Exploring Information Entropy-driven Interaction and Hierarchical Edge Perception for Lightweight ORSI Salient Object Detection

Yihua Tu, Wenqi Si, Gongyang Li, *Member, IEEE*, Chengjun Han, Yun Sui, and Weisi Lin, *Fellow, IEEE*

***Abstract*—Existing Salient Object Detection in Optical Remote Sensing Image (ORSI-SOD) methods are often cumbersome and require expensive computational resources, making them inconvenient to deploy on aircraft. Lightweight ORSI-SOD methods alleviate this issue. However, the cluttered background and varied object shapes of ORSIs cause the saliency maps generated by lightweight methods to have blurry boundaries or even be interfered by backgrounds. In this paper, we propose a novel lightweight ORSI-SOD network via exploring *I*nformation Entropy-driven Interaction and *H*ierarchical *E*dge *P*erception, *i.e.*, *iHEPNet*. Our iHEPNet employs an information interaction and edge perception strategy to mitigate the issues in lightweight methods. Specifically, iHEPNet consists of a lightweight backbone and a saliency predictor, seamlessly bridged by two important modules. The Information Entropy-driven Detail-Context Interaction Module (iDCIM) connected to the backbone is responsible for intra-layer information interaction. iDCIM measures the information entropy of features along the channel dimension to decouple detail features (*e.g.*, textures and edges) from context features (*e.g.*, smooth regions), and promote them to mutually interact and fuse. Following the iDCIM, the Hierarchical Edge Perception Module (HEPM) further explores the inter-layer edge information. HEPM comprehensively aggregates the differences across multi-layer features to generate an edge attention map for sharpening object boundaries, thereby facilitating the predictor to produce boundary-preserving saliency maps. Extensive experiments on three public ORSI-SOD benchmarks demonstrate that our iHEPNet outperforms the vast majority of existing lightweight methods with fewer parameters (merely 1.99M). Furthermore, it exhibits highly competitive performance even compared to conventional normal-size methods. The code and results of our method are available at https://github.com/MathLee/iHEPNet.**

***Index Terms*—Salient object detection, information entropy, detail-context Interaction, hierarchical edge perception.**

I. Introduction

Salient Object Detection in Optical Remote Sensing Images (ORSI-SOD) aims to automatically identify and extract the most visually distinctive objects from complex aerial scenes [1]–[4]. Unlike natural scene images [5]–[8], ORSIs are inherently characterized by unique properties, such as top-down perspective, arbitrary object orientations, cluttered background, varied object shapes and scales, low contrast, and diverse topologies [9]–[11]. These properties make ORSI-SOD more challenging than conventional SOD tasks [12]–[18]. Consequently, directly transferring the existing relevant SOD methods to ORSIs fails to achieve satisfactory performance. To this end, we are dedicated to developing a specialized ORSI-SOD method.

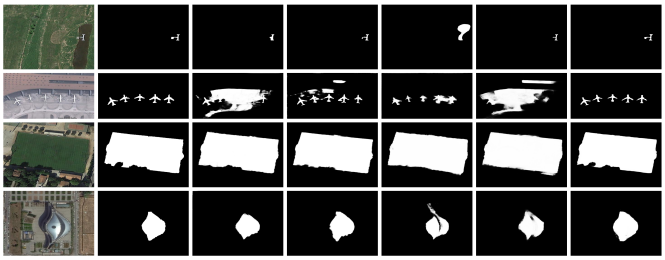

Fig. 1. Saliency maps of four state-of-the-art lightweight ORSI-SOD methods (*i.e.*, SeaNet [19], MEANet [20], SggNet [21], and LiteSalNet [22]) and our method. All methods are edge-based methods, but our method has the clearest edges and the fewest count of parameters. Please zoom-in for details.

To adapt to the characteristics of ORSIs, a series of specialized ORSI-SOD methods have been proposed [11], [23]–[27]. Benefiting from robust backbone networks [28]–[31] and elaborately designed modules, these methods have achieved remarkable performance. However, this high performance typically comes at the cost of high computational overhead, making them computationally intensive and hard to deploy on airborne devices. It is precisely the heavy backbones and complicated modules that make existing ORSI-SOD methods overly complex. To alleviate these issues, researchers shift their attention to lightweight ORSI-SOD, which no longer solely prioritizes model performance, but rather strikes an optimal balance between accuracy and computational efficiency [19], [32], [33].

As a pioneer in lightweight ORSI-SOD, CorrNet [32] directly replaces the regular convolution in VGG [28] with the depthwise separable convolution, yielding a lightweight but effective backbone. Coupled with the specific feature correlation module and decoder blocks, CorrNet reduces the count of model parameters to the sub-5M level (*i.e.*, 4.09M) for the first time. Following in the footsteps of CorrNet, numerous lightweight ORSI-SOD methods [19]–[22], [33]–[36] adopt lightweight pre-trained networks, such as MobileNet-V2 [37], RepVGG [38], and MobileViT-S [39], as backbones,

Yihua Tu, Wenqi Si, Gongyang Li, Chengjun Han, and Yun Sui are with the School of Communication and Information Engineering, and the Key Laboratory of Specialty Fiber Optics and Optical Access Networks, Shanghai University, Shanghai 200444, China (e-mail: yihuatu@shu.edu.cn; swq@shu.edu.cn; ligongyang@shu.edu.cn; monxxcn@gmail.com; yun_sui@shu.edu.cn).

Weisi Lin is with the School of Computer Science and Engineering, Nanyang Technological University, Singapore 639798 (e-mail: wslin@ntu.edu.sg).

Corresponding author: Gongyang Li.

gradually reducing the count of model parameters to the sub-3M level. To mitigate the inherent inadequacies of lightweight backbones in feature extraction, edge information is widely utilized in lightweight ORSI-SOD methods to preserve the object integrity of saliency maps [19]–[22]. However, due to the restricted extraction of edge information from solely individual or adjacent layers by existing lightweight ORSI-SOD methods (*e.g.*, SeaNet [19], MEANet [20], SggNet [21], and LiteSalNet [22]), their saliency maps exhibit blurry boundaries or suffer from background interference, as shown in Fig. 1.

Inspired by the above observation, in this paper, we further explore edge information by developing an information interaction and edge perception strategy to address these issues. This strategy extracts edge cues in a hierarchical manner while seamlessly facilitating information interaction. Furthermore, considering that the currently used lightweight backbones [37]–[39] still possess a relatively large count of parameters, we adopt the more compact MobileViT-XS [39] as the backbone. By coupling it with our strategy, we propose a novel lightweight ORSI-SOD network, named *iHEPNet*, which is built upon *I*nformation Entropy-driven Interaction and *H*ierarchical *E*dge *P*erception. iHEPNet achieves comparable performance to previous state-of-the-art lightweight ORSI-SOD methods while reducing the count of parameters, driving the overall count of parameters down to the sub-2M level (merely 1.99M). As illustrated in Fig. 1, compared to other edge-based lightweight counterparts, the saliency maps generated by iHEPNet are the most accurate, exhibiting the sharpest and most complete boundaries.

Specifically, we implement our strategy through the Information Entropy-driven Detail-Context Interaction Module (iDCIM) and the Hierarchical Edge Perception Module (HEPM). The iDCIM utilizes information entropy to quantify the information content of a feature. Since information entropy effectively reflects the disorder of image textures, a higher value indicates more complex textures, whereas a lower value denotes smoother regions [40]. Therefore, guided by the channel-wise information entropy, the features are decoupled into detail and context representations, which subsequently undergo mutual interactions. As for the HEPM, it hierarchically captures edge differences between the current feature and higher-level features. It then aggregates multi-level edge cues to generate an effective edge attention map for sharpening object boundaries. By embedding these two modules into an encoder-decoder framework, iHEPNet can handle challenging scenarios in ORSIs, *e.g.*, objects with complex geometric edges, similar backgrounds, and tiny objects.

Our main contributions are summarized as follows:

- We propose *iHEPNet*, a novel lightweight method with the information interaction and edge perception strategy for ORSI-SOD. iHEPNet has merely 1.99M parameters and 2.69G floating point operations. Nevertheless, it outperforms the vast majority of lightweight methods and is even comparable to normal-size methods.
- We propose an Information Entropy-driven Detail-Context Interaction Module, which decouples the basic features into detail features and context features based on channel-wise information entropy, and promotes the intra-layer detail-context interaction.
- We propose a Hierarchical Edge Perception Module, which combines edge attention and object attention to sharpen object boundaries. The edge attention, which fully exploits inter-layer edge information, is generated by aggregating the differences across multi-layer features.

II. Related Work

A. Normal-size ORSI-SOD Method

High detection accuracy is the goal pursued by ORSI-SOD methods. With the emergence of new technologies in the field of deep learning, numerous high-performance ORSI-SOD methods have been proposed [1]–[3], [11], [23], [41]–[45]. Through the seamless combination of robust backbones [28]–[31] and elaborately designed modules, these methods adapt to the unique characteristics of optical remote sensing scenarios, pushing detection accuracy to a remarkable height.

Edge cues are widely used in ORSI-SOD to preserve object boundaries. There are various ways to use edge cues. Typically, edges are integrated into a multi-task learning framework [44], [46], [47], which collaborates with pixel-level object annotation to supervise the outputted prediction map, thereby accelerating model convergence. Beyond imposing edge supervision on the final prediction, edge supervision can also be leveraged to enhance the discriminability of intermediate features, driving the design of various edge-aware modules [1], [11], [23], [42], [44]–[48]. In addition to the aforementioned explicit edge extraction via supervision, edge information can also be captured through the filtering operation [49], which is flexible and produces rich edge representations [45], [48].

Semantic cues also play an important role in ORSI-SOD. Such semantic cues can be used to roughly locate the position of objects, which acts as a basis for accurate detection. They are usually extracted from high-level basic features, and then employed to guide the fusion of low-level features [2], [41], [50], [51]. This process is usually accompanied by the adjacent fusion strategy. As the name implies, rather than focusing solely on the intra-fusion of single-level features, the adjacent fusion strategy emphasizes exploiting the complementarity between adjacent-level features [24], [52]–[54]. The joint attention [24], [52] and cross-attention [53], [54] serve as commonly used mechanisms to execute the adjacent fusion.

The above techniques are conventionally implemented in combination with an encoder-decoder framework. Recently, some researchers have introduced denoising diffusion models [55], [56] into ORSI-SOD, which produce saliency maps through a generative diffusion process [3], [43], [57]–[59]. This represents a new paradigm that extracts conditional information from ORSIs and then injects it into the denoising network to iteratively optimize the noise mask. This diffusion-based paradigm breaks through the bottleneck of the traditional encoder-decoder framework, improving the detection accuracy by a substantial margin.

However, the aforementioned ORSI-SOD methods neglect the model complexity, rendering the deployment of high-performance models on aircraft exceedingly difficult. Notably, due to the inherent shortcomings of denoising diffusion

models, diffusion-based methods require T inference steps to produce a saliency map, resulting in a particularly high computational overhead. Therefore, we propose iHEPNet to achieve an optimal balance between performance and model complexity. With merely 1.99M parameters and 2.67G floating point operations (FLOPs), iHEPNet achieves performance comparable to conventional normal-size methods.

B. Lightweight ORSI-SOD Method

In addition to high detection accuracy, the balance between performance and model complexity is also the goal pursued by recent lightweight ORSI-SOD methods. Existing lightweight ORSI-SOD methods primarily achieve model lightweighting by replacing cumbersome backbones and developing efficient modules. For instance, Li *et al*. [32] replaced the regular convolutions of the last three blocks in vanilla VGG with the depthwise separable convolutions. This effort yielded CorrNet, the first lightweight ORSI-SOD method with merely 4.09M parameters. CorrNet inspires subsequent lightweight methods to adopt lightweight backbones, significantly reducing the count of parameters.

Specifically, Luo *et al*. [33] employed MobileNet-V2 [37] as the backbone, and explored the multi-scale attention fusion, resulting in SAFINet with 3.12M parameters. Li *et al*. [34] adopted RepVGG-A0 [38] as the backbone to accelerate the inference speed, enabling SOLNet to achieve an inference speed of 161 fps and 6.52M parameters. In addition to the lightweight convolutional neural networks, Han *et al*. [35] utilized the lightweight transformer (*i.e*., MobileViT-S [39]) as the backbone, and modeled global and local relationships to improve the discriminative ability, resulting in RAMENet with 5.18M parameters and impressive performance.

However, an inherent drawback of lightweight backbones lies in their limited feature extraction capability. Inspired by normal-sized methods, the popular edge information is introduced into lightweight ORSI-SOD methods to compensate for this drawback. Li *et al*. [19] extracted the adjacent edge features through the filtering operation, and aligned them via L2 loss. Similarly, Liu *et al*. [21] also employed the filtering operation to capture edge features from integrated features, and further imposed an edge supervision. Liang *et al*. [20] directly employed the edge supervision to extract the edge-embedded attention from the individual-layer feature. Ai *et al*. [22] constructed a triple-task learning framework jointly supervised by saliency, edge, and skeleton.

The above edge-based methods extract restricted edge information from individual or adjacent layers. Moreover, the lightweight backbones used still carry a relatively high parameter burden. To this end, we propose iHEPNet with merely 1.99M parameters and superior performance. In iHEPNet, we employ the highly compact MobileViT-XS as the backbone, drastically reducing the parameter burden. Additionally, we develop HEPM to capture valuable edge information in a hierarchical manner, enriching the diversity of edge representations. Driven by these tailored designs, iHEPNet surpasses most state-of-the-art lightweight counterparts with merely 1.99M parameters.

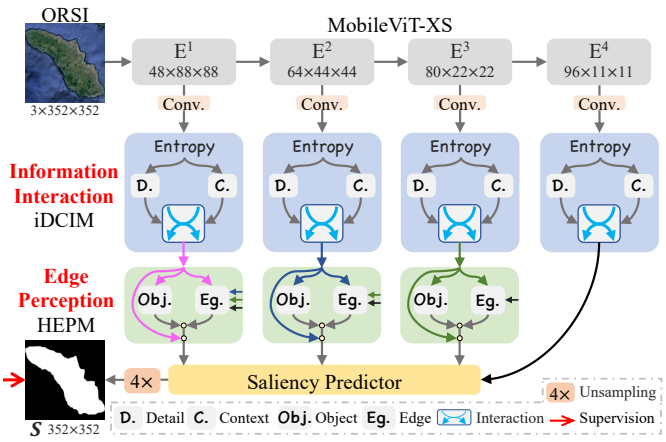

Fig. 2. The overall architecture of the proposed lightweight iHEPNet. iHEPNet comprises a backbone (*i.e*., MobileViT-XS [39]), four Information Entropy-driven Detail-Context Interaction Modules (iDCIMs), three Hierarchical Edge Perception Modules (HEPMs), and a saliency predictor. Initially, an input ORSI is fed into the MobileViT-XS backbone to extract basic features of four layers. Subsequently, these features undergo information interaction in iDCIMs, followed by edge perception in HEPMs. Finally, the saliency predictor aggregates the refined features to infer the final saliency map $\boldsymbol{S}$.

III. Methodology

A. Network Overview

We illustrate the architecture of our lightweight iHEPNet in Fig. 2. iHEPNet employs the MobileViT-XS [39] as the feature extractor, and modifies the partial decoder [60] as the saliency predictor, resulting in the commonly used encoder-decoder architecture. Moreover, the Information Entropy-driven Detail-Context Interaction Modules (iDCIMs) and the Hierarchical Edge Perception Modules (HEPMs) seamlessly bridge the feature extractor and the saliency predictor.

Given an input ORSI with the size of $3\times352\times352$, the MobileViT-XS backbone extracts four-level basic features, which are denoted as $\{\boldsymbol{f}_{\mathrm{b}}^{i} \in \mathbb{R}^{c_i\times h_i\times w_i}\}_{i=1}^{4}$, where $c_i \in \{48, 64, 80, 96\}$ and $h_i/w_i = \frac{352}{2^{i+1}}$. These basic features are respectively processed by a convolutional layer to unify their channel dimensions, generating updated representations $\{\boldsymbol{f}_{\mathrm{b}}^{i} \in \mathbb{R}^{c\times h_i\times w_i}\}_{i=1}^{4}$ with $c = 32$. Guided by our specific information interaction and edge perception strategy, we enhance the information interaction of $\boldsymbol{f}_{\mathrm{b}}^{i}$ in iDCIMs, and undergo edge perception in HEPMs. Concretely, the iDCIM decouples $\boldsymbol{f}_{\mathrm{b}}^{i}$ into detail and context representations, and enhances the intra-layer detail-context interaction, generating $\{\boldsymbol{f}_{\mathrm{dci}}^{i} \in \mathbb{R}^{c\times h_i\times w_i}\}_{i=1}^{4}$. Subsequently, the HEPM hierarchically extracts inter-layer edge cues from $\{\boldsymbol{f}_{\mathrm{dci}}^{i}\}_{i=1}^{4}$ to acquire precise edge attention maps and collaboratively utilizes object attention maps to refine $\{\boldsymbol{f}_{\mathrm{dci}}^{i}\}_{i=1}^{3}$, producing discriminative features $\{\boldsymbol{f}_{\mathrm{hep}}^{i} \in \mathbb{R}^{c\times h_i\times w_i}\}_{i=1}^{3}$. Finally, the saliency predictor aggregates $\boldsymbol{f}_{\mathrm{dci}}^{4}$ and $\{\boldsymbol{f}_{\mathrm{hep}}^{i}\}_{i=1}^{3}$ to generate the final saliency map, denoted as $\boldsymbol{S} \in [0, 1]^{1\times352\times352}$.

B. Information Entropy-driven Detail-Context Interaction Module

Due to the high-altitude top-down perspective and vast coverage of ORSIs, objects often appear as large and holistic

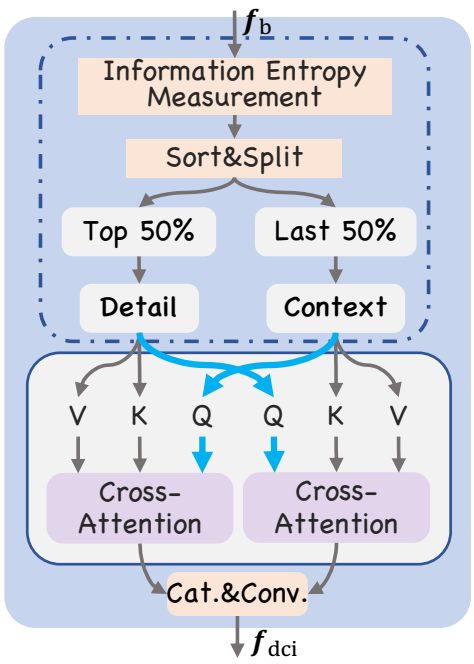

Fig. 3. Illustration of the Information Entropy-driven Detail-Context Interaction Module.

blocks with obvious macro features. This inherent nature poses a significant challenge for segmenting delicate structures and sharp boundaries of the objects. Consequently, we investigate whether macro features and fine-grained features within the basic features $\boldsymbol{f}_{\mathrm{b}}^{i}$ can be explicitly decoupled. Motivated by the fact that information entropy effectively characterizes the spatial disorder of image textures [40], we propose the Information Entropy-driven Detail-Context Interaction Module to decouple these two types of features in the channel dimension, and subsequently facilitate explicit feature interaction. The structure of our iDCIM is illustrated in Fig. 3. In the following, we elaborate on iDCIM in two parts, *i.e.*, the information entropy-driven feature decoupling and the detail-context interaction.

1) Information Entropy-driven Feature Decoupling: The input of iDCIM is $\boldsymbol{f}_{\mathrm{b}} \in \mathbb{R}^{32\times h\times w}$ (index is omitted for brevity). We directly perform the information entropy measurement[1] on $\boldsymbol{f}_{\mathrm{b}}$ to evaluate the information entropy for each channel of $\boldsymbol{f}_{\mathrm{b}}$, generating entropy values $\{\boldsymbol{e}^n\}_{n=1}^{32}$. Then, we sort $\{\boldsymbol{e}^n\}_{n=1}^{32}$ in descending order and obtain $\{\hat{\boldsymbol{e}}^n\}_{n=1}^{32}$. Subsequently, the sorted $\{\hat{\boldsymbol{e}}^n\}_{n=1}^{32}$ is evenly partitioned into two groups, *i.e.*, the top 50% $\{\hat{\boldsymbol{e}}^n\}_{n=1}^{16}$ and the last 50% $\{\hat{\boldsymbol{e}}^n\}_{n=17}^{32}$. A higher value indicates more complex textures, whereas a lower value denotes smoother regions. Therefore, based on this division of entropy values, $\boldsymbol{f}_{\mathrm{b}}$ is decoupled into the detail features $\boldsymbol{f}_{\mathrm{d}} \in \mathbb{R}^{16\times h\times w}$ (corresponding to the top 50% $\{\hat{\boldsymbol{e}}^n\}_{n=1}^{16}$) and the context features $\boldsymbol{f}_{\mathrm{c}} \in \mathbb{R}^{16\times h\times w}$ (corresponding to the last 50% $\{\hat{\boldsymbol{e}}^n\}_{n=17}^{32}$). The detail features are the fine-grained features, such as intricate textures and edges, while the context features are the macro features, such as the smooth regions.

[1]We calculate the feature entropy through a pipeline of absolute transformation, spatial flattening, probability normalization, and entropy calculation.

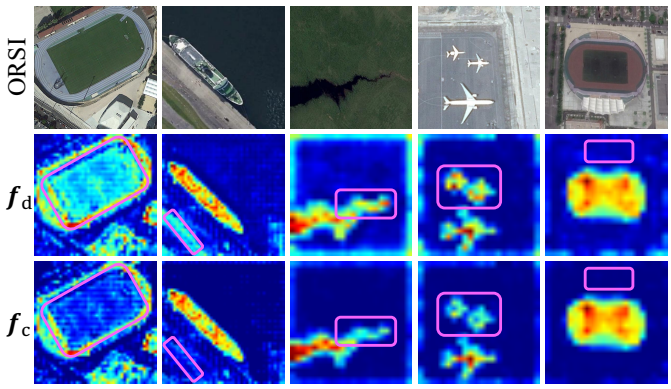

Fig. 4. Feature visualization of $\boldsymbol{f}_{\mathrm{d}}$ and $\boldsymbol{f}_{\mathrm{c}}$.

The above operations are parameter-free, and thus introduce negligible computational cost to the overall model.

To intuitively demonstrate the effect of our information entropy-driven feature decoupling, we visualize $\boldsymbol{f}_{\mathrm{d}}$ and $\boldsymbol{f}_{\mathrm{c}}$ in Fig. 4. As observed, $\boldsymbol{f}_{\mathrm{d}}$ exhibits a more extensive activation region, which demonstrates that $\boldsymbol{f}_{\mathrm{d}}$ contains richer information, such as more details and textures. Conversely, $\boldsymbol{f}_{\mathrm{c}}$ displays a more consistent and noise-free activation across the background regions, indicating that the background regions of $\boldsymbol{f}_{\mathrm{c}}$ are substantially smoother.

2) Detail-Context Interaction: Following the decoupling, we perform feature interaction to bridge the gap between fine-grained details and global context. Here, we adopt the cross-self-attention [31] to implement the detail-context interaction. We first project $\boldsymbol{f}_{\mathrm{d}}$ to $\{\boldsymbol{Q}_{\mathrm{d}}, \boldsymbol{K}_{\mathrm{d}}, \boldsymbol{V}_{\mathrm{d}}\}$, and $\boldsymbol{f}_{\mathrm{c}}$ to $\{\boldsymbol{Q}_{\mathrm{c}}, \boldsymbol{K}_{\mathrm{c}}, \boldsymbol{V}_{\mathrm{c}}\}$. Then, we interchange $\boldsymbol{Q}_{\mathrm{d}}$ and $\boldsymbol{Q}_{\mathrm{c}}$, and perform cross-self-attention on $\{\boldsymbol{Q}_{\mathrm{c}}, \boldsymbol{K}_{\mathrm{d}}, \boldsymbol{V}_{\mathrm{d}}\}$ and $\{\boldsymbol{Q}_{\mathrm{d}}, \boldsymbol{K}_{\mathrm{c}}, \boldsymbol{V}_{\mathrm{c}}\}$, respectively. The above process can be formulated as:

$$\{\boldsymbol{Q}_{\mathrm{d}}, \boldsymbol{K}_{\mathrm{d}}, \boldsymbol{V}_{\mathrm{d}}\} = \mathrm{Proj}(\boldsymbol{f}_{\mathrm{d}}), \tag{1}$$

$$\{\boldsymbol{Q}_{\mathrm{c}}, \boldsymbol{K}_{\mathrm{c}}, \boldsymbol{V}_{\mathrm{c}}\} = \mathrm{Proj}(\boldsymbol{f}_{\mathrm{c}}), \tag{2}$$

$$\boldsymbol{f}_{\mathrm{d-c}} = \mathrm{CrossAtt}(\boldsymbol{Q}_{\mathrm{c}}, \boldsymbol{K}_{\mathrm{d}}, \boldsymbol{V}_{\mathrm{d}}), \tag{3}$$

$$\boldsymbol{f}_{\mathrm{c-d}} = \mathrm{CrossAtt}(\boldsymbol{Q}_{\mathrm{d}}, \boldsymbol{K}_{\mathrm{c}}, \boldsymbol{V}_{\mathrm{c}}), \tag{4}$$

where $\mathrm{Proj}(\cdot)$ is the project operation implemented through convolutional layers, and $\mathrm{CrossAtt}(\cdot)$ is the cross-self-attention. Notably, the cross-self-attention performs spatial self-attention on the corresponding channels to maintain efficiency.

In this bidirectional interaction process, the query $\boldsymbol{Q}_{\mathrm{d}}$ generated from detail features retrieves key $\boldsymbol{Q}_{\mathrm{c}}$ and value $\boldsymbol{Q}_{\mathrm{c}}$ from the context features. This operation allows the details to locate semantic anchors within the context, preventing detail cues from activating cluttered background noise. The query $\boldsymbol{Q}_{\mathrm{c}}$ from the context features simultaneously retrieves key $\boldsymbol{Q}_{\mathrm{d}}$ and value $\boldsymbol{Q}_{\mathrm{d}}$ from the detail features. This operation enables the context to seek precise boundary constraints within the details, preventing the over-smoothing of salient object boundaries.

In the final stage of iDCIM, we concatenate $\boldsymbol{f}_{\mathrm{d-c}}$ and $\boldsymbol{f}_{\mathrm{c-d}}$, and employ a convolutional layer to integrate them, yielding the output of iDCIM, *i.e.*, $\boldsymbol{f}_{\mathrm{dci}}$. Overall, iDCIM is highly

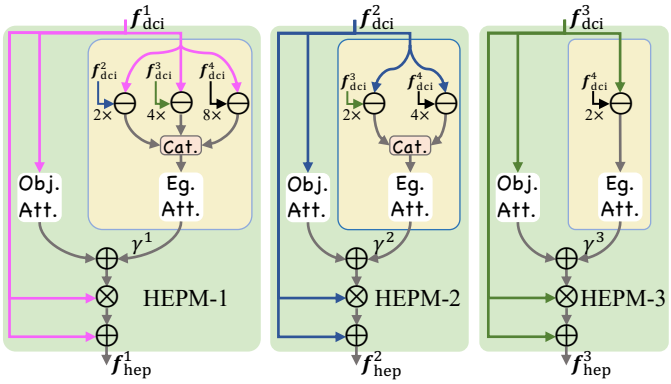

Fig. 5. Illustration of Hierarchical Edge Perception Module at three levels, *i.e.*, HEPM-1, HEPM-2, and HEPM-3. $\ominus$ is the element-wise subtraction with the absolute value function, $\oplus$ is the element-wise summation, and $\otimes$ is the element-wise multiplication.

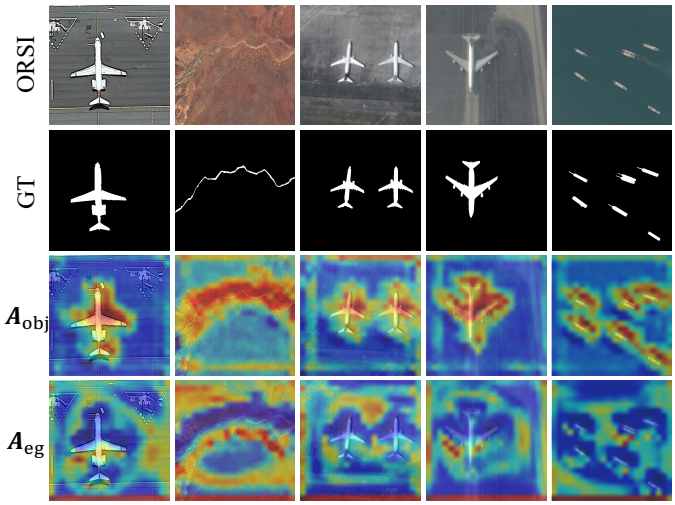

Fig. 6. Feature visualization of $\boldsymbol{A}_{\text{obj}}$ and $\boldsymbol{A}_{\text{eg}}$.

lightweight. It bridges the gap between macroscopic scene understanding and microscopic structural preservation, effectively enabling iHEPNet to adapt to complex ORSI scenarios.

C. Hierarchical Edge Perception Module

The issue of blurry boundaries is prevalent in existing lightweight ORSI-SOD methods, even those leveraging edge information [19]–[22]. These methods typically extract edge information from individual or adjacent layers. However, edge information is indispensable for precise saliency detection, *how to fully exploit and utilize edge information remains a challenging problem*. Consequently, we propose the Hierarchical Edge Perception Module to extract edge attention across multiple layers, and integrate it with object attention to produce discriminative features. The structure of three HEPMs is illustrated in Fig. 5. Since the structure of HEPM varies slightly across different levels, we take HEPM-1 as an example to elaborate from three parts, *i.e.*, the hierarchical edge extraction, the attention integration, and the feature modulation.

1) Hierarchical Edge Extraction: The input of HEPM-1 is $\{\boldsymbol{f}_{\text{dci}}^{i}\}_{i=1}^{4}$. We first unsample the size of $\{\boldsymbol{f}_{\text{dci}}^{i}\}_{i=2}^{4}$ to match $\boldsymbol{f}_{\text{dci}}^{1}$, yielding $\{\hat{\boldsymbol{f}}_{\text{dci}}^{i}\}_{i=2}^{4}$. Subsequently, we extract edge information from the pairs $\{\boldsymbol{f}_{\text{dci}}^{1}, \hat{\boldsymbol{f}}_{\text{dci}}^{2}\}$, $\{\boldsymbol{f}_{\text{dci}}^{1}, \hat{\boldsymbol{f}}_{\text{dci}}^{3}\}$, and $\{\boldsymbol{f}_{\text{dci}}^{1}, \hat{\boldsymbol{f}}_{\text{dci}}^{4}\}$, respectively. Taking $\{\boldsymbol{f}_{\text{dci}}^{1}, \hat{\boldsymbol{f}}_{\text{dci}}^{4}\}$ as an example, we compute the absolute difference between them to generate the edge features $\boldsymbol{f}_{\text{eg}}^{4}$. Since the original $\boldsymbol{f}_{\text{dci}}^{4}$ is a higher-level feature than $\boldsymbol{f}_{\text{dci}}^{1}$, the upsampled $\hat{\boldsymbol{f}}_{\text{dci}}^{4}$ contains fewer details than $\boldsymbol{f}_{\text{dci}}^{1}$. Thus, the above operation can capture edge cues [49]. Finally, we concatenate the multi-level edge features, and perform the edge attention operation (implemented via the spatial attention [61]) to yield an edge attention map $\boldsymbol{A}_{\text{eg}}$. The above process can be formulated as:

$$\boldsymbol{A}_{\text{eg}} = \text{SA}\left(\text{Cat}(|\boldsymbol{f}_{\text{dci}}^{1} \ominus \hat{\boldsymbol{f}}_{\text{dci}}^{2}|, |\boldsymbol{f}_{\text{dci}}^{1} \ominus \hat{\boldsymbol{f}}_{\text{dci}}^{3}|, |\boldsymbol{f}_{\text{dci}}^{1} \ominus \hat{\boldsymbol{f}}_{\text{dci}}^{4}|)\right), \tag{5}$$

where $\text{SA}(\cdot)$ is the spatial attention and $\text{Cat}(\cdot)$ is the concatenation. In this way, $\boldsymbol{A}_{\text{eg}}$ incorporates edge cues of various granularities, enabling a more comprehensive characterization of object boundaries.

2) Attention Integration: In addition to edge information, object information is the cornerstone for accurate saliency detection. Therefore, we introduce an additional object attention branch (also implemented via the spatial attention) to capture the object attention map $\boldsymbol{A}_{\text{obj}}$ from $\boldsymbol{f}_{\text{dci}}^{1}$. Notably, we visualize $\boldsymbol{A}_{\text{obj}}$ and $\boldsymbol{A}_{\text{eg}}$ in Fig. 6. We can observe a distinct discrepancy between $\boldsymbol{A}_{\text{obj}}$ and $\boldsymbol{A}_{\text{eg}}$. Specifically, $\boldsymbol{A}_{\text{obj}}$ exhibits a high response primarily toward the salient object regions, whereas $\boldsymbol{A}_{\text{eg}}$ shows a visible response specifically targeted at the object boundaries. Subsequently, we perform a weighted fusion on $\boldsymbol{A}_{\text{obj}}$ and $\boldsymbol{A}_{\text{eg}}$ to generate the joint attention map $\boldsymbol{A}_{\text{joint}}$ as follows:

$$\boldsymbol{A}_{\text{joint}} = \boldsymbol{A}_{\text{obj}} + \gamma^{1} \cdot \boldsymbol{A}_{\text{eg}}, \tag{6}$$

where γ^{1} is a learnable parameter and $\gamma^{1} \in (0, 1)$. This adaptive integration enables the HEPM to learn an optimal attention map from the data, which facilitates precise object localization while simultaneously sharpening the object boundaries.

3) Feature Modulation: We transfer the informative cues from $\boldsymbol{A}_{\text{joint}}$ to the input $\boldsymbol{f}_{\text{dci}}^{1}$. Specifically, we first apply $\boldsymbol{A}_{\text{joint}}$ as a modulation factor for $\boldsymbol{f}_{\text{dci}}^{1}$. Then, following the residual learning paradigm, the modulated features are added back to $\boldsymbol{f}_{\text{dci}}^{1}$, resulting in the output of HEPM-1, *i.e.*, $\boldsymbol{f}_{\text{hep}}^{1}$. The above process can be formulated as:

$$\boldsymbol{f}_{\text{hep}}^{1} = (\boldsymbol{A}_{\text{joint}} \otimes \boldsymbol{f}_{\text{dci}}^{1}) \oplus \boldsymbol{f}_{\text{dci}}^{1}. \tag{7}$$

As shown in Fig. 5, following a similar procedure, HEPM-2 produces $\boldsymbol{f}_{\text{hep}}^{2}$ and HEPM-3 produces $\boldsymbol{f}_{\text{hep}}^{3}$. Overall, we believe that our hierarchical manner possesses a greater potential to excavate informative edge features, which facilitates subsequent saliency inference. Additionally, the computational cost and parameter count of HEPM are minimal, ensuring that it does not impose a burden on iHEPNet.

D. Saliency Predictor and Training Loss Function

As shown in Fig. 2, the inputs of our saliency predictor are $\boldsymbol{f}_{\text{dci}}^{4} \in \mathbb{R}^{32\times11\times11}$, $\boldsymbol{f}_{\text{hep}}^{3} \in \mathbb{R}^{32\times22\times22}$, $\boldsymbol{f}_{\text{hep}}^{2} \in \mathbb{R}^{32\times44\times44}$, and $\boldsymbol{f}_{\text{hep}}^{1} \in \mathbb{R}^{32\times88\times88}$. Motivated by the robust inference

TABLE I

QUANTITATIVE EVALUATIONS AND MODEL COMPLEXITY ANALYSES AGAINST STATE-OF-THE-ART ORSI-SOD METHODS CROSS THREE BENCHMARK DATASETS. ↑ AND ↓ DENOTE THAT HIGHER AND LOWER VALUES ARE PREFERRED, RESPECTIVELY. THE TOP TWO RESULTS IN EACH METRIC ARE HIGHLIGHTED IN **RED** AND **BLUE**.

Methods	Backbone	Input size	Param↓ (M)	FLOPs↓ (G)	Speed↑ (fps)	EORSSD [10]				ORSSD [9]				ORSI-4199 [11]			
						$S_\alpha\uparrow$	$F_\beta^{\max}\uparrow$	$E_\xi^{\max}\uparrow$	$\mathcal{M}\downarrow$	$S_\alpha\uparrow$	$F_\beta^{\max}\uparrow$	$E_\xi^{\max}\uparrow$	$\mathcal{M}\downarrow$	$S_\alpha\uparrow$	$F_\beta^{\max}\uparrow$	$E_\xi^{\max}\uparrow$	$\mathcal{M}\downarrow$
Normal-size ORSI-SOD Methods																	
MJRBM$_{22}$ [11]	VGG	352^2	43.54	95.7	32	.9197	.8656	.9646	.0099	.9204	.8842	.9623	.0163	.8593	.8493	.9311	.0374
EMFINet$_{22}$ [23]	VGG	256^2	107.26	480.9	25	.9290	.8720	.9711	.0084	.9366	.9002	.9737	.0109	.8675	.8584	.9340	.0330
MCCNet$_{22}$ [1]	VGG	256^2	67.65	112.8	95	.9327	.8904	.9755	.0066	.9437	.9155	.9800	.0087	.8746	.8690	.9413	.0316
HFANet$_{22}$ [62]	ResNet+PVT	448^2	60.53	68.3	26	.9380	.8876	.9740	.0070	.9399	.9112	.9770	.0092	.8767	.8700	.9431	.0314
ERPNet$_{23}$ [42]	ResNet	224^2	56.48	87.2	50	.9210	.8632	.9603	.0089	.9254	.8974	.9710	.0135	.8670	.8553	.9290	.0357
ACCoNet$_{23}$ [24]	VGG	256^2	102.55	179.95	81	.9290	.8837	.9727	.0074	.9437	.9149	.9796	.0088	.8775	.8686	.9412	.0314
GeleNet$_{23}$ [53]	PVT	352^2	25.45	11.66	30	.9376	.8923	.9828	.0064	.9469	.9254	.9860	.0079	**.8862**	.8842	.9544	**.0264**
GLGCNet$_{23}$ [48]	PVT	352^2	25.15	9.8	21	.9375	.8924	.9803	.0055	.9488	.9236	.9864	.0071	.8839	.8808	.9508	.0274
MIRGNet$_{24}$ [52]	VGG	256^2	78.72	136.2	46.9	.9383	**.8930**	.9789	.0056	.9455	.9192	.9812	.0081	-	-	-	-
TSCNet$_{24}$ [2]	VGG	256^2	103.56	116.82	59	.9376	.8900	.9765	.0061	.9428	.9198	.9850	.0081	.8783	.8771	.9486	.0295
SFANet$_{24}$ [50]	Res2Net	256^2	25.1	7.7	-	.9349	.8833	.9769	.0058	.9453	.9192	.9830	.0077	.8761	.8710	.9447	.0292
RAGRNet$_{24}$ [46]	Res2Net	256^2	35.6	17.8	31	.9361	.8852	.9785	.0057	**.9507**	.9242	.9861	**.0066**	.8811	.8811	.9492	.0284
ADSTNet$_{24}$ [63]	Res2Net+T	256^2	62.09	27.72	40	.9311	.8804	.9769	.0065	.9379	.9124	.9807	.0086	.8710	.8698	.9433	.0318
PRNet$_{24}$ [25]	PVT	352^2	20.8	8.47	21	.9276	.8684	.9784	.0054	.9459	.9177	.9848	.0075	**.8873**	.8819	.9527	.0272
BCARNet$_{25}$ [54]	ResNet	352^2	24	7	-	.9361	.8871	.9761	.0054	.9465	.9196	.9833	.0071	.8757	.8689	.9407	.0306
MRBINet$_{25}$ [47]	Res2Net	256^2	32.4	42.8	9.0	.9351	.8852	.9766	.0056	.9474	.9199	.9851	.0069	.8824	.8800	.9489	.0268
SAANet$_{25}$ [51]	Res2Net	352^2	43.06	43.31	17	.9397	.8918	.9801	.0057	.9483	.9226	.9840	.0078	.8806	.8773	.9455	.0302
MTPNet$_{25}$ [44]	SwinT	352^2	55.0	-	91	.9385	**.8930**	.9808	**.0050**	**.9538**	**.9330**	**.9916**	**.0066**	**.8862**	**.8873**	**.9553**	.0268
DPU-Former$_{25}$ [26]	DPU-Former	352^2	44.20	32.51	-	**.9401**	**.8930**	**.9816**	.0056	.9412	.9263	.9868	**.0062**	.8833	**.8877**	**.9547**	**.0263**
RoCAFENet$_{25}$ [27]	PVT	352^2	31.66	24.15	19.96	**.9437**	**.8983**	**.9844**	**.0053**	.9497	**.9272**	**.9873**	.0074	.8847	.8845	.9465	.0267
Lightweight ORSI-SOD Methods																	
CorrNet$_{22}$ [32]	LFE-VGG	256^2	4.09	21.09	100	.9289	.8778	.9696	.0083	.9380	.9129	.9790	.0098	.8623	.8560	.9330	.0366
SeaNet$_{23}$ [19]	MobileNet-V2	288^2	2.76	1.66	96	.9208	.8649	.9710	.0073	.9260	.8942	.9767	.0105	.8772	.8653	.9426	.0308
SAFINet$_{24}$ [33]	MobileNet-V2	288^2	3.12	7.63	-	.9267	.8799	.9732	.0065	.9401	.9106	.9786	.0086	.8583	.8573	.9367	.0340
MEANet$_{24}$ [20]	MobileNet-V2	352^2	3.27	9.62	33	.9282	.8766	.9708	.0070	.9340	.9033	.9768	.0098	.8708	.8662	.9417	.0312
SOLNet$_{25}$ [34]	RepVGG-A0	256^2	6.52	8.14	161	.9171	.8609	.9623	.0078	.9284	.9012	.9734	.0111	.7901	.7708	.8848	.0592
SggNet$_{25}$ [21]	MobileNet-V2	288^2	2.70	1.38	108	.9278	.8871	.9762	.0068	.9342	.9030	.9758	.0111	.8563	.8461	.9337	.0351
LiteSalNet$_{25}$ [22]	MobileNet-V2	256^2	3.90	7.35	35	.9273	.8877	.9743	.0063	.9372	.9124	.9789	.0090	.8521	.8385	.9273	.0383
RAMENet$_{25}$ [35]	MobileViT-S	352^2	5.18	8.72	92	**.9419**	**.8971**	**.9822**	**.0050**	**.9489**	**.9229**	**.9862**	**.0072**	**.8776**	**.8758**	**.9503**	**.0277**
iHEPNet (Ours)	MobileViT-XS	352^2	1.99	2.69	220	**.9388**	**.8914**	**.9832**	**.0046**	**.9435**	**.9154**	**.9820**	**.0084**	**.8825**	**.8776**	**.9506**	**.0274**

capability of the triple-input partial decoder [60], we adapt its architecture into a four-input saliency predictor. Specifically, we follow the fusion paradigm of the original decoder and integrate an extra level of features. To maintain efficiency, depthwise separable convolutions are employed throughout our saliency predictor, facilitating a lightweight design. Finally, the initial saliency map generated by the predictor is processed through a 4× upsampling layer to produce the final saliency map $\boldsymbol{S} \in [0,1]^{1\times352\times352}$.

We employ a comprehensive training loss function to supervise our iHEPNet, which integrates the weighted binary cross-entropy loss (ℓ_{wbce}), the weighted intersection-over-union loss (ℓ_{wiou}), and the F-measure loss (ℓ_{fm}). These components provide multi-dimensional supervision from complementary perspectives. ℓ_{wbce} focuses on pixel-level accuracy by assigning higher weights to hard pixels. ℓ_{wiou} emphasizes global structural integrity and shape consistency. ℓ_{fm} directly optimizes iHEPNet towards a higher F-measure, effectively bridging the gap between training objectives and the evaluation metric. The training loss function ℓ_{train} is formulated as:

$$\ell_{\text{train}} = \ell_{\text{wbce}}(\boldsymbol{S},\boldsymbol{G}) + \ell_{\text{wiou}}(\boldsymbol{S},\boldsymbol{G}) + \ell_{\text{fm}}(\boldsymbol{S},\boldsymbol{G}), \quad (8)$$

where $\boldsymbol{G} \in \{0,1\}^{1\times352\times352}$ denotes the Ground Truth (GT).

IV. EXPERIMENTS

A. Experimental Setup

1) Datasets: We adopt three ORSI-SOD datasets, named ORSSD [9][2], EORSSD [10][3], and ORSI-4199 [11][4], to conduct experiments. The ORSSD dataset has 800 ORSIs and their corresponding GTs, among which 600 images form the training set and the remaining 200 form the test set. The EORSSD dataset has 2000 ORSIs and their corresponding GTs, among which 1400 images form the training set and the remaining 600 form the test set. The ORSI-4199 dataset has 4199 ORSIs and their corresponding GTs, among which 2000 images form the training set and the remaining 2199 form the test set. These three datasets are trained and tested independently, not jointly.

2) Implementation Details: All experiments are conducted on PyTorch using an NVIDIA RTX 3090 GPU (24GB memory). For training and test, the input size of iHEPNet is 352×352. Accordingly, in the training phase, we first resize all ORSIs, and then perform data augmentation on them, such as rotation and flipping. We adopt the pre-trained parameters of MobileViT-XS to effectively extract basic features. All parameters of iHEPNet are optimized using the Adam optimizer

[2]https://li-chongyi.github.io/proj_optical_saliency.html
[3]https://github.com/rmcong/DAFNet_TIP20
[4]https://github.com/wchao1213/ORSI-SOD

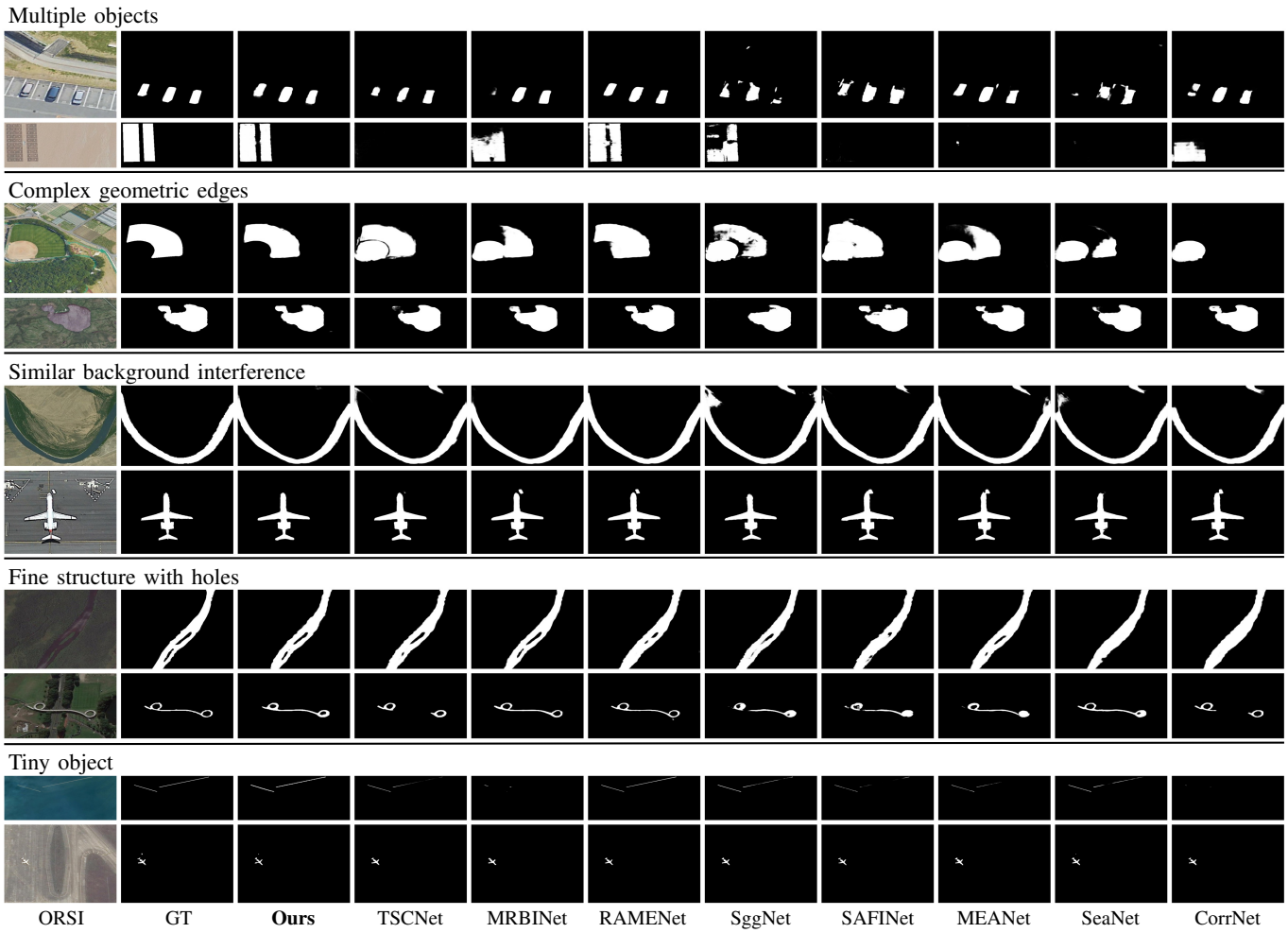

Fig. 7. Visual comparisons with two normal-size ORSI-SOD methods with similar performance (TSCNet and MRBINet) and six lightweight ORSI-SOD methods (RAMENet, SggNet, SAFINet, MEANet, SeaNet, and CorrNet).

with an initial learning rate of $2e^{-4}$ and a batch size of 16 until convergence. Notably, the initial learning rate is decayed by a factor of 10 every 40 epochs.

3) Evaluation Metrics: We adopt the accuracy and the model complexity metrics to comprehensively evaluate the model. The accuracy metrics[5] include S-measure (S_α, α = 0.5) [64], maximum F-measure ($F_\beta^{\max}$, β^2 = 0.3) [65], maximum E-measure ($E_\xi^{\max}$) [66], and mean absolute error ($\mathcal{M}$). The model complexity metrics include parameter count, computational cost, and inference speed (excluding I/O time).

B. Comparison with State-of-the-arts

For a comprehensive comparison, we compare our iHEPNet against two types of state-of-the-art methods on three ORSI-SOD datasets, *i.e.*, EORSSD, ORSSD, and ORSI-4199. The first type contains 20 normal-size ORSI-SOD methods, including MJRBM [11], EMFINet [23], MCCNet [1], HFANet [62], ERPNet [42], ACCoNet [24], GeleNet [53], GLGCNet [48], MIRGNet [52], TSCNet [2], SFANet [50], RAGRNet [46], ADSTNet [63], PRNet [25], BCARNet [54], MRBINet [47], SAANet [51], MTPNet [44], DPU-Former [26], and RoCAFENet [27]. The second type comprises 8 lightweight ORSI-SOD methods, including CorrNet [32], SeaNet [19], SAFINet [33], MEANet [20], SOLNet [34], SggNet [21], LiteSalNet [22], and RAMENet [35]. We collect the saliency maps of the aforementioned methods through their open-source code repositories and public benchmarks[6].

1) Quantitative Comparison: We summarize the comprehensive quantitative results in Tab. I. On the EORSSD dataset, our iHEPNet achieves the best M of 0.0046 among all methods, and the highest E_ξ^{max} of 0.9832 among all lightweight methods. It even outperforms some normal-size methods, *e.g.*, MTPNet [44] and MRBINet [47], demonstrating superior performance. On the ORSSD dataset, our iHEPNet maintains competitive performance with an S_α of 0.9435. It significantly surpasses early lightweight methods, such as SeaNet [19] and MEANet [20], across all four metrics. On the ORSI-4199 datasets, our iHEPNet demonstrates robust generalization on this large-scale dataset, achieving the highest F_β^{max} (0.8776) and E_ξ^{max} (0.9506) in the lightweight methods. Notably, its F_β^{max} even exceeds that of several recent normal-size methods, *e.g.*, BCARNet [54] and SAANet [51].

2) Model Complexity Comparison: We also report the model complexity results in Tab. I. The primary advantage

[5]https://github.com/MathLee/MatlabEvaluationTools

[6]https://github.com/MathLee/ORSI-HRSI-SOD_Summary

TABLE II
ABLATION STUDIES ON EVALUATING THE INDIVIDUAL CONTRIBUTION OF EACH MODULE IN IHEPNET. THE BEST ONE IN EACH METRIC IS **BOLD**.

Baseline	iDCIM	HEPM	Param (K)	FLOPs (M)	EORSSD [10]			
					$S_\alpha \uparrow$	$F_\beta^{\max} \uparrow$	$E_\xi^{\max} \uparrow$	$\mathcal{M} \downarrow$
✓			1973.59	2594.22	0.9263	0.8763	0.9760	0.0058
✓	✓		1986.14 **+12.55**	2677.29 **+83.07**	0.9372	0.8884	0.9805	0.0051
✓		✓	1978.27 **+4.68**	2611.92 **+17.70**	0.9369	0.8886	0.9798	0.0052
✓	✓	✓	1990.82 **+17.23**	2694.99 **+100.77**	**0.9388**	**0.8914**	**0.9832**	**0.0046**

of iHEPNet lies in its exceptional efficiency, characterized by its ultra-lightweight architecture and remarkable inference speed. Specifically, our iHEPNet occupies the lowest parameter count among all methods at merely 1.99M. This is less than 40% of the size of the most recent lightweight RAMENet (5.18M), and only a tiny fraction of those used by normal-size methods like MCCNet (67.65M). In terms of computational cost, iHEPNet requires only 2.69G FLOPs. Compared to CorrNet, it achieves superior performance while utilizing only 12.7% of its computational resources (2.69G vs 21.09G). Furthermore, iHEPNet sets a new benchmark for speed, reaching 220 fps. This is significantly faster than the fastest lightweight method, SOLNet (161 fps), and nearly 7 times faster than the typical normal-size method, MJRBM (32 fps). In summary, the comparative results demonstrate that iHEPNet successfully achieves high-performance saliency detection within an extremely constrained resource budget.

3) Visual Comparison: To intuitively demonstrate the superiority of iHEPNet, we provide a visual comparison with eight state-of-the-art methods across five challenging scenarios, as illustrated in Fig. 7. When dealing with scenes containing multiple objects, most lightweight competitors either miss small objects or generate adhesive predictions that fail to distinguish individual objects. In contrast, iHEPNet accurately identifies all objects with clear separations. For objects possessing complex geometric edges, iHEPNet produces the sharpest and most complete saliency maps. While other methods suffer from boundary blurring (*e.g.*, SAFINet and MEANet), our HEPM effectively preserves geometric integrity. In scenarios where the objects are highly indistinct from the background (*i.e.*, similar background interference), iHEPNet exhibits superior discriminative power. It effectively suppresses background clutter, whereas methods like MEANet and SeaNet produce significant false positives. For objects with internal gaps (*i.e.*, fine structure with holes), iHEPNet demonstrates exceptional microscopic structural preservation. Detecting the tiny object is a classic challenge for lightweight method due to feature dilution. Leveraging its multi-scale feature integration, iHEPNet precisely localizes the tiny object without losing its distinct shape details. In summary, iHEPNet effectively balances macro-understanding with micro-preservation, producing more accurate saliency maps than recent lightweight and even certain normal-size methods.

TABLE III
ABLATION STUDIES ON EVALUATING THE RATIONALITY OF EACH COMPONENT IN IDCIM. THE BEST ONE IN EACH METRIC IS **BOLD**.

Variant	EORSSD [10]			
	$S_\alpha \uparrow$	$F_\beta^{\max} \uparrow$	$E_\xi^{\max} \uparrow$	$\mathcal{M} \downarrow$
w/ DirectSplit	0.9338	0.8858	0.9780	0.0060
Mean	0.9352	0.8866	0.9798	0.0053
Maximum	0.9338	0.8831	0.9770	0.0057
Variance	0.9355	0.8874	0.9802	0.0057
CSA→SSA	0.9344	0.8856	0.9788	0.0052
CrossAtt→IndepAtt	0.9336	0.8831	0.9778	0.0054
Information Entropy (Ours)	**0.9372**	**0.8884**	**0.9805**	**0.0051**

C. Ablation Studies

We conduct comprehensive ablation studies on the EORSSD dataset using the same configurations to evaluate the contribution of components in our iHEPNet. In particular, we investigate the individual contribution of modules, the structure of iDCIM and HEPM, and the training loss function.

1) The Individual Contribution of Each Module in iHEPNet: We implement the information interaction and edge perception strategy through iDCIM and HEPM. As reported in Tab. II, we provide three variants to evaluate the individual contribution of iDCIM and HEPM. Notably, in the Baseline variant, we remove all four iDCIMs and three HEPMs from iHEPNet, and directly transfer the updated $\{\boldsymbol{f}_{\mathrm{b}}^i\}_{i=1}^4$ to the saliency predictor.

Specifically, the Baseline achieves an S_α of 0.9263 and an M of 0.0058 with 1973.59K parameters. The integration of the iDCIM yields a substantial performance leap to 0.9372 in S_α, and reduces the pixel-level error to 0.0051, while incurring a negligible parameter count of only 12.55K. This indicates that the information entropy-driven decoupling and intra-layer interaction of iDCIM effectively compensate for the limited feature extraction capabilities of the lightweight backbone. Similarly, the introduction of the HEPM enhances F_β^{max} to 0.8886 with a mere 4.68K parameter increase, proving that the hierarchical edge perception mechanism is highly efficient at capturing structural boundaries without burdening iHEPNet. When both modules are embedded into the complete iHEPNet, iHEPNet reaches its peak performance across all metrics, particularly achieving a top-tier S_α of 0.9388 and a minimal

TABLE IV
ABLATION STUDIES ON EVALUATING THE RATIONALITY OF EACH COMPONENT IN HEPM. THE BEST ONE IN EACH METRIC IS **BOLD**.

Variant	EORSSD [10]			
	$S_\alpha \uparrow$	$F_\beta^{\max} \uparrow$	$E_\xi^{\max} \uparrow$	$\mathcal{M} \downarrow$
w/ CurLayer	0.9349	0.8860	0.9791	0.0060
w/ AdjLayers	0.9353	0.8859	0.9777	0.0057
w/o EdgeAtt	0.9341	0.8859	0.9772	0.0062
w/o ObjAtt	0.9334	0.8853	0.9786	0.0058
$\gamma = 1$	0.9352	0.8837	0.9788	0.0056
$\gamma = 0.5$	0.9356	0.8872	0.9793	0.0053
w/ MultiLayers (Ours)	**0.9369**	**0.8886**	**0.9798**	**0.0052**

TABLE V
ABLATION STUDIES ON EVALUATING THE EFFECTIVENESS OF THE TRAINING LOSS FUNCTIONS. THE BEST ONE IN EACH METRIC IS **BOLD**.

ℓ_{wbce}	ℓ_{wiou}	ℓ_{fm}	EORSSD [10]			
			$S_\alpha \uparrow$	$F_\beta^{\max} \uparrow$	$E_\xi^{\max} \uparrow$	$\mathcal{M} \downarrow$
✓			0.9287	0.8758	0.9789	0.0064
✓	✓		0.9370	0.8879	0.9811	0.0049
✓	✓	✓	**0.9388**	**0.8914**	**0.9832**	**0.0046**

M of 0.0046. Crucially, this significant performance gain is realized with a total parameter increase of only 17.23K and a modest computational increment of 100.77M FLOPs compared to the Baseline.

2) The Rationality of Each Component in iDCIM: The core of iDCIM lies in the feature decoupling and the feature interaction, where the former is based on the channel-wise information entropy. As reported in Tab. III, we provide six variants to evaluate the effectiveness of these critical components: 1) directly splitting $\boldsymbol{f}_{\text{b}}^i$ evenly along the channel dimension (*i.e.*, *w/ DirectSplit*), 2) adopting the channel-wise mean values for feature decoupling (*i.e.*, Mean), 3) adopting the channel-wise maximum values for feature decoupling (*i.e.*, Maximum), 4) adopting the channel-wise variance values for feature decoupling (*i.e.*, Variance), 5) replacing the spatial self-attention with the channel self-attention in the feature interaction (*i.e.*, CSA→SSA), and 6) replacing the cross-self-attention with the independent self-attention without $\boldsymbol{Q}$ exchanging (*i.e.*, CrossAtt→IndepAtt).

Specifically, comparing the complete iDCIM (*i.e.*, "Information Entropy") with *w/ DirectSplit*, our entropy-driven feature decoupling method improves S_α from 0.9338 to 0.9372 and reduces M from 0.0060 to 0.0051. Although other statistics-based feature decoupling methods, such as "Mean", "Maximum", and "Variance", show better results than *w/ DirectSplit*, they are still inferior to our "Information Entropy". This indicates that information entropy provides a more robust criterion for distinguishing detail features from context features. Furthermore, "CSA→SSA" results in

TABLE VI
ABLATION STUDIES ON EVALUATING THE FLEXIBILITY OF OUR IHEPNET ON DIFFERENT BACKBONES AND INPUT SIZES. THE TOP TWO RESULTS IN EACH METRIC ARE HIGHLIGHTED IN **RED** AND **BLUE**.

Methods	Backbone	Input size	Param↓ (M)	FLOPs↓ (G)	EORSSD [10]			
					$S_\alpha \uparrow$	$F_\beta^{\max} \uparrow$	$E_\xi^{\max} \uparrow$	$\mathcal{M} \downarrow$
CorrNet [32]	LFE-VGG	256^2	4.09	21.09	.9289	.8778	.9696	.0083
SOLNet [34]	RepVGG-A0	256^2	6.52	8.14	.9171	.8609	.9623	.0078
MEANet [20]	MobileNet-V2	352^2	3.27	9.62	.9282	.8766	.9708	.0070
SeaNet [19]	MobileNet-V2	288^2	2.76	1.66	.9208	.8649	.9710	.0073
SAFINet [33]	MobileNet-V2	288^2	3.12	7.63	.9267	.8799	.9732	.0065
SggNet [21]	MobileNet-V2	288^2	2.70	1.38	.9278	.8871	.9762	.0068
LiteSalNet [22]	MobileNet-V2	256^2	3.90	7.35	.9273	.8877	.9743	.0063
RAMENet [35]	MobileViT-S	352^2	5.18	8.72	**.9419**	**.8971**	**.9822**	**.0050**
iHEPNet (Ours)	MobileViT-XS	352^2	1.99	2.69	**.9388**	**.8914**	**.9832**	**.0046**
iHEPNet (Ours)	MobileNet-V2	288^2	2.31	1.84	.9307	.8782	.9799	.0055
iHEPNet (Ours)	MobileNet-V2	256^2	2.31	1.43	.9287	.8766	.9755	.0056

a performance drop across all metrics (*e.g.*, $E_\xi^{\max}$ decreases to 0.9788), demonstrating that spatial self-attention is more effective than channel self-attention in modeling the intra-layer dependencies. "CrossAtt→IndepAtt" has a significant performance decline, indicating that the cross-self-attention-based detail-context interaction is a core operation in iDCIM.

3) The Rationality of Each Component in HEPM: The core of HEPM lies in the edge attention and the object attention, where the former extracts edge cues from multiple layers in a hierarchical manner. As reported in Tab. IV, we provide six variants to evaluate the effectiveness of these critical components: 1) extracting edge cues from the current layer rather than multiple layers (*i.e.*, *w/ CurLayer*). 2) extracting edge cues from adjacent layers rather than multiple layers (*i.e.*, *w/ AdjLayers*), 3) removing the edge attention from HEPM (*i.e.*, *w/o EdgeAtt*), 4) removing the object attention from HEPM (*i.e.*, *w/o ObjAtt*), 5) fixing the learnable γ to 1 (*i.e.*, $\gamma = 1$), and 6) fixing the learnable γ to 0.5 (*i.e.*, $\gamma = 0.5$).

Specifically, comparing our *w/ MultiLayers* with *w/ CurLayer* and *w/ AdjLayers*, *w/ MultiLayers* is comprehensively superior to both *w/ CurLayer* and *w/ AdjLayers*, indicating that extracting edge cues from multiple layers facilitates better structural perception. Furthermore, the removal of either the edge attention or the object attention results in a performance decline. Notably, *w/o EdgeAtt* yields the highest M of 0.0062, indicating that explicit edge cues are vital for precise boundary sharpening. Similarly, *w/o ObjAtt* shows a decrease in S_α to 0.9334, confirming that the object attention is essential for maintaining regional integrity and suppressing background noise. These results collectively demonstrate that the synergy between hierarchical multi-layer edge extraction and the dual-attention mechanism is critical for generating high-quality saliency maps. In addition, changing the learnable parameter γ to the fixed parameter hinders the module from performing more effective attention integration, and leads to performance degradation for both $\gamma = 1$ and $\gamma = 0.5$. We report the learned parameters, where γ^1 is 0.4935, γ^2 is 0.4689, and γ^3 is 0.5046. The fixed parameter 0.5 is closer to the above learned parameters than 1, therefore $\gamma = 0.5$ has better performance than $\gamma = 1$.

4) The Effectiveness of the Training Loss Function: To

TABLE VII
GENERALIZATION EXPERIMENTS ON THE COD TASK WITH STATE-OF-THE-ART METHODS ON TWO COD BENCHMARKS. THE TOP TWO RESULTS IN EACH METRIC ARE HIGHLIGHTED IN **RED** AND **BLUE**.

Method	Backbone	Input size	Param↓ (M)	FLOPs↓ (G)	CAMO [67]				COD10K [68]			
					$S_\alpha\uparrow$	$F_\beta^{\max}\uparrow$	$E_\xi^{\max}\uparrow$	$\mathcal{M}\downarrow$	$S_\alpha\uparrow$	$F_\beta^{\max}\uparrow$	$E_\xi^{\max}\uparrow$	$\mathcal{M}\downarrow$
SLSR [69]	ResNet	352^2	25.21	57.90	.7872	.7532	.8538	.0801	.8044	.7315	.8918	.0368
FEDER [70]	ResNet	352^2	44.13	64.04	.8021	.7887	.8731	.0712	.8223	.7678	.9050	.0316
RISNet [71]	ResNet	704^2	26.58	51.08	**.8695**	**.8617**	**.9298**	**.0504**	**.8733**	**.8369**	**.9378**	**.0247**
BGNet [72]	Res2Net	416^2	58.38	77.80	.8116	.7987	.8817	.0734	.8307	.7737	.9115	.0326
SINetV2 [73]	Res2Net	352^2	12.31	27.98	.8201	.8011	.8950	.0705	.8151	.7519	.9060	.0368
UTGR [74]	PT	473^2	121.33	37.54	.7847	.7541	.8545	.0859	.8178	.7422	.8908	.0354
MGL [75]	SwinT	384^2	63.55	443.56	.7718	.7334	.8419	.0894	.8109	.7332	.8895	.0367
CamoFormer [76]	PVT	384^2	71.40	84.03	.8169	.8008	.8850	.0671	.8381	.7864	**.9300**	**.0291**
RUN [77]	PVT	384^2	79.15	331.68	.8042	.7862	.8730	.0698	.8228	.7665	.9045	.0308
DSAM [78]	Hiera&PVT	1024^2	121.93	743.97	**.8306**	**.8336**	**.9204**	**.0610**	**.8442**	**.8046**	.9298	.0329
iHEPNet (Ours)	MobileViT-XS	352^2	1.99	2.69	.7904	.7736	.8682	.0771	.8151	.7505	.8968	.0348

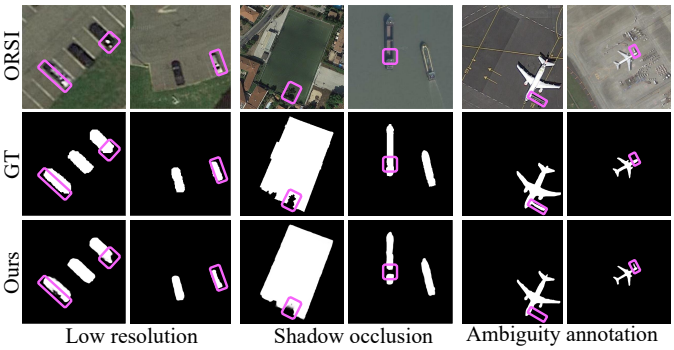

Fig. 8. Failure cases of our iHEPNet.

investigate the effectiveness of each component in our training loss function, we provide two variants. As reported in Tab. V, using the standard ℓ_{wbce} as a baseline yields an S_α of 0.9287 and an M of 0.0064. The introduction of ℓ_{wiou} significantly boosts performance across all metrics, with F_β^{max} increasing from 0.8758 to 0.8879 and the M dropping to 0.0049, which demonstrates that ℓ_{wiou} effectively guides our iHEPNet to focus on global structural integrity. Finally, incorporating ℓ_{fm} further refines the predicted saliency maps to achieve the best results in all metrics, reaching a peak S_α of 0.9388 and a minimal M of 0.0046. This consistent improvement confirms that the combination of these three loss functions enables a comprehensive supervision strategy that successfully balances pixel-level accuracy and structural consistency.

5) The Flexibility of Our iHEPNet on Different Backbones and Input Sizes: To verify the flexibility of our iHEPNet, we conduct ablation experiments by deploying our iHEPNet on the commonly used MobileNet-V2 under varying input sizes (*i.e.* 288^2 and 256^2). We report the quantitative results in Tab. VI. When embedded with the lightweight MobileViT-XS at a 352^2 size, our iHEPNet requires extremely low parameters and burdens of merely 1.99M and 2.69G FLOPs, yet it achieves the top-tier $E_\xi^{\max}$ of 0.9832 and the absolute lowest $\mathcal{M}$ of 0.0046. This significantly surpasses RAMENet, which carries substantially heavier computational burdens. When evaluated with the MobileNet-V2 backbone at a 288^2 size, our iHEPNet consistently exhibits superior performance compared to SeaNet, SAFINet, and SggNet, but with fewer parameters at 2.31M. Similarly, under the input size of 256^2, our iHEPNet not only outperforms LiteSalNet, but also has fewer parameters and computational burdens than LiteSalNet. These empirical results soundly demonstrate that our iHEPNet possesses exceptional flexibility and architectural scalability, enabling it to adapt to different backbones and input sizes while sustaining highly robust performance.

6) The Generalization Experiments on Camouflaged Object Detection: To verify the generalization of our iHEPNet, we conduct experiments on the challenging Camouflaged Object Detection (COD) task [67], [68], [79]. We adopt the training sets of the CAMO dataset [67] and the COD10K dataset [68] to jointly train our iHEPNet, and evaluate it on the test sets of both datasets. We report the quantitative results of our iHEPNet and state-of-the-art COD methods, including SLSR [69], FEDER [70], RISNet [71], BGNet [72], SINetV2 [73], UTGR [74], MGL [75], CamoFormer [76], RUN [77], and DSAM [78], in Tab. VII.

We can observe that although our iHEPNet lags behind most COD methods, it achieves superior performance over SLSR, UTGR, and MGL using only 1.6%–7.9% of their parameter count. This demonstrates that our iHEPNet achieves a good balance between model complexity and performance. The inferior performance on the COD task arises because our iHEPNet is specifically designed for the SOD task. In SOD, salient objects can be distinguished from the background. Accordingly, we propose iDCIM and HEPM tailored for the SOD task. However, in COD, camouflaged objects are often difficult to differentiate from the background, especially in terms of object boundaries, which causes these two modules to fail. In addition, our lightweight backbone has limited feature extraction capability. The combination of these factors results in our iHEPNet not performing outstandingly on the COD task, yet it achieves the optimal trade-off between performance and model complexity.

D. Failure Cases and Future Work

Despite the remarkable performance and architectural robustness demonstrated by our iHEPNet across various scenarios, it still encounters performance bottlenecks under certain extreme conditions. As shown in Fig. 8, we summarize the typical failure cases of our iHEPNet into three types. The first one is the low resolution. In cases characterized by severe resolution degradation (the first two columns), the fine-grained visual cues of objects become heavily blurred. Although our iHEPNet incorporates an edge perception mechanism to sharpen object contours, the heavily pixelated structures inherently restrict the discriminative capability of the edge features, leading to incomplete extractions at boundaries. The second one is the shadow occlusion. Illumination variations, particularly severe shadow occlusions (the middle two columns), pose another hard challenge. Strong shadows cast by adjacent objects or ship hulls alter the original textural features of the salient objects. Consequently, our iHEPNet suffers from structural over-segmentation or object-background confusion, failing to precisely decouple the salient object from its associated shadow field. The last one is the ambiguity annotation. Semantic ambiguity in GTs (the last two columns) hinders precise prediction. For complex airport scenes, specific ancillary structures, such

as passenger boarding bridges, exhibit inconsistent annotations across different images, being labeled as salient foreground in some samples but discarded as background in others. Such annotation inconsistencies introduce optimization orientation conflicts during the training process, causing iHEPNet to produce ambiguous prediction results.

In future work, we will focus on two major directions. First, we aim to integrate lightweight low-level vision enhancement modules, such as super-resolution and image deshadowing sub-networks, directly into the feature extraction stage of iHEPNet. This end-to-end synergy will effectively restore pixel-level granularity and eliminate occlusions before dense prediction. Second, we plan to explore robust learning paradigms, such as noise-tolerant learning or uncertainty-aware estimation, to automatically mitigate the adverse effects of ambiguous annotations, thereby further strengthening the perceptual reliability of our iHEPNet.

V. Conclusion

In this paper, we provide an in-depth exploration of the utility of edge information and present iHEPNet, an ultra-lightweight ORSI-SOD framework. To achieve a compact architecture, iHEPNet adopts the efficient MobileViT-XS as its backbone. We develop the information interaction and edge perception strategy in iHEPNet to compensate for the limited representational capacity inherent in MobileViT-XS. Specifically, we employ iDCIM to decouple detail features from context features in an information entropy-driven manner, followed by an efficient intra-layer detail-context interaction. Building upon these enriched features, we leverage HEPM to capture an edge attention map in a hierarchical manner, which is different from conventional edge-based lightweight methods. By combining the edge attention map with the object attention map, HEPM effectively sharpens object boundaries. Finally, a lightweight saliency predictor is employed to infer high-precision saliency maps. To the end, our lightweight iHEPNet achieves top-tier performance with minimal parameter count (1.99M) and low computational cost (2.69G FLOPs).

References

[1] 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, pp. 1–13, 2022.

[2] G. Li, Z. Bai, and Z. Liu, "Texture-semantic collaboration network for ORSI salient object detection," *IEEE Trans. Circuits Syst. II-Express Briefs*, vol. 71, no. 4, pp. 2464–2468, Apr. 2024.

[3] J. Han, J. Sun, F. Wang, F. Sun, and H. Li, "ORSIDiff: Diffusion model for salient object detection in optical remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 63, pp. 1–15, 2025.

[4] B. Wan, R. Cong, X. Zhou, H. Fang, C. Lv, and S. Kwong, "G2HFNet: Geogran-aware hierarchical feature fusion network for salient object detection in optical remote sensing images," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 36, no. 5, pp. 6472–6484, 2026.

[5] J. He, K. Fu, X. Liu, and Q. Zhao, "Samba: A unified mamba-based framework for general salient object detection," in *Proc. IEEE CVPR*, Jun. 2025, pp. 25 314–25 324.

[6] K. Fu, D.-P. Fan, G.-P. Ji, Q. Zhao, J. Shen, and C. Zhu, "Siamese network for RGB-D salient object detection and beyond," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 9, pp. 5541–5559, 2022.

[7] Y. Jiang, X. Li, K. Fu, and Q. Zhao, "Transformer-based light field salient object detection and its application to autofocus," *IEEE Trans. Image Process.*, vol. 33, pp. 6647–6659, 2024.

[8] K. Fu, D.-P. Fan, G.-P. Ji, and Q. Zhao, "JL-DCF: Joint learning and densely-cooperative fusion framework for RGB-D salient object detection," in *Proc. IEEE CVPR*, Jun. 2020, pp. 3049–3059.

[9] 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.

[10] 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.

[11] Z. Tu, C. Wang, C. Li, M. Fan, H. Zhao, and B. Luo, "ORSI salient object detection via multiscale joint region and boundary model," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–13, 2022.

[12] J. Chen, G. Li, Z. Zhang, L. Chang, and D. Zeng, "STENet: Superpixel token enhancing network for RGB-D salient object detection," *IEEE Trans. Multimedia*, 2026. [Online]. Available: 10.1109/TMM.2026.3680601

[13] S. Shi, G. Li, R. Cong, S. Xiao, and W. Lin, "Diffusion-driven RGB-D salient object detection with temporal modulation," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 36, no. 8, pp. 11 831–11 843, 2026.

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

[15] 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.

[16] Z. Tu, Y. Ma, C. Li, J. Tang, and B. Luo, "Edge-guided non-local fully convolutional network for salient object detection," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 31, no. 2, pp. 582–593, Feb. 2021.

[17] 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.

[18] J. Tang, D. Fan, X. Wang, Z. Tu, and C. Li, "RGBT salient object detection: Benchmark and a novel cooperative ranking approach," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 30, no. 12, pp. 4421–4433, Dec. 2020.

[19] 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, pp. 1–12, 2023.

[20] 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, p. 121778, Mar. 2024.

[21] 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. 816, Feb. 2025.

[22] Z. Ai, H. Luo, and J. Wang, "A lightweight multistream framework for salient object detection in optical remote sensing," *IEEE Trans. Geosci. Remote Sens.*, vol. 63, pp. 1–15, Mar. 2025.

[23] X. Zhou, K. Shen, Z. Liu, C. Gong, J. Zhang, and C. Yan, "Edge-aware multiscale feature integration network for salient object detection in optical remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–15, 2022.

[24] 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.

[25] 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, May 2024.

[26] Y. Sun, J. Yan, J. Qian, C. Xu, J. Yang, and L. Luo, "Dual-perspective united transformer for object segmentation in optical remote sensing images," *Proc. IJCAI*, 2025.

[27] Y. Wang, P. Zheng, R. Zhao, and L. Wang, "Robust salient object detection in optical remote sensing images via multiscale contextual attention and feature enhancement," *IEEE J. Sel. Top. Appl. Earth Observ. Remote Sens.*, pp. 1–20, Sept. 2025.

[28] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. ICLR*, May 2015, pp. 1–14.

[29] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE CVPR*, Jun. 2016, pp. 770–778.

[30] W. Wang, E. Xie, X. Li, D.-P. Fan, K. Song, D. Liang, T. Lu, P. Luo, and L. Shao, "Pyramid vision transformer: A versatile backbone for dense prediction without convolutions," in *Proc. IEEE ICCV*, Oct. 2021, pp. 548–558.

[31] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. NeurIPS*, Dec. 2017, pp. 6000–6010.

[32] G. Li, Z. Liu, Z. Bai, W. Lin, and H. Ling, "Lightweight salient object detection in optical remote sensing images via feature correlation," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–12, 2022.

[33] H. Luo, J. Wang, and B. Liang, "Spatial attention feedback iteration for lightweight salient object detection in optical remote sensing images," *IEEE J. Sel. Top. Appl. Earth Observ. Remote Sens.*, vol. 17, pp. 13 809–13 823, Jul. 2024.

[34] 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, pp. 1–14, 2025.

[35] J. Han, F. Sun, Y. Hou, J. Sun, and H. Li, "Exploring a lightweight and efficient network for salient object detection in ORSI," *IEEE Trans. Geosci. Remote Sens.*, vol. 63, pp. 1–14, Jul. 2025.

[36] G. Li, S. Shi, Y. Wu, W. Lin, and Z. Bai, "Lightweight ORSI salient object detection via frequency and mutual assistance attention," *IEEE Trans. Geosci. Remote Sens.*, vol. 64, pp. 1–12, 2026.

[37] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "MobileNetV2: Inverted residuals and linear bottlenecks," in *Proc. IEEE CVPR*, Jun. 2018, pp. 4510–4520.

[38] X. Ding, X. Zhang, N. Ma, J. Han, G. Ding, and J. Sun, "RepVGG: Making VGG-style ConvNets great again," in *Proc. IEEE CVPR*, Jun. 2021, pp. 13 728–13 737.

[39] S. Mehta and M. Rastegari, "MobileViT: Light-weight, general-purpose, and mobile-friendly vision transformer," in *Proc. ICLR*, Apr. 2021, pp. 1–13.

[40] R. M. Haralick, K. Shanmugam, and I. Dinstein, "Textural features for image classification," *IEEE Trans. Syst. Man Cybern.*, vol. SMC-3, no. 6, pp. 610–621, 1973.

[41] Z. Huang, H. Chen, B. Liu, and Z. Wang, "Semantic-guided attention refinement network for salient object detection in optical remote sensing images," *Remote Sens.*, vol. 13, no. 11, p. 2163, May 2021.

[42] X. Zhou, K. Shen, L. Weng, R. Cong, B. Zheng, J. Zhang, and C. Yan, "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.

[43] C. Zeng, J. Zhang, and S. Kwong, "Learning conditional diffusion transformer for salient object detection in optical remote sensing images," *IEEE Trans. Cybern.*, vol. 56, no. 8, pp. 4255–4268, 2026.

[44] H. Luo, J. He, and S. Yang, "Prompt-driven multi-task learning with task tokens for ORSI salient object detection," *IEEE J. Sel. Top. Appl. Earth Observ. Remote Sens.*, pp. 1–15, Sept. 2025.

[45] X. Fang, Q. Wang, Q. Chen, and G. Li, "Interactive edge awareness network for salient object detection in optical remote sensing images," *Expert Syst. Appl.*, vol. 316, p. 131879, Jun. 2026.

[46] J. Zhao, Y. Jia, L. Ma, and L. Yu, "Recurrent adaptive graph reasoning network with region and boundary interaction for salient object detection in optical remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, pp. 1–20, Jul. 2024.

[47] Y. Jia, J. Zhao, L. Ma, and L. Yu, "Multistrategy region and boundary interaction network for salient object detection in optical remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 63, pp. 1–16, Jul. 2025.

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

[49] M. Feng, H. Lu, and E. Ding, "Attentive feedback network for boundary-aware salient object detection," in *Proc. IEEE CVPR*, Jun. 2019, pp. 1623–1632.

[50] Y. Quan, H. Xu, R. Wang, Q. Guan, and J. Zheng, "ORSI salient object detection via progressive semantic flow and uncertainty-aware refinement," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, pp. 1–13, Jan. 2024.

[51] Y. Ge, T. Liang, J. Ren, M. He, H. Bi, and Q. Zhang, "Semantic awareness aggregation for salient object detection in remote sensing images," *Eng. Appl. Artif. Intell.*, vol. 160, p. 111837, Nov. 2025.

[52] J. Zhao, Y. Jia, L. Ma, and L. Yu, "Multilevel interactive reverse-guided network for salient object detection in optical remote sensing images," *IEEE J. Sel. Top. Appl. Earth Observ. Remote Sens.*, vol. 17, pp. 12 983–12 999, Jul. 2024.

[53] G. Li, Z. Bai, Z. Liu, X. Zhang, and H. Ling, "Salient object detection in optical remote sensing images driven by transformer," *IEEE Trans. Image Process.*, vol. 32, pp. 5257–5269, Sept. 2023.

[54] 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, pp. 1–12, Aug. 2025.

[55] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," in *Proc. NeurIPS*, Dec. 2020, pp. 6840–6851.

[56] J. Song, C. Meng, and S. Ermon, "Denoising diffusion implicit models," in *Proc. ICLR*, May 2021, pp. 1–12.

[57] Y. Hou and T. Li, "DiffORSINet: Salient object detection in optical remote sensing images via conditional diffusion model," *IEEE Trans. Geosci. Remote Sens.*, vol. 63, pp. 1–13, 2025.

[58] Y. Shui, Y. Chang, and Z. Wang, "Fixation-guided diffusion for salient object detection in optical remote sensing images," *IEEE Geosci. Remote Sens. Lett.*, vol. 23, pp. 1–5, 2026.

[59] J. Yang, B. Zhong, Q. Liang, Y. Tan, H. Xia, and S. Song, "Mamba-driven diffusion model for salient object detection in optical remote sensing images," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 36, no. 5, pp. 6095–6107, 2026.

[60] Z. Wu, L. Su, and Q. Huang, "Cascaded partial decoder for fast and accurate salient object detection," in *Proc. IEEE CVPR*, Jun. 2019, pp. 3902–3911.

[61] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "CBAM: Convolutional block attention module," in *Proc. ECCV*, Sept. 2018, pp. 3–19.

[62] Q. Wang, Y. Liu, Z. Xiong, and Y. Yuan, "Hybrid feature aligned network for salient object detection in optical remote sensing imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–15, 2022.

[63] J. Zhao, Y. Jia, L. Ma, and L. Yu, "Adaptive dual-stream sparse transformer network for salient object detection in optical remote sensing images," *IEEE J. Sel. Top. Appl. Earth Observ. Remote Sens.*, vol. 17, pp. 5173–5192, Feb. 2024.

[64] M.-M. Cheng and D.-P. Fan, "Structure-measure: A new way to evaluate foreground maps," *Int. J. Comput. Vis.*, vol. 129, no. 9, pp. 2622–2638, Sept. 2021.

[65] R. Achanta, S. Hemami, F. Estrada, and S. Susstrunk, "Frequency-tuned salient region detection," in *Proc. IEEE CVPR*, Jun. 2009, pp. 1597–1604.

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

[67] T.-N. Le, T. V. Nguyen, Z. Nie, M.-T. Tran, and A. Sugimoto, "Anabranch network for camouflaged object segmentation," *Comput. Vis. Image Und.*, vol. 184, pp. 45–56, 2019.

[68] D.-P. Fan, G.-P. Ji, G. Sun, M.-M. Cheng, J. Shen, and L. Shao, "Camouflaged object detection," in *Proc. IEEE CVPR*, 2020, pp. 2774–2784.

[69] Y. Lv, J. Zhang, Y. Dai, A. Li, B. Liu, N. Barnes, and D.-P. Fan, "Simultaneously localize, segment and rank the camouflaged objects," in *Proc. IEEE CVPR*, 2021, pp. 11 586–11 596.

[70] C. He, K. Li, Y. Zhang, L. Tang, Y. Zhang, Z. Guo, and X. Li, "Camouflaged object detection with feature decomposition and edge reconstruction," in *Proc. IEEE CVPR*, 2023, pp. 22 046–22 055.

[71] L. Wang, J. Yang, Y. Zhang, F. Wang, and F. Zheng, "Depth-aware concealed crop detection in dense agricultural scenes," in *Proc. IEEE CVPR*, 2024, pp. 17 201–17 211.

[72] Y. Sun, S. Wang, C. Chen, and T.-Z. Xiang, "Boundary-guided camouflaged object detection," in *Proc. IJCAI*, 2022, pp. 1335–1341.

[73] D.-P. Fan, G.-P. Ji, M.-M. Cheng, and L. Shao, "Concealed object detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 10, pp. 6024–6042, 2022.

[74] F. Yang, Q. Zhai, X. Li, R. Huang, A. Luo, H. Cheng, and D.-P. Fan, "Uncertainty-guided transformer reasoning for camouflaged object detection," in *Proc. IEEE ICCV*, 2021, pp. 4126–4135.

[75] Q. Zhai, X. Li, F. Yang, C. Chen, H. Cheng, and D.-P. Fan, "Mutual graph learning for camouflaged object detection," in *Proc. IEEE CVPR*, 2021, pp. 12 992–13 002.

[76] B. Yin, X. Zhang, D.-P. Fan, S. Jiao, M.-M. Cheng, L. Van Gool, and Q. Hou, "CamoFormer: Masked separable attention for camouflaged object detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 46, no. 12, pp. 10 362–10 374, 2024.

[77] C. He, R. Zhang, F. Xiao, C. Fang, L. Tang, Y. Zhang, L. Kong, D.-P. Fan, K. Li, and S. Farsiu, "RUN: Reversible unfolding network for concealed object segmentation," in *Proc. ICML*, 2025, pp. 22 853–22 864.

[78] Z. Yu, X. Zhang, L. Zhao, Y. Bin, and G. Xiao, "Exploring deeper! Segment anything model with depth perception for camouflaged object detection," in *Proc. ACM MM*, 2024, pp. 4322–4330.

[79] X. Zhang, T. Xiao, G.-P. Ji, X. Wu, K. Fu, and Q. Zhao, "Explicit motion handling and interactive prompting for video camouflaged object detection," *IEEE Trans. Image Process.*, vol. 34, pp. 2853–2866, 2025.