Noname manuscript No.
(will be inserted by the editor)

IPDiff: Diffusion-driven ORSI Salient Object Detection with Information Reconstruction and Multi-Prior Guidance

Gongyang Li[1] · Zhen Bai[2] · Runmin Cong[3] · Dan Zeng[1] · Weisi Lin[4] · Xiao-Ping Zhang[5]

Received: date / Accepted: date

Abstract Existing Salient Object Detection in Optical Remote Sensing Image (ORSI-SOD) methods mainly adopt the static inference strategy, which uses fixed trained model parameters for saliency inference in the testing phase. This means that even if the generated saliency map has errors, it cannot be further optimized. In this paper, we propose the novel *IPDiff*, a *Diff*usion-driven ORSI-SOD method with *I*nformation Reconstruction and Multi-*P*rior Guidance. We build IPDiff based on a unique dynamic optimization strategy, which endows IPDiff with the ability to iteratively optimize saliency maps with a dynamic parameter. Specifically, we formulate ORSI-SOD as a conditional diffusion problem in IPDiff. IPDiff first extracts informative conditional priors from ORSIs, including the saliency prior and the hierarchical priors, in the prior network with the assistance of the information reconstruction-driven attention module. The saliency prior can provide positional information of salient objects, while the hierarchical priors can provide specific detail and semantic information of salient objects. Under the guidance of these priors, IPDiff then iteratively denoises random noise as the timestep dynamically changes in the denoising network, generating saliency maps that are close to ground truths. Notably, we simultaneously supervise IPDiff in both spatial and spectral domains through a hybrid loss function to achieve efficient network training. Comprehensive experiments on public ORSSD, EORSSD, and ORSI-4199 datasets demonstrate that our proposed IPDiff achieves the best performance compared to 46 state-of-the-art methods. Our code and results will be released once the paper is accepted. The code and results of our method are available at https://github.com/MathLee/IPDiff.

Keywords Salient object detection · Optical remote sensing image · Dynamic optimization strategy · Information reconstruction · Multiple priors

✉ Runmin Cong
E-mail: rmcong@sdu.edu.cn

✉ Dan Zeng
E-mail: dzeng@shu.edu.cn

Gongyang Li
E-mail: ligongyang@shu.edu.cn

Zhen Bai
E-mail: bz536476@163.com

Weisi Lin
E-mail: wslin@ntu.edu.sg

Xiao-Ping Zhang
E-mail: xpzhang@ieee.org

[1] School of Communication and Information Engineering, Shanghai University, Shanghai, China

[2] Department of Medical Equipment, the First Affiliated Hospital of Zhengzhou University, Zhengzhou, China

[3] School of Control Science and Engineering, Shandong University, Jinan, China

[4] School of Computer Science and Engineering, Nanyang Technological University, Singapore

[5] Shenzhen Key Laboratory of Ubiquitous Data Enabling, Tsinghua Shenzhen International Graduate School, Tsinghua University, Shenzhen, China

1 Introduction

Optical Remote Sensing Images (ORSIs) refer to images captured by cameras mounted on spacecraft, drones,

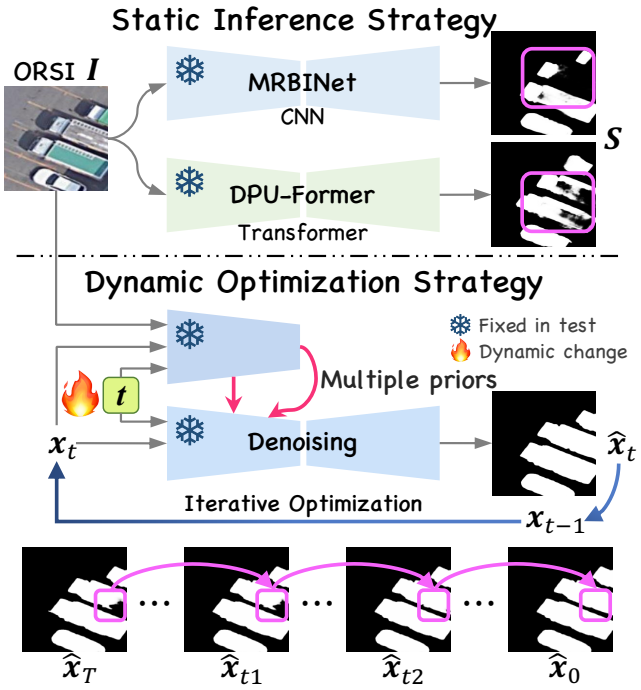

Fig. 1: Two strategies in ORSI-SOD. Existing static inference strategy uses fixed trained model parameters for saliency inference, which is formulated as $\boldsymbol{S} = \phi_\theta(\boldsymbol{I})$ with fixed trained parameters θ. $\phi(\cdot)$ represents the ORSI-SOD model, such as CNN-based MRBINet (Jia et al, 2025) and transformer-based DPU-Former (Sun et al, 2025b). Our unique dynamic optimization strategy iteratively optimizes saliency maps as t dynamically changes, which is formulated as $\widehat{\boldsymbol{x}}_t = \psi_\theta(\boldsymbol{x}_t, t, \boldsymbol{I})\&\widehat{\boldsymbol{x}}_t \overset{\text{Sample}}{\longrightarrow} \boldsymbol{x}_{t-1}$ and $\boldsymbol{S} = \widehat{\boldsymbol{x}}_0$ with fixed trained parameters θ and dynamically changing t. $\psi(\cdot)$ represents our IPDiff.

airplanes, *etc*, and have the characteristics of strong intuitiveness, high spatial resolution, and rich spectral information (Li et al, 2025b). ORSI processing plays an important role in various fields, such as agriculture, military, oceanography, ecological protection, and geological exploration (Wu et al, 2025). Salient Object Detection (SOD) aims to pop out the most attractive objects in images or videos (Wang et al, 2017; Peng et al, 2021; Chen et al, 2021; Sun et al, 2022; Li et al, 2023d; Hu et al, 2024; Lin et al, 2024; Wu et al, 2026; Zhou et al, 2026; Shi et al, 2026; Chen et al, 2026; Li et al, 2026; Liu et al, 2023). It is a fundamental topic in the computer vision community. Recently, ORSI-SOD (Li et al, 2019; Zhang et al, 2021; Tu et al, 2022; Li et al, 2023a; Gu et al, 2024; Jia et al, 2025; Sun et al, 2025b) has become a hot topic. It can quickly locate eye-catching objects in ORSIs, and serves as the cornerstone for ORSI understanding and interpretation.

ORSI-SOD has made breakthroughs in succession with the assistance of deep learning technologies, such as Convolutional Neural Networks (CNNs) (Simonyan and Zisserman, 2015; He et al, 2016) and transformers (Wang et al, 2022b; Liu et al, 2021d). The existing ORSI-SOD methods can be divided into four categories, including CNN-based methods (Li et al, 2019; Huang et al, 2021; Zhang et al, 2021; Zhou et al, 2022; Tu et al, 2022; Li et al, 2022b; Zhou et al, 2023; Li et al, 2023b; Gu et al, 2024; Zhao et al, 2024b; Li et al, 2024; Zhao et al, 2024c; Quan et al, 2024; Gu et al, 2025; Jia et al, 2025), transformer-based methods (Bai et al, 2023; Li et al, 2023a; Dong et al, 2024; Sun et al, 2024; Teng et al, 2025; Sun et al, 2025b; Meng et al, 2025; Sun et al, 2025c; Xie et al, 2025), hybrid backbone-based methods (Wang et al, 2022a; Zhao et al, 2024a; Wang et al, 2025; Li et al, 2025a), and lightweight methods (Li et al, 2022a, 2023c; Luo et al, 2024; Cheng et al, 2024; Liu et al, 2025; Han et al, 2025; Li et al, 2025c). As the name implies, the CNN-based methods typically adopt CNNs, such as VGG (Simonyan and Zisserman, 2015) and ResNet (He et al, 2016), as backbones, and explore edge cues (Zhou et al, 2022; Tu et al, 2022; Zhao et al, 2024c), multi-level interaction (Zhang et al, 2021; Li et al, 2023b; Zhao et al, 2024b), and multi-input architecture (Li et al, 2019; Zhou et al, 2022). The transformer-based methods typically adopt transformers, such as pyramid vision transformer (Wang et al, 2022b) and Swin Transformer (Liu et al, 2021d), as backbones, and explore the global-local-global scheme (Bai et al, 2023), the global-to-local paradigm (Li et al, 2023a; Teng et al, 2025), and the global-local integration strategy (Sun et al, 2024, 2025b; Xie et al, 2025). While hybrid backbone-based methods simultaneously use CNNs and transformers. Different from the above three categories, the lightweight methods aim to achieve a balance between performance and model complexity for practical applications.

The above four categories of ORSI-SOD methods can be summarized as the predictive framework, mainly training a prediction model to learn to infer saliency values from a large amount of ORSI data. After training, they adopt the static inference strategy for saliency inference, *i.e.*, using fixed trained model parameters to infer saliency maps once. This strategy has an obvious drawback, *i.e.*, the erroneous saliency values once occurring cannot be corrected. As shown in Fig. 1, the erroneous region (pink boxes) in the saliency maps generated by CNN-based MRBINet (Jia et al, 2025) and transformer-based DPU-Former (Sun et al, 2025b) will persist. Different from the predictive framework in ORSI-SOD, researchers develop the generative adversarial framework for SOD in Natural Scene Images

(NSI-SOD) (Ji et al, 2018; Wu et al, 2020). In the training phase, they train both the generative model and the discriminative model to improve the ability of the generative model to generate high-quality saliency maps. However, after training, they use the generative model with fixed training parameters to infer saliency values once, which also belongs to the static inference strategy. In addition, the generative adversarial framework has training instability and mode collapse issues, which limit its application in SOD.

The generative adversarial framework is eye-catching, inspiring us to explore a generation solution for ORSI-SOD. We have noticed that recent diffusion models (Ho et al, 2020; Nichol and Dhariwal, 2021; Song et al, 2021; Rombach et al, 2022; Hoogeboom et al, 2023) may be a suitable generative solution. Diffusion models are originally intended for image and video generation topics. Its core principle lies in simulating the processes of noise diffusion and reverse denoising to generate realistic data from random noise. Moreover, it controls the generation process through dynamic timesteps. Therefore, we attempt to solve ORSI-SOD from a generative perspective based on the diffusion model, and propose a dynamic optimization strategy to break through the limitations of the previous static inference strategy. The strategy performs denoising with fixed trained parameters and dynamically changing timesteps in the testing phase, which means *the saliency inference is no longer a one-time occurrence*. With this strategy, we propose the novel IPDiff, a diffusion-driven ORSI-SOD method with information reconstruction and multi-prior guidance. As shown in Fig. 1, the dynamic optimization strategy endows our IPDiff with the ability to iteratively optimize saliency maps, *i.e.*, the missing truck carriage in the saliency map is completely segmented out with the dynamic change of timesteps.

Unlike the vanilla diffusion model that directly generates images from noise, we regard ORSI-SOD as a conditional diffusion problem. Our IPDiff treats ORSIs as condition information to guide denoising from noise. In particular, IPDiff consists of a prior network and a denoising network. The prior network extracts specific conditional priors from ORSIs, including the saliency prior and the hierarchical priors. To extract the useful content of priors, we propose the Information Reconstruction-driven Attention Module (IRAM) to adaptively reconstruct features in the spectral domain. Subsequently, these multiple priors are sequentially injected into the denoising network. In the denoising network, the saliency prior stabilizes the position of salient objects, and then the hierarchical priors hierarchically enrich the details and semantics of salient objects. To mitigate the negative impact of the noisy mask, we propose the Information Perturbation Module (IPM) to enhance the anti-interference capability of the denoising network. Notably, IPM is only equipped in the training phase. With all components working together, our IPDiff can generate accurate saliency maps through iterative optimization as the timestep dynamically changes, showing good adaptation to the complex and variable scenes of ORSIs.

Our main contributions are summarized as follows:

- We propose a novel diffusion-driven ORSI-SOD framework based on the unique dynamic optimization strategy, namely *IPDiff*, which differs from previous methods based on the static inference strategy. IPDiff formulates ORSI-SOD as a conditional diffusion problem, which first extracts conditional priors from ORSIs as guidance, and then iteratively denoises random noise to generate saliency maps close to ground truths.
- We propose the IRAM to achieve robust enhancement on basic features, generating informative conditional priors. IRAM reconstructs information in the spectral domain through adaptive spectrum decoupling and information aggregation, and produces the attention map from the reconstructed information to enhance features.
- We propose a Multi-Prior Guidance Denoising Network to optimize saliency maps step-by-step by denoising random noise under the guidance of the saliency prior and the hierarchical priors. Notably, our denoising network has strong feature representation and anti-interference capabilities, as it is equipped with multiple IPMs in the training phase.

2 Related Work

2.1 Salient Object Detection in Optical Remote Sensing Images

Recently, ORSI-SOD has developed rapidly and occupies an important position in the field of SOD. This is attributed to the repeated breakthroughs in deep learning technologies, such as CNNs (Simonyan and Zisserman, 2015; He et al, 2016), transformers (Wang et al, 2022b; Liu et al, 2021d), and attention mechanisms (Woo et al, 2018; Vaswani et al, 2017). With the application of various technologies, the challenges of ORSI-SOD have been overcome one by one, and its performance has gradually improved. At present, existing ORSI-SOD methods can be divided into four categories. The first three categories are classified according to the differences in the backbones used, namely CNN-based methods, transformer-based methods, and hybrid

backbone-based methods. The last category focuses on the complexity of the model, called lightweight methods.

CNN-based ORSI-SOD methods (Li et al, 2019; Huang et al, 2021; Zhang et al, 2021; Zhou et al, 2022; Tu et al, 2022; Li et al, 2022b; Zhou et al, 2023; Li et al, 2023b; Gu et al, 2024; Zhao et al, 2024b; Li et al, 2024; Zhao et al, 2024c; Quan et al, 2024; Gu et al, 2025; Jia et al, 2025) account for a large proportion, due to the classic architecture of CNNs (*i.e.*, VGG (Simonyan and Zisserman, 2015) and ResNet (He et al, 2016)) and their powerful feature extraction capabilities. In this category, researchers employed edge cues to outline the boundaries of salient objects, improving the fine-grained details of salient objects (Zhou et al, 2023, 2022; Li et al, 2022b; Tu et al, 2022; Zhao et al, 2024c; Jia et al, 2025). While Huang *et al.* (Huang et al, 2021) and Quan *et al.* (Quan et al, 2024) exploited high-level semantic cues to accurately locate salient objects and reduce omissions of salient objects. The multi-level interaction is a popular strategy. It aims to explore the complementarity of information at different granularities among multi-level features of various scales to better represent salient objects (Zhang et al, 2021; Li et al, 2023b; Zhao et al, 2024b; Li et al, 2024; Gu et al, 2025, 2024). Different from the above methods, Li *et al.* (Li et al, 2019) and Zhou *et al.* (Zhou et al, 2022) used the multi-input architecture to directly extract multi-scale features from multiple ORSIs of different scales to adapt to complex ORSI scenes.

Transformer-based ORSI-SOD method (Bai et al, 2023; Li et al, 2023a; Dong et al, 2024; Sun et al, 2024; Teng et al, 2025; Sun et al, 2025b; Meng et al, 2025; Sun et al, 2025c; Xie et al, 2025) is a rising star. It utilizes transformers (Wang et al, 2022b; Liu et al, 2021d) to establish long-range dependencies of ORSIs, extracting the global information of ORSIs. Based on the global information, the global-local-global scheme (Bai et al, 2023), the global-to-local paradigm (Li et al, 2023a; Teng et al, 2025), and the global-local integration strategy (Sun et al, 2024, 2025b; Xie et al, 2025) have been proposed successively to interact global and local information from different perspectives. These methods overcame the limitations of CNN-based methods which cannot model global information, further improving detection performance. Moreover, to address the problem of complex orientations in ORSIs, researchers extracted directional cues from global information by using convolutions with multiple directions to better perceive and determine the directions of salient objects (Li et al, 2023a; Teng et al, 2025; Sun et al, 2025c).

Hybrid backbone-based methods (Wang et al, 2022a; Zhao et al, 2024a; Wang et al, 2025; Li et al, 2025a) draw on the advantages of different types of backbones and jointly extract comprehensive features. There are three ways to integrate different types of backbones. The first way integrates CNN blocks and transformer blocks in a single-stream structure. Wang *et al.* (Wang et al, 2022a) followed the first way to model local and global context at different levels from ORSIs. The second way uses CNN and transformer in parallel in a dual-stream structure. Zhao *et al.* (Zhao et al, 2024a) and Wang *et al.* (Wang et al, 2025) followed the second way to extract local and global information and fuse them. The last way is also a dual-stream structure. The first stream is the same as the first way, and the second stream integrates CNN blocks and Mamba blocks (Liu et al, 2024). Li *et al.* (Li et al, 2025a) developed the complex backbone to leverage the global modeling capabilities of transformer and the local processing advantages of Mamba, enriching the feature extraction manner of ORSI-SOD.

Lightweight ORSI-SOD methods (Li et al, 2022a, 2023c; Luo et al, 2024; Cheng et al, 2024; Liu et al, 2025; Han et al, 2025; Li et al, 2025c) generally use lightweight backbones and develop lightweight modules to achieve efficient SOD. As a pioneer, Li *et al.* (Li et al, 2022a) modified the vanilla VGG to a lightweight VGG backbone, greatly reducing model parameters to 4.09M. Subsequently, MobileNet series (Sandler et al, 2018; Howard et al, 2019) dominated the lightweight ORSI-SOD category. MobileNet-V2 (Sandler et al, 2018) was combined with the multi-level collaboration strategy (Li et al, 2023c; Liu et al, 2025) and the attention recursion mechanism (Luo et al, 2024), further reducing the model parameters and the computational load. Cheng *et al.* (Cheng et al, 2024) combined MobileNet-V3 (Howard et al, 2019) with the multi-level collaboration strategy, achieving competitive performance with 0.5G FLOPs. Differently, Han *et al.* (Han et al, 2025) introduced MobileViT (Mehta and Rastegari, 2021) to enhance the feature extraction, improving performance by increasing a few parameters. Li *et al.* (Li et al, 2025c) introduced RepVGG structure (Ding et al, 2021) to accelerate the inference speed to 161 fps.

Although the four aforementioned categories of methods have greatly promoted the development of ORSI-SOD, they all adopted the static inference strategy for one-time saliency inference. The inherent drawback of this strategy results in saliency maps inferred by existing methods being unoptimizable. This places high demands on the design of ORSI-SOD methods, requiring them to have strong adaptability to various scenes of ORSIs so that they can infer satisfactory saliency maps at one time. Obviously, this is hard. To break through the current situation, we introduce the dy-

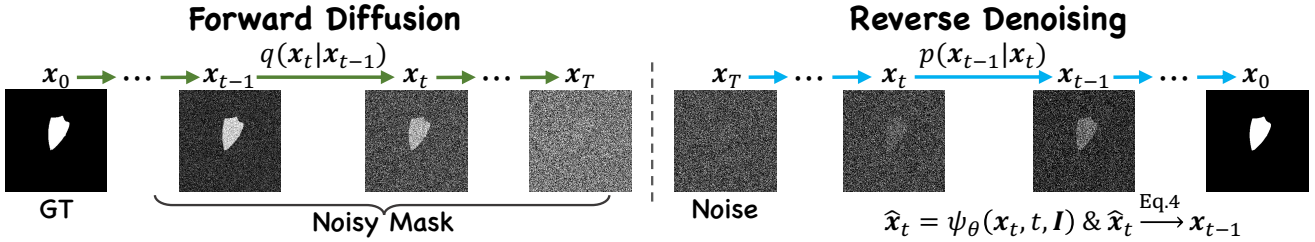

Fig. 2: Illustration of the forward diffusion process and the reverse denoising process.

namic optimization strategy into our ORSI-SOD solution, *i.e.*, IPDiff. The dynamic optimization strategy can perform multiple iterative inferences in the testing phase based on dynamically changing timesteps and fixed trained parameters. The erroneous saliency values in the previous inference can be corrected in the current inference. In this way, our IPDiff can continuously optimize saliency maps through multiple iterative inferences, being able to generate accurate saliency maps.

2.2 Denoising Diffusion Models

Denoising diffusion model is inspired by nonequilibrium thermodynamics (Ho et al, 2020). It learn the data distribution by simulating the two-way process of *gradual addition of noise* and *step-by-step removal of noise*, and ultimately generates realistic data from random noise, such as images and videos. Since its proposal, diffusion models have been optimized in multiple aspects, such as sampling acceleration, generation speed, synthesis resolution, scalability, and text-to-image generation (Nichol and Dhariwal, 2021; Song et al, 2021; Rombach et al, 2022; Hoogeboom et al, 2023).

The powerful generation capability of diffusion models has led to their application in some visual tasks, such as aerial semantic segmentation (Toker et al, 2024), cardiac ultrasound segmentation (Vyver et al, 2025), steel surface defect detection (Tai et al, 2025), and fixation prediction (Aydemir et al, 2024). Researchers promoted performance improvement by synthesizing training data through diffusion models, *i.e.*, using diffusion models for data augmentation. Obviously, these methods face the problems of data quality and data authenticity. We believe that the application of diffusion models in the above visual topics has not deviated from the data generation capability inherent to diffusion models.

Differently, some researchers used diffusion models to treat visual tasks, such as medical image segmentation (Wolleb et al, 2022; Wu et al, 2023, 2024), edge detection (Ye et al, 2024), change detection (Wen et al, 2024), and camouflaged object detection (Sun et al, 2025a), as image-to-image translation tasks. They modified the vanilla diffusion framework to a conditional diffusion framework, that is, extracting conditional information from inputs and inserting it into the denoising process to generate corresponding outputs. This way of using diffusion models provides fresh research directions for the computer vision community. The research focus will be on *how to effectively utilize conditional information to guide the denoising process* or *how to design an effective denoising network*.

Since the diffusion model has the ability to optimize the output according to the timestep, we handle ORSI-SOD with the diffusion model in this paper. Considering existing issues in ORSI-SOD, we regard ORSI-SOD as a conditional diffusion problem in our IPDiff. Existing methods based on conditional diffusion for other tasks (Wolleb et al, 2022; Wu et al, 2023, 2024; Ye et al, 2024; Wen et al, 2024; Sun et al, 2025a) usually only extract one type of conditional information from the input, which is insufficient. Differently, we extract multiple conditional priors from the input ORSI in our prior network, including the saliency prior and the hierarchical priors. The saliency prior is an initial saliency map generated from the input ORSI, and can provide the position information of salient objects to the encoder of our denoising network. The hierarchical priors are the different scale features of the input ORSI, and can provide specific detail and semantic information of salient objects to the encoder of our denoising network. Under the guidance of these priors, in our denoising network, the encoder can thoroughly perceive the current noisy mask and effectively guide the decoder to recover salient objects.

3 Methodology

3.1 Preliminaries

In general, the denoising diffusion framework (Ho et al, 2020) includes one process for *gradual addition of noise* (*i.e.*, the forward diffusion process) and one process for *step-by-step removal of noise* (*i.e.*, the reverse denoising process). As shown in Fig. 2, in the forward diffusion process, the ground truth (GT) $\boldsymbol{x}_0$ is gradually

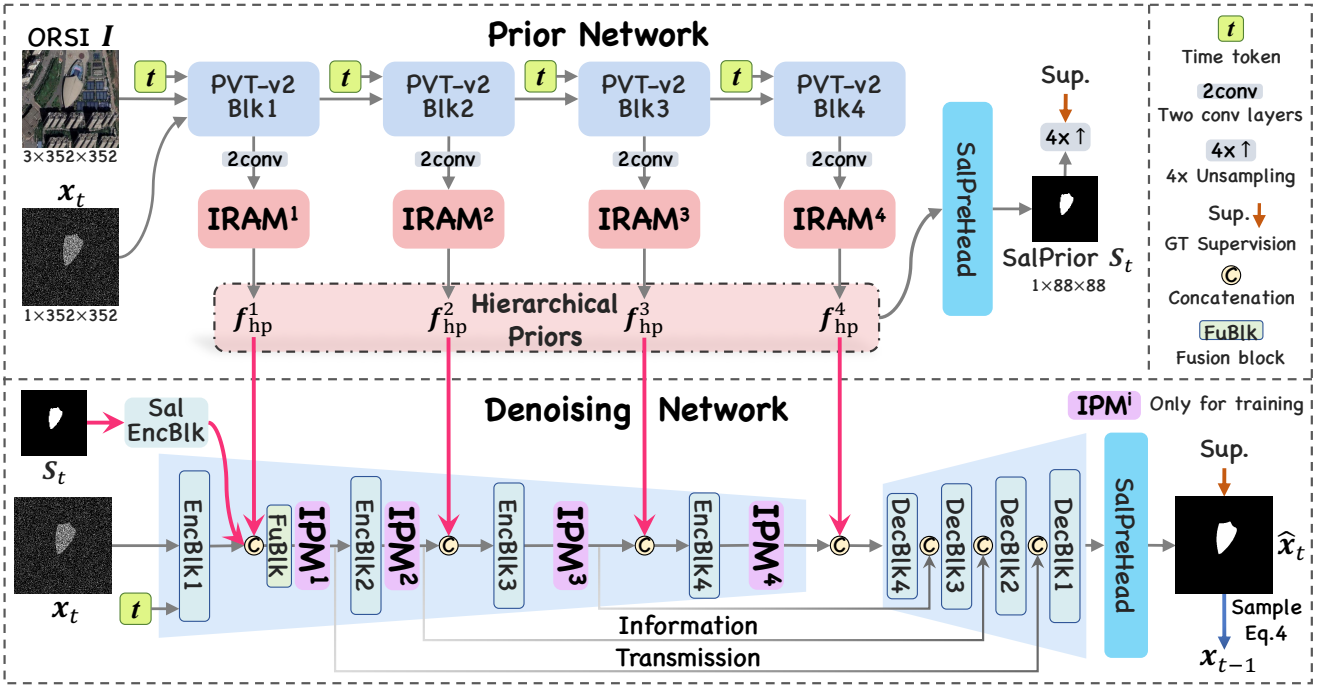

Fig. 3: The overall framework of the proposed IPDiff. IPDiff consists of a prior network and a denoising network. First, the prior network encodes specific conditional priors from ORSIs, including the saliency prior $\boldsymbol{S}_t$ and the hierarchical priors $\{\boldsymbol{f}^i_{\text{hp}}\}^4_{i=1}$, with the assistance of PVT-v2 and IRAMs. Then, under the guidance of these multiple priors, the denoising network denoises the noisy mask $\boldsymbol{x}_t$ to recover the clear one. Notably, we only equip IPMs in the encoder of the denoising network in the training phase to enhance the anti-interference capability of the denoising network.

noised over the timestep $t \in \{1, 2, ..., T\}$, obtaining a set of noisy masks $\{\boldsymbol{x}_t\}^T_{t=1}$. On the contrary, in the reverse denoising process, the random Gaussian noise $\boldsymbol{x}_T$ is gradually clear until it is recovered to the original data $\boldsymbol{x}_0$.

1) Forward Diffusion Process: The forward diffusion process is a Markov noising process, producing the noisy mask $\boldsymbol{x}_t$ as follows:

$$q(\boldsymbol{x}_t|\boldsymbol{x}_{t-1}) = \mathcal{N}(\boldsymbol{x}_t; \sqrt{1-\beta_t}\boldsymbol{x}_{t-1}, \beta_t\mathbf{I}), \tag{1}$$

where $\mathcal{N}(\cdot)$ is the Gaussian distribution, $\beta_t \in (0, 1)$ is the noise variance schedule (Ho et al, 2020), $\mathbf{I}$ is the identity matrix, and $\beta_t\mathbf{I}$ forms the covariance matrix. Therefore, through the propagation of Markov chain, starting from $\boldsymbol{x}_0$ (*i.e.*, GT), we can obtain $\boldsymbol{x}_t$ as follows:

$$q(\boldsymbol{x}_t|\boldsymbol{x}_0) = \prod_{s=1}^{t} q(\boldsymbol{x}_s|\boldsymbol{x}_{s-1}) = \mathcal{N}\left(\boldsymbol{x}_t; \sqrt{\bar{\alpha}_t}\boldsymbol{x}_0, (1-\bar{\alpha}_t)\mathbf{I}\right), \tag{2}$$

where $\bar{\alpha}_t = \prod^t_{s=1} \alpha_s = \prod^t_{s=1}(1-\beta_s)$. Through the above forward diffusion process, we can get a set of noisy masks $\{\boldsymbol{x}_t\}^T_{t=1}$ for training.

2) Reverse Denoising Process: The goal of the reverse denoising process is to model the posterior distribution $p(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t)$ as follows:

$$p(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t) = \mathcal{N}(\boldsymbol{x}_{t-1}; \mu(\boldsymbol{x}_t, t), \sigma^2_t\mathbf{I}), \tag{3}$$

where $\mu(\boldsymbol{x}_t, t)$ is the mean of the Gaussian distribution, formulated as $\mu(\boldsymbol{x}_t, t) = \frac{\sqrt{\alpha_t}(1-\bar{\alpha}_{t-1})}{1-\bar{\alpha}_t}\boldsymbol{x}_t + \frac{\sqrt{\bar{\alpha}_{t-1}}(1-\alpha_t)}{1-\bar{\alpha}_t}\widehat{\boldsymbol{x}}_0$, and σ^2_t is the variance of the Gaussian distribution, formulated as $\sigma^2_t = \frac{1-\bar{\alpha}_{t-1}}{1-\bar{\alpha}_t}\beta_t$.

To achieve this goal of the reverse denoising process, a network is designed to predict $\widehat{\boldsymbol{x}}_0$ from $\boldsymbol{x}_t$. Here, for convenience, in subsequent expressions, we use $\widehat{\boldsymbol{x}}_t$ instead of $\widehat{\boldsymbol{x}}_0$. $\mu(\boldsymbol{x}_t, t)$ is updated as follows:

$$\mu(\boldsymbol{x}_t, t) = \frac{\sqrt{\alpha_t}\,(1-\bar{\alpha}_{t-1})}{1-\bar{\alpha}_t}\boldsymbol{x}_t + \frac{\sqrt{\bar{\alpha}_{t-1}}\,(1-\alpha_t)}{1-\bar{\alpha}_t}\widehat{\boldsymbol{x}}_t. \tag{4}$$

Specifically, we achieve the network in our IPDiff. Since our IPDiff is conditioned on the input ORSI, we formulate it as follows:

$$\widehat{\boldsymbol{x}}_0/\widehat{\boldsymbol{x}}_t = \psi_\theta(\boldsymbol{x}_t, t, \boldsymbol{I}), \tag{5}$$

where $\psi(\cdot)$ is our IPDiff, θ is its parameters, and $\boldsymbol{I}$ represents the input ORSI. Thus, starting from the random

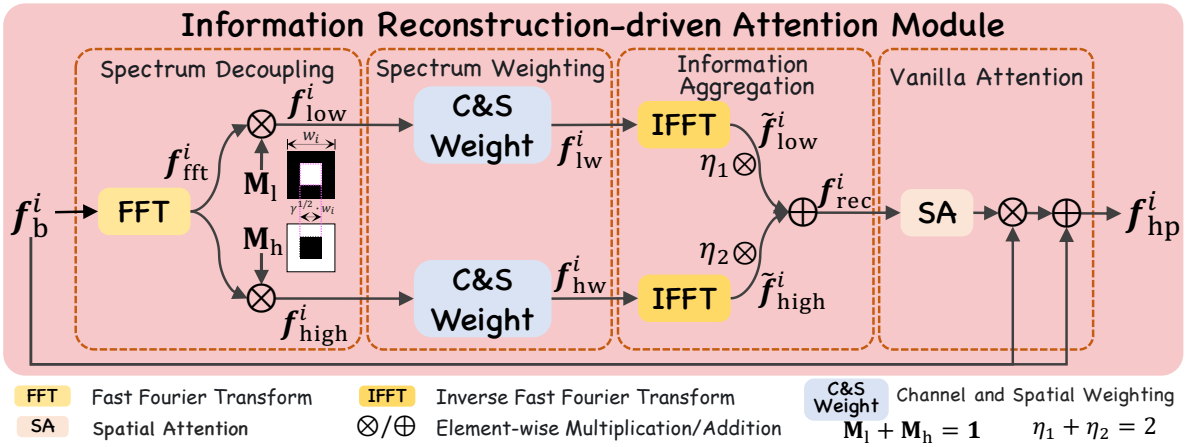

Fig. 4: Illustration of the Information Reconstruction-driven Attention Module.

Gaussian noise $\boldsymbol{x}_T \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, our IPDiff can generate a set of $\widehat{\boldsymbol{x}}_t$ to be sampled for generating corresponding $\boldsymbol{x}_{t-1}$, achieving progressive denoising of $\boldsymbol{x}_T$ to recover to the clear $\boldsymbol{x}_0$.

3.2 Network Overview

As shown in Fig. 3, the proposed IPDiff (*i.e.*, $\psi_\theta(\boldsymbol{x}_t, t, \boldsymbol{I})$) consists of a prior network and a denoising network. As their name suggests, the prior network is responsible for prior extraction, while the denoising network is responsible for noisy mask denoising.

Concretely, in the prior network, we adopt PVT-v2 (Wang et al, 2022b) as the backbone, whose inputs are the ORSI $\boldsymbol{I} \in \mathbb{R}^{3\times352\times352}$ and the noisy mask $\boldsymbol{x}_t \in \mathbb{R}^{1\times352\times352}$. In the four blocks of PVT-v2, each block is embedded with the time token $\boldsymbol{t}$ of an appropriate size. The time token $\boldsymbol{t}$ is derived from the timestep t. After each block, we use two convolutional layers to adjust the channel number of its output features to 256. Thus, we get four-level basic features denoted as $\{\boldsymbol{f}_{\mathrm{b}}^i \in \mathbb{R}^{c\times h_i\times w_i}\}_{i=1}^4$, where $c = 256$ and $h_i/w_i = \frac{352}{2^{i+1}}$. $\boldsymbol{f}_{\mathrm{b}}^i$ is sent to IRAM to reconstruct information in the spectral domain to extract informative content, generating a conditional prior $\boldsymbol{f}_{\mathrm{hp}}^i \in \mathbb{R}^{c\times h_i\times w_i}$. And four-level conditional priors form the hierarchical priors $\{\boldsymbol{f}_{\mathrm{hp}}^i\}_{i=1}^4$. Moreover, we extract the saliency prior $\boldsymbol{S}_t \in [0,1]^{1\times h_1\times w_1}$ $(h_1/w_1 = 88)$ from the integration of the hierarchical priors through a saliency prediction head, *i.e.*, SalPreHead.

Then, the saliency prior $\boldsymbol{S}_t$ and the hierarchical priors $\{\boldsymbol{f}_{\mathrm{hp}}^i\}_{i=1}^4$ are injected into the denoising network. The input of the denoising network is the noisy mask $\boldsymbol{x}_t$, and the time token $\boldsymbol{t}$ is only embedded in the first encoder block. $\boldsymbol{S}_t$ is integrated into the initial stage of the encoder, while $\{\boldsymbol{f}_{\mathrm{hp}}^i\}_{i=1}^4$ are hierarchically integrated into the encoder. Four IPMs are embedded into the encoder to improve the feature representation ability, and they only appear in the training phase. The output features of the encoder are transferred to the decoder through the information transmission. Finally, $\widehat{\boldsymbol{x}}_t \in [0,1]^{1\times352\times352}$ is predicted through a SalPreHead. According to Eq. 4, $\boldsymbol{x}_{t-1}$ is sampled from $\widehat{\boldsymbol{x}}_t$, and proceeds the next iteration of optimization. After completing all T iterations, we adopt $\widehat{\boldsymbol{x}}_0$ as the final saliency map $\boldsymbol{S}_{\mathrm{final}} \in [0,1]^{1\times352\times352}$.

3.3 Information Reconstruction-driven Attention Module

As is well known, objects in ORSIs have large size differences and varied shooting angles (Li et al, 2019; Zhang et al, 2021; Tu et al, 2022), which may lead to significant differences in the spatial representation of the same type of objects. This brings great challenges to ORSI-SOD. In addition, ORSIs often contain a large amount of spatial redundant information due to their large-scale coverage, such as large areas of uniform vegetation, water bodies, and sandy land (Li et al, 2019; Zhang et al, 2021; Tu et al, 2022). This may lead to excessive attention to these redundant regions when processing ORSIs in the spatial domain, resulting in a waste of computing resources. To alleviate the above issues, we propose the Information Reconstruction-driven Attention Module to reconstruct information in the spectral domain.

In the spectral domain, size differences essentially correspond to spectral differences, *i.e.*, large-size objects correspond to low-frequency components, while small-size objects correspond to high-frequency components. Shooting angle variation can be described by the rotation or translation characteristics of the spec-

trum. Thus, in spectral domain reconstruction, the robustness of the model's perception to size and shooting angle variation can be enhanced through adaptive weighting of spectral components, such as adjusting the weights of low and high frequencies. In addition, in the spectral domain, redundant information mostly corresponds to low-frequency components with concentrated energy, while key information (such as salient regions and edges) corresponds to specific high-frequency components. Through spectral domain reconstruction, the model can specifically focus on the spectral components containing key information, reduce the ineffective learning of redundant low frequencies, and improve the efficiency of model and the sensitivity to salient regions. Therefore, our IRAM reconstructs information through adaptive spectrum decoupling and adaptive information aggregation, so as to extract useful content in a flexible and learnable manner. We illustrate the detailed structure of IRAM in Fig. 4. In the following, we introduce IRAM in detail from four parts, *i.e.*, spectrum decoupling, spectrum weighting, information aggregation, and vanilla attention.

1) Spectrum Decoupling: The input of our IRAM is $\boldsymbol{f}_{\mathrm{b}}^{i} \in \mathbb{R}^{c\times h_i\times w_i}$. We transform $\boldsymbol{f}_{\mathrm{b}}^{i}$ into the spectral domain, generating $\boldsymbol{f}_{\mathrm{fft}}^{i} \in \mathbb{R}^{c\times h_i\times w_i}$ as follows:

$$\boldsymbol{f}_{\mathrm{fft}}^{i} = \mathrm{FFT}(\boldsymbol{f}_{\mathrm{b}}^{i}), \tag{6}$$

where $\mathrm{FFT}(\cdot)$ is the fast Fourier transform. Then, we decouple $\boldsymbol{f}_{\mathrm{fft}}^{i}$ into low-frequency components and high-frequency components.

Concretely, we adopt two binary masks $\{\mathbf{M}_{\mathrm{l}}, \mathbf{M}_{\mathrm{h}}\} \in \{0,1\}^{1\times h_i\times w_i}$ to achieve spectrum decoupling. As the spectrum decoupling part shown in Fig. 4, we set the central square area of $\mathbf{M}_{\mathrm{l}}$ to 1, and the other areas to 0. The hyperparameter $\gamma \in (0,1)$ controls the size of the central square area. We formulate $\mathbf{M}_{\mathrm{l}}$ as follows:

$$\mathbf{M}_{\mathrm{l}}^{h,w}(\gamma) = \begin{cases} 1, & h \in [(1-\gamma^{1/2})\frac{h_i}{2}, (1+\gamma^{1/2})\frac{h_i}{2}] \\ & \&w \in [(1-\gamma^{1/2})\frac{w_i}{2}, (1+\gamma^{1/2})\frac{w_i}{2}], \\ 0, & \text{otherwise.} \end{cases} \tag{7}$$

Following this, we can obtain $\mathbf{M}_{\mathrm{h}}$ as follows:

$$\mathbf{M}_{\mathrm{h}} = \mathbf{1} - \mathbf{M}_{\mathrm{l}}, \tag{8}$$

where $\mathbf{1} \in \{1\}^{1\times h_i\times w_i}$. These two binary masks are multiplied to $\boldsymbol{f}_{\mathrm{fft}}^{i}$ respectively to achieve spectrum decoupling, generating low-frequency components $\boldsymbol{f}_{\mathrm{low}}^{i} \in \mathbb{R}^{c\times h_i\times w_i}$ and high-frequency components $\boldsymbol{f}_{\mathrm{high}}^{i} \in \mathbb{R}^{c\times h_i\times w_i}$.

In practice, it is difficult to achieve an optimal manual decoupling of low-frequency components and high-frequency components for ORSI-SOD. Therefore, we adopt an adaptive way to decouple the spectrum, that is, we set the hyperparameter γ as a learnable hyperparameter. In this way, our adaptive spectral decoupling can learn from ORSI data the appropriate spectral decoupling hyperparameter that adapts to complex ORSI scenes.

2) Spectrum Weighting: Since different types of information, such as background, shape, objects, and edge, are distributed in specific frequency bands of low-frequency components and high-frequency components (Shan et al, 2021), we continue to weight the obtained low-frequency components and high-frequency components. Here, we perform the classical channel attention and spatial attention (Woo et al, 2018) on $\boldsymbol{f}_{\mathrm{low}}^{i}$ and $\boldsymbol{f}_{\mathrm{high}}^{i}$, respectively, to achieve spectrum weighting, generating $\boldsymbol{f}_{\mathrm{lw}}^{i} \in \mathbb{R}^{c\times h_i\times w_i}$ and $\boldsymbol{f}_{\mathrm{hw}}^{i} \in \mathbb{R}^{c\times h_i\times w_i}$. We formulate the above spectrum weighting as follows:

$$\begin{aligned} \boldsymbol{f}_{\mathrm{lw}}^{i} &= \mathrm{SA}(\mathrm{CA}(\boldsymbol{f}_{\mathrm{low}}^{i})), \\ \boldsymbol{f}_{\mathrm{hw}}^{i} &= \mathrm{SA}(\mathrm{CA}(\boldsymbol{f}_{\mathrm{high}}^{i})), \end{aligned} \tag{9}$$

where $\mathrm{CA}(\cdot)$ and $\mathrm{SA}(\cdot)$ are channel attention and spatial attention, respectively.

Enhancing low-frequency and high-frequency components in both channel and spatial dimensions is conducive to highlighting objects in specific frequency bands. This is a simple yet effective method.

3) Information Aggregation: Here, we perform the information aggregation on the weighted low-frequency and high-frequency components to achieve the information reconstruction. We convert $\boldsymbol{f}_{\mathrm{lw}}^{i}$ and $\boldsymbol{f}_{\mathrm{hw}}^{i}$ back to the spatial domain, generating $\tilde{\boldsymbol{f}}_{\mathrm{low}}^{i} \in \mathbb{R}^{c\times h_i\times w_i}$ and $\tilde{\boldsymbol{f}}_{\mathrm{high}}^{i} \in \mathbb{R}^{c\times h_i\times w_i}$ as follows:

$$\begin{aligned} \tilde{\boldsymbol{f}}_{\mathrm{low}}^{i} &= \mathrm{IFFT}(\boldsymbol{f}_{\mathrm{lw}}^{i}), \\ \tilde{\boldsymbol{f}}_{\mathrm{high}}^{i} &= \mathrm{IFFT}(\boldsymbol{f}_{\mathrm{hw}}^{i}), \end{aligned} \tag{10}$$

where $\mathrm{IFFT}(\cdot)$ is the inverse fast Fourier transform. Compared to the original $\boldsymbol{f}_{\mathrm{low}}^{i}$ and $\boldsymbol{f}_{\mathrm{high}}^{i}$, $\tilde{\boldsymbol{f}}_{\mathrm{low}}^{i}$ and $\tilde{\boldsymbol{f}}_{\mathrm{high}}^{i}$ have stronger representation capabilities.

At this time, we choose an adaptive way to aggregate $\tilde{\boldsymbol{f}}_{\mathrm{low}}^{i}$ and $\tilde{\boldsymbol{f}}_{\mathrm{high}}^{i}$ instead of simply aggregating them, which helps reconstruct information that is conducive to ORSI-SOD. We set a learnable hyperparameter η_1 as the weight of $\tilde{\boldsymbol{f}}_{\mathrm{low}}^{i}$, and another learnable hyperparameter η_2 as the weight of $\tilde{\boldsymbol{f}}_{\mathrm{high}}^{i}$, reconstructing $\boldsymbol{f}_{\mathrm{rec}}^{i} \in \mathbb{R}^{c\times h_i\times w_i}$ as follows:

$$\boldsymbol{f}_{\mathrm{rec}}^{i} = \eta_1 \otimes \tilde{\boldsymbol{f}}_{\mathrm{low}}^{i} + \eta_2 \otimes \tilde{\boldsymbol{f}}_{\mathrm{high}}^{i}, \tag{11}$$

where $\{\eta_1, \eta_1\} \in (0,2)$ and $\eta_1 + \eta_2 = 2$, and $\otimes$ is the element-wise multiplication. Adaptive information aggregation is a soft aggregation that is more suitable for

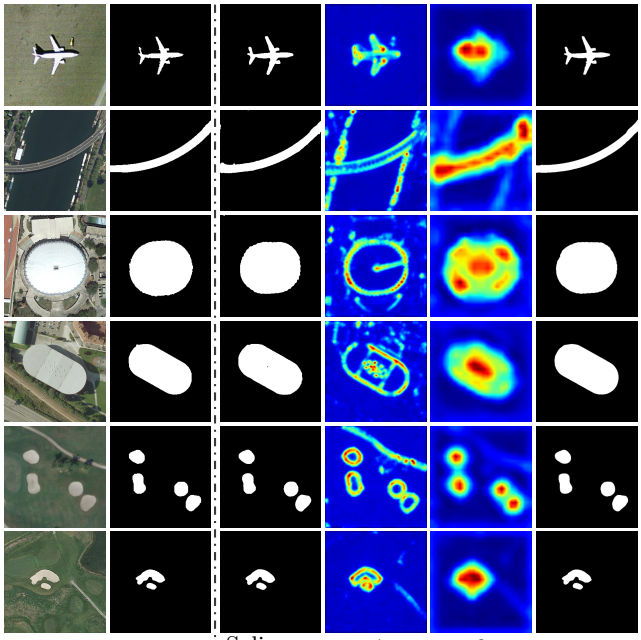

Fig. 5: Visualization of the saliency prior and the hierarchical priors ($\boldsymbol{f}^1_{\text{hp}}$ and $\boldsymbol{f}^3_{\text{hp}}$).

deep learning-based models than hard (or fixed) information aggregation (*i.e.*, $\eta_1 = \eta_2 = 1$). In this way, the reconstructed $\boldsymbol{f}^i_{\text{rec}}$ is discriminative.

4) Vanilla Attention: Finally, based on the vanilla spatial attention (Woo et al, 2018), we adopt the reconstructed $\boldsymbol{f}^i_{\text{rec}}$ to enhance the original input $\boldsymbol{f}^i_{\text{b}}$, generating the output of IRAM $\boldsymbol{f}^i_{\text{hp}} \in \mathbb{R}^{c\times h_i\times w_i}$ as follows:

$$\boldsymbol{f}^i_{\text{hp}} = (\text{SA}(\boldsymbol{f}^i_{\text{rec}}) \otimes \boldsymbol{f}^i_{\text{b}}) \oplus \boldsymbol{f}^i_{\text{b}}, \tag{12}$$

where $\oplus$ is the element-wise addition.

In summary, the adaptive spectrum decoupling and adaptive information aggregation endow IRAM with the ability to extract useful information from the basic features. Its output $\boldsymbol{f}^i_{\text{hp}}$ is also referred to as the conditional prior. As shown in Fig. 3, by using IRAMs to process $\{\boldsymbol{f}^i_{\text{b}}\}^4_{i=1}$, we obtain four-level conditional priors, forming the hierarchical priors $\{\boldsymbol{f}^i_{\text{hp}}\}^4_{i=1}$. The saliency prior originates from the hierarchical priors $\{\boldsymbol{f}^i_{\text{hp}}\}^4_{i=1}$. Therefore, IRAM serves as the cornerstone of multiple priors and plays a vital role in the prior network.

3.4 Multi-Prior Guidance Denoising Network

The denoising network is the core of the reverse denoising process. To improve its effectiveness, we inject it with specially extracted conditional priors from ORSIs, *i.e.*, the saliency prior and the hierarchical priors. In Fig. 5, we intuitively visualize the saliency prior and the hierarchical priors ($\boldsymbol{f}^1_{\text{hp}}$ and $\boldsymbol{f}^3_{\text{hp}}$). We observe that the saliency prior indeed provides accurate positional guidance, which helps to localize the main object regions during the reverse denoising process. For hierarchical priors, $\boldsymbol{f}^1_{\text{hp}}$ is the shallow-level prior, exhibiting clear and rich details and textural information (*e.g.*, object edges and fine structures). In contrast, the deep-level prior $\boldsymbol{f}^3_{\text{hp}}$ presents strong semantic information, with responses concentrated on the object regions while effectively suppressing background interference. We name our denoising network the multi-prior guidance denoising network. The detailed structure of the multi-prior guidance denoising network is illustrated at the bottom of Fig. 3.

The multi-prior guidance denoising network is built on the encoder-decoder architecture. Its main input is the noisy mask $\boldsymbol{x}_t$ accompanied by the time token $\boldsymbol{t}$. We first adopt an encoder block (*i.e.*, EncBlk1[1]) to extract basic features $\boldsymbol{f}^1_{\text{en}} \in \mathbb{R}^{c\times h_1\times w_1}$ from $\boldsymbol{x}_t$. Since the saliency prior $\boldsymbol{S}_t$ is a coarse saliency map, it contains positional information of salient objects. We adopt a saliency encoder block (*i.e.*, SalEncBlk[2]) to extract such positional feature $\boldsymbol{f}_{\text{sal}} \in \mathbb{R}^{c\times h_1\times w_1}$ from $\boldsymbol{S}_t$, and fuse it with $\boldsymbol{f}^1_{\text{en}}$ at the initial stage of the encoder. Our approach of injecting the saliency prior to the denoising network is unique, and it helps to stabilize the position of objects in complex ORSI scenes. Since $\boldsymbol{x}_t$ and $\boldsymbol{S}_t$ are essentially grayscale images, the information that can be extracted is relatively limited. Therefore, we also incorporate $\boldsymbol{f}^1_{\text{hp}}$, which can provide detail information, into them to achieve feature fusion through a fusion block (*i.e.*, FuBlk[3]), generating $\boldsymbol{f}^1_{\text{fu}} \in \mathbb{R}^{c\times h_1\times w_1}$. Then, we arrange three EncBlks[4] to continuously extract basic features of different levels in the encoder, getting $\{\boldsymbol{f}^i_{\text{en}} \in \mathbb{R}^{c\times h_i\times w_i}\}^4_{i=2}$. In this process, $\{\boldsymbol{f}^i_{\text{hp}}\}^4_{i=2}$ are hierarchically injected into the encoder, providing specific detail and semantic information of objects.

The stability of feature extraction is undoubtedly important for denoising a noisy mask. We introduce IPM into the above encoder of the multi-prior guidance denoising network, and arrange it after FuBlk, EncBlk2, EncBlk3, and EncBlk4, generating the corresponding perturbation feature $\boldsymbol{f}^i_{\text{ipm}} \in \mathbb{R}^{c\times h_i\times w_i}$. We formulate IPM as follows:

$$\boldsymbol{f}^i_{\text{ipm}} = \begin{cases} \boldsymbol{f}^i_{\text{fu}} \otimes \mathbf{M}^i_{\text{p}}, & i = 1, \\ \boldsymbol{f}^i_{\text{en}} \otimes \mathbf{M}^i_{\text{p}}, & i = 2, 3, 4, \end{cases} \tag{13}$$

[1] EncBlk1 consists of a 7×7 convolutional layer, a ResNet block, and a 3×3 convolutional layer.

[2] SalEncBlk consists of a 3×3 convolutional layer.

[3] FuBlk consists of a 3×3 convolutional layer.

[4] EncBlk2, EncBlk3, and EncBlk4 each consist of a 3×3 convolutional layer.

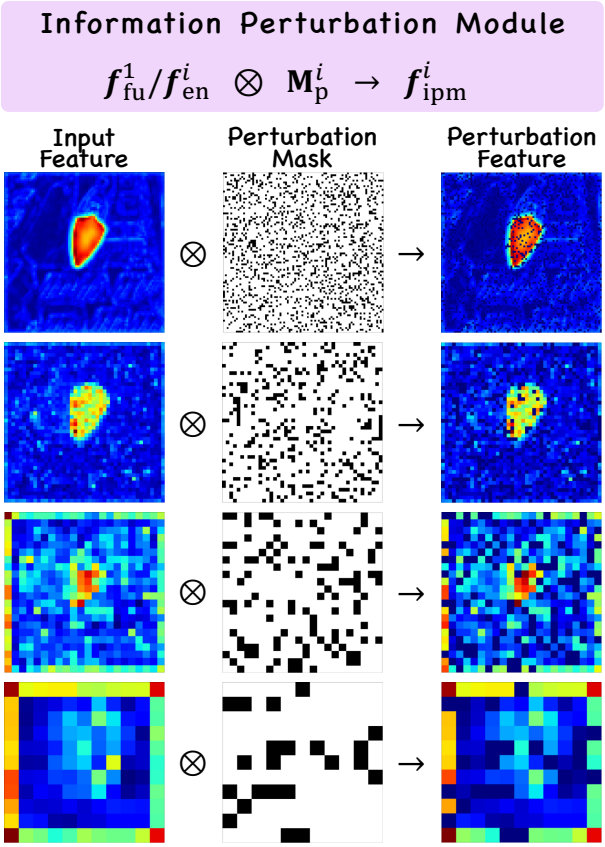

Fig. 6: Visualization of operations and features in the Information Perturbation Module. From top to bottom are features in IPM^1 to IPM^4, respectively.

where $\mathbf{M}^i_{\mathrm{p}}$ is the perturbation mask, belonging to $\{0,1\}^{c\times h_i\times w_i}$, and the proportion of 0 (*i.e.*, perturbation rate) in $\mathbf{M}^i_{\mathrm{p}}$ is $r \in [0,1]$. We visualize features in IPM with $r = 20\%$ in Fig. 6. The perturbation mask masks part of the information in the input features through multiplication. This operation enables the encoder to adaptively recover the lost information during the feature extraction process, thereby naturally improving the feature representation and anti-interference capabilities of the encoder. Notably, IPMs are only embedded into the encoder in the training phase, and will not disturb the testing phase.

The decoder corresponds to the encoder. It contains four blocks (*i.e.*, DecBlk[5]). Since the priors containing position, detail, and semantic information are hierarchically injected into the encoder, we introduce an information transmission to hierarchically transfer them to the decoder to fully utilize this information. With the help of SalPreHead, our denoising network outputs the optimized $\widehat{\boldsymbol{x}}_t$. $\widehat{\boldsymbol{x}}_t$ is used to sample $\boldsymbol{x}_{t-1}$ for the next iterative optimization until $\widehat{\boldsymbol{x}}_0$ is obtained as the final saliency map $\boldsymbol{S}_{\mathrm{final}}$. As shown in the last column of Fig. 5, benefiting from the positional guidance of the saliency prior and the multi-scale detail and semantic cues provided by the hierarchical priors, the final saliency map $\boldsymbol{S}_{\mathrm{final}}$ exhibits high consistency with the ground truth.

In summary, our multi-prior guidance denoising network can achieve stable denoising with the assistance of multiple priors and IPMs. Specifically, the operation of IPM is relatively simple, but its improvement in feature representation and anti-interference capabilities of our denoising network is significant. With the collaboration of all components, our denoising network can effectively resist the noise and interference specific to optical imaging (such as illumination changes, atmospheric scattering, and cloud occlusion) and has good adaptability to complex ORSI scenes.

3.5 Spatial-Spectral Collaborative Alignment-based Hybrid Loss Function

As illustrated in Fig. 3, our IPdiff has two items that need to be supervised in the training phase, *i.e.*, $\widehat{\boldsymbol{x}}_t \in [0,1]^{1\times 352\times 352}$ and $\boldsymbol{S}_t \in [0,1]^{1\times 88\times 88}$. Different previous methods that only supervise the network in the spatial domain (Li et al, 2019; Jia et al, 2025; Sun et al, 2025b), we simultaneously align $\widehat{\boldsymbol{x}}_t$ and $\boldsymbol{S}_t$ with GTs in both spatial and spectral domains to improve the efficiency of network training. Accordingly, we construct a spatial-spectral collaborative alignment-based hybrid loss function $L_{\mathrm{spa\&spe}}$ as follows:

$$L_{\mathrm{spa\&spe}} = \underbrace{L_{\mathrm{spa-base}} + L_{\mathrm{spa-edge}}}_{\text{Spatial}} + \underbrace{L_{\mathrm{spe}}}_{\text{Spectral}}. \tag{14}$$

For the spatial item, we not only retain the traditional $L_{\mathrm{spa-base}}$ including the weighted binary cross-entropy loss (ℓ_{wbce}) and the weighted intersection-over-union loss (ℓ_{wiou}), but also introduce the edge loss $L_{\mathrm{spa-edge}}$ to directly focus on edges of salient objects through the binary cross-entropy loss (ℓ_{bce}). We formulate $L_{\mathrm{spa-base}}$ and $L_{\mathrm{spa-edge}}$ as follows:

$$L_{\mathrm{spa-base}} = \ell_{\mathrm{wbce}}(\boldsymbol{S},\boldsymbol{G}) + \ell_{\mathrm{wiou}}(\boldsymbol{S},\boldsymbol{G}), \tag{15}$$

$$L_{\mathrm{spa-edge}} = \ell_{\mathrm{bce}}(|\boldsymbol{S} - \mathrm{AP}(\boldsymbol{S})|, |\boldsymbol{G} - \mathrm{AP}(\boldsymbol{G})|), \tag{16}$$

where $\boldsymbol{S}$ is predicted saliency map, $\boldsymbol{G}$ is GT, and $\mathrm{AP}(\cdot)$ is a 3×3 average pooling layer with stride of 1 and padding of 1.

[5] DecBlk1 is composed of a sequence of a 3×3 convolutional layer, a 2× upsampling layer, a 3×3 convolutional layer, another 2× upsampling layer, and a final 3×3 convolutional layer. DecBlk2, DecBlk3, and DecBlk4 each consist of two 3×3 convolutional layers followed by a 2× upsampling layer.

For the spectral item L_{spe}, we separate the real and imaginary parts of the predicted saliency map in the spectral domain, and adopt the structural similarity index loss (ℓ_{ssim}) to align these two parts with their corresponding parts of GT, respectively. We formulate L_{spe} as follows:

$$L_{\text{spe}} = \big[\underbrace{\ell_{\text{ssim}}\big(\text{Re}(\text{FFT}(\boldsymbol{S})), \text{Re}(\text{FFT}(\boldsymbol{G}))\big)}_{\text{Real}} + \underbrace{\ell_{\text{ssim}}\big(\text{Im}(\text{FFT}(\boldsymbol{S})), \text{Im}(\text{FFT}(\boldsymbol{G}))\big)}_{\text{Imaginary}}\big]/2, \tag{17}$$

where $\text{Re}(\cdot)$ and $\text{Im}(\cdot)$ mean the real part and the imaginary part, respectively. L_{spe} provides a new perspective for network training, which can increase the diversity of supervision on the traditional spatial supervision.

We adopt the hybrid loss function $L_{\text{spa\&spe}}$ to align $\widehat{\boldsymbol{x}}_t$ and $\boldsymbol{S}_t$ with GTs, and formulate the total loss function L_{total} as follows:

$$L_{\text{total}} = L_{\text{spa\&spe}}(\widehat{\boldsymbol{x}}_t, \boldsymbol{G}) + 0.5 \cdot L_{\text{spa\&spe}}(\text{Up}(\boldsymbol{S}_t), \boldsymbol{G}), \tag{18}$$

where $\text{Up}(\cdot)$ is the upsampling operation, and $\boldsymbol{G}$ belongs to $\{0,1\}^{1\times352\times352}$. Notably, we set the coefficient of $\boldsymbol{S}_t$ to 0.5, which is smaller than that of $\widehat{\boldsymbol{x}}_t$. This is because $\boldsymbol{S}_t$ with a small size tends to have a larger loss value, and a smaller coefficient can make the training focus more on the final output of IPDiff $\widehat{\boldsymbol{x}}_t$.

4 Experiments

4.1 Experimental Setup

1) Datasets: We conduct experiments on three commonly used ORSI-SOD datasets, *i.e.*, ORSSD (Li et al, 2019), EORSSD (Zhang et al, 2021), and ORSI-4199 (Tu et al, 2022) datasets. The ORSSD dataset[6] is a small dataset, containing 800 ORSIs and GTs, among which 600 images form the training set and 200 images form the test set. The EORSSD dataset[7] extends the ORSSD dataset to 2,000 ORSIs and GTs, among which 1,400 images form the training set and 600 images form the test set. The ORSI-4199 dataset[8] is a big dataset, containing 4,199 ORSIs and GTs, among which 2,000 images form the training set and 2,199 images form the test set. Following the traditional mode of ORSI-SOD (Li et al, 2023c,a; Quan et al, 2024), we train and test our IPDiff separately on each of the three datasets.

2) Implementation Details: We conduct experiments using PyTorch framework and an NVIDIA RTX 3090 GPU. The input size of our IPDiff is set to 352×352. We initialize the backbone (*i.e.*, PVT-v2 (Wang et al, 2022b)) of our prior network with the pre-trained parameters. We rotate and flip inputs for data augmentation. We adopt the AdamW optimizer to conduct network training for 150 epochs with an initial learning rate of $1e^{-4}$ and a batch size of 16. The perturbation rate r of IPM is set to 20%. For the denoising diffusion process, we set the total timestep T of our IPDiff to 10.

3) Evaluation Metrics: We use four quantitative evaluation metrics from an evaluation tool[9] to assess the performance of our IPDiff and all compared methods on ORSSD, EORSSD, and ORSI-4199 datasets, including S-measure (S_α, $\alpha = 0.5$) (Cheng and Fan, 2021), maximum F-measure ($F_\beta^{\max}$, $\beta^2 = 0.3$) (Achanta et al, 2009), maximum E-measure ($E_\xi^{\max}$) (Fan et al, 2018), and mean absolute error (MAE, $\mathcal{M}$). The first three metrics are better when they are larger, while the last one is better when it is smaller.

In addition, we adopt the parameter count, the computational cost, and the inference speed (without I/O time) to evaluate the model complexity. The first two metrics are better when they are smaller, while the last one is better when it is larger.

4.2 Comparison with State-of-the-arts

We compare our IPDiff with 46 state-of-the-art NSI-SOD methods, ORSI-SOD methods, and diffusion-driven segmentation methods on the EORSSD, ORSSD, and ORSI-4199 datasets. NSI-SOD methods include R3Net (Deng et al, 2018), PoolNet (Liu et al, 2019), EGNet (Zhao et al, 2019), GCPA (Chen et al, 2020), MINet (Pang et al, 2020), ITSD (Zhou et al, 2020), GateNet (Zhao et al, 2020), CSNet (Gao et al, 2020), SAMNet (Liu et al, 2021c), HVPNet (Liu et al, 2021b), SUCA (Li et al, 2021), PA-KRN (Xu et al, 2021), VST (Liu et al, 2021a), DPORTNet (Liu et al, 2022), DNTD (Fang et al, 2022), and ICON (Zhuge et al, 2023). ORSI-SOD methods include LVNet (Li et al, 2019), DAFNet (Zhang et al, 2021), SAR-Net (Huang et al, 2021), MJRBM (Tu et al, 2022), EMFINet (Zhou et al, 2022), CorrNet (Li et al, 2022a), MCCNet (Li et al, 2022b), HFANet (Wang et al, 2022a), ERPNet (Zhou et al, 2023), SeaNet (Li et al, 2023c), ACCoNet (Li et al, 2023b), GeleNet (Li et al, 2023a), GLGCNet (Bai et al, 2023), MIRGNet (Zhao

[6] https://li-chongyi.github.io/proj_optical_saliency.html
[7] https://github.com/rmcong/DAFNet_TIP20
[8] https://github.com/wchao1213/ORSI-SOD
[9] https://github.com/MathLee/MatlabEvaluationTools

Table 1: Quantitative and model complexity comparisons with state-of-the-art relevant methods on EORSSD, ORSSD, and ORSI-4199 datasets. ↑ indicates that the larger the better, while ↓ the opposite. We mark the best result in **bold** and the second best result in *italic*.

Methods	Type	Input Size	Param (M)↓	FLOPs (G)↓	Speed (fps)↑	EORSSD (Zhang et al, 2021)				ORSSD (Li et al, 2019)				ORSI-4199 (Tu et al, 2022)			
						$S_\alpha\uparrow$	$F_\beta^{\max}\uparrow$	$E_\xi^{\max}\uparrow$	$\mathcal{M}\downarrow$	$S_\alpha\uparrow$	$F_\beta^{\max}\uparrow$	$E_\xi^{\max}\uparrow$	$\mathcal{M}\downarrow$	$S_\alpha\uparrow$	$F_\beta^{\max}\uparrow$	$E_\xi^{\max}\uparrow$	$\mathcal{M}\downarrow$
R3Net$_{18}$ (Deng et al, 2018)	CN	300^2	56.1	47.5	2	.8184	.7498	.9483	.0171	.8141	.7456	.8913	.0399	.8142	.7847	.8880	.0401
PoolNet$_{19}$ (Liu et al, 2019)	CN	400×300	53.6	123.4	25	.8207	.7545	.9292	.0210	.8403	.7706	.9343	.0358	.8271	.8010	.8964	.0541
EGNet$_{19}$ (Zhao et al, 2019)	CN	380×320	108.0	291.9	9	.8601	.7880	.9570	.0110	.8721	.8332	.9731	.0216	.8464	.8267	.9161	.0440
GCPA$_{20}$ (Chen et al, 2020)	CN	320^2	67.1	54.3	23	.8869	.8347	.9524	.0102	.9026	.8687	.9509	.0168	-	-	-	-
MINet$_{20}$ (Pang et al, 2020)	CN	320^2	47.5	146.3	12	.9040	.8344	.9442	.0093	.9040	.8761	.9545	.0144	-	-	-	-
ITSD$_{20}$ (Zhou et al, 2020)	CN	228^2	17.1	54.5	16	.9050	.8523	.9556	.0106	.9050	.8735	.9601	.0165	-	-	-	-
GateNet$_{20}$ (Zhao et al, 2020)	CN	384^2	100.0	108.3	25	.9114	.8566	.9610	.0095	.9186	.8871	.9664	.0137	-	-	-	-
CSNet$_{20}$ (Gao et al, 2020)	CN	224^2	0.14	0.7	38	.8364	.8341	.9535	.0169	.8910	.8790	.9628	.0186	.8241	.8124	.9096	.0524
SAMNet$_{21}$ (Liu et al, 2021c)	CN	336^2	1.33	0.5	44	.8622	.7813	.9421	.0132	.8761	.8137	.9478	.0217	.8409	.8249	.9186	.0432
HVPNet$_{21}$ (Liu et al, 2021b)	CN	336^2	1.23	1.1	26	.8734	.8036	.9482	.0110	.8610	.7938	.9320	.0225	.8471	.8295	.9201	.0419
SUCA$_{21}$ (Li et al, 2021)	CN	256^2	117.7	56.4	24	.8988	.8229	.9520	.0097	.8989	.8484	.9584	.0145	.8794	.8692	.9438	.0304
PA-KRN$_{21}$ (Xu et al, 2021)	CN	600^2	141.1	617.7	16	.9192	.8639	.9616	.0104	.9239	.8890	.9680	.0139	.8491	.8415	.9280	.0382
VST$_{21}$ (Liu et al, 2021a)	TN	224^2	44.1	23.2	23	.9208	.8716	.9743	.0067	.9365	.9095	.9810	.0094	.8790	.8717	.9481	.0281
DPORTNet$_{22}$ (Liu et al, 2022)	CN	352^2	18.9	60.4	16	.8960	.8363	.9423	.0150	.8827	.8309	.9214	.0220	.8094	.7789	.8759	.0569
DNTD$_{22}$ (Fang et al, 2022)	CN	224^2	28.8	8.1	-	.8957	.8189	.9378	.0113	.8698	.8231	.9286	.0217	.8444	.8310	.9158	.0425
ICON$_{23}$ (Zhuge et al, 2023)	TN	352^2	65.7	61.8	34	.9185	.8622	.9687	.0073	.9256	.8939	.9704	.0116	.8752	.8763	.9521	.0282
LVNet$_{19}$ (Li et al, 2019)	CO	128^2	207.0	-	1	.8630	.7794	.9254	.0146	.8815	.8263	.9456	.0207	-	-	-	-
DAFNet$_{21}$ (Zhang et al, 2021)	CO	128^2	29.3	68.5	26	.9166	.8614	**.9861**	.0060	.9191	.8928	.9771	.0113	-	-	-	-
SARNet$_{21}$ (Huang et al, 2021)	CO	336^2	25.9	129.7	47	.9240	.8719	.9620	.0099	.9134	.8850	.9557	.0187	-	-	-	-
MJRBM$_{22}$ (Tu et al, 2022)	CO	352^2	43.5	95.7	32	.9197	.8656	.9646	.0099	.9204	.8842	.9623	.0163	.8593	.8493	.9311	.0374
EMFINet$_{22}$ (Zhou et al, 2022)	CO	256^2	107.3	480.9	25	.9290	.8720	.9711	.0084	.9366	.9002	.9737	.0109	.8675	.8584	.9340	.0330
CorrNet$_{22}$ (Li et al, 2022a)	CO	256^2	4.09	21.1	100	.9289	.8778	.9696	.0083	.9380	.9129	.9790	.0098	.8623	.8560	.9330	.0366
MCCNet$_{22}$ (Li et al, 2022b)	CO	256^2	67.6	112.8	95	.9327	.8904	.9755	.0066	.9437	.9155	.9800	.0087	.8746	.8690	.9413	.0316
HFANet$_{22}$ (Wang et al, 2022a)	HO	448^2	60.5	68.3	26	.9380	.8876	.9740	.0070	.9399	.9112	.9770	.0092	.8767	.8700	.9431	.0314
ERPNet$_{23}$ (Zhou et al, 2023)	CO	224^2	56.5	87.2	50	.9210	.8632	.9603	.0089	.9254	.8974	.9710	.0135	.8670	.8553	.9290	.0357
SeaNet$_{23}$ (Li et al, 2023c)	CO	288^2	2.76	1.7	96	.9208	.8649	.9710	.0073	.9260	.8942	.9767	.0105	.8772	.8653	.9426	.0308
ACCoNet$_{23}$ (Li et al, 2023b)	CO	256^2	102.5	179.9	81	.9290	.8837	.9727	.0074	.9437	.9149	.9796	.0088	.8775	.8686	.9412	.0314
GeleNet$_{23}$ (Li et al, 2023a)	TO	352^2	25.4	11.7	30	.9376	.8923	*.9828*	.0064	.9469	.9254	.9860	.0079	.8862	.8842	.9544	.0264
GLGCNet$_{23}$ (Bai et al, 2023)	TO	352^2	25.1	9.8	21	.9375	.8924	.9803	.0055	.9488	.9236	.9864	.0071	.8839	.8808	.9508	.0274
MIRGNet$_{24}$ (Zhao et al, 2024b)	CO	256^2	78.7	136.2	46	.9383	*.8930*	.9789	.0056	.9455	.9192	.9812	.0081	-	-	-	-
SAFINet$_{24}$ (Luo et al, 2024)	CO	288^2	3.12	7.6	-	.9267	.8799	.9732	.0065	.9401	.9106	.9786	.0086	-	-	-	-
TSCNet$_{24}$ (Li et al, 2024)	CO	256^2	103.6	116.8	59	.9376	.8900	.9765	.0061	.9428	.9198	.9850	.0081	.8783	.8771	.9486	.0295
RAGRNet$_{24}$ (Zhao et al, 2024c)	CO	256^2	35.6	17.8	31	.9361	.8852	.9785	.0057	*.9507*	.9242	.9861	.0066	.8811	.8811	.9492	.0284
SFANet$_{24}$ (Quan et al, 2024)	CO	256^2	25.1	7.7	-	.9349	.8833	.9769	.0058	.9453	.9192	.9830	.0077	.8761	.8710	.9447	.0292
UDCNet-R$_{24}$ (Sun et al, 2024)	CO	352^2	72.3	101.2	43	.9310	.8821	.9774	.0056	.9497	.9239	.9850	.0068	.8802	.8808	.9515	.0266
TLCKDNet$_{24}$ (Dong et al, 2024)	TO	256^2	50.0	31.7	32	.9350	.8843	.9788	.0056	.9421	.9114	.9794	.0082	-	-	-	-
PRNet$_{24}$ (Gu et al, 2024)	TO	352^2	20.8	8.5	21	.9276	.8684	.9784	*.0054*	.9459	.9177	.9848	.0075	.8873	.8819	.9527	.0272
ADSTNet$_{24}$ (Zhao et al, 2024a)	HO	256^2	62.1	27.7	40	.9311	.8804	.9769	.0065	.9379	.9124	.9807	.0086	.8710	.8698	.9433	.0318
SOLNet$_{25}$ (Li et al, 2025c)	CO	256^2	6.52	8.1	161	.9171	.8609	.9623	.0078	.9284	.9012	.9734	.0111	-	-	-	-
SggNet$_{25}$ (Liu et al, 2025)	CO	288^2	2.70	1.4	108	.9278	.8871	.9762	.0068	.9342	.9030	.9758	.0111	-	-	-	-
BCARNet$_{25}$ (Gu et al, 2025)	CO	352^2	24.0	7.0	-	.9361	.8871	.9761	*.0054*	.9465	.9196	.9833	.0071	.8757	.8689	.9407	.0306
MRBINet$_{25}$ (Jia et al, 2025)	CO	256^2	32.4	42.8	9	.9351	.8852	.9766	.0056	.9474	.9199	.9851	.0069	.8824	.8800	.9489	.0268
DPU-Former$_{25}$ (Sun et al, 2025b)	TO	352^2	44.2	32.5	43	*.9401*	*.8930*	.9816	.0056	.9412	*.9263*	*.9868*	*.0062*	.8833	*.8877*	*.9547*	.0263
EnsemDiff$_{22}$ (Wolleb et al, 2022)	Diff	224^2	124.0	190.2	0.03	.8516	.7690	.9063	.0173	.8795	.8189	.9343	.0245	.8051	.7583	.8681	.0558
MedSegDiffv2$_{24}$ (Wu et al, 2024)	Diff	256^2	139.7	260.6	0.03	.7429	.5739	.7918	.0342	.7727	.6457	.8225	.0508	.7490	.6835	.8202	.0754
CamoDiff$_{25}$ (Sun et al, 2025a)	Diff	384^2	71.7	60.3	8	.9314	.8890	.9801	.0113	.9343	.9215	.9863	.0187	*.8891*	.8804	.9539	*.0250*
IPDiff (Ours)	Diff	352^2	82.6	70.4	4	**.9461**	**.9033**	**.9861**	**.0044**	**.9557**	**.9362**	**.9915**	**.0054**	**.8907**	**.8885**	**.9572**	**.0237**

CN/TN: CNN-/Transformer-based NSI-SOD method, CO/TO: CNN-/Transformer-based ORSI-SOD method.
HO: Hybrid backbone-based ORSI-SOD method, Diff: Diffusion-based method.
We reserve two decimal places for the parameter count of lightweight methods.

et al, 2024b), SAFINet (Luo et al, 2024), TSCNet (Li et al, 2024), RAGRNet (Zhao et al, 2024c), SFANet (Quan et al, 2024), UDCNet-R (Sun et al, 2024), TLCKDNet (Dong et al, 2024), PRNet (Gu et al, 2024), ADSTNet (Zhao et al, 2024a), SOLNet (Li et al, 2025c), SggNet (Liu et al, 2025), BCARNet (Gu et al, 2025), MRBINet (Jia et al, 2025), and DPU-Former (Sun et al, 2025b). Diffusion-driven segmentation methods include EnsemDiff (Wolleb et al, 2022), MedSegDiffv2 (Wu et al, 2024), and CamoDiff (Sun et al, 2025a). The saliency maps for the above methods are obtained from the authors or by running public source codes.

1) Quantitative and Model Complexity Comparison: In Tab. 1, we report the quantitative and model complexity comparison results of our IPDiff and other

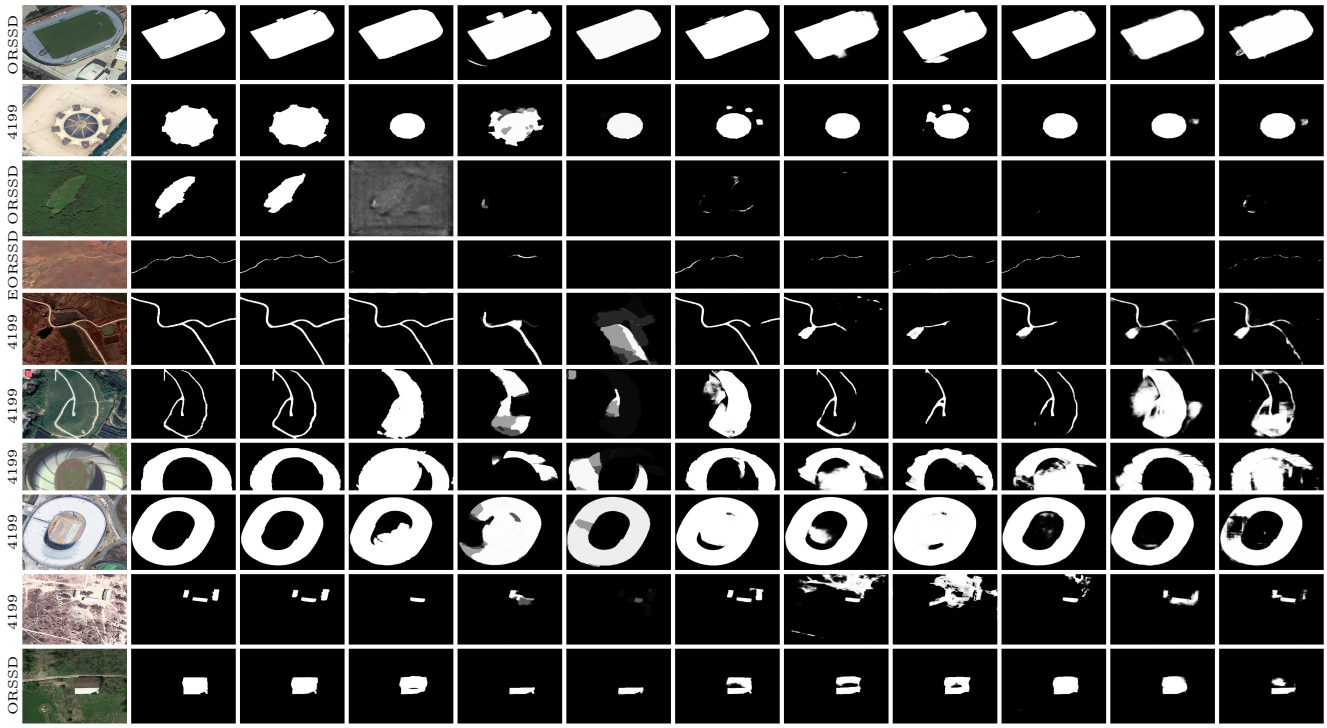

Fig. 7: Qualitative comparisons with nine representative state-of-the-art methods. The source dataset is shown on the left side of the ORSI. We abbreviate the ORSI-4199 dataset as 4199.

46 compared methods on the EORSSD, ORSSD, and ORSI-4199 datasets. Overall, our method outperforms all compared methods in all four evaluation metrics. On the EORSSD dataset, our S_α is one of the only two methods that exceed 0.94, one of which is 0.9401 for DPU-Former (Sun et al, 2025b) and the other is 0.9461 for our IPDiff. Our $F_\beta^{\max}$ is the only one that exceeds 0.90, reaching 0.9033, which is 1.03% higher than 0.8930 of the second-highest MIRGNet (Zhao et al, 2024b) and DPU-Former (Sun et al, 2025b). Our $\mathcal{M}$ is the only one below 0.005, reaching 0.0044, which is 0.001 lower than 0.0054 of the second-lowest PRNet (Gu et al, 2024) and BCARNet (Gu et al, 2025). On the ORSSD dataset, our S_α is one of the only two methods that exceed 0.95, one of which is 0.9507 for RAGRNet (Zhao et al, 2024c) and the other is 0.9557 for our IPDiff. Our $F_\beta^{\max}$ is the only one that exceeds 0.93, reaching 0.9362, which is 0.99% higher than 0.9263 of the second-highest DPU-Former (Sun et al, 2025b). Our $E_\xi^{\max}$ is the only one that exceeds 0.99, reaching an amazing 0.9915. Our $\mathcal{M}$ is the only one below 0.006, reaching 0.0054. On the ORSI-4199 dataset, our S_α is the only one that exceeds 0.89, reaching 0.8907. Moreover, compared to NSI-SOD methods, specialized ORSI-SOD methods show more excellent performance. Diffusion-driven segmentation methods perform on par with some specialized ORSI-SOD methods, achieving impressive performance. But they are inferior to our specifically designed IPDiff for ORSI-SOD.

Specifically, we also provide a detailed analysis of our dynamic optimization strategy and different representative paradigms. Most existing ORSI-SOD methods, such as GeleNet and DPU-Former, are typical one-shot baselines. These methods map an input to the saliency map in a single forward pass. As shown in Tab. 1, although DPU-Former and GeleNet achieve competitive results, our IPDiff outperforms them across all metrics, demonstrating the necessity of an optimization strategy. Notably, their advantage lies in efficiency. PRNet and GLGCNet are multi-stage refinement baselines, which progressively generate saliency maps within the decoder and perform gradual refinement. However, unlike our IPDiff which treats optimization as a dynamic denoising process, these methods rely on static architectural stages. The experimental results show that IPDiff significantly surpasses PRNet and GLGCNet. This indicates that the diffusion-style dynamic optimization strategy is inherently more effective than conventional multi-stage refinement.

Regarding the model complexity, the parameter count and computational cost of our IPDiff are 82.6M and 70.4G FLOPs, respectively, achieving an inference speed of 4 fps with a 352×352 input. Obviously, our IPDiff has a significant disadvantage in inference speed.

Table 2: Ablation study on the effectiveness of each component in IRAM. InfRec, SpeDec, SpeWei, and InfAgg are the abbreviations for information reconstruction, spectrum decoupling, spectrum weighting, and information aggregation, respectively. The best result of each metric is shown in **bold**.

No.	SA	Direct InfRec	SpeDec		SpeWei		InfAgg		EORSSD (Zhang et al, 2021)			
			Fix γ	Adp γ	LowFre	HighFre	Fix $\{\eta_1, \eta_2\}$	Adp $\{\eta_1, \eta_2\}$	$S_\alpha \uparrow$	$F_\beta^{\max} \uparrow$	$E_\xi^{\max} \uparrow$	$\mathcal{M} \downarrow$
1	✓								.9375	.8952	.9793	.0072
2	✓	✓							.9448	.9004	.9822	**.0040**
3	✓		✓		✓		✓		.9434	.8998	.9827	.0046
4	✓		✓			✓	✓		.9444	.8993	.9811	.0043
5	✓		✓		✓	✓	✓		.9453	.9010	.9828	.0047
6	✓		✓		✓	✓		✓	.9450	.9019	.9839	.0046
7	✓			✓	✓	✓	✓		.9458	.9007	.9831	.0045
8	✓			✓	✓	✓		✓	**.9461**	**.9033**	**.9861**	.0044

Fix means fixed hyperparameter, while Adp means adaptive hyperparameter.

This is because our IPDiff is built on a diffusion-based architecture, which has an inherent requirement of T denoising steps in the reverse denoising process. But it achieves a better balance between performance and efficiency as compared to two diffusion-based methods, *i.e.*, EnsemDiff and MedSegDiffv2. Notably, although its iterative nature results in a higher computational cost than most one-shot methods, our IPDiff remains lighter than several heavy CNN-based competitors in parameter count, *e.g.*, EGNet, ACCoNet, and TSCNet.

2) Qualitative Comparison: In Fig. 7, we show the saliency maps generated from our IPDiff and nine representative state-of-the-art methods, including three diffusion-driven segmentation methods (CamoDiff, MedSegDiffv2, and EnsemDiff), four ORSI-SOD methods (DPU-Former, MRBINet, BCARNet, and RAGRNet), and two NSI-SOD methods (ICON and VST). There are ten cases in Fig. 7, totaling six challenging ORSI scenes from three datasets. The first scene is objects with fine structures (*i.e.*, 1st and 2nd cases). Our IPDiff outlines fine structures, while other methods only locate the position of objects. The second one is the low-contrast scene (*i.e.*, 3rd and 4th cases). Most methods lose low-contrast objects, but our IPDiff accurately highlights them. The third one is objects with complex topology (*i.e.*, 5th and 6th cases), with road being a typical representative. The complex topology causes many methods to fail, but our IPDiff handles it well. The fourth one is hollow objects (*i.e.*, 7th and 8th cases), with gymnasium being a typical representative. Most methods fail to properly segment the hollow, but our method segments the gymnasium with a complete hollow. The fifth one is the overexposed scene (*i.e.*, 9th case). The special ORSI scene caused by optical imaging makes objects difficult to distinguish, resulting in objects in the saliency maps of many methods being stuck together. The last one is objects composed of regions with significant appearance differences (*i.e.*, last case). Significant appearance differences put some methods in a dilemma.

4.3 Ablation Studies

We conduct comprehensive ablation experiments to assess the effectiveness of each part of our IPDiff on the EORSSD dataset. Specifically, we evaluate 1) the effectiveness of each component in IRAM, 2) the rationality of each component in the denoising network, 3) the influence of the perturbation rate r in IPM, 4) the influence of the number of IPMs, 5) the importance of each component of the hybrid loss function, 6) the effectiveness of the dynamic optimization strategy, 7) the influence of the timestep T, and 8) the generalization ability evaluation.

1) Effectiveness of Each Component in IRAM: IRAM achieves effective information reconstruction in the spectral domain through spectrum decoupling, spectrum weighting, and information aggregation. Specifically, we perform spectrum decoupling and information aggregation in an adaptive way with learnable hyperparameters. Here, we conduct variants to gradually verify the effectiveness of each component in IRAM. We report the quantitative ablation results in Tab. 2.

First, we verify the effectiveness of information reconstruction in the spectral domain. As shown in No.1 of Tab. 2, we only keep the vanilla attention and discard the three parts for information reconstruction. The performance degradation is obvious, *i.e.*, 0.86% in S_α and 0.81% in $F_\beta^{\max}$, which proves the effectiveness of in-

Table 3: Ablation study on the rationality of each component in the denoising network. The best result of each metric is shown in **bold**.

Models	EORSSD (Zhang et al, 2021)			
	$S_\alpha \uparrow$	$F_\beta^{\max} \uparrow$	$E_\xi^{\max} \uparrow$	$\mathcal{M} \downarrow$
w/o SalPrior	.9414	.8935	.9788	.0051
w/ IntegratedPrior	.9407	.8936	.9823	.0046
w/o IPM	.9446	.8994	.9811	.0050
w/o InfTrans	.9396	.8915	.9800	.0050
Ours	**.9461**	**.9033**	**.9861**	**.0044**

Table 4: Ablation study on the influence of the perturbation rate r in IPM. The best result of each metric is shown in **bold**.

Perturbation rate r	EORSSD (Zhang et al, 2021)			
	$S_\alpha \uparrow$	$F_\beta^{\max} \uparrow$	$E_\xi^{\max} \uparrow$	$\mathcal{M} \downarrow$
0%	.9446	.8994	.9811	.0050
10%	.9428	.9001	.9846	.0045
20% **(Ours)**	**.9461**	**.9033**	.9861	**.0044**
30%	.9453	.9028	**.9862**	.0046
40%	.9431	.9022	.9858	.0045
50%	.9438	.8987	.9824	.0045
60%	.9428	.8974	.9837	**.0044**

formation reconstruction in the spectral domain. Moreover, we verify the necessity of spectrum decoupling for information reconstruction. As shown in No.2 of Tab. 2, we directly perform spectrum weighting on the entire spectrum without decoupling low-frequency and high-frequency components, and then reconstruct information, *i.e.*, Direct InfRec. Spectrum weighting improves performance, but is suboptimal due to the absence of spectrum decoupling.

Then, we verify the effectiveness of spectrum weighting. As shown in No.3∼5 of Tab. 2, we decouple the spectrum and aggregate information in a fixed way, and provide three combinations of weighting low-frequency and high-frequency components. Abandoning either high-frequency weighting or low-frequency weighting is inferior to weighting both low-frequency and high-frequency components. This indicates that weighting both low-frequency and high-frequency components is more conducive to mining useful information for ORSI-SOD.

Last, we verify the effectiveness of adaptive hyperparameters for spectrum decoupling and information aggregation. As shown in No.5∼8 of Tab. 2, we provide four combinations of γ and $\{\eta_1, \eta_2\}$, *i.e.*, fixed γ and fixed $\{\eta_1, \eta_2\}$, fixed γ and adaptive $\{\eta_1, \eta_2\}$, adaptive γ and fixed $\{\eta_1, \eta_2\}$, and adaptive γ and adaptive $\{\eta_1, \eta_2\}$. Adaptive hyperparameters obviously help our IPDiff better decouple the spectrum and reconstruct information that can adapt to the complex ORSI scenes.

2) Rationality of Each Component in the Denoising Network: To investigate the individual contribution of components in our multi-prior guidance denoising network, we design four variants: 1) removing the saliency prior, *i.e.*, *w/o SalPrior*, 2) changing the injection way of hierarchical priors to integrated prior injection, *i.e.*, *w/ IntegratedPrior*, 3) removing IPM in the training phase, *i.e.*, *w/o IPM*, and 4) removing information transmission to prohibit information from being transferred to the decoder hierarchically, *i.e.*, *w/o InfTrans*. We report the quantitative ablation results in Tab. 3.

Without the saliency prior to provide positional information, the performance of *w/o SalPrior* degrades quite a lot, with a drop of nearly 1.0% in $F_\beta^{\max}$. *w/ IntegratedPrior* is the same as CamoDiff (Sun et al, 2025a), which first integrates hierarchical priors into a comprehensive prior and then injects it into the highest level of the encoder of the denoising network. However, this injection approach is not as effective as hierarchically injecting priors into the encoder for extracting valid features. *w/o IPM* reduces the feature extraction and anti-interference capabilities of our denoising network, resulting in a 0.5% decrease in $E_\xi^{\max}$. Surprisingly, the performance degradation of *w/o InfTrans* is quite severe (*i.e.*, 1.18% in $F_\beta^{\max}$), which indicates that transferring the information extracted in the encoder to the decoder can improve the decoder's perception of salient objects.

3) Influence of the Perturbation Rate r in IPM: The perturbation rate r in IPM determines how much information is discarded. We conduct ablation experiments to experimentally analyze its influence. As shown in Tab. 4, when the perturbation rate r is 20%, our IPDiff performs best, and is higher than the variant that does not discard information (*i.e.*, r is 0%). We observe that as more and more information is discarded (with r increasing from 20% to 60%), the performance deteriorates increasingly. This is consistent with our intuitive understanding, that is, the more information is discarded, the more difficult it is to reconstruct features. Therefore, the feature extraction and anti-interference capabilities will be weakened.

4) Influence of the Number of IPMs: In the denoising network, we hierarchically integrate four IPMs into its encoder during the training phase to enhance its anti-interference capability. To evaluate the influence of the number of IPMs on the anti-interference capabil-

Table 5: Ablation study on the influence of the number of IPMs. The best result of each metric is shown in **bold**.

Number of IPMs	EORSSD (Zhang et al, 2021)			
	$S_\alpha \uparrow$	$F_\beta^{\max} \uparrow$	$E_\xi^{\max} \uparrow$	$\mathcal{M} \downarrow$
0	.9446	.8994	.9811	.0050
1	.9432	.8995	.9819	.0047
2	.9441	.8984	.9823	.0047
3	.9454	**.9033**	.9848	.0046
4 **(Ours)**	**.9461**	**.9033**	**.9861**	**.0044**

Table 6: Ablation study on the importance of each component of the hybrid loss function. The best result of each metric is shown in **bold**.

Losses	EORSSD (Zhang et al, 2021)			
	$S_\alpha \uparrow$	$F_\beta^{\max} \uparrow$	$E_\xi^{\max} \uparrow$	$\mathcal{M} \downarrow$
$L_{\text{spa}-\text{base}}$	.9395	.8924	.9803	.0066
$L_{\text{spa}-\text{base}}$+$L_{\text{spa}-\text{edge}}$ (L_{spa})	.9410	.8964	.9824	**.0044**
$L_{\text{spa}-\text{base}}$+$L_{\text{spe}}$	.9416	.8960	.9821	.0064
L_{spa}+L_{spe} ($L_{\text{spa\&spe}}$)	**.9461**	**.9033**	**.9861**	**.0044**

ity, we conduct variants with 0, 1, 2, and 3 IPMs. We report the experimental results in Tab. 5. As observed, the performance generally increases monotonically with the number of IPMs. However, the performance in S_α and $F_\beta^{\max}$ exhibits slight fluctuations when the number of IPMs is low (1 or 2). We believe this is because the IPM introduces a local feature masking (*i.e.*, local perturbation) mechanism to simulate noise and interference. When only a limited number of IPMs are equipped, the local perturbation may exceed the reconstruction capacity of the denoising network at that specific stage, thereby causing a slight drop in performance. As the number of IPMs increases to 3 and 4, we observe a consistent improvement across all metrics. This trend indicates a synergistic effect across the hierarchical stages of our denoising network. By equipping all four encoder blocks with IPMs, the denoising network is forced to learn robust feature representations across multiple scales, ranging from fine-grained textures to high-level semantics. Consequently, this configuration achieves peak performance with an S_α of 0.9461 and an $\mathcal{M}$ of 0.0044, endowing our denoising network with a comprehensive anti-interference capability.

5) Importance of Each Component of the Hybrid Loss Function: As formulated in Eq. 14, our hybrid loss function $L_{\text{spa\&spe}}$ consists of three components, including $L_{\text{spa}-\text{base}}$, $L_{\text{spa}-\text{edge}}$, and L_{spe}, to supervise the predicted saliency map in both spatial and spectral domains. To investigate the individual contribution of components in our $L_{\text{spa\&spe}}$, we design three loss variants for network training: 1) only using the traditional $L_{\text{spa}-\text{base}}$, 2) only supervising the predicted saliency map in the spatial domain, and 3) combining $L_{\text{spa}-\text{base}}$ and L_{spe}. We report the quantitative ablation results in Tab. 6. The traditional $L_{\text{spa}-\text{base}}$ is suboptimal. With the help of edge loss $L_{\text{spa}-\text{edge}}$, L_{spa} and $L_{\text{spa\&spe}}$ both achieve performance improvements, especially with a reduction of 33.33% and 31.25% respectively in $\mathcal{M}$. By additionally applying unique spectral domain alignment, three metrics (*i.e.*, S_α, $F_\beta^{\max}$, and $E_\xi^{\max}$) significantly improve. With all three losses training together, our IPDiff achieves the best performance.

Table 7: The performance of the t-th step saliency maps ($\widehat{\boldsymbol{x}}_t$). The best result of each metric is shown in **bold**.

t-th step	EORSSD (Zhang et al, 2021)			
	$S_\alpha \uparrow$	$F_\beta^{\max} \uparrow$	$E_\xi^{\max} \uparrow$	$\mathcal{M} \downarrow$
$\widehat{\boldsymbol{x}}_9$	.9452	.9020	.9851	.0050
$\widehat{\boldsymbol{x}}_8$	.9453	.9023	.9851	.0049
$\widehat{\boldsymbol{x}}_7$	.9455	.9025	.9850	.0049
$\widehat{\boldsymbol{x}}_6$	.9456	.9026	.9850	.0048
$\widehat{\boldsymbol{x}}_5$	.9462	.9027	.9850	.0048
$\widehat{\boldsymbol{x}}_4$	.9460	.9026	.9850	.0048
$\widehat{\boldsymbol{x}}_3$	.9460	.9029	.9851	.0047
$\widehat{\boldsymbol{x}}_2$	**.9464**	.9032	.9856	.0046
$\widehat{\boldsymbol{x}}_1$	.9462	.9032	.9858	.0045
$\widehat{\boldsymbol{x}}_0$ **(Ours)**	.9461	**.9033**	**.9861**	**.0044**

6) Effectiveness of the Dynamic Optimization Strategy: The dynamic optimization strategy is the core of our IPDiff. To verify the effectiveness of this strategy, we report the performance of saliency maps across all T (T=10) steps, *i.e.*, $\{\widehat{\boldsymbol{x}}_t\}_{t=0}^{9}$, in Tab. 7. We observe that $F_\beta^{\max}$ exhibits a consistent upward trend, while $\mathcal{M}$ continuously decreases, indicating a steady improvement in model performance across the dynamic optimization process. For S_α, although slight fluctuations are observed, it generally follows an increasing trajectory, peaking at $\widehat{\boldsymbol{x}}_2$ (0.9464) and maintaining a superior value at $\widehat{\boldsymbol{x}}_0$ (0.9461) compared to the initial $\widehat{\boldsymbol{x}}_9$ (0.9452). $E_\xi^{\max}$ remains stable in the early stages, and shows a marked increase during the final steps. Overall, all four metrics show a clear performance enhancement from $\widehat{\boldsymbol{x}}_9$ to $\widehat{\boldsymbol{x}}_0$, demonstrating the effectiveness of the dynamic optimization strategy.

To intuitively verify the effectiveness of the dynamic optimization strategy, we provide a timestep-wise analysis from $\widehat{\boldsymbol{x}}_9$ to $\widehat{\boldsymbol{x}}_0$ in Fig. 8. By decomposing the prediction errors into false positives (red) and false negatives

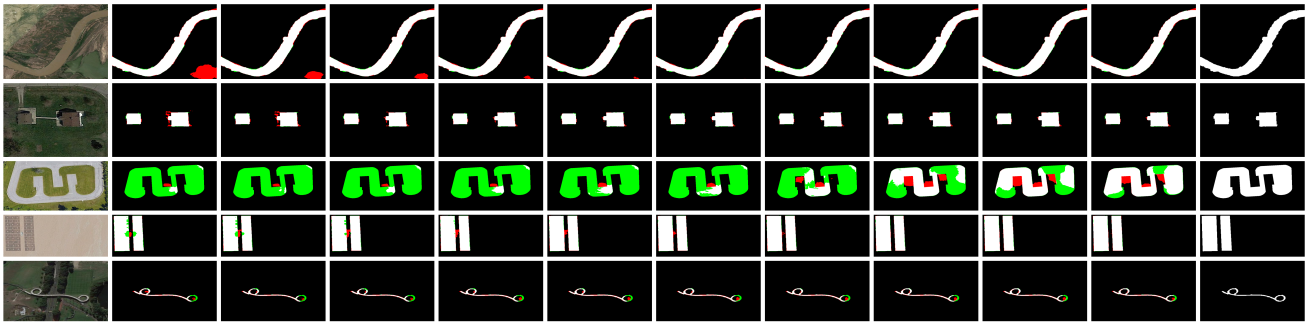

Fig. 8: Visualization of the dynamic optimization from $\widehat{x}_9$ to $\widehat{x}_0$ in our IPDiff. In these saliency maps, we highlight the true positives in white, the true negatives in black, the false positives in red, and the false negatives in green. Please zoom in for details.

Table 8: Ablation study on the influence of the timestep T. The best result of each metric is shown in **bold**.

Timestep T	Speed (fps)	EORSSD (Zhang et al, 2021)			
		$S_\alpha \uparrow$	$F_\beta^{\max} \uparrow$	$E_\xi^{\max} \uparrow$	$\mathcal{M} \downarrow$
1	24.3	.9395	.8974	.9821	.0054
5	7.9	.9455	.9029	.9845	**.0044**
10 **(Ours)**	4.0	**.9461**	**.9033**	**.9861**	**.0044**
15	2.8	.9419	.8977	.9818	.0045
20	2.1	.9413	.8967	.9832	.0049
50	0.9	.9412	.8989	.9844	**.0044**
100	0.4	.9419	.8965	.9839	.0049

(green), we observe a clear progressive correction process. First, the strategy demonstrates a remarkable capability in background noise suppression, *i.e.*, false positive reduction. As shown in the first and second cases, the initial saliency maps ($\widehat{x}_9$) often contain red regions caused by complex background interference, such as riverside textures or shadows. Through the iterative optimization, these false positives are gradually pruned and nearly disappear by $\widehat{x}_4$. Second, the progressive recovery of object details (*i.e.*, false negative reduction) is highly evident. In the third and fourth cases, IPDiff initially fails to capture the internal structure (marked in green). However, as the timestep t evolves, the green areas are systematically replaced by white pixels (*i.e.*, true positives). This proves that the dynamic optimization strategy allows IPDiff to recover the missing parts and maintain internal consistency within objects. Finally, the strategy can sharpen object boundaries. Across all cases, particularly the slender road in the last case, the coarse and fragmented predictions in the early stages ($\widehat{x}_9$-$\widehat{x}_7$) are refined into relatively crisp and continuous boundaries in $\widehat{x}_0$. The above cases demonstrate that our dynamic optimization strategy effectively polishes the saliency maps to achieve higher accuracy.

7) Influence of the Timestep T: The timestep T is a critical hyperparameter of our IPDiff. To evaluate the influence of the timestep T, we additionally conduct dedicated experiments using $T \in \{1, 5, 15, 20, 50, 100\}$, while keeping all other settings constant. We report the experimental results and inference speed (fps) in Tab. 8. We observe that increasing T from 1 to 10 yields consistent improvements across all metrics. For instance, S_α improves from 0.9395 to 0.9461. Although the inference speed decreases from 24.3 fps to 4.0 fps, $T = 10$ provides a high-quality prediction while maintaining a decent inference speed. Notably, the performance does not increase linearly with the number of timesteps. When T exceeds 10, the metrics, such as *e.g.*, $F_\beta^{\max}$ and $E_\xi^{\max}$, begin to saturate or even fluctuate slightly, while the computational cost increases substantially (*e.g.*, the inference speed drops to merely 0.4 fps at $T = 100$). This indicates that our diffusion-based dynamic optimization strategy does not perform better with a longer timestep, and its error correction capability reaches a plateau rather than improving indefinitely with additional timesteps. Therefore, based on these analyses, we select $T = 10$ as the default setting for IPDiff.

8) Generalization Ability Evaluation: To further evaluate the generalization and robustness of our IPDiff, we conduct cross-dataset testing and zero-shot testing, as reported in Tab. 9. We compare our IPDiff with several representative methods, including ACCoNet, GeleNet, TSCNet, and DPU-Former.

For cross-dataset testing, we train IPDiff on one ORSI-SOD dataset and test on another, *i.e.*, ORSI-4199→EORSSD and ORSSD→EORSSD. Compared with state-of-the-art competitors, our IPDiff achieves superior performance. Notably, in the ORSI-4199→EORSSD task, IPDiff outperforms the second-best method, GeleNet, by 2.12% in S_α and

Table 9: Generalization ability evaluation with state-of-the-art representative methods on the EORSSD dateset, the 360-SOD dateset, and the 360-SSOD dateset. The first dataset is used for cross-dataset testing, while the last two datasets are 360° omnidirectional image SOD datasets used for zero-shot testing. The best result of each metric is shown in **bold**.

Methods	Cross-dataset Testing								Zero-shot Testing							
	ORSI-4199→EORSSD				ORSSD→EORSSD				ORSSD→360-SOD				ORSSD→360-SSOD			
	$S_\alpha\uparrow$	$F_\beta^{\max}\uparrow$	$E_\xi^{\max}\uparrow$	$\mathcal{M}\downarrow$	$S_\alpha\uparrow$	$F_\beta^{\max}\uparrow$	$E_\xi^{\max}\uparrow$	$\mathcal{M}\downarrow$	$S_\alpha\uparrow$	$F_\beta^{\max}\uparrow$	$E_\xi^{\max}\uparrow$	$\mathcal{M}\downarrow$	$S_\alpha\uparrow$	$F_\beta^{\max}\uparrow$	$E_\xi^{\max}\uparrow$	$\mathcal{M}\downarrow$
ACCoNet$_{23}$	.8249	.7375	.8565	.0640	.8871	.8117	.9229	.0128	.5281	.2121	.6402	.0860	.5099	.1957	.6039	.0987
GeleNet$_{23}$	.8716	.7883	.9097	.0258	.8927	.8205	.9374	.0154	.5639	.3002	.6808	.0764	.5247	.2458	.6061	.0889
TSCNet$_{24}$	.8307	.7902	.8974	.0658	.8823	.8112	.9233	.0215	.4845	.1771	.5971	.1455	.4717	.1529	.5849	.1366
DPU-Former$_{25}$	.8558	.8058	.9068	.0445	.8923	**.8341**	**.9415**	.0134	.5423	.2814	.6313	.1393	.5242	.2627	.6248	.1297
IPDiff (Ours)	**.8928**	**.8284**	**.9354**	**.0146**	**.8960**	.8338	.9411	**.0113**	**.6026**	**.3775**	**.7366**	**.0638**	**.5544**	**.3177**	**.6933**	**.0879**

4.01% in $F_\beta^{\max}$. This demonstrates that IPDiff effectively captures invariant structural features across different optical remote sensing platforms and sensors. We also observe that IPDiff and DPU-Former each have their own advantages in the ORSSD→EORSSD task. This is mainly because the EORSSD dataset is a direct extension of the ORSSD dataset, resulting in a highly similar data distribution. This indicates that while both models fit well to in-distribution data, the unique advantage of IPDiff lies in its superior generalization to more challenging and unseen senses.

We further challenge IPDiff by performing zero-shot testing on 360° omnidirectional images. Specifically, we directly applied the weights trained on the ORSSD dataset to the 360-SOD dataset (Li et al, 2020) and the 360-SSOD dataset (Ma et al, 2020), which are two 360° omnidirectional image SOD datasets, *i.e.*, ORSSD→360-SOD and ORSSD→360-SSOD. This is a rigorous task due to the significant domain gap between top-down remote sensing views and panoramic perspectives. Remarkably, IPDiff maintains high performance, significantly surpassing all compared methods. For instance, on the 360-SOD dataset, IPDiff achieves a 9.61% improvement in $F_\beta^{\max}$ over DPU-Former. These results indicate that our information perturbation and multi-prior guidance enable the IPDiff to learn the intrinsic manifold of saliency beyond specific imaging domains.

5 Conclusion

In this paper, we propose IPDiff, a novel diffusion-driven ORSI-SOD framework. IPDiff follows a unique dynamic optimization strategy to iteratively optimize saliency maps with the dynamically changing timestep in the testing phase. In IPDiff, we formulate ORSI-SOD as a conditional diffusion problem. Specifically, we introduce a prior network to extract the saliency prior and the hierarchical priors from ORSIs as conditional priors. In the prior network, we employ IRAMs to adaptively reconstruct information in the spectral domain to increase the useful content of multiple priors. Then, we inject the above priors into the carefully designed denoising network. In the denoising network, the noisy mask is combined with the saliency prior and the hierarchical priors successively to absorb specific position, detail, and semantic information of salient objects. IPMs are equipped in the denoising network in the training phase to stabilize feature representation and provide anti-interference capability. As the timestep dynamically changes, the noisy mask is iteratively optimized to recover a clear saliency map. Comprehensive experiments, including quantitative and qualitative comparisons and ablation studies, demonstrate the leading position of IPDiff in the ORSI-SOD community and the effectiveness of its components.

Data Availability Statement

The datasets used in this work are all publicly available. ORSSD is available at https://li-chongyi.github.io/proj_optical_saliency.html. EORSSD is available at https://github.com/rmcong/DAFNet_TIP20. ORSI-4199 is available at https://github.com/wchao1213/ORSI-SOD.

References

Achanta R, Hemami S, Estrada F, Susstrunk S (2009) Frequency-tuned salient region detection. In: Proc. IEEE CVPR, pp 1597–1604

Aydemir B, Bhattacharjee D, Zhang T, Salzmann M, Süsstrunk S (2024) Data augmentation via latent diffusion for saliency prediction. In: Proc. ECCV, pp 360–377

Bai Z, Li G, Liu Z (2023) Global-local-global context-aware network for salient object detection in optical remote sensing images. ISPRS J Photogramm Remote Sens 198:184–196

Chen H, Li Y, Deng Y, Lin G (2021) Cnn-based rgb-d salient object detection: Learn, select, and fuse. Int J Comput Vis 129(7):2076–2096

Chen J, Li G, Zhang Z, Chang L, Zeng D (2026) STENet: Superpixel token enhancing network for RGB-D salient object detection. IEEE Trans Multimedia URL 10.1109/TMM.2026.3680601

Chen Z, Xu Q, Cong R, Huang Q (2020) Global context-aware progressive aggregation network for salient object detection. In: Proc. AAAI, pp 10599–10606

Cheng B, Liu Z, Wang Q, Shen T, Fu C, Tian A (2024) Lightweight progressive multilevel feature collaborative network for remote sensing image salient object detection. IEEE Trans Geosci Remote Sens 62:1–17

Cheng MM, Fan DP (2021) Structure-measure: A new way to evaluate foreground maps. Int J Comput Vis 129(9):2622–2638

Deng Z, Hu X, Zhu L, Xu X, Qin J, Han G, Heng PA (2018) R^3Net: Recurrent residual refinement network for saliency detection. In: Proc. IJCAI, pp 684–690

Ding X, Zhang X, Ma N, Han J, Ding G, Sun J (2021) RepVGG: Making VGG-style ConvNets great again. In: Proc. IEEE CVPR, pp 13728–13737

Dong P, Wang B, Cong R, Sun HH, Li C (2024) Transformer with large convolution kernel decoder network for salient object detection in optical remote sensing images. Comput Vis Image Und 240:103917

Fan DP, Gong C, Cao Y, Ren B, Cheng MM, Borji A (2018) Enhanced-alignment measure for binary foreground map evaluation. In: Proc. IJCAI, pp 698–704

Fang C, Tian H, Zhang D, Zhang Q, Han J, Han J (2022) Densely nested top-down flows for salient object detection. Sci China Inf Sci 65(8):1–14

Gao SH, Tan YQ, Cheng MM, Lu C, Chen Y, Yan S (2020) Highly efficient salient object detection with 100k parameters. In: Proc. ECCV, pp 702–721

Gu S, Song Y, Zhou Y, Bai Y, Yang X, He Y (2024) PRNet: Parallel refinement network with group feature learning for salient object detection in optical remote sensing images. IEEE Geosci Remote Sens Lett 21:1–5

Gu Y, Chen S, Sun X, Ji J, Zhou Y, Ji R (2025) Optical remote sensing image salient object detection via bidirectional cross-attention and attention restoration. Pattern Recognit 164:1–12

Han J, Sun F, Hou Y, Sun J, Li H (2025) Exploring a lightweight and efficient network for salient object detection in ORSI. IEEE Trans Geosci Remote Sens 63:1–14

He K, Zhang X, Ren S, Sun J (2016) Deep residual learning for image recognition. In: Proc. IEEE CVPR, pp 770–778

Ho J, Jain A, Abbeel P (2020) Denoising diffusion probabilistic models. In: Proc. NeurIPS, pp 6840–6851

Hoogeboom E, Heek J, Salimans T (2023) Simple diffusion: end-to-end diffusion for high resolution images. In: Proc. ICML, pp 13213–13232

Howard A, Sandler M, Chen B, Wang W, Chen LC, Tan M, Chu G, Vasudevan V, Zhu Y, Pang R, Adam H, Le Q (2019) Searching for MobileNetV3. In: Proc. IEEE ICCV, pp 1314–1324

Hu X, Sun F, Sun J, Wang F, Li H (2024) Cross-modal fusion and progressive decoding network for RGB-D salient object detection. Int J Comput Vis 132(8):3067–3085

Huang Z, Chen H, Liu B, Wang Z (2021) Semantic-guided attention refinement network for salient object detection in optical remote sensing images. Remote Sens 13(11):2163

Ji Y, Zhang H, Wu QJ (2018) Saliency detection via conditional adversarial image-to-image network. Neurocomputing 316:357–368

Jia Y, Zhao J, Ma L, Yu L (2025) Multistrategy region and boundary interaction network for salient object detection in optical remote sensing images. IEEE Trans Geosci Remote Sens 63:1–16

Li C, Cong R, Hou J, Zhang S, Qian Y, Kwong S (2019) Nested network with two-stream pyramid for salient object detection in optical remote sensing images. IEEE Trans Geosci Remote Sens 57(11):9156–9166

Li G, Liu Z, Bai Z, Lin W, Ling H (2022a) Lightweight salient object detection in optical remote sensing images via feature correlation. IEEE Trans Geosci Remote Sens 60:5617712

Li G, Liu Z, Lin W, Ling H (2022b) Multi-content complementation network for salient object detection in optical remote sensing images. IEEE Trans Geosci Remote Sens 60:5614513

Li G, Bai Z, Liu Z, Zhang X, Ling H (2023a) Salient object detection in optical remote sensing images driven by transformer. IEEE Trans Image Process 32:5257–5269

Li G, Liu Z, Zeng D, Lin W, Ling H (2023b) Adjacent context coordination network for salient object detection in optical remote sensing images. IEEE Trans Cybern 53(1):526–538

Li G, Liu Z, Zhang X, Lin W (2023c) Lightweight salient object detection in optical remote-sensing images via semantic matching and edge alignment. IEEE Trans Geosci Remote Sens 61:5601111

Li G, Bai Z, Liu Z (2024) Texture-semantic collaboration network for ORSI salient object detection. IEEE Trans Circuits Syst II-Express Briefs 71(4):2464–2468

Li G, Shi S, Wu Y, Lin W, Bai Z (2026) Lightweight ORSI salient object detection via frequency and mutual assistance attention. IEEE Trans Geosci Remote Sens 64:5617112

Li J, Su J, Xia C, Tian Y (2020) Distortion-adaptive salient object detection in 360° omnidirectional images. IEEE J Sel Top Signal Process 14(1):38–48

Li J, Pan Z, Liu Q, Wang Z (2021) Stacked U-shape network with channel-wise attention for salient object detection. IEEE Trans Multimedia 23:1397–1409

Li J, Ji W, Zhang M, Piao Y, Lu H, Cheng L (2023d) Delving into calibrated depth for accurate RGB-D salient object detection. Int J Comput Vis 131(4):855–876

Li J, Wang Z, Xu N, Zhang C (2025a) TSFANet: Trans-Mamba hybrid network with semantic feature alignment for remote sensing salient object detection. Remote Sens 17(11):1902

Li Y, Li X, Dai Y, Hou Q, Liu L, Liu Y, Cheng MM, Yang J (2025b) LSKNet: A foundation lightweight backbone for remote sensing. Int J Comput Vis 133(3):1410–1431

Li Z, Miao Y, Li X, Li W, Cao J, Hao Q, Li D, Sheng Y (2025c) Speed-oriented lightweight salient object detection in optical remote sensing images. IEEE Trans Geosci Remote Sens 63:1–14

Lin J, Zhu L, Shen J, Fu H, Zhang Q, Wang L (2024) ViDSOD-100: A new dataset and a baseline model for RGB-D video salient object detection. Int J Comput Vis 132(11):5173–5191

Liu J, Dian R, Li S, Liu H (2023) SGFusion: A saliency guided deep-learning framework for pixel-level image fusion. Inf Fusion 91:205–214

Liu J, He J, Chen H, Yang R, Huang Y (2025) A lightweight semantic- and graph-guided network for advanced optical remote sensing image salient object detection. Remote Sens 17(5):816

Liu JJ, Hou Q, Cheng MM, Feng J, Jiang J (2019) A simple pooling-based design for real-time salient object detection. In: Proc. IEEE CVPR, pp 3912–3921

Liu N, Zhang N, Wan K, Shao L, Han J (2021a) Visual saliency transformer. In: Proc. IEEE ICCV, pp 4702–4712

Liu Y, Gu YC, Zhang XY, Wang W, Cheng MM (2021b) Lightweight salient object detection via hierarchical visual perception learning. IEEE Trans Cybern 51(9):4439–4449

Liu Y, Zhang XY, Bian JW, Zhang L, Cheng MM (2021c) SAMNet: Stereoscopically attentive multi-scale network for lightweight salient object detection. IEEE Trans Image Process 30:3804–3814

Liu Y, Zhang D, Liu N, Xu S, Han J (2022) Disentangled capsule routing for fast part-object relational saliency. IEEE Trans Image Process 31:6719–6732

Liu Y, Tian Y, Zhao Y, Yu H, Xie L, Wang Y, Ye Q, Jiao J, Liu Y (2024) VMamba: Visual state space model. Proc NeurIPS 37:103031–103063

Liu Z, Lin Y, Cao Y, Hu H, Wei Y, Zhang Z, Lin S, Guo B (2021d) Swin Transformer: Hierarchical vision transformer using shifted windows. In: Proc. IEEE ICCV, pp 9992–10002

Luo H, Wang J, Liang B (2024) Spatial attention feedback iteration for lightweight salient object detection in optical remote sensing images. IEEE J Sel Top Appl Earth Observ Remote Sens 17:13809–13823

Ma G, Li S, Chen C, Hao A, Qin H (2020) Stage-wise salient object detection in 360○ omnidirectional image via object-level semantical saliency ranking. IEEE Trans Vis Comput Graph 26(12):3535–3545

Mehta S, Rastegari M (2021) MobileViT: Light-weight, general-purpose, and mobile-friendly vision transformer. In: Proc. ICLR, pp 1–13

Meng L, Li H, Han H, Xu M, Wu J, Hou S, Duan W (2025) Progressive enhancement of foreground features for salient object detection in optical remote sensing images. IEEE J Sel Top Appl Earth Observ Remote Sens 18:7572–7591

Nichol AQ, Dhariwal P (2021) Improved denoising diffusion probabilistic models. In: Proc. ICML, pp 8162–8171

Pang Y, Zhao X, Zhang L, Lu H (2020) Multi-scale interactive network for salient object detection. In: Proc. IEEE CVPR, pp 9410–9419

Peng P, Yang KF, Luo FY, Li YJ (2021) Saliency detection inspired by topological perception theory. Int J Comput Vis 129(8):2352–2374

Quan Y, Xu H, Wang R, Guan Q, Zheng J (2024) ORSI salient object detection via progressive semantic flow and uncertainty-aware refinement. IEEE Trans Geosci Remote Sens 62:1–13

Rombach R, Blattmann A, Lorenz D, Esser P, Ommer B (2022) High-resolution image synthesis with latent diffusion models. In: Proc. IEEE CVPR, pp 10674–10685

Sandler M, Howard A, Zhu M, Zhmoginov A, Chen LC (2018) MobileNetV2: Inverted residuals and linear bottlenecks. In: Proc. IEEE CVPR, pp 4510–4520

Shan L, Li X, Wang W (2021) Decouple the high-frequency and low-frequency information of images for semantic segmentation. In: Proc. IEEE ICASSP,

pp 1805–1809

Shi S, Li G, Cong R, Xiao S, Lin W (2026) Diffusion-driven RGB-D salient object detection with temporal modulation. IEEE Trans Circuits Syst Video Technol URL 10.1109/TCSVT.2026.3684803

Simonyan K, Zisserman A (2015) Very deep convolutional networks for large-scale image recognition. In: Proc. ICLR, pp 1–14

Song J, Meng C, Ermon S (2021) Denoising diffusion implicit models. In: Proc. ICLR, pp 1–12

Sun K, Chen Z, Lin X, Sun X, Liu H, Ji R (2025a) Conditional diffusion models for camouflaged and salient object detection. IEEE Trans Pattern Anal Mach Intell 47(4):2833–2848

Sun P, Zhang W, Li S, Guo Y, Song C, Li X (2022) Learnable depth-sensitive attention for deep RGB-D saliency detection with multi-modal fusion architecture search. Int J Comput Vis 130(11):2822–2841

Sun Y, Yang J, Luo L (2024) United domain cognition network for salient object detection in optical remote sensing images. IEEE Trans Geosci Remote Sens 62:1–14

Sun Y, Yan J, Qian J, Xu C, Yang J, Luo L (2025b) Dual-perspective united transformer for object segmentation in optical remote sensing images. In: Proc. IJCAI, pp 1909–1917

Sun Y, Zhao H, Zhou J (2025c) DKETFormer: Salient object detection in optical remote sensing images based on discriminative knowledge extraction and transfer. Neurocomputing 625:129558

Tai Y, Huang Z, Peng T, Zhang Z (2025) DefFiller: mask-conditioned generation with diffusion prior for saliency-based defect detection. Vis Comput 41:10947–10962

Teng Y, Guo Z, Wang Y, Wang L, Zheng P (2025) Direction-aware attention and semantic guidance network for salient object detection in optical remote sensing images. In: Proc. ICMR, pp 1264–1272

Toker A, Eisenberger M, Cremers D, Leal-Taixé L (2024) SatSynth: Augmenting image-mask pairs through diffusion models for aerial semantic segmentation. In: Proc. IEEE CVPR, pp 27685–27695

Tu Z, Wang C, Li C, Fan M, Zhao H, Luo B (2022) ORSI salient object detection via multiscale joint region and boundary model. IEEE Trans Geosci Remote Sens 60:1–13

Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser Ł, Polosukhin I (2017) Attention is all you need. In: Proc. NeurIPS, pp 6000–6010

Vyver GVD, Lenz AT, Smistad E, Olaisen SH, Grenne B, Holte E, Dalen H, Løvstakken L (2025) Generative augmentations for improved cardiac ultrasound segmentation using diffusion models. arXiv preprint arXiv:250220100

Wang J, Jiang H, Yuan Z, Cheng MM, Hu X, Zheng N (2017) Salient object detection: A discriminative regional feature integration approach. Int J Comput Vis 123(2):251–268

Wang Q, Liu Y, Xiong Z, Yuan Y (2022a) Hybrid feature aligned network for salient object detection in optical remote sensing imagery. IEEE Trans Geosci Remote Sens 60:1–15

Wang W, Xie E, Li X, Fan DP, Song K, Liang D, Lu T, Luo P, Shao L (2022b) PVT v2: Improved baselines with pyramid vision transformer. Comput Vis Media 8:415–424

Wang Z, Li R, Wang X, Xu N, You Z (2025) PIFR-Net: Position information guided feature reconstruction network for salient object detection in remote sensing images. IEEE J Sel Top Appl Earth Observ Remote Sens 18:17787–17804

Wen Y, Ma X, Zhang X, Pun MO (2024) GCD-DDPM: A generative change detection model based on difference-feature-guided ddpm. IEEE Trans Geosci Remote Sens 62:1–16

Wolleb J, Sandkühler R, Bieder F, Valmaggia P, Cattin PC (2022) Diffusion models for implicit image segmentation ensembles. In: Proc. MIDL, vol 172, pp 1336–1348

Woo S, Park J, Lee JY, Kweon IS (2018) CBAM: Convolutional block attention module. In: Proc. ECCV, pp 3–19

Wu J, Fu R, Fang H, Zhang Y, Yang Y, Xiong H, Liu H, Xu Y (2023) MedSegDiff: Medical image segmentation with diffusion probabilistic model. In: Proc. MIDL, vol 227, pp 1623–1639

Wu J, Ji W, Fu H, Xu M, Jin Y, Xu Y (2024) MedSegDiff-V2: Diffusion-based medical image segmentation with transformer. In: Proc. AAAI, pp 6030–6038

Wu K, Zhang Y, Ru L, et al (2025) A semantic-enhanced multi-modal remote sensing foundation model for Earth observation. Nat Mach Intell URL https://doi.org/10.1038/s42256-025-01078-8

Wu Y, Liu Z, Zhou X (2020) Saliency detection using adversarial learning networks. J Vis Commun Image Represent 67:102761

Wu Z, Wen J, Shen L, Fan X, Xu Y, Yang J, Zhang D (2026) Weakly supervised salient object detection with text supervision. Int J Comput Vis 134:74

Xie Y, Liu S, Chen H, Cao S, Zhang H, Feng D, Wan Q, Zhu J, Zhu Q (2025) Localization, balance, and affinity: A stronger multifaceted collaborative salient object detector in remote sensing images. IEEE Trans Geosci Remote Sens 63:1–17

Xu B, Liang H, Liang R, Chen P (2021) Locate globally, segment locally: A progressive architecture with knowledge review network for salient object detection. In: Proc. AAAI, pp 3004–3012

Ye Y, Xu K, Huang Y, Yi R, Cai Z (2024) Diffusionedge: Diffusion probabilistic model for crisp edge detection. In: Proc. AAAI, pp 6675–6683

Zhang Q, Cong R, Li C, Cheng MM, Fang Y, Cao X, Zhao Y, Kwong S (2021) Dense attention fluid network for salient object detection in optical remote sensing images. IEEE Trans Image Process 30:1305–1317

Zhao J, Liu JJ, Fan DP, Cao Y, Yang J, Cheng MM (2019) EGNet: Edge guidance network for salient object detection. In: Proc. IEEE ICCV, pp 8779–8788

Zhao J, Jia Y, Ma L, Yu L (2024a) Adaptive dual-stream sparse transformer network for salient object detection in optical remote sensing images. IEEE J Sel Top Appl Earth Observ Remote Sens 17:5173–5192

Zhao J, Jia Y, Ma L, Yu L (2024b) Multilevel interactive reverse-guided network for salient object detection in optical remote sensing images. IEEE J Sel Top Appl Earth Observ Remote Sens 17:12983–12999

Zhao J, Jia Y, Ma L, Yu L (2024c) Recurrent adaptive graph reasoning network with region and boundary interaction for salient object detection in optical remote sensing images. IEEE Trans Geosci Remote Sens 62:1–20

Zhao X, Pang Y, Zhang L, Lu H, Zhang L (2020) Suppress and balance: A simple gated network for salient object detection. In: Proc. ECCV, pp 35–51

Zhou H, Xie X, Lai JH, Chen Z, Yang L (2020) Interactive two-stream decoder for accurate and fast saliency detection. In: Proc. IEEE CVPR, pp 9138–9147

Zhou W, Tang B, Cong R, Jiang Q (2026) Turbidity-similarity decoupling: Feature-consistent mutual learning for underwater salient object detection. IEEE Trans Image Process 35:495–510

Zhou X, Shen K, Liu Z, Gong C, Zhang J, Yan C (2022) Edge-aware multiscale feature integration network for salient object detection in optical remote sensing images. IEEE Trans Geosci Remote Sens 60:1–15

Zhou X, Shen K, Weng L, Cong R, Zheng B, Zhang J, Yan C (2023) Edge-guided recurrent positioning network for salient object detection in optical remote sensing images. IEEE Trans Cybern 53(1):539–552

Zhuge M, Fan DP, Liu N, Zhang D, Xu D, Shao L (2023) Salient object detection via integrity learning. IEEE Trans Pattern Anal Mach Intell 45(3):3738–3752