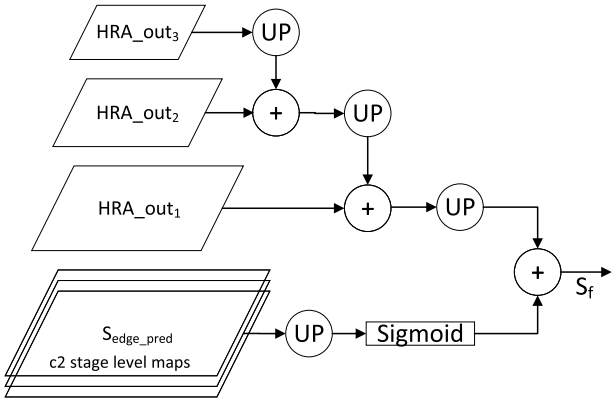

FIGURE 6. Feature pyramid network.

is applied to the coarse saliency prediction to align it with the corresponding backbone stage resolution, followed by a Sigmoid function, which serves as a gating mechanism for holistic reverse attention. Finally, obtain attention guided feature maps from CNN and Transformer streams using equations 18 and 19.

$$\mathrm{HRA}_{\mathrm{c\_di}} = f_{\mathrm{c\_di}} \odot (1 - (\mathit{Sigmoid}(\mathit{UpSample}(\mathrm{F}_{\mathrm{c\_sal}})))). \tag{18}$$

$$\mathrm{HRA}_{\mathrm{t\_di}} = f_{\mathrm{t\_di}} \odot (1 - (\mathit{Sigmoid}(\mathit{UpSample}(\mathrm{F}_{\mathrm{t\_sal}})))). \tag{19}$$

To select the commonalities from two features $\mathrm{HRA}_{\mathrm{c\_di}}$ and $\mathrm{HRA}_{\mathrm{c\_di}}$, element-wise addition operation is performed, which merges the information from dual streams.

$$\mathrm{HRA\_out}_{\mathrm{i}} = \mathrm{HRA}_{\mathrm{t\_di}} + \mathrm{HRA}_{\mathrm{c\_di}}. \tag{20}$$

D. FEATURE PYRAMID NETWORK

The architecture used in FPN is shown in Fig. 6. We combine the shallow and deep features in Feature Pyramid Network (FPN) module. Shallow features comprise of edge saliency information and deep features capture saliency attentive maps. These features are illustrated in Fig. 7.

We obtain edge details of all prominent objects represented as S_{edge_pred}. The contextual saliency at highest level contains coarse salient objects predictions. At mid level more information about object shape is added. At low level contextual prediction is less precise. The high level features are upsampled and combined with adjacent low level feature. Therefore, progressive refinement from top to bottom only contain statutory salient regions. The final saliency map $\mathrm{S_f}$ is obtained using equation 21.

$$\mathrm{S_f} = \mathit{Sigmoid}(\mathit{UP}(\mathrm{S_{edge\_pred}})) + \mathit{UP}(\mathrm{HRA\_out_1} + \mathit{UP}(\mathrm{HRA\_out_2} + \mathit{UP}(\mathrm{HRA\_out_3})))). \tag{21}$$

E. LOSS FUNCTION

Total loss is defined in equation 22.

$$\mathcal{L}_{\mathrm{total}} = \mathcal{L}_{\mathrm{f}}(\mathrm{S_f}) + \mathcal{L}_{\mathrm{c}}(\mathrm{F_{c\_sal}}) + \mathcal{L}_{\mathrm{t}}(\mathrm{F_{t\_sal}}) + \mathcal{L}_{\mathrm{e}}(\mathrm{S_{edge\_pred}}). \tag{22}$$

where, $\mathcal{L}_{\mathrm{f}}$, $\mathcal{L}_{\mathrm{c}}$ and $\mathcal{L}_{\mathrm{t}}$ are binary cross entropy loss. $\mathcal{L}_{\mathrm{e}}$ is smooth L1 loss.

IV. EXPERIMENT

A. EXPERIMENT PROTOCOL

1) DATASETS

ORSSD [24] is the first optical RSI dataset comprising of 800 ORSI samples (600 train images + 200 test images) with pixel-wise annotations. Published in 2019, it has become a widely used benchmark dataset for the ORSI-SOD task in research. ORSSD comprises a diverse set of salient objects in optical remote-sensing images, including ships, cars, airplanes, playgrounds, rivers, islands, bridges, and more. The images, sourced from Google Earth and other remote-sensing platforms, feature a combination of natural and man-made structures. The dataset poses challenges such as cluttered and complex backgrounds, varying object scales, and a wide range of object types. EORSSD [25] dataset consists of 2000 image samples (1400 train images + 600 test images). EORSSD features a wide variety of salient objects, such as buildings, roads, ships, aircraft, cars, water bodies, and islands. Analysis of the dataset shows that ships and aircraft are the most prevalent, comprising roughly 22.2% and 21.5% of the images, respectively. A key challenge in EORSSD is detecting small objects, as many salient targets occupy only a small portion of the image. In fact, 84.65% of the scenes contain objects covering less than 10% of the image area, and 39% feature objects occupying less than 1%. ORSI-4199 [14] dataset has 4199 images (2000 train images + 2199 test images). These datasets are diverse in nature including multiple spatial resolutions, object(s) with cast shadow, multiple categories of objects, terrain of various kinds, and varied lightning conditions etc. We evaluate our proposed model on these datasets. To remain consistent with

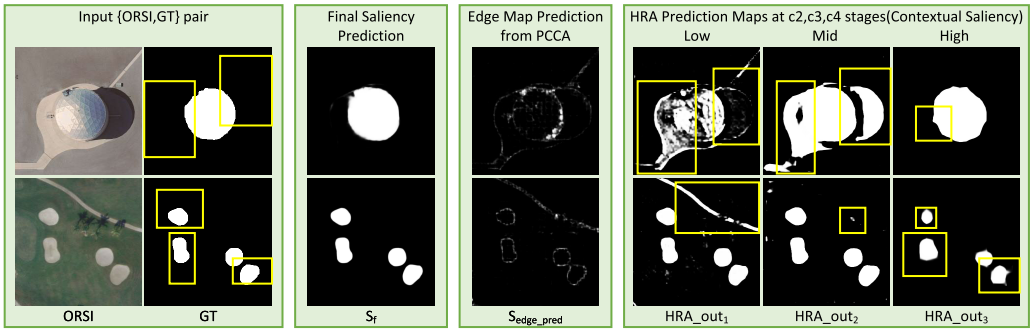

FIGURE 7. Edge and semantic cues at various stages prior to the application of the feature pyramid network (FPN). The bounding boxes represents redundant or missing details.

prior work, we choose train samples of ORSSD and EORSSD datasets for training. We applied flipping and rotation to augment the training datasets. Specifically, each original sample was transformed to produce seven additional samples. This not only increases the total number of training samples (11200 train samples of EORSSD) but also improves the model's robustness. For optical remote sensing images, object locations and sizes can vary significantly and are not always centered. Therefore, these augmentation strategies help the model generalize better to diverse spatial distributions

2) EVALUATION METRICS

For quantitative evaluation and comparison with competing models following five metrics are employed: Structural measure (Smeasure) [26], maximum F-measure (Fmax) [27], maximum E-measure (Emax) [28], Mean Absolute Error (MAE) [29], and P-R curve under different combination of precision and recall scores [30], [31]

3) IMPLEMENTATION DETAILS

The machine used to train proposed model is equipped with a single NVIDIA Tesla P100 GPU with 16 GB of memory, and an Intel® Xeon® CPU @2.00 GHz with 32 GB of RAM. We adopt PyTorch frame work to implement proposed model. Before training, images are resized to 384×384 resolution. The Conformer backbone is initialized using a pre-trained model trained on the ImageNet dataset, while all remaining parameters are initialized with PyTorch's default settings. Model use batch size of 4. We use Adam optimizer with learning rate of 5e-5 and weight decay of 5e-4. We implement ReduceLROnPlateau scheduler to adaptively lower the learning rate by half if loss does not improve by 5 epochs. The propose model converges after 50 epochs.

B. COMPARISON WITH STATE-OF-THE-ART MODELS

We conduct quantitative as well as qualitative comparison of proposed model with 17 state-of-the-art RSI-SOD models, including SggNet [8], BCARNet [15], PRNet [10], DSINet [11], SOLNet [12], MJRBM [14], ASNet [16], CSI-RF [9], EMHANet [17], MCIC [18], HFCNet [19], ISAANet [20], CorrNet [21], MEANet [22], LVNet [24], DAFNet [25] and MMWNet [23]. The saliency maps and quantitative results of these methods are provided by respective authors.

1) QUANTITATIVE COMPARISON

The quantitative comparison of proposed approach with the SOTA ORSI-SOD models on two benchmark datasets is provided in Table 3. It is evident that our proposed model outperforms existing approaches on all four metrics. Specifically, our proposed model shows 0.4% improvement in $Average - S_\alpha$ measure, 0.25% improvement in $Average - F_\beta^m$ measure and 11.57% decrease in mean absolute error value over two datasets. We have also presented F-measure curves under 255 thresholds and Precision-Recall (PR) curves of proposed and existing models in Fig. 8. The F-measure curve of our proposed model illustrates a large area under the curve, which signifies better outcome as compared to SOTA models. Furthermore, the graphical analysis of PR curves demonstrates that our proposed model achieves strong performance. Improved performance is shown in the PR curve by closeness to the top-right corner.

2) STATISTICAL ANALYSIS OF MODEL PERFORMANCE

To assess whether the observed quantitative improvements of our proposed model,provided in Table 3, are statistically significant compared to existing baseline methods, we performed paired t-tests and Wilcoxon signed-rank tests using per-image MAE values across 200 images from the ORSSD dataset. The results are presented in Table 4. The paired t-test evaluates whether the mean difference in MAE between our proposed model and a baseline model is statistically different from zero. A p-value below 0.05 (in several cases such as BCARNet, DSINet, LGIPNet and PRNet) indicates that the mean performance difference is significant. The Wilcoxon signed-rank test is a non-parametric alternative that does not assume normality. It evaluates whether the median difference in MAE across paired samples is zero, and is sensitive to consistent differences across individual images. A p-value below 0.05 (in all rows) indicates that our model consistently outperforms the baseline, even if the mean improvement is small.

TABLE 3. Quantitative comparison of ORSI-SOD models on ORSSD and EORSSD datasets. The best results are highlighted in **red**, the second-best results in **blue**, and a '-' indicates that the result is not available.

Method	Year	Publication	Backbone	ORSSD				EORSSD			
				$S_\alpha\uparrow$	$F_\beta^m\uparrow$	$E_\zeta^m\uparrow$	MAE↓	$S_\alpha\uparrow$	$F_\beta^m\uparrow$	$E_\zeta^m\uparrow$	MAE↓
LVNet	2019	IEEE TGRS	-	0.8815	0.8414	-	0.0207				
DAFNet	2020	IEEE TIP	VGG-16	0.9191	0.9174	-	0.0125	0.9167	0.8922	-	0.0060
MJRBM	2021	IEEE TGRS	VGG-16	-	0.9025	-	0.0145	-	0.8701	-	0.0099
CorrNet	2022	IEEE TGRS	LFE-VGG	0.9380	0.9129	0.9790	0.0098	0.9289	0.8778	0.9696	0.0083
CSI-RF	2023	IEEE GRSL	Res2Net-50	0.9445	0.9183	0.9813	0.0084	0.9342	0.8843	0.9763	0.0061
SOLNet	2024	IEEE TGRS	VGG-16	0.9284	0.8946	0.9689	0.0111	0.9171	0.8513	0.9577	0.0078
MMWNet	2024	IEEE TGRS	VGG-16	0.9481	0.9188	0.8917	0.0075	0.9384	0.8907	0.9756	0.0063
PRNet	2024	IEEE TGRS	EfficientNetB4	0.9468	0.9013	0.9824	**0.0068**	0.9269	0.8510	0.9732	**0.0053**
ASNet	2024	IEEE TGRS	Hybrid	0.9441	0.9172	0.9803	0.0081	0.9345	0.8959	0.9783	0.0055
MCIC	2024	IEEE TGRS	VGG-16	0.9433	0.9135	0.9801	0.0090	0.9373	0.8868	0.9765	0.0070
HFCNet	2024	IEEE TGRS	VGG-16	**0.9521**	0.9247	**0.9885**	0.0073	0.9407	0.8864	0.9793	0.0054
ISAANet	2024	IEEE TGRS	MobileNet-V3	0.9429	0.9166	0.9836	0.0077	0.9351	0.8869	0.9761	0.0063
DSINet	2024	VC	Hybrid	-	**0.9262**	0.9881	0.0080	**0.9507**	0.8947	**0.9841**	0.0056
MEANet	2024	ESWA	MobileNet-V2	0.9340	0.8934	0.9730	0.0098	0.9282	**0.9658**	0.9282	0.0070
SggNet	2025	RS	MobileNet-V2	0.9342	0.9032	0.9759	0.0111	0.9279	0.8770	0.9762	0.0068
BCARNet	2025	PR	ResNet-50	0.9466	0.9082	0.9833	0.0071	0.9360	0.8794	0.9761	0.0057
EMHANet	2025	IEEE Access	MobileNet-V3	0.9342	0.8827	0.9756	0.0096	0.9274	0.8552	0.9685	0.0069
Ours	2025	IEEE Access	Conformer	**0.9585**	**0.9308**	**0.9898**	**0.0061**	**0.9517**	**0.9660**	**0.9844**	**0.0046**

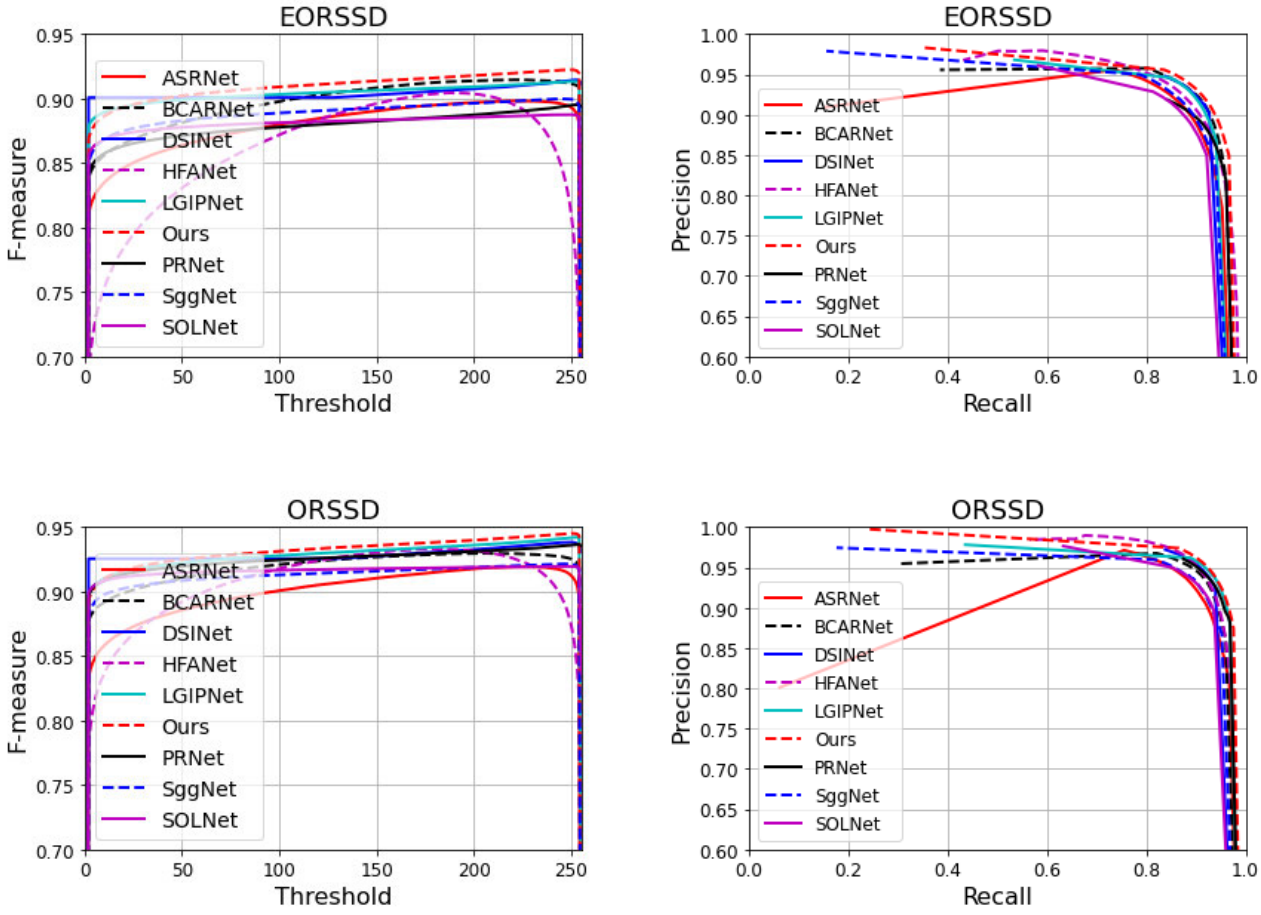

FIGURE 8. F-measure curves (first column) and Precision-Recall (PR) curves (second column) across ORSSD and EORSSD datasets.

TABLE 4. Statistical analysis of model performance. *Significance is determined at $p < 0.05$.

Baseline Model	Paired t-test (t-stat, p-value)	Wilcoxon Test (W, p-value)	Significance*
ASRNet	-0.740, 0.460	8342.0, 0.0372	Significant (**Wilcoxon**)
BCARNet	-2.476, 0.0141	3293.0, 1.66e-16	Significant (**Both**)
DSINet	-2.206, 0.0285	4884.0, 2.91e-10	Significant (**Both**)
HFANet	-1.376, 0.170	6270.0, 3.98e-06	Significant (**Wilcoxon**)
LGIPNet	-2.180, 0.0305	4260.0, 1.61e-12	Significant (**Both**)
PRNet	-2.210, 0.0283	3583.0, 3.00e-15	Significant (**Both**)
SOLNet	-0.628, 0.531	7265.0, 0.000678	Significant (**Wilcoxon**)

3) QUALITATIVE COMPARISON

For qualitative comparison, various challenging scenarios from published work [14], [15], [16], [17], [18], [19], [20] are selected. In Fig. 9, visual representation of the selected challenging scenes are presented.

a. **Irregular topology:** Specifically, irregular topology is key feature of optical remote sensing images. We have presented two examples of irregular salient region structure in first two rows of Fig. 9. The complete structure is captured by our proposed model while existing SOTA models fail to do so.
b. **Multiple small objects in cluttered background:** Another challenging scene is illustrated in third row of Fig. 9, where there are multiple small objects in cluttered background. Our proposed model identified the context saliency and captured the relevant details.
c. **Objects with low contrast:** Similarly, in fourth and fifth rows of Fig. 9, our proposed model captures the salient object in low contrast scenarios very effectively.
d. **Object with shadow:** Object with shadow is presented in sixth row of Fig. 9, where our proposed model outperforms existing approaches.
e. **Large object:** In seventh row of Fig. 9, the selected sample consist of large object with internal color variation. As compared to existing models, our proposed model capture the whole object without any gaps.
f. **Objects characterized by intra-object color variation:** In eighth row of Fig. 9, an object containing heterogeneous color regions is presented. In this challenging scenario, our proposed model captures the diverse color patterns and outperforms the SOTA models.
g. **Objects of different scales:** Another challenging example having multiple objects with different scales are presented in last row of Fig. 9. Here, SOTA models fail to capture all salient objects, whereas, our proposed scheme recognizes all salient objects.

C. ABLATION STUDIES

To evaluate the effectiveness of various modules of proposed model, we carry out a comprehensive ablation study on ORSSD dataset. In particular, we analyze the effectiveness of treating shallow and deep features distinctively. We provide ablation studies to analyze: 1) The contribution of Parallel Convolution-Channel Attention (PCCA) for edge guidance using only shallow features. 2) The contribution of Holistic Reverse Attention (HRA) for contextual saliency using deep features and 3) The contribution of Conformer Network as backbone feature extractor. The experimental setup, including parameters and datasets, follows Section IV.

TABLE 5. Quantitative results about effectiveness of edge guidance using shallow features. The best results are highlighted in red.

Method	ORSSD Dataset			
	$S_\alpha\uparrow$	$F_\beta^m\uparrow$	$E_\zeta^m\uparrow$	MAE$\downarrow$
Ours without PCCA	0.9293	0.9264	0.9833	0.0121
Ours	**0.9585**	**0.9308**	**0.9898**	**0.0061**
Method	EORSSD Dataset			
	$S_\alpha\uparrow$	$F_\beta^m\uparrow$	$E_\zeta^m\uparrow$	MAE$\downarrow$
Ours without PCCA	0.922	0.9616	0.977	0.0106
Ours	**0.9517**	**0.966**	**0.9844**	**0.0046**

1) ABLATION STUDY DEMONSTRATING THE CONTRIBUTION OF PARALLEL CONVOLUTION-CHANNEL ATTENTION (PCCA) FOR EDGE GUIDANCE USING ONLY SHALLOW FEATURES

To study the contribution of Parallel Convolution-Channel Attention (PCCA) and the effectiveness of edge guidance using low level details, we train a variant of proposed model, in which PCCA is removed. The quantitative results presented in Table 5 on ORSSD show that removing PCCA results in performance decline in all four evaluation metrics. Specifically, 3.05% drop in S_α, 0.47% drop in F_β^m and 0.66% drop in E_ζ^m. Moreover, 98% increase in mean absolute error also indicate degraded performance. Additionally, results on EORSSD dataset show performance decline of 3.2%\0.45% in $S_\alpha \backslash F_\beta^m$ metrics. E_ζ^m value also dropped from 0.9844 to 0.977 without PCCA module. Pixel-level inaccuracies become more prominent without PCCA, as evidenced by a 56.6% increase in MAE in EORSSD.

The visualization results given in Fig. 10, show that edge guidance plays a critical role in obtaining accurate saliency map.

2) ABLATION STUDY DEMONSTRATING THE CONTRIBUTION OF HOLISTIC REVERSE ATTENTION (HRA) FOR CONTEXTUAL SALIENCY USING DEEP FEATURES

Deep features extracts the global semantics of an image, thereby, highlighting the contextual saliency only. Our proposed model apply Holistic Reverse Attention (HRA) to accomplish this using high level features. To demonstrate the effectiveness of HRA, we obtain an ablated version of proposed model by removing HRA module. The results presented in Table 6 show the contribution of HRA in proposed strategy. Significant decrease up to 6.4% in $AverageS_\alpha$, 4.2% in $AverageF_\beta^m$ and 3% in $AverageE_\zeta^m$ is observed in the ablated variant. Furthermore, increase in the pixel-level inaccuracy in the ablated version is quantified by

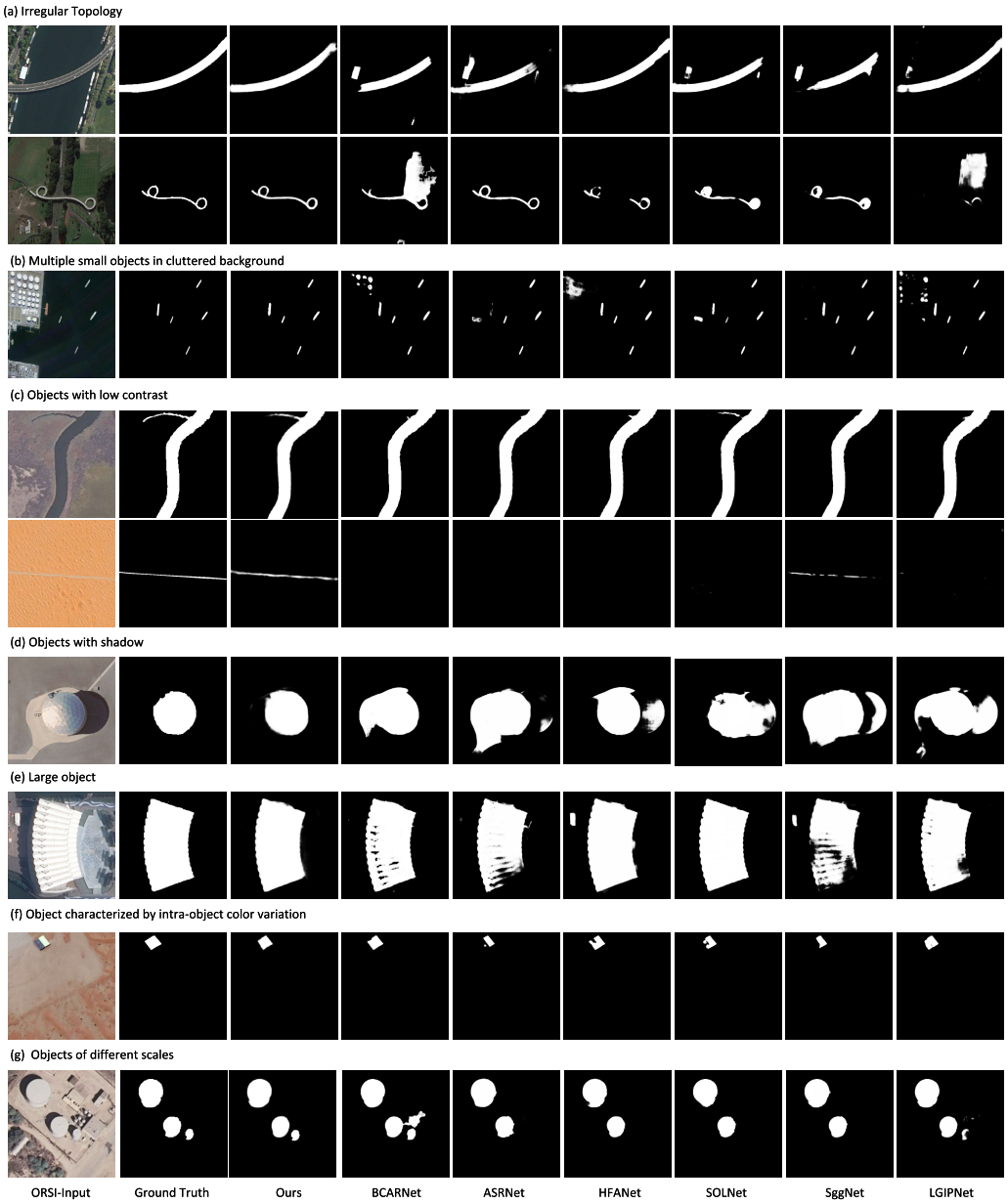

FIGURE 9. Visual comparison of our proposed model and SOTA models.

TABLE 6. Quantitative results about effectiveness of contextual saliency using deep features. The best results are highlighted in red.

Method	ORSSD Dataset			
	$S_\alpha\uparrow$	$F_\beta^m\uparrow$	$E_\zeta^m\uparrow$	MAE$\downarrow$
Ours without HRA	0.9007	0.8926	0.9618	0.0206
Ours	**0.9585**	**0.9308**	**0.9898**	**0.0061**
Method	EORSSD Dataset			
	$S_\alpha\uparrow$	$F_\beta^m\uparrow$	$E_\zeta^m\uparrow$	MAE$\downarrow$
Ours without HRA	0.8939	0.9278	0.9564	0.0191
Ours	**0.9517**	**0.9660**	**0.9844**	**0.0046**

increase in MAE from 0.0061 to 0.0206 in ORSSD dataset and 0.0046 to 0.0191 in EORSSD dataset. Fig. 11 shows that without HRA, model is unable to predict accurately.

3) ABLATION STUDY ABOUT CONTRIBUTION OF CONFORMER NETWORK AS BACKBONE FEATURE EXTRACTOR

To demonstrate the role of Conformer Network as backbone feature extractor, the heat maps extracted at different stages of Conformer network is illustrated in Fig. 12. The CNN stream extracts low-level details while Transformer stream extracts high-level details. Furthermore, the response of both streams at shallow and deep features are distinct which verifies the rational behind processing them separately. The deepest feature map obtained from Transformer stream emphasizes only on salient region. Shallow features in the CNN stream highlight edge information, while deeper features concentrate on salient regions and suppress shadow influence. At the

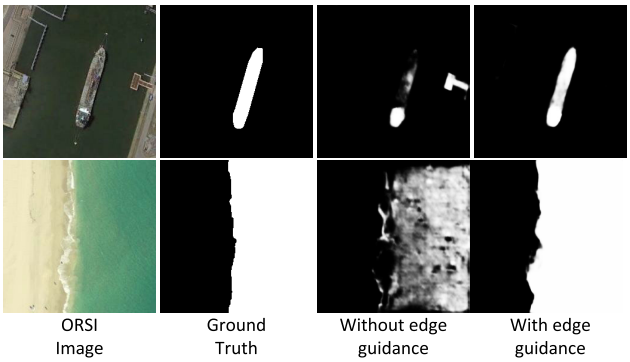

FIGURE 10. Ablation study about edge guidance using only shallow features.

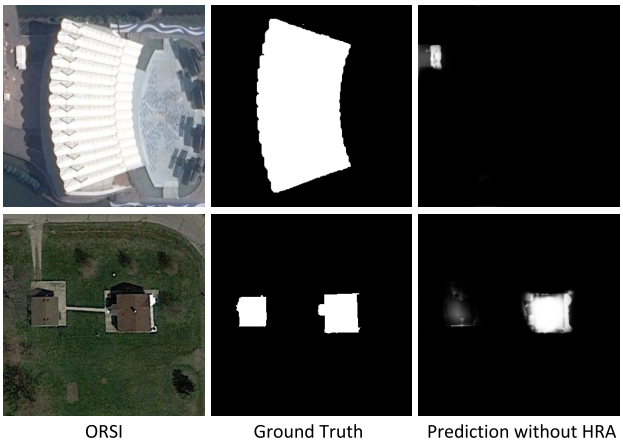

FIGURE 11. Ablation study about contextual saliency using deep features.

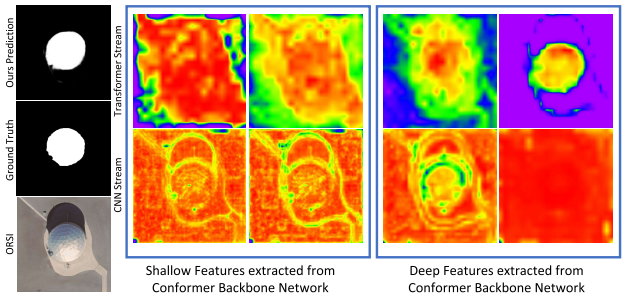

FIGURE 12. Ablation study about conformer network as backbone feature extractor.

deepest level, CNN representations become less effective, as the model inherently emphasizes local details rather than global context.

4) REPLACEMENT OF HOLISTIC REVERSE ATTENTION (HRA) WITH CONVOLUTIONAL BLOCK ATTENTION MODULE (CBAM)

HRA is a task-specific attention module designed to focus on missing regions in the saliency map. To evaluate the effectiveness of HRA, we train a variant of the proposed model in which HRA is replaced with CBAM, while keeping all other settings unchanged. The results are presented in Table 7. It is observed that *Average*S_α is decreased upto 2.39%, *Average* F_β^m upto 1.4% and *Average* E_ζ^m upto 1.17% in the ablated variant. Furthermore, 38% increase is observed in average mean absolute error.

TABLE 7. Quantitative results about replacement of holistic reverse attention (HRA) with convolutional block attention module (CBAM). The best results are highlighted in red.

Method	ORSSD Dataset			
	$S_\alpha\uparrow$	$F_\beta^m\uparrow$	$E_\zeta^m\uparrow$	MAE$\downarrow$
HRA replaced with CBAM	0.9317	0.9236	0.9738	0.0093
Ours	**0.9585**	**0.9308**	**0.9898**	**0.0061**
Method	EORSSD Dataset			
	$S_\alpha\uparrow$	$F_\beta^m\uparrow$	$E_\zeta^m\uparrow$	MAE$\downarrow$
HRA replaced with CBAM	0.9339	0.9468	0.9774	0.008
Ours	**0.9517**	**0.9660**	**0.9844**	**0.0046**

V. DISCUSSION AND LIMITATIONS

Tailored for ORSI-SOD, our method excels at managing the distinct complexities present in optical remote sensing images. Various aspects of the proposed approach underscore its robustness and efficacy. Along with that we have also provided accuracy vs efficiency tradeoff and failure case analysis in this section.

A. LEVERAGING CONFORMER FOR RICH FEATURE EXTRACTION

Our approach is motivated by the need to separately exploit shallow and deep features because shallow features encode local structural details, whereas deep features capture broader, high-level semantic information. Standard CNN backbones often underperform in extracting effective deep features, which has led many models to adopt hybrid architectures [11], [16]. In contrast, we utilize a Conformer with feature coupling units, enabling efficient capture of both shallow and deep representations without requiring a hybrid network.

B. EDGE-GUIDED ATTENTION FOR ENHANCED PRECISION

Many existing models leverage edge-guided attention to improve boundary preservation and enhance saliency detection [15], [16], [19]. Unlike standard edge-guided modules, in our proposed PCCA, transformer-extracted features from the Conformer are used for channel attention, providing global structural context of salient objects. Simultaneously, CNN-extracted features from the Conformer are employed for parallel convolution and pooling operations. Five context-weighted parallel operations are fused to produce the final edge-map. This design effectively integrates global and local information, enhancing the robustness and accuracy of our model.

TABLE 8. Performance trade-off analysis. The best results are highlighted in red.

Method	Backbone	Params (M) ↓	FLOPs (G) ↓	Inference speed (FPS) ↑	Avg-S_α↑	Avg-MAE↓
CorrNet	VGG-16	4.09	21.09	100	0.933	0.00905
MEANet	MobileNetV2	3.27	9.62	115	0.933	0.0084
Ours	Conformer	90.2	200.34	33	0.9551	0.0053

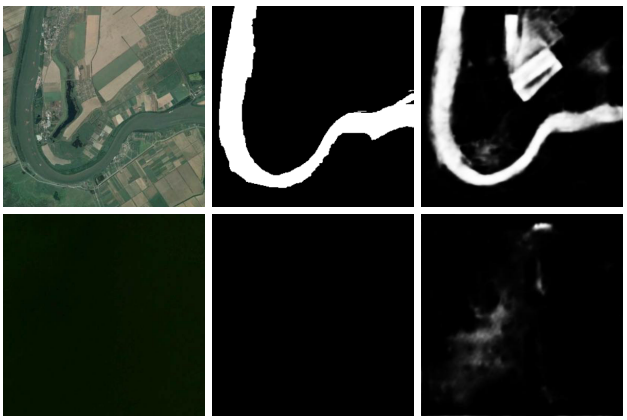

FIGURE 13. Failure cases.

C. GLOBAL SEMANTIC ENCODING

Salient objects in optical RSIs are often small, camouflaged, or scale-varying. The HRA module uses semantic context from deep features to boost the detection accuracy.

D. PERFORMANCE TRADE-OFF ANALYSIS

Although our proposed model achieves superior accuracy across multiple evaluation metrics, this performance improvement introduces a trade-off in terms of computational speed. In Table 8, we have provided a comprehensive comparison of complexity verses accuracy of proposed model with existing lightweight approaches. Our model requires 90.2 million parameters, 200.34G FLOPs and exhibits inference speed of 33 frames per second(FPS). Our proposed model shows 36.9% improvement in average mean absolute error and 2.37% increase in average s-measure, when compared with most efficient lightweight SOTA model. This performance–speed trade-off indicates that while our model is well-suited for accuracy-critical RSI-SOD applications. Further optimization,such as incorporating lightweight backbone, pruning, or knowledge distillation based strategies are needed to improve its practical usability.

E. FAILURE CASE ANALYSIS

In some optical remote sensing images either salient and non-salient objects are indistinguishable or salient object is completely absent. These two complex scenarios are presented in Fig. 13. Since our model is designed to capture intra-object color variations, as presented in Fig. 9. Consequently, the model struggles to differentiate between salient and non salient regions in first row of Fig. 13. In second row of Fig. 13, although the salient object is completely absent, however, proposed model captures the slight color variations, thereby producing wrong result. In future, we will address this issue by learning more abstract semantic information.

VI. CONCLUSION

This work introduce a novel approach to detect salient object(s) from optical remote sensing images. The proposed model overcomes the major shortcomings of existing methods by incorporating two novel modules: Parallel Convolution-Channel Attention (PCCA), designed to enhance shallow feature representation and Holistic Reverse Attention (HRA), designed to refine deep feature learning. Furthermore, we carefully chose the Conformer network as our backbone, which consists of dual streams—CNN and Transformer—and integrated a Feature Coupling Unit (FCU) after each stage to enhance the complementarities between the two streams. Our PCCA module defines the edge details of salient objects in supervised manner. While HRA is responsible for capturing global topological structure to emphasize salient region only. Our model shows average increase of ≈0.4% (S_α) and ≈0.26% (F_β^m) across two benchmark datasets. Moreover, an average decrease of ≈11.57% in mean absolute error is observed, showcasing the supremacy of proposed model. Further investigation may also consider multimodal integration, domain adaptation, and real-time deployment to broaden the applicability of the proposed approach.

REFERENCES

[1] C. Chen, M. Song, S. Pang, and C. Peng, "Adapting generic RGB-D salient object detection for specific traffic scenarios," *IEEE Trans. Intell. Transp. Syst.*, vol. 26, no. 8, pp. 12329–12343, Aug. 2025.

[2] L. Wei and Z. Zhu, "Modal-aware interaction network for RGB-D salient object detection," *IEEE Trans. Instrum. Meas.*, vol. 74, pp. 1–12, 2025.

[3] S. Kanwal and I. A. Taj, "Incomplete RGB-D salient object detection: Conceal, correlate and fuse," *Pattern Recognit.*, vol. 155, Nov. 2024, Art. no. 110700.

[4] X. Zhou, Z. Li, and T. Tong, "Medical image segmentation and saliency detection through a novel color contextual extractor," in *Proc. Int. Conf. Artif. Neural Netw.*, Feb. 2023, pp. 457–468.

[5] P. Zhang, T. Zhuo, W. Huang, K. Chen, and M. Kankanhalli, "Online object tracking based on CNN with spatial–temporal saliency guided sampling," *Neurocomputing*, vol. 257, pp. 115–127, Sep. 2017.

[6] G. Liu and D. Fan, "A model of visual attention for natural image retrieval," in *Proc. Int. Conf. Inf. Sci. Cloud Comput. Companion*, Dec. 2013, pp. 728–733.

[7] N. C. Garcia, P. Morerio, and V. Murino, "Cross-modal learning by hallucinating missing modalities in RGB-D vision," in *Proc. Multimodal Scene Underst.* Amsterdam, The Netherlands: Elsevier, 2019, pp. 383–401.

[8] J. Liu, J. He, H. Chen, R. Yang, and Y. Huang, "A lightweight Semantic- and graph-guided network for advanced optical remote sensing image salient object detection," *Remote Sens.*, vol. 17, no. 5, p. 861, Feb. 2025.

[9] 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, pp. 1–5, 2023.

[10] S. Gu, Y. Song, Y. Zhou, Y. Bai, X. Yang, and Y. He, "PRNet: Parallel refinement network with group feature learning for salient object detection in optical remote sensing images," *IEEE Geosci. Remote Sens. Lett.*, vol. 21, pp. 1–5, 2024.
[11] Y. Ge, T. Liang, J. Ren, J. Chen, and H. Bi, "Enhanced salient object detection in remote sensing images via dual-stream semantic interactive network," *Vis. Comput.*, vol. 41, no. 7, pp. 5153–5169, May 2025.
[12] Z. Li, Y. Miao, X. Li, W. Li, J. Cao, Q. Hao, D. Li, and Y. Sheng, "Speed-oriented lightweight salient object detection in optical remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 63, 2025, Art. no. 5601014.
[13] X. Zhang, Y. Yu, D. Li, and Y. Wang, "Progressive self-prompting segment anything model for salient object detection in optical remote sensing images," *Remote Sens.*, vol. 17, no. 2, p. 342, Jan. 2025.
[14] 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.
[15] Y. Gu, S. Chen, X. Sun, J. Ji, Y. Zhou, and R. Ji, "Optical remote sensing image salient object detection via bidirectional cross-attention and attention restoration," *Pattern Recognit.*, vol. 164, Aug. 2025, Art. no. 111478.
[16] R. Yan, L. Yan, G. Geng, Y. Cao, P. Zhou, and Y. Meng, "ASNet: Adaptive semantic network based on transformer—CNN for salient object detection in optical remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, 2024, Art. no. 5608716.
[17] Q. Tang, Z. Wang, X. Wang, and S.-W. Zhang, "EMHANet: Lightweight salient object detection for remote sensing images via edge-aware multiscale feature fusion," *IEEE Access*, vol. 13, pp. 89164–89178, 2025.
[18] K. Huang, N. Li, J. Huang, and C. Tian, "Exploiting memory-based cross-image contexts for salient object detection in optical remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, 2024, Art. no. 5614615.
[19] Y. Liu, M. Xu, T. Xiao, H. Tang, Y. Hu, and L. Nie, "Heterogeneous feature collaboration network for salient object detection in optical remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, 2024, Art. no. 5635114.
[20] Z. Yao and W. Gao, "Iterative saliency aggregation and assignment network for efficient salient object detection in optical remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, 2024, Art. no. 5633213.
[21] G. Li, Z. Liu, Z. Bai, W. Lin, and H. Ling, "Lightweight salient object detection in optical remote sensing images via feature correlation," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5617712.
[22] B. Liang and H. Luo, "MEANet: An effective and lightweight solution for salient object detection in optical remote sensing images," *Expert Syst. Appl.*, vol. 238, Mar. 2024, Art. no. 121778.
[23] L. Di, B. Zhang, and Y. Wang, "Multiscale and multidimensional weighted network for salient object detection in optical remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, 2024, Art. no. 5625114.
[24] 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.
[25] Q. Zhang, R. Cong, C. Li, M.-M. Cheng, Y. Fang, X. Cao, Y. Zhao, and S. Kwong, "Dense attention fluid network for salient object detection in optical remote sensing images," *IEEE Trans. Image Process.*, vol. 30, pp. 1305–1317, 2021.
[26] D.-P. Fan, M.-M. Cheng, Y. Liu, T. Li, and A. Borji, "Structure-measure: A new way to evaluate foreground maps," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 4558–4567.
[27] R. Achanta, S. Hemami, F. Estrada, and S. Susstrunk, "Frequency-tuned salient region detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 1597–1604.
[28] 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*.
[29] F. Perazzi, P. Krahenbuhl, Y. Pritch, and A. Hornung, "Saliency filters: Contrast based filtering for salient region detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2012, pp. 733–740.
[30] R. Cong, J. Lei, H. Fu, F. Porikli, Q. Huang, and C. Hou, "Video saliency detection via sparsity-based reconstruction and propagation," *IEEE Trans. Image Process.*, vol. 28, no. 10, pp. 4819–4831, Oct. 2019.
[31] R. Cong, J. Lei, C. Zhang, Q. Huang, X. Cao, and C. Hou, "Saliency detection for stereoscopic images based on depth confidence analysis and multiple cues fusion," *IEEE Signal Process. Lett.*, vol. 23, no. 6, pp. 819–823, Jun. 2016.
[32] 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.
[33] T. Wellmann, A. Lausch, E. Andersson, S. Knapp, C. Cortinovis, J. Jache, S. Scheuer, P. Kremer, A. Mascarenhas, R. Kraemer, A. Haase, F. Schug, and D. Haase, "Remote sensing in urban planning: Contributions towards ecologically sound policies?" *Landscape Urban Planning*, vol. 204, Dec. 2020, Art. no. 103921.
[34] E. Duraklı and E. Aptoula, "Domain generalized object detection for remote sensing images," in *Proc. 31st Signal Process. Commun. Appl. Conf. (SIU)*, Jul. 2023, pp. 1–4.
[35] Z. Peng, W. Huang, S. Gu, L. Xie, Y. Wang, J. Jiao, and Q. Ye, "Conformer: Local features coupling global representations for visual recognition," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 357–366.
[36] 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.
[37] L. Zhang, Y. Liu, and J. Zhang, "Saliency detection based on self-adaptive multiple feature fusion for remote sensing images," *Int. J. Remote Sens.*, vol. 40, no. 22, pp. 8270–8297, Nov. 2019.
[38] Z. Liu, D. Zhao, Z. Shi, and Z. Jiang, "Unsupervised saliency model with color Markov chain for oil tank detection," *Remote Sens.*, vol. 11, no. 9, p. 1089, May 2019.
[39] L. Li, H. Li, and P. Ren, "Underwater image captioning via attention mechanism based fusion of visual and textual information," *Inf. Fusion*, vol. 123, Nov. 2025, Art. no. 103269.
[40] X. Zeng, M. Xu, Y. Hu, H. Tang, Y. Hu, and L. Nie, "Adaptive edge-aware semantic interaction network for salient object detection in optical remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5617416.
[41] 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.
[42] M. Xu, S. Wang, Y. Hu, H. Tang, R. Cong, and L. Nie, "Cross-model nested fusion network for salient object detection in optical remote sensing images," *IEEE Trans. Cybern.*, vol. 55, no. 11, pp. 5332–5345, Nov. 2025.
[43] X. Dong, J. Wang, and B. Dong, "Salient object detection in optical remote sensing images based on hybrid edge fusion perception," *Digit. Signal Process.*, vol. 165, Oct. 2025, Art. no. 105332.
[44] D. Ouyang, S. He, G. Zhang, M. Luo, H. Guo, J. Zhan, and Z. Huang, "Efficient multi-scale attention module with cross-spatial learning," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Jun. 2023, pp. 1–5.
[45] X. Cai, Q. Lai, Y. Wang, W. Wang, Z. Sun, and Y. Yao, "Poly kernel inception network for remote sensing detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2024, pp. 27706–27716.
[46] Q.-L. Zhang and Y.-B. Yang, "SA-Net: Shuffle attention for deep convolutional neural networks," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Jun. 2021, pp. 2235–2239.
[47] Y. Xu, X. Yu, J. Zhang, L. Zhu, and D. Wang, "Weakly supervised RGB-D salient object detection with prediction consistency training and active scribble boosting," *IEEE Trans. Image Process.*, vol. 31, pp. 2148–2161, 2022.

SAMRA KANWAL received the Ph.D. degree in electrical engineering from the Capital University of Science and Technology, Islamabad, Pakistan. She is currently an Assistant Professor with the Department of Computer Software Engineering, National University of Science and Technology, Islamabad. Her research interests include pattern recognition, image processing, and computer vision applications.

NAZAR WAHEED received the Ph.D. degree in engineering and information technology from the University of Technology Sydney (UTS), Australia. He is currently a Distinguished Cybersecurity Educator and Researcher with over 15 years of experience across prominent academic institutions. He is also a Faculty Member with the Higher Colleges of Technology (HCT), United Arab Emirates. He has been a fellow of the Higher Education Academy (FHEA), since 2018. He has significantly contributed to curriculum development and has published extensively in reputable international journals and conferences. He actively participates in global collaborative research initiatives aimed at advancing cybersecurity knowledge and practices. His professional career is defined by integrity, dedication, and a passion for fostering innovation and excellence in education and research. His research interests include cybersecurity, the IoT security, networking, and privacy.

BUSHRA RASHID received the B.S. degree in electrical engineering from Wah Engineering College, Wah Cantt, Pakistan, in 2013, and the M.S. and Ph.D. degrees in electrical engineering from Comsats University, Islamabad, Wah Campus, Pakistan. She is currently an Assistant Professor with the Biomedical Engineering Department, King Faisal University, KSA. Her research interests include resource allocation in wireless communication systems, wireless sensor networks, and application of optimization methods to engineering problems.

NAYEF ALQAHTANI (Member, IEEE) received the Ph.D. degree in electrical and computer engineering from Oakland University, Rochester Hills, MI, USA, in 2021. He is currently an Assistant Professor with King Faisal University, Saudi Arabia. His research interests include signal and control systems, intelligent systems, digital signal and image processing, smart grid, and cyber physical systems.

ALI ALQAHTANI received the Ph.D. degree in computer engineering from Oakland University, Rochester Hills, MI, USA, in 2020. He is currently an Assistant Professor with Najran University (NU). His research interests include machine learning in general and deep learning in image and signal processing, wireless vehicular networks (VANETs), wireless sensor networks, and cyber-physical systems.