Phase Congruency-Guided Cross-Scale Contextual Fusion Network for Salient Object Detection in Optical Remote Sensing Images

Junfang Jiang , Wanjin Wang , Xiaohui Lin * , Pingping Miao , Lina Gao , Mingzhu Xu *

Posted Date: 27 March 2026

doi: 10.20944/preprints202603.2230.v1

Keywords: salient object detection (SOD); optical remote sensing images (ORSI); phase congruency; dynamic spatial attention; residual connection

 doi:10.20944/preprints202603.2230.v1

Article

Phase Congruency-Guided Cross-Scale Contextual Fusion Network for Salient Object Detection in Optical Remote Sensing Images

Junfang Jiang [1], Wanjin Wang [2], Xiaohui Lin [3,*], Pingping Miao [3], Lina Gao [4] and Mingzhu Xu [3,*]

[1] College of Oceanic and Atmospheric Sciences, Ocean University of China, Songling Road, Qingdao 266100, China
[2] China Energy International Group Company Limited, Building 208, Ciyunsi Beili, Chaoyang District, Beijing 100025, China
[3] School of Software, Shandong University, Shunhua Road, Jinan 250101, China
[4] College of Intelligent Systems Science and Engineering, Harbin Engineering University, Nantong Road, Harbin 150001, China
* Correspondence: xhlin109@163.com (X.L.); xumingzhu@sdu.edu.cn (M.X.)

Abstract

In recent years, salient object detection in optical remote sensing images (ORSI-SOD) has garnered increasing research attention. However, in practical applications, issues such as blurred target edges under low contrast and complex background interference continue to restrict the accuracy and robustness of detection. To address these problems, this paper proposes the Phase Congruency-Guided Cross-Scale Contextual Fusion Network (PCFNet). Specifically, we design a novel Phase Congruency Enhanced (PCE) Module to solve the problem of low contrast between targets and backgrounds. It acquires multi-scale phase features via Fourier decomposition, fuses them with Transformer shallow features and uses a tailored loss weighting mechanism to weight phase congruency learning for better PCE module adaptation. To address complex background interference, we design a novel Dynamic Residual Fusion (DRF) Module. It leverages dynamic spatial attention and residual connections to refine multi-scale features and enables the model to accurately capture effective target features under complex background interference. Experiments on ORSSD, EORSSD, and ORSI4199 benchmarks show that PCFNet outperforms 23 state-of the-art methods in core metrics, and ablation studies further confirm the effectiveness of each module.

Keywords: salient object detection(SOD); optical remote sensing images(ORSI); phase congruency; dynamic spatial attention; residual connection

1. Introduction

Salient object detection (SOD) is a fundamental task in computer vision that mimics human visual attention mechanisms to automatically identify and segment the most visually prominent regions in images [1]. In recent years, advances in deep learning have fueled substantial progress in SOD for natural scene images (NSI-SOD), with successful applications in image retrieval, object recognition, visual tracking and various other vision-related tasks [2–5]. At the same time, SOD for optical remote sensing images (ORSI-SOD) has garnered growing interest. It delivers essential prior information for downstream tasks like environmental change monitoring, aviation navigation, underwater detection, and urban planning [6–8]. Unlike natural scene images (NSI) captured by handheld cameras, optical remote sensing images (ORSI) refer to color imagery obtained by satellite or aerial sensors (with a wavelength range of 400–760 nm), leading to inherent distinctions between NSI and ORSI. Specifically, NSI typically exhibit fixed viewing angles, consistent object scales, and relatively simple backgrounds, which facilitate high foreground–background contrast and clear object boundaries. In contrast, ORSI often involve diverse imaging perspectives, large variations in object scale, and heterogeneous land cover compositions. These characteristics commonly lead to low contrast edges and severe background

 doi:10.20944/preprints202603.2230.v1

clutter. These notable differences make mature NSI–SOD methods perform subpar when directly applied to ORSI–SOD scenarios. This highlights an urgent need for specialized approaches to fully adapt to the unique characteristics of ORSI, ensuring reliable and accurate salient object detection.

In the field of ORSI–SOD, a key challenge lies in the low contrast of target objects, often caused by uneven illumination, atmospheric effects, or long imaging distances. Although CNNs excel at local feature extraction, limited receptive fields and reliance on intensity variations make them struggle to capture sufficient discriminative cues in low contrast regions [9,10]. Transformers capture global contextual dependencies well, but patch level tokenization often blurs fine edge details, especially in low contrast scenarios [11,12].

To better recover edge details in low contrast scenes, some recent works explore frequency domain representations to enhance structural cues. However, these methods mainly rely on magnitude spectra and pay limited attention to phase information [13–15]. Phase Congruency (PC) measures the agreement of phase angles across multiple frequencies, which makes it naturally insensitive to intensity contrast while remaining highly sensitive to edges and textures. Motivated by this property, we propose a novel Phase Congruency Enhanced Module (PCE). It injects PC cues into a Transformer backbone to strengthen target feature representations in low contrast scenes. To further amplify the effect of enhanced phase features on key regions [16], we also design a loss weighting mechanism (pc_weight) for subsequent loss supervision. It accurately strengthens the feature expression of low–contrast targets and significantly improves detection accuracy.

Another challenge arises from severe background interference. ORSI backgrounds often contain abundant target–irrelevant components such as clouds, vegetation, water bodies, and terrain shadows. These elements exhibit visual similarity to salient objects and are therefore easily confused with true targets, causing significant interference in accurate recognition and segmentation. Inspired by the idea of "residual attention guidance" in the field of image reconstruction and cross–attention dynamic fusion in the field of autonomous driving [17–19], we design a novel Dynamic Residual Fusion Module(DRF). It's designed to address complex background interference in ORSI–SOD. Its core lies in inter–scale cross–scale feature integration. The module first fuses shallow and deep features, then uses dynamic spatial attention and channel attention to purify the fused features. This highlights task relevant regions and suppresses redundant background. In addition, residual connections are used during fusion to preserve fine grained structural details and prevent weak edges from being lost under complex interference. As a result, DRF performs both feature fusion and feature purification, improving robust foreground segmentation under severe background interference.

The main contributions are as follows:

- We propose a novel end–to–end network PCFNet. It integrates frequency domain phase enhancement with dynamic cross scale feature refinement to reduce blurred target boundaries under low contrast and improve robustness under complex background interference.
- We design a novel PC module based on Fourier decomposition theory. It fuses multi scale phase features with shallow Transformer features to solve blurred target perception caused by low contrast and compensate for local detail loss.
- We design a novel DRF module. It integrates dynamic spatial attention and residual connections to achieve complementary fusion and effective selection of multi scale features, suppressing complex background interference while preserving salient structures.
- We conduct extensive experiments on three benchmark datasets (ORSSD, EORSSD, ORSI4199). Through quantitative comparison with 23 SOTA methods, ablation experiments, and qualitative analysis of complex scenes, the effectiveness and robustness of the proposed network and core modules are fully demonstrated.

 doi:10.20944/preprints202603.2230.v1

2. Related Work

2.1. Salient Object Detection for NSI

Early SOD methods for Natural Scene Images (NSI) relied on handcrafted features, which only utilized low-level visual cues such as color contrast [20]. Due to the lack of high-level semantic capture capability, these methods exhibited insufficient robustness in complex scenes. With the rise of deep learning, encoder-decoder architectures became the mainstream. However, early deep learning models had limitations in modeling long-range semantic relationships. Luo et al. [21] confirmed that this limitation would lead to semantic incoherence when processing large-scale targets, making it difficult to adapt to the requirements of complex scenes.

The introduction of self-attention mechanisms achieved a performance breakthrough. Such methods can directly model long-range semantic dependencies by calculating the correlation weights of all pixel pairs. For example, the DETR framework proposed by Carion et al. [22] transformed SOD into a set prediction task, abandoning anchor boxes to realize end-to-end segmentation and greatly simplifying the process. Liu et al.'s [23] Swin Transformer adopted a shifted window mechanism, which reduced computational complexity while retaining global modeling capability, adapting to high-resolution image scenarios. To fuse global semantics and local details, Xie et al. [24] designed a dual-branch Pyramid Grafting Network, achieving feature complementarity through cross-attention guidance.

Although numerous high-performance solutions have been developed for NSI-SOD, the particularities of ORSI in terms of target-motion blur, scale range, and imaging perspective determine that these methods cannot be directly adapted to ORSI-SOD tasks. Nevertheless, validated technical frameworks and design ideas still provide important references for research on ORSI-SOD.

2.2. Salient Object Detection for ORSI

In light of the distinctive properties of ORSI, many researchers have introduced innovative ideas and designed dedicated, effective algorithms for ORSI-SOD.

Li et al. [25] pointed out that the unique characteristics of ORSI put forward higher requirements for the adaptability of detection algorithms. Early methods combining handcrafted features and threshold segmentation were only applicable to small-scale simple scenes. With the rise of deep learning, U-Net and its variants [26] became the mainstream. However, the local modeling limitation of CNNs is more prominent in remote sensing scenarios, which easily leads to missed detection of dense targets or incomplete segmentation of targets with complex topological structures.

Deep learning promoted the development of efficiency and various optimization strategies have been targeted core problems. Zeng et al. [27] designed an adaptive edge-aware semantic interaction network, which improved the boundary accuracy of irregular targets such as rivers and edges. Some studies [28] also used boundary enhancement loss functions to optimize segmentation integrity. Gao et al. [29] proposed the Adaptive Spatialization Transformer, which enhances salient region representation by non-uniform token allocation and suppresses background redundancy. Li et al. [30] proposed a global context relation-guided network to capture spatial dependencies and aggregate multi-level features. Cheng et al. [31] proposed the Multimodal-Guided Transformer, which fuses RGB and depth information for heterogeneous feature collaboration network adopted the design of "global semantic modeling and local detail calibration" and achieved performance break-through through authoritative calibration. Despite the progress of the aforementioned studies, core challenges such as weak features of low-contrast targets and complex backgrounds in ORSI-SOD remain unsolved.

2.3. Frequency-Domain Analysis

Frequency-domain analysis provides a complementary perspective to spatial feature extraction by capturing target structure and texture information, and it has been widely applied in computer vision tasks. For example, feature enhancement networks based on discrete cosine transform and multi-scale

 doi:10.20944/preprints202603.2230.v1

analysis models using wavelet transform both improve target detail by enhancing medium-high frequency components [32].

As a classic frequency-domain analysis method, Phase Congruency (PC) is based on Fourier decomposition theory. It quantifies the phase synchronization of multi-scale frequency components and inherently offers advantages of illumination robustness and complete detail preservation. However, in ORSI-SOD, the application of frequency domain analysis and phase congruency still requires further exploration.

3. Proposed Method

In this section, we will introduce the proposed PCFNet in detail. Section 3.1 presents an overview of the general framework of PCFNet. Section 3.2 outlines the Transformer-based feature extractor. Sections 3.3and 3.4 describe the core components PCE and DRF in detail, respectively. Section 3.5 introduces the decoder selected for the model. Finally, Section 3.6 explains the loss function used in the model.

3.1. Framework Overview

As shown in Figure 1, the proposed PCFNet adopts an encoder-decoder architecture. The input image passes sequentially through four Swin-B blocks to extract features at four scales. At the shallowest level, the PCE module compensates for target detail loss caused by low contrast through phase congruency computation. Subsequently, deeper-level features are refined by the DRF module, which fuses and optimizes multi-scale features to suppress background interference while preserving salient structures. Finally, the hierarchical features are aggregated by the SGAED decoder to produce a high-resolution saliency map.

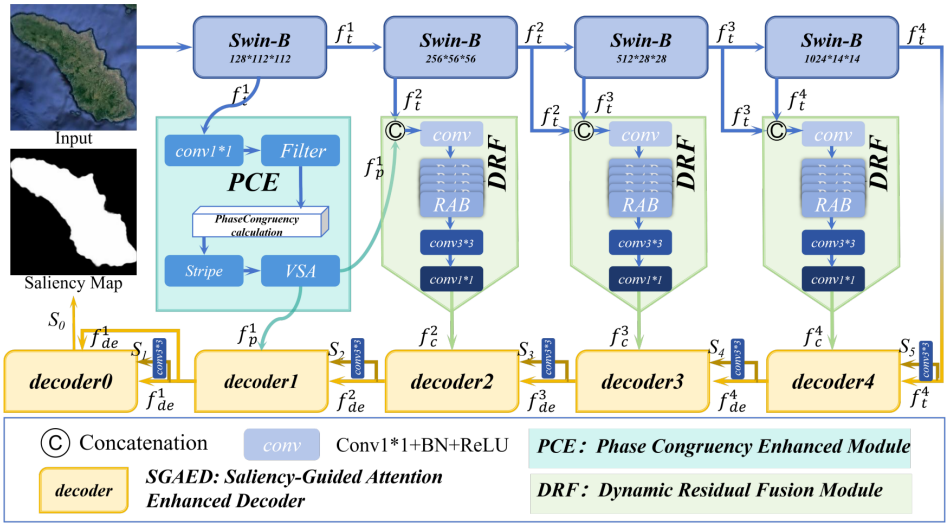

Figure 1. Overall architecture of the proposed network. The network consists of four components: the Swin-B feature extractor, the PCE module,the DRF module, and the SGAED decoder.

3.2. Swin Transformer-based Feature Extractor

Following existing studies, we select the classical Swin-B as the basic feature extractor [23,25]. Its hierarchical representation and window based local self attention are well suited to the dense prediction requirements of ORSI-SOD. Different from the original Swin-B architecture designed for general vision tasks, this paper based adaptive optimization targeting the characteristics of remote sensing images. We retain the four-level feature extraction structure of the backbone network, and adjust the channel dimensions of deep layers to balance semantic representation capability and computational cost. The Swin-B in Figure 1 corresponds to the transformer-based encoder subnetwork, which consists of four feature extraction blocks. We take the output of each block as the side-output feature map, defined as $\mathbf{f}_t^i \in \mathbb{R}^{C_i \times H_i \times W_i} (i \in \{1,2,3,4\})$. For the input image, the channel numbers $C_{1,2,3,4}$ of each feature

 doi:10.20944/preprints202603.2230.v1

map are $\{128, 256, 512, 1024\}$, and the resolutions $H_{1,2,3,4} = W_{1,2,3,4}$ correspond to $\{112, 56, 28, 14\}$ respectively.

3.3. Phase Congruency Enhanced Module

ORSI often suffer from edge blurring and false target detection in low-contrast scenes, which impairs the discriminability of salient targets. To address this issue, we design a Phase Congruency Enhanced (PCE) Module (as shown in Figure 2), which is deployed after the first stage Swin-B feature extractor (for the feature map $\mathbf{f}_t^1$). Through the synergy of frequency-domain phase feature analysis and spatial-domain feature enhancement, the PCE module strengthens fine-grained local features of remote sensing targets, providing high-quality inputs for subsequent cross-scale feature fusion.

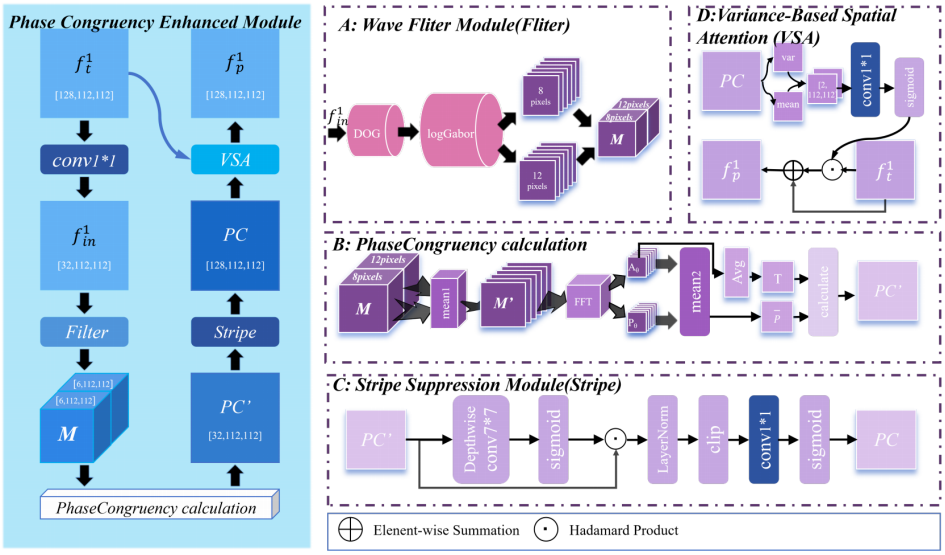

Figure 2. Internal structure of PCE, consisting of Wave Filter Module(A), Phase Congruency Calculation(B), Stripe Suppression Module(C) and Variance-Based Spatial Attention(D)

1)A: The Wave Filter Module serves as the frequency-domain response extraction unit of the PCE module, responsible for capturing multi-scale and multi-directional edge information of targets. First, the first-stage feature $\mathbf{f}_t^1$ output by Swin-B is compressed to 32 channels via a 1×1 convolution to reduce computational complexity:

$$\mathbf{f}_{in}^1 = \mathrm{Conv}_{1\times1}(\mathbf{f}_t^1) \tag{1}$$

Subsequently, $\mathbf{f}_{in}^1$ is fed into the Wave Filter Module. It is successively processed by the Difference of Gaussians (DoG) filtering and the multi-scale multi-orientation LogGabor filtering. The module applies the same filtering operation to all 32 channels. For each channel, the LogGabor filter bank is constructed with 2 scales (wavelengths of 8 pixels and 12 pixels) and 6 evenly distributed orientations($0°, 30°, 60°, 90°, 120°, 150°$). By this means, a total of 12 frequency-domain response maps(M) are generated, which cover edge texture features of targets at different scales and directions. They provide basic inputs for subsequent phase analysis.

2)B: The Phase Congruency Calculation module is the core functional unit of the PCE module. Based on the visual cognition law, image features concentrate at positions where the phase superposition of Fourier harmonics is maximum [34]. Therefore, this property can be exploited to achieve accurate localization of target features. After feeding M into this module, we first perform scale wise averaging. For each of the six orientations, we average the two scales with equal weights of 0.5 to obtain the refined M′. Here, each of the six waves corresponds to one of the 6 orientations (denoted by θ, where $\theta = 1, 2, \ldots, 6$). Then, we decompose each response map of the refined six waves into amplitude components $\mathbf{A}_{\theta,c}$ (c denotes the number of channels, $c = 1, 2, \ldots, 32$) and phase components $\mathbf{P}_{\theta,c}$ via Fast Fourier Transform (FFT) for each channel. We calculate the amplitude-weighted mean phase $\bar{P}_c$ of the phase components:

 doi:10.20944/preprints202603.2230.v1

$$\bar{P}_c = \frac{\sum_{\theta=1}^{6}(\mathbf{A}_{\theta,c} \odot \mathbf{P}_{\theta,c})}{\sum_{\theta=1}^{6} \mathbf{A}_{\theta,c} + \varepsilon} \tag{2}$$

where ε is a small constant to avoid division by zero. Then, in the "calculate" step, we compute the phase deviation and phase congruency via Eq. (3) to (5), where Eq. (3) is dedicated to phase deviation calculation and Eqs. (4)(5) for phase congruency evaluation.

$$\Delta\phi_{\theta,c} = \cos(\mathbf{P}_{\theta,c} - \bar{P}_c) - \left|\sin(\mathbf{P}_{\theta,c} - \bar{P}_c)\right| \tag{3}$$

$$\mathrm{PC}'_c = \frac{\sum_{\theta=1}^{6} \max(\mathbf{A}_{\theta,c} \odot (\Delta\phi_{\theta,c} - T_c), 0)}{\sum_{\theta=1}^{6} \mathbf{A}_{\theta,c} + \varepsilon} \tag{4}$$

$$\mathrm{PC}' = \mathrm{Concat}\left(\mathrm{PC}'_1, \mathrm{PC}'_2, \ldots, \mathrm{PC}'_{32}\right) \tag{5}$$

where $T_c = 0.5 \cdot \mathrm{Avg}_{3\times3}(\bar{\mathbf{A}}_{\theta,c})$ denotes the local amplitude mean threshold. Here, $\bar{\mathbf{A}}_\theta$ is obtained by averaging the amplitude spectrum $\mathbf{A}_\theta$ over the orientation dimension ($\bar{\mathbf{A}}_\theta = \frac{1}{6}\sum_{\theta=1}^{6} \mathbf{A}_\theta$). $\mathrm{Avg}_{3\times3}(\cdot)$ represents 3×3 average pooling. $\max(\cdot, 0)$ filters out invalid phase contributions. Notably, the above operations are performed independently for each of the 32 channels in $\mathbf{f}_{in}^1$. Thus, the final dimension of PC′ is $32 \times 112 \times 112$, with a value range of $[0, 1]$. A larger value indicates a higher probability that the position is a target feature.

3)C: Stripe Suppression Module has a dual function of eliminating stripe artifacts and recovering channel dimension. As shown in Figure 3, (a) and (c) are feature maps exhibiting stripe artifacts caused by phase congruency computation without the stripe suppression module, while (b) and (d) are the clean feature maps obtained after applying the module, with stripe interference clearly eliminated. After getting the clean feature maps, it restores the feature channel dimension to be consistent with that of the input feature. This lays a solid foundation for subsequent spatial attention enhancement. The input of this module is the preliminary phase congruency map PC′ and the specific processing process is as follows:

$$\begin{cases} \tilde{\mathrm{PC}} = \mathrm{PC}' \odot \sigma\left(\mathrm{DepthwiseConv}_{7\times7}(\mathrm{PC}')\right) \\ \mathrm{PC} = \sigma\left(\mathrm{Conv}_{1\times1}\left(\mathrm{Clamp}\left(\mathrm{LayerNorm}\left(\tilde{\mathrm{PC}}\right)\right)\right)\right) \end{cases} \tag{6}$$

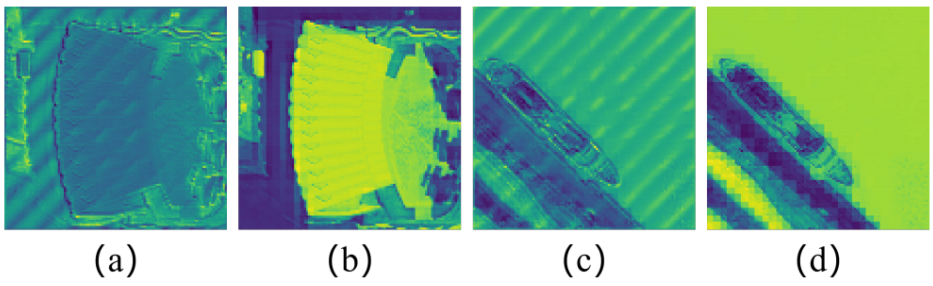

Figure 3. (a)(c) are without the Stripe Suppression Module and (b)(d) are with it. (a)(b) form the first example pair and (c)(d) form the second example pair.

$\sigma(\cdot)$ represents sigmoid activation function. First, the 7×7 depthwise separable convolution is performed on PC′ to capture the response of stripe artifact regions. And the suppression mask is generated by the Sigmoid function. The mask is multiplied with the original PC′ in an element-wise manner to suppress the response of stripe regions. Then, the layer normalization is used to standardize the feature distribution. And the value clipping is adopted to limit the dynamic range of feature values$[-3, 3]$, completing the artifact elimination and feature stabilization. Subsequently, the 1×1 convolution is applied to the artifact-removed features to restore the channel dimension from 32 to 128.

 doi:10.20944/preprints202603.2230.v1

Finally, the Sigmoid function is used to map the feature values to the range of $[0, 1]$, generating the phase congruency weight map PC.

4)D: The Variance-Based Spatial Attention is the output enhancement unit of the PCE module, used to improve the spatial discriminability of features. We use variance because in ORSI, boundaries or texture regions of salient objects usually exhibit noticeable gray-level jumps, leading to high local variance. High variance typically reflects the contours or locations of salient objects, while low variance corresponds to smooth regions. After feeding PC into this module, we first calculate the mean $\mu(\text{PC})$ and variance $\sigma(\text{PC})$ across all channels for each spatial position. And we concatenate them and compress the dimension via a 1×1 convolution to generate spatial attention weights. Then we perform element-wise multiplication with f_t^1 and add a residual connection to avoid semantic loss. Finally the enhanced feature f_p^1 is outputted, as shown in Eq. (7):

$$f_p^1 = f_t^1 \odot \sigma(\text{Conv}_{1\times1}([\mu(\text{PC}), \sigma(\text{PC})])) + f_t^1 \tag{7}$$

The enhanced feature f_p^1 with significantly strengthened target information is fed into the subsequent DRF module for cross-scale feature fusion.

3.4. Dynamic Residual Fusion (DRF) Module

The Dynamic Residual Fusion (DRF) Module is designed to alleviate the insufficient cross-scale feature fusion and suppress complex background interference. It leverages the residual attention mechanism to gradually achieve sufficient interaction between shallow and deep features in both channel and spatial dimensions. As shown in Figure 4, the DRF module consists of a feature concatenation unit, Residual Dual Attention Blocks (RABs), and a global residual connection. In the DRF, the feature concatenation unit is used to fuse cross-scale features. The dual attention mechanism is employed to filter out cross-scale features. And the residual connection ensures the preservation of detailed information and gradient stability during the feature enhancement process.

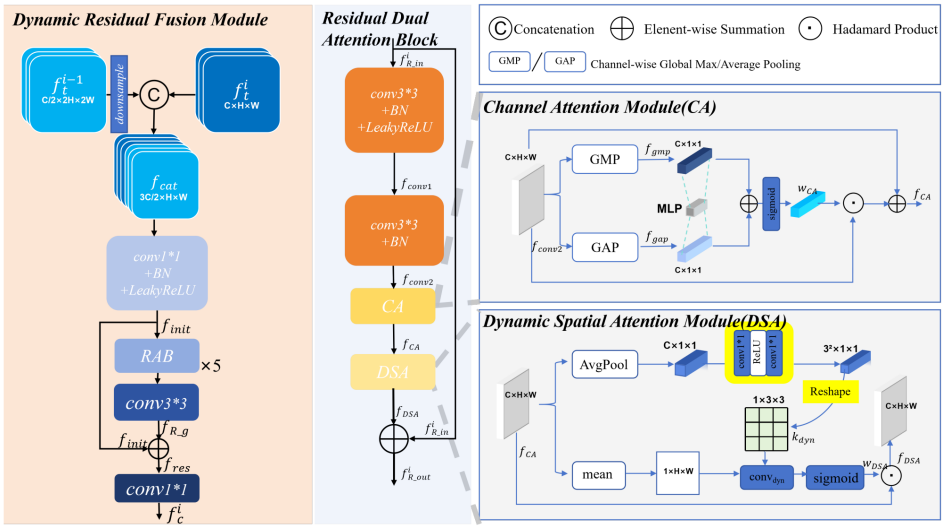

Figure 4. Internal structure of DRF. Five RABs are connected in serie.

The input of DRF includes two cross-scale features f_t^{i-1} and f_t^i. First, f_t^{i-1} (after downsampling) and f_t^i are concatenated along the channel dimension to obtain the feature f_{cat}. Then a 1×1 convolution followed by BN and LeakyReLU is used to compress the channel dimension, yielding the initial feature f_{init}.

$$f_{init} = \text{LeakyReLU}(\text{BN}(\text{Conv}_{1\times1}(f_{cat}))) \tag{8}$$

Subsequently, f_{init} is fed into a residual attention group composed of 5 stacked RABs. Enlightened by the gradient residual attention network [17], which effectively aggregates fine-grained features via stacked residual dense blocks. We choose 5 RABs as intra-group units by balancing the feature

 doi:10.20944/preprints202603.2230.v1

enhancement performance, computational overhead and gradient stability. The validity of this configuration is demonstrated in subsequent ablation experiments. For each RAB, the input feature is first preprocessed by two 3×3 convolutions. The first layer is followed by BN and LeakyReLU, and the second layer only retains BN, resulting in the feature f_{conv2}.

$$f_{conv1} = \text{LeakyReLU}(\text{BN}(\text{Conv}_{3\times3}(f_{R,in}))) \tag{9}$$

$$f_{conv2} = \text{BN}(\text{Conv}_{3\times3}(f_{conv1})) \tag{10}$$

Based on f_{conv2}, Channel Attention (CA) Module is first used to generate channel weights. Global max pooling (GMP) and global average pooling (GAP) along the channel dimension are performed on f_{conv2} to obtain f_{gmp} and f_{gap}. These two features are fed into a shared MLP for dimension reduction and restoration, then summed up and activated by sigmoid to generate the weight w_{CA}. The weight is element-wise multiplied with f_{conv2}.

$$f_{gmp} = \text{GMP}_c(f_{conv2}), \quad f_{gap} = \text{GAP}_c(f_{conv2}) \tag{11}$$

$$w_{CA} = \sigma\big(\text{MLP}(f_{gmp}) + \text{MLP}(f_{gap})\big) \tag{12}$$

$$f_{CA} = f_{conv2} \odot w_{CA} + f_{conv2} \tag{13}$$

Then, the Dynamic Spatial Attention (DSA) Module is applied to f_{CA} to achieve spatial feature enhancement. First, global average pooling is performed on the input feature f_{CA}, followed by two successive 1×1 convolutions and ReLU activation. Then, a 3×3 convolution kernel is generated through the Reshape operation. The module is dynamic in that the kernel shape remains 3×3, while its weight parameters are not fixed preset values. Instead, the weight parameters are adaptively calculated from the independent input features of each individual sample. Each sample in a batch is endowed with a unique set of 3×3 convolution kernel weights, which can adapt to the spatial feature pattern of the sample itself in a targeted manner. This is the core embodiment of the "dynamic" property in our DSA module. Meanwhile, the channel-wise mean operation is conducted on f_{CA} to obtain a single-channel spatial feature map. The sample-specific 3×3 dynamic convolution kernels are used to perform dynamic convolution on this single-channel spatial feature map. After that, the spatial attention weight w_{DSA} is generated by the Sigmoid activation function. Finally, the original feature f_{CA} is multiplied element wise by the spatial attention weight w_{DSA} to obtain the enhanced feature f_{DSA}.

$$k_{dyn} = \text{Reshape}(\text{Conv}_{1\times1}(\text{ReLU}(\text{Conv}_{1\times1}(\text{AvgPool}(f_{CA}))))) \tag{14}$$

$$w_{DSA} = \sigma\Big(\text{Conv}_{dyn}(\text{mean}(f_{CA}))\Big) \tag{15}$$

$$f_{DSA} = f_{CA} \odot w_{DSA} \tag{16}$$

The output of each RAB is added to the input feature via a local residual connection.

$$\begin{aligned} f_{R\text{-}g} = f^{5}_{R\text{-}out} = \text{Conv}_{3\times3}\big(&RAB\_5(RAB\_4 \\ &(RAB\_3(RAB\_2(RAB\_1(f_{init})))))\big) \end{aligned} \tag{17}$$

$$f^{i}_{R\text{-}out} = f_{DSA} + f^{i}_{R\text{-}in} \ (i = 1, 2, 3, 4, 5) \tag{18}$$

 doi:10.20944/preprints202603.2230.v1

$$f_{res} = f_{R\text{-}g} + f_{init} \tag{19}$$

$$f_c^i = \mathrm{Conv}_{1\times1}(f_{res}) \tag{20}$$

The output of five serially connected RABs is integrated into $f_{R\text{-}g}$ through a 3×3 convolution, then added to f_{init} via a global residual connection to obtain f_{res}. Finally, a 1×1 convolution is used to adjust the channel dimension, making it consistent with the f_t^i, yielding the fused feature f_c^i of DRF.

3.5. Decoder

To ensure the integrity of the network architecture and achieve accurate mapping from high-level fused features to pixel-level saliency maps, this paper directly adopts the Saliency-Guided Attention Enhanced Decoder (SGAED) from the literature as the decoding module of the network [31]. This decoder can adapt to the feature format output by the encoder in this paper without additional modifications. Its core function is to receive the three groups of high-level fused features ($\mathbf{f}_c^i \in \mathbb{R}^{256\times\frac{112}{2^{i-1}}\times\frac{112}{2^{i-1}}}, i = 2,3,4$) from the DRF module and the enhanced shallow feature ($\mathbf{f}_p^1 \in \mathbb{R}^{128\times112\times112}$) from the PCE module. In terms of the specific process, f_t^4 feature and $S_5(14 \times 14)$ generated from the f_t^4 feature are fed into decoder4, followed by decoder4 to decoder1 that gradually perform feature upsampling and detail refinement (decoder4 does not perform upsampling, $\mathbf{f}_{de}^i \in \mathbb{R}^{C\times\frac{112}{2^{i-1}}\times\frac{112}{2^{i-1}}}, S_i(\frac{112}{2^{i-1}} \times \frac{112}{2^{i-1}}), i = 2,3,4, \mathbf{f}_{de}^2 \in \mathbb{R}^{128\times112\times112}$). Specifically, decoder0 is designed to optimize the $S_1(112 \times 112)$ features at the resolution of 112×112 for generating the high-detail $S_0(224 \times 224)$, which supplies fine-grained supervision. Through repeated computation, key features are strengthened and representation capability is improved, while preserving the integrity of the five SGAED components in the original framework. Finally, all saliency maps are interpolated to 224×224 and used as the supervised output and final predictions.

3.6. Loss Function

The Loss constructed in this paper combines CE Loss and IoU Loss [35,36]. And we design a novel weighting mechanism for the phase congruency map output by the phase congruency enhancement module.

1) Saliency Supervision We adopt CE Loss and IoU Loss to perform basic supervision on the last 4 layers of intermediate saliency maps, which is formulated as follows:

$$\mathcal{L}_B = \sum_{i=2}^{5} \left(\mathcal{L}_{CE}(\mathrm{Up}_{k_i}(S_i), G) + \mathcal{L}_{\mathrm{IoU}}(\mathrm{Up}_{k_i}(S_i), G)\right) \tag{21}$$

where S_i represents the i-th saliency prediction map ($i = 2,3,4,5$), G is the ground-truth label map, and $\mathrm{Up}_{k_i}(\cdot)$ denotes the upsampling operation with an upsampling factor of k_i. The value of k_i is defined as:

$$k_i = \begin{cases} 2^4, & \text{if } i = 5 \\ 2^i, & \text{if } i \leq 4 \end{cases} \tag{22}$$

2) Phase Congruency Map Weighting (pc_weight) Existing loss functions fail to focus on salient structural regions in ORSI, so we design a phase congruency map weighting mechanism, referred to as pc_weight. It strengthens supervision on key regions and mitigates the impacts of weak features of low contrast targets. It should be noted that we only apply this weighted loss to S_0 and S_1. Because the phase congruency enhancement module is embedded only in stage 1 of feature extraction, and its effect is limited to the shallow feature stage.

The pc_map refers to the single channel phase congruency map derived by averaging the 128 channels phase congruency map PC generated by Eq. (6) in Section 3.3 above. It reflects the distribution characteristics of salient structures such as edges and textures in the image. First, the original pc_map

 doi:10.20944/preprints202603.2230.v1

is interpolated to the size of the current prediction map and normalized to obtain the pixel-level weight ω (pc_weight), calculated as:

$$\omega = \frac{\mathrm{Up}_s(\text{pc\_map}) - \min(\mathrm{Up}_s(\text{pc\_map}))}{\max(\mathrm{Up}_s(\text{pc\_map})) - \min(\mathrm{Up}_s(\text{pc\_map})) + \beta} \tag{23}$$

where $\mathrm{Up}_s(\cdot)$ is the upsampling operation to interpolate pc_map to the size of the i-th prediction map. $\beta = 10^{-8}$ is used to avoid denominator being zero. For the first two layers of high-resolution intermediate saliency maps ($S_i, i = 0, 1$), this weight is introduced to weight the loss, formulated as:

$$\begin{aligned} \mathcal{L}^{(i)}_{\text{weighted}} = \mathcal{L}_{CE}\big(\mathrm{Up}_{k_i}(S_i) \odot \omega, G \odot \omega\big) \\ + \mathcal{L}_{\text{IoU}}\big(\mathrm{Up}_{k_i}(S_i) \odot \omega, G \odot \omega\big) \end{aligned} \tag{24}$$

ω is the normalized pixel-level phase congruency weight map (ranging in $[0, 1]$), $\mathrm{Up}_{k_i}(\cdot)$ denotes the upsampling operation with an upsampling factor of k_i, and $\odot$ represents the element-wise multiplication operation.

3) Overall Loss Function Finally, the overall loss function $\mathcal{L}_{\text{total}}$ of the model is formulated as follows:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_B + \sum_{i \in \{0,1\}} \mathcal{L}^{(i)}_{\text{weighted}} \tag{25}$$

4. Experiments

4.1. Experimental Settings

1) Dataset We adopt three mainstream benchmark datasets for ORSI-SOD, including ORSSD [37], EORSSD [38] and ORSI4199 [39]. These datasets cover scenes of varying complexity to comprehensively test the model's detection performance.

ORSSD, the first public dataset for remote sensing salient object detection, with 800 pixel-annotated images (600 training, 200 testing). It covers regular scenes, verifying the model's basic performance.

EORSSD is an extension of ORSSD, with 2000 images (1400 training, 600 testing). It adds targets with complex backgrounds and irregular structures, increasing detection difficulty to test the model's adaptability to complex scenes.

ORSI4199 is a large-scale challenging dataset, with 4199 high-precision annotated images (2000 training, 2199 testing). It includes multi-attribute complex targets (large, small, low-contrast) in realistic scenes, verifying the model's generalization ability.

2) Implementation Details To verify the proposed network's performance in ORSI-SOD, experiments are implemented via the PyTorch framework on a workstation with NVIDIA RTX 3090 Ti GPUs, using Python 3.8, PyTorch 1.11.0 and related libraries. During training, images are augmented with random flip, rotation and Gaussian blur, resized to 224×224. The Adam optimizer (initial learning rate 7×10^{-5}, $\beta_1 = 0.9$, $\beta_2 = 0.999$) is used, with learning rate adjusted via StepLR (step size=20 epochs, decay factor=0.5). Batch size is 8 and the model is trained for 100 epochs.

3) Evaluation Metrics Four mainstream metrics are adopted in the experiments to comprehensively evaluate the model performance.

S-Measure (S_α) [40]: It measures the structural similarity between the saliency map and the ground truth, integrating object-level (S_{oj}) and region-level (S_{re}) similarity, with the formula as follows:

$$S_\alpha = \alpha \times S_{oj} + (1 - \alpha) \times S_{re} \tag{26}$$

where the balance weight α is set to 0.5 by default.

 doi:10.20944/preprints202603.2230.v1

F-Measure (F_β) [20]: It is a comprehensive metric that balances precision (Pr) and recall (Re), calculated by:

$$F_\beta = \frac{(1+\beta^2) \times Pr \times Re}{\beta^2 \times Pr + Re} \tag{27}$$

Following the mainstream setting, β is set to 0.3.

E-measure (E_ξ) [41]: It considers both pixel-level correspondence and image-level statistical information simultaneously, with the formula:

$$E_\xi = \frac{1}{H \times W} \sum_{x=1}^{H} \sum_{y=1}^{W} \phi(x,y) \tag{28}$$

where H and W represent the height and width of the saliency map, and $\phi(x,y)$ denotes the enhanced alignment function.

MAE (M) [42]: It calculates the average pixel deviation between the saliency map and the ground truth, defined as:

$$\text{MAE} = \frac{1}{H \times W} \sum_{x=1}^{H} \sum_{y=1}^{W} |S(x,y) - G(x,y)| \tag{29}$$

where $S(x,y)$ and $G(x,y)$ are the pixel values of the saliency map and the ground truth at position (x,y), respectively.

4.2. Comparison with SOTA Methods

1) Comparison Methods Our proposed model (ours) and 23 state-of-the-art (SOTA) models are compared across all three benchmark datasets. The compared methods encompass a diverse range of models: RRWR [20] and RCRR [43] are two conventional NSI-SOD models. EGNet [44], MINet [45], and GatedNet [46] are three deep NSI-SOD models.

LVNet [37], DAFNet [38], MCCNet [47], CorrNet [48], ASTTNet [28], MJRBM [49], RRNet [40], EMFINet [50], ERPNet [44], ACCoNet [51], AESINet [27], DCCNet [52], LSHNet [53], MCPNet [54] are fourteen deep ORSI-SOD models. HFANet [55], ADSTNet [56], HFCNet [31], CMNFNet [57] are four Hybrid ORSI-SOD models. Table 1 lists all the quantitative results, which were generated by running the corresponding open-source codes provided by the author and adopting the default parameter configurations, or through calculations based on publicly accessible saliency maps.

2) Quantitative Comparisons and Discussions As shown in Table 1, our method (ours) performs prominently in key metrics across ORSSD, EORSSD, and ORSI4199 datasets. Overall, it outperforms most state-of-the-art methods. On the ORSSD dataset, our method ranks first in S_α, F_β, E_ξ, and MAE (0.9540, 0.9305, 0.9888, 0.0071 respectively). Compared with CMNFNet, it achieves a 0.65% improvement in S_α and a 1.16% improvement in F_β. On the EORSSD dataset, our method ranks first in F_β(0.8943), E_ξ (0.9843) and MAE (0.0048), and second in S_α (slightly behind HFCNet-R), while maintaining more balanced overall performance. On the ORSI4199 dataset, our method ranks first in two metrics, with S_α(0.8858) and F_β (0.8859). In summary, our method demonstrates competitive-ness across all datasets, verifying its effectiveness in ORSI-SOD.

3) Qualitative Comparisons and Discussions We select representative visual examples from the ORSSD, EORSSD, and ORSI4199 datasets. These examples are grouped by scene type (as shown in Figure 5) to qualitatively compare our method with state-of-the-art models.

 doi:10.20944/preprints202603.2230.v1

Table 1. Quantitative comparison of our method with 23 methods proposed by other researchers on the ORSSD, EORSSD and ORSI4199 datasets. The symbol "↑" indicates that a higher value is better for the metric, while "↓" indicates that a lower value is better. The top three results are highlighted in red, blue, and green, respectively. Results of some models may be unavailable for partial datasets, which are indicated as "-".

Method	Publication	Type	ORSSD				EORSSD				ORSI4199			
			$S_\alpha\uparrow$	$F_\beta\uparrow$	$E_\xi\uparrow$	MAE↓	$S_\alpha\uparrow$	$F_\beta\uparrow$	$E_\xi\uparrow$	MAE↓	$S_\alpha\uparrow$	$F_\beta\uparrow$	$E_\xi\uparrow$	MAE↓
RRWR	2015 CVPR	T-NSI	0.6835	0.5590	0.7649	0.1324	0.5992	0.3993	0.6894	0.1677	0.6416	0.5407	0.7116	0.1717
RCRR	2018 TIP	T-NSI	0.6849	0.5591	0.7651	0.1277	0.6007	0.3995	0.6882	0.1644	0.6491	0.548	0.7192	0.1637
ASTTNet	2023 TGRS	T-ORSI	0.9347	0.9060	0.9794	0.0094	0.9253	0.8741	0.9580	0.006	0.8827	0.8788	0.9512	0.0273
EGNet	2019 ICCV	C-NSI	0.8721	0.8332	0.9731	0.0216	0.8601	0.7880	0.9570	0.0110	0.8516	0.8371	0.9241	0.0385
MINet	2020 CVPR	C-NSI	0.9040	0.8761	0.9545	0.0144	0.9040	0.8344	0.9442	0.0093	0.8116	0.7988	0.8961	0.0504
GatedNet	2020 ECCV	C-NSI	0.9186	0.8871	0.9664	0.0137	0.9114	0.8566	0.9610	0.0095	0.8545	0.8450	0.9256	0.0393
LVNet-V	2019 TGRS	C-ORSI	0.8815	0.8263	0.9456	0.0207	0.8630	0.7794	0.9254	0.0146	-	-	-	-
DAFNet-V	2021 TIP	C-ORSI	0.9191	0.8928	0.9771	0.0113	0.9166	0.8614	0.9861	0.0060	0.8492	0.8348	0.9181	0.0422
MCCNet-V	2021 TGRS	C-ORSI	0.9437	0.9155	0.9800	0.0087	0.9327	0.8904	0.9755	0.0066	-	-	-	-
CorrNet-V	2022 TGRS	C-ORSI	0.9380	0.9129	0.9790	0.0098	0.9289	0.8778	0.9696	0.0083	0.8626	0.8560	0.9333	0.0366
MJRBM-R	2022 TGRS	C-ORSI	0.9211	0.8885	0.9686	0.0145	0.9091	0.8555	0.9655	0.0099	0.8582	0.8511	0.9343	0.0372
RRNet-R	2022 TGRS	C-ORSI	0.9339	0.9011	0.9722	0.0113	0.9266	0.8743	0.9665	0.0082	0.8585	0.8500	0.9286	0.0367
EMFINet-R	2022 TGRS	C-ORSI	0.9432	0.9155	0.9813	0.0095	0.9319	0.8742	0.9712	0.0075	0.8712	0.8636	0.9403	0.0313
ERPNet-R	2023 TCYB	C-ORSI	0.9352	0.9036	0.9738	0.0114	0.9252	0.8743	0.9665	0.0082	0.8636	0.8528	0.9292	0.0388
ACCoNet-R	2023 TCYB	C-ORSI	0.9428	0.9149	0.9819	0.0087	0.9302	0.8821	0.9759	0.0067	0.8805	0.8688	0.9424	0.032
AESINet-R	2023 TGRS	C-ORSI	0.9455	0.9160	0.9814	0.0085	0.9347	0.8792	0.9757	0.0064	0.8755	0.8726	0.9459	0.0305
DCCNet	2024 LGRS	C-ORSI	0.9417	0.9168	0.9805	0.0092	0.9345	0.8887	0.9761	0.0067	0.8705	0.8619	0.9348	0.0347
LSHNet	2024 TGRS	C-ORSI	0.9491	0.9200	0.9824	0.0075	0.9370	0.8643	0.9761	0.0064	0.8759	0.8758	0.9462	0.0299
MCPNet	2024 TGRS	C-ORSI	0.9433	0.9135	0.9807	0.0090	0.9373	0.8868	0.9765	0.0070	0.8736	0.8667	0.9402	0.0324
HFANet-R	2022 TGRS	H-ORSI	0.9399	0.9117	0.9770	0.0092	0.9380	0.8876	0.9740	0.0071	0.8767	0.8700	0.9431	0.0314
ADSTNet-R	2024 JSTARS	H-ORSI	0.9379	0.9124	0.9807	0.0086	0.9311	0.8804	0.9769	0.0065	0.8710	0.8698	0.9433	0.0318
HFCNet-R	2024 TGRS	H-ORSI	0.9521	0.9247	0.9885	0.0073	0.9407	0.8864	0.9793	0.0054	0.8838	0.8833	0.9539	0.0277
CMNFNet	2025 TCYB	H-ORSI	0.9475	0.9189	0.9832	0.0078	0.9377	0.8851	0.9774	0.0063	0.8774	0.8752	0.9885	0.0301
ours	-	T-ORSI	0.9540	0.9305	0.9888	0.0071	0.9393	0.8943	0.9843	0.0048	0.8858	0.8859	0.9531	0.0279

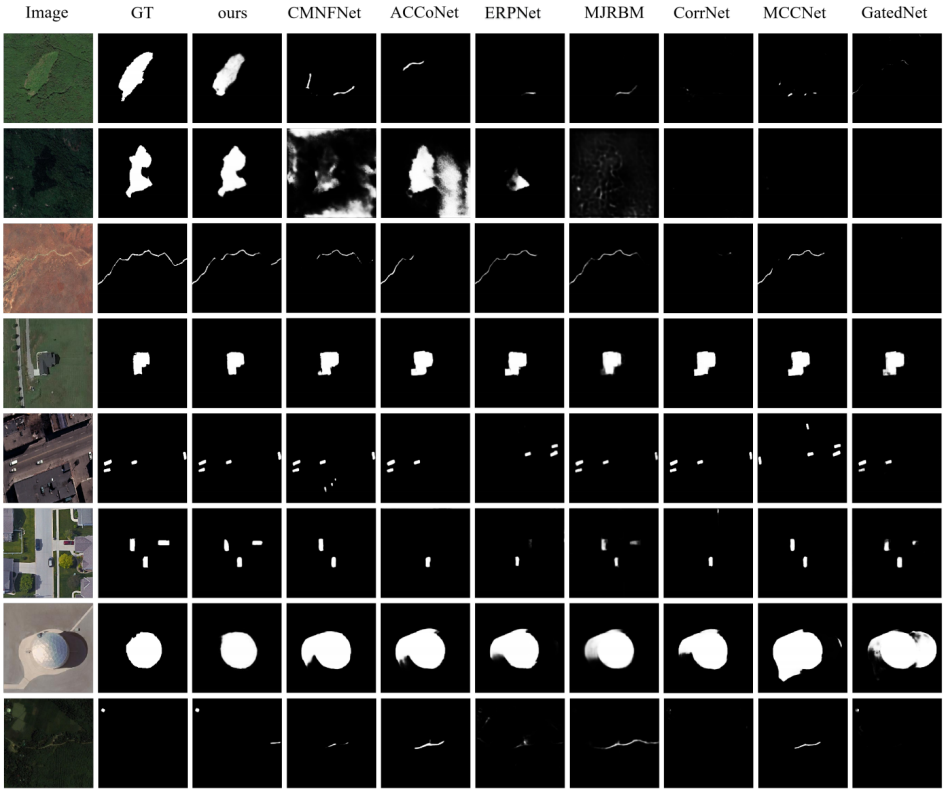

Figure 5. Qualitative comparisons of our method with seven representative SOTA methods in eight challenging scenarios. Rows 1, 2, 3, 4: low-contrast scenes, Rows 5, 6, 7, 8: complex background interference scenes.

For the first four rows representing low-contrast scenes, the target and background exhibit high texture similarity. In the 1st row, this similarity causes blurred edges in the saliency maps of ACCoNet and ERPNet. In the 2, 3, 4 rows, the subtle grayscale difference between target and background

 doi:10.20944/preprints202603.2230.v1

leads to missing targets or blurred adhesion in the saliency maps of CMNFNet and MJRBM. In contrast, our method enhances the discriminability of target features in low-contrast regions through the PCE module, producing saliency maps that accurately restore target contours and closely match the ground truth. For the last four rows representing complex background interference scenes, the complex background easily interferes with target detection. In Rows 5 and 6 with complex building scenes, models such as CMNFNet and MCCNet tend to introduce background noise spots. In contrast, our model separates the target from the background more clearly and remains robust to irrelevant background interference. In the 7th row featuring a dome-shaped scene, GatedNet suffers from edge distortion due to background element interference. While our dual module effectively resists redundant background interference and accurately aligns with the true target contour. In the 8th row, the target is extremely small and easily confused with the background, causing methods like CorrNet and MCCNet to miss the target or falsely activate background regions. In contrast, our method can detect it correctly.

4.3. Ablation Study

To verify the effectiveness of the proposed modules, core components, and loss functions, we conduct ablation experiments in this section. The ablations are designed at the module, component, and loss function levels. We further analyze the contribution of each part to model performance with visualization results.

1) Ablation study between different modules To verify the effectiveness of the proposed PCE and DRF modules, we designed four model variants on the ORSSD and EORSSD datasets: a) Base (Swin-B Encoder and SGAED Decoder), b) Base+PCE, c) Base+DRF, d) Base+DRF+PCE (ours).

The quantitative results (Table 2) demonstrate that: On the ORSSD dataset, the Base performs the worst. After integrating PCE, S_α is improved by 0.74%. When adding DRF, S_α increases by 0.64% while the F_β and E_ξ are slightly improved. Our method, which combines both modules, not only enhances the target discriminability in low-contrast regions via PCE but also aligns cross-scale features and filters background interference via DRF. Compared with the Baseline, our method achieves a 0.99% improvement in S_α and a 1.40% improvement in F_β. Consistent trends are observed on the EORSSD dataset, where our method improves S_α and F_β by 0.67% and 2.73% respectively compared with the Baseline, confirming the targeted value of the two modules in addressing typical issues of optical remote sensing images.

Table 2. Ablation study between different modules on ORSSD and EORSSD.

No.	Base	PCE	DRF	ORSSD			EORSSD		
				$S_\alpha \uparrow$	$F_\beta \uparrow$	$E_\xi \uparrow$	$S_\alpha \uparrow$	$F_\beta \uparrow$	$E_\xi \uparrow$
1	✓			0.9441	0.9165	0.9666	0.9326	0.8670	0.9589
2	✓	✓		0.9511	0.9215	0.9686	0.9361	0.8706	0.9582
3	✓		✓	0.9505	0.9267	0.9868	0.9354	0.8901	0.9806
4	✓	✓	✓	**0.9540**	**0.9305**	**0.9888**	**0.9393**	**0.8943**	**0.9843**

In Figure 6, the role of our modules is further demonstrated: In the scene of the 1st row, the Base result exhibits target edge diffusion due to background interference. Adding DRF makes the predicted contours more compact and well defined, suggesting that DRF suppresses redundant background responses. In the 2nd row with a low contrast linear target, the Base result shows discontinuities. Schemes with PCE preserve the target shape more completely, indicating PCE's ability to enhance low contrast features. In the 3rd row, schemes with a single module lose details. Combining DRF and PCE preserves the target shape much better, showing their synergy in suppressing interference and enhancing features for more accurate segmentation.

 doi:10.20944/preprints202603.2230.v1

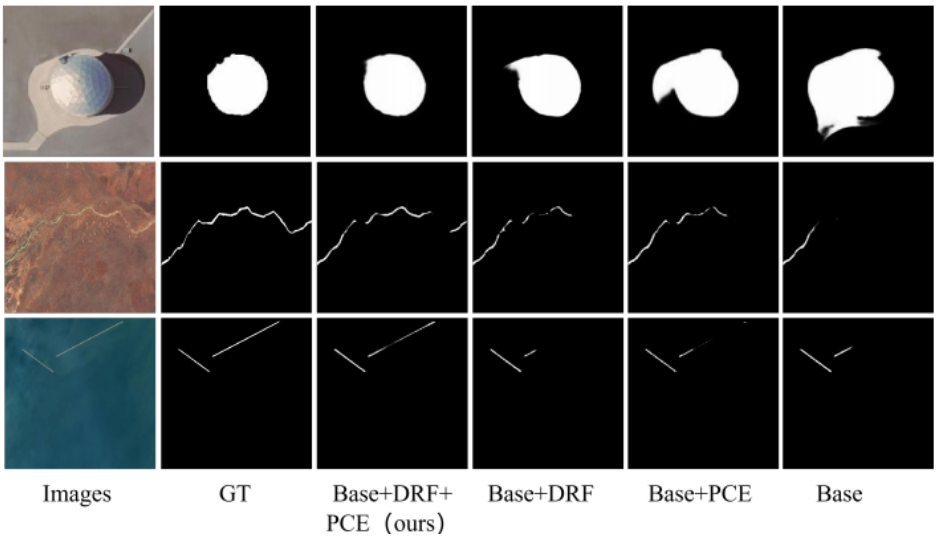

Figure 6. Ablation study between different modules.

To verify the enhancement effect of PCE on features extracted by Swin Transformer Stage 1, we generate feature heatmaps with a Tkinter based feature map viewer, as shown in Figure 7. The differences between the original Stage 1 heatmaps and the enhanced ones are clear, reflecting improved target details and saliency. In Row 1, the original Stage 1 features show small brightness differences between target and background and blurry contours. After enhancement, the target region becomes brighter and more distinguishable. In Row 2, the original features contain only sparse dot like highlights. After enhancement, the target region becomes more uniform and its shape is more complete and better separated from the background. In Row 3, the original edge responses are discontinuous. After enhancement, continuous highlight bands appear along the target edges. These observations indicate that PCE improves target–background separability in low contrast scenes and strengthens edge integrity, providing a stronger detail basis for subsequent cross-scale fusion.

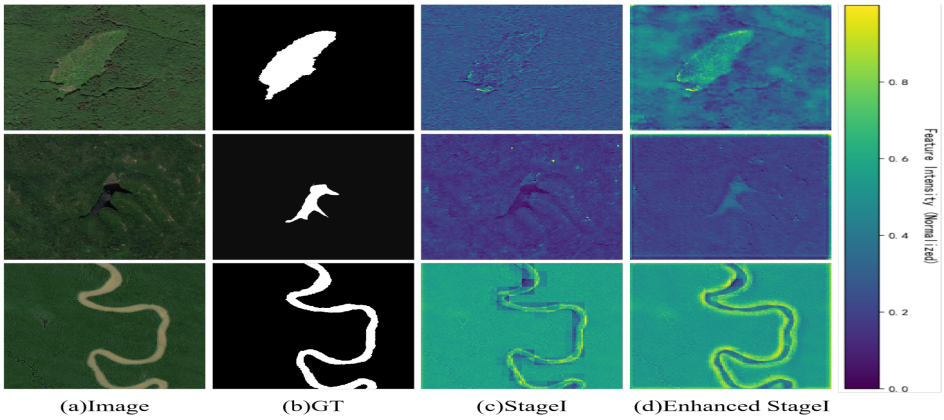

Figure 7. Heatmap of the enhancement effect of the PCE module on the Stage 1 features of the Swin Transformer.

2) Ablation study within the modules: To verify the effectiveness of the core components in the PCE and DRF modules, we designed ablation variants with key components removed. And we analyzed their performance on the ORSSD and EORSSD datasets.

For the PCE module, we designed two variants: w/o PC calculation (we removed the Filter, Phase Congruency Calculation, Stripe and only retained the VSA) and w/o VSA (PCE without VSA). As shown in Table 3(a), either removing the VSA or only having VSA leads to performance degradation. As shown in Figure 8(a), in the 1st row, removing these two parts results in the loss of texture details inside the target. In the 2nd row, "ours" accurately reproduces the target shape, while "w/o VSA" suffers from detail loss and background interference. This indicates that the phase congruency

 doi:10.20944/preprints202603.2230.v1

calculation enhances target edges and textures while suppressing background noise, and VSA focuses on the target region to strengthen effective feature responses. The two parts work synergistically, serving as the core of the PCE module for low-contrast feature enhancement, and are indispensable for performance improvement.

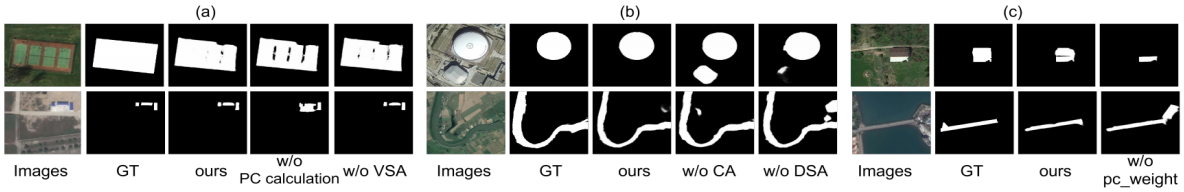

Figure 8. Visualization of ablation study results within modules: (a) ablation experiments inside the PCE module, (b) ablation experiments inside the DRF module, and (c) ablation experiments on the loss function.

Table 3. Ablation study within the modules on ORSSD and EORSSD datasets.

Model variants	ORSSD			EORSSD		
	$S_\alpha \uparrow$	$F_\beta \uparrow$	$E_\xi \uparrow$	$S_\alpha \uparrow$	$F_\beta \uparrow$	$E_\xi \uparrow$
(a) Ablation study in PCE						
ours	**0.9540**	**0.9305**	**0.9888**	**0.9393**	**0.8943**	**0.9843**
w/o PC calculation	0.9491	0.9239	0.9834	0.9328	0.8875	0.9788
w/o VSA	0.9491	0.9246	0.9840	0.9381	0.8917	0.9819
(b) Ablation study in DRF						
ours	**0.9540**	**0.9305**	**0.9888**	**0.9393**	**0.8943**	**0.9843**
w/o CA	0.9489	0.9294	0.9806	0.9330	0.8866	0.9797
w/o DSA	0.9509	0.9262	0.9874	0.9364	0.8892	0.9783

For the DRF module, we designed two variants: w/o CA (DRF without CA) and w/o DSA (DRF without DSA). The quantitative results (Table 3(b)) indicate that the performance decreases after removing CA or DSA. For dome shaped building targets in Figure 8(b), the saliency maps of w/o CA contain background clutter, and the target contours of w/o DSA are distorted. The river scene is similar in situation. The complete DRF module generates saliency maps with clear and intact shapes, verifying that the synergy between CA and DSA effectively resolves background challenges.

In this paper, we select 5 stacked RABs for the DRF module, which is the optimal choice based on ablation experiments. Table 4 shows that with the number of RABs from 3 to 5 improves all three core metrics (S_α, F_β, E_ξ) on the ORSSD dataset. This suggests that more RABs strengthen feature fusion and redundancy filtering. When the number of RABs exceeds 5, parameter redundancy causes gradient attenuation, and all metrics show a decline. Therefore, the stacked structure of 5 RABs enables the model to achieve the best performance.

Table 4. Ablation study on the number of RAB.

Model variants	$S_\alpha \uparrow$	$F_\beta \uparrow$	$E_\xi \uparrow$
3 RAB	0.9457	0.9196	0.9824
4 RAB	0.9488	0.9219	0.9828
5 RAB	**0.9540**	**0.9305**	**0.9888**
6 RAB	0.9509	0.9276	0.9869
7 RAB	0.9531	0.9274	0.9863

3) The impact of our pc_weight on the loss function: This section verifies the effectiveness of pc_weight. In the original design, saliency map S_0 and S_1 use pc-weight combination of CE Loss

 doi:10.20944/preprints202603.2230.v1

and IOU Loss, while S_2–S_5 adopt ordinary CE and IOU Loss. Removing pc_weight in this ablation experiment means that S_0 and S_1 also use only ordinary CE and IOU Loss (consistent with S_2–S_5). The experimental results are shown in Table 5 and the Figure 8(c). From Table 5, it can be observed that after removing pc_weight, only the basic losses of CE Loss and IOU Loss remain. S_α of ORSSD decreases from 0.9540 to 0.9484, and F_β drops from 0.9305 to 0.9244. This indicates that pc_weight effectively improves the model performance. The visualization of the ablation experiment for the proposed pc_weight is presented in Figure 8(c). For the low-contrast building targets, the prediction map of *w/o* pc_weight loses the prominent structures of the buildings, which demonstrates that pc_weight strengthens the detail supervision of low-contrast targets. For bridge targets, the prediction map of *w/o* pc_weight exhibits misinformation, proving that pc_weight constrains the structural integrity of the targets.

Table 5. The impact of our pc_weight on the loss function.

No.	Baseline	pc_weight	ORSSD			EORSSD		
			$S_\alpha \uparrow$	$F_\beta \uparrow$	$E_\xi \uparrow$	$S_\alpha \uparrow$	$F_\beta \uparrow$	$E_\xi \uparrow$
1	✓		0.9484	0.9244	0.9844	0.9323	0.8854	0.9641
2	✓	✓	**0.9540**	**0.9305**	**0.9888**	**0.9393**	**0.8943**	**0.9843**

4.4. Computational Complexity Analysis

1) Complexity Changes from Module Superposition The model in this paper uses TF and SGAED as the Baseline. After adding the PCE module, FLOPs increase by 6.67G, and Params increase by 0.53M. The PCE module only introduces lightweight frequency-domain filter operations, resulting in minimal complexity overhead. After adding the DRF module: FLOPs increase by 21G, and Params increase by 13M. The multi-group RAB attention blocks in DRF are the main source of complexity, but they ensure the effectiveness of cross-scale fusion.

2) Complexity Comparison with Existing Methods Our proposed model (ours) has 126.94G FLOPs and 117.29M parameters. For high-performance methods such as EMFINet (176.87G) and ACCoNet (184.50G), our model achieves better performance while reducing the complexity by around 30%. For HFCNet, our model achieves better performance with fewer parameters and comparable FLOPs. For models with lower complexity (e.g., AESINet), although our model has slightly higher complexity, its performance is significantly better (see Section 4.2 for details).

 doi:10.20944/preprints202603.2230.v1

Table 6. Model Computational Complexity Comparison: (a) complexity analysis of the proposed modules, (b) comparison of complexity with some state-of-the-art methods.

(a)		
Models	FLOPs	Params
TF	71.44 G	86.64 M
TF+SGAED	99.27 G	103.76 M
TF+SGAED+PCE	105.94 G (↑6.67)	104.29 M (↑0.53)
TF+SGAED+PCE+DRF(Ours)	126.94 G (↑21)	117.29 M (↑13)
(b)		
Models	FLOPs	Params
MCCNet	117.15 G	67.65 M
EMFINet	176.87 G	95.09 M
ERPNet	131.63 G	77.19 M
ACCQNet	184.50 G	102.55 M
AESINet	53.42 G	41.05 M
ASTTNet	43.12 G	23.35 M
ADSTNet	62.09 G	27.72 M
HFCNet	120.41 G	140.75 M
ours	126.94 G	117.29 M

5. Conclusion

In this paper, we propose a novel PCFNet to tackle the complex target detection challenge in ORSI-SOD tasks. We embed PCE and DRF into the Swin Transformer backbone to build a unified pipeline that combines frequency domain detail enhancement, cross scale semantic fusion, and refined supervision. Specifically, PCE extracts frequency-domain phase features, which breaks the limitation of traditional spatial enhancement that relies on brightness differences. It improves the detail representation capability of targets in low-contrast scenes. The dual attention block (RAB) designed in DRF fuses channel and dynamic spatial attention to resolve problems of complex background interference. Extensive experiments on 3 datasets validate the superior performance of PCFNet.

Data Availability Statement: The data will be made publicly available upon acceptance of the paper.

Conflicts of Interest: The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work presented in jithis manuscript.

References

1. A. Borji, M. Cheng, H. Jiang and J. Li, "Salient Object Detection: A Benchmark," *IEEE Transactions on Image Processing*, vol. 24, no. 12, pp. 5706–5722, Dec. 2015.
2. W. Wang, Q. Lai, H. Fu, J. Shen, H. Ling and R. Yang, "Salient Object Detection in the Deep Learning Era: An In-Depth Survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 6, pp. 3239–3259, June 2022.
3. P. Zhang, T. Zhuo, W. Huang, K. Chen, and M. Kankanhalli, "Online object tracking based on CNN with spatial–temporal saliency guided sampling," *Neurocomputing*, vol. 257, pp. 115–127, Sep. 2017.
4. L. Gao, B. Liu, P. Fu, M. Xu and J. Li, "Visual Tracking via Dynamic Saliency Discriminative Correlation Filter," *Applied Intelligence*, vol. 52, no. 6, pp. 5897–5911, 2022.
5. X. Song, H. Lin, H. Wen, B. Hou, M. Xu and L. Nie, "A Comprehensive Survey on Composed Image Retrieval," *ACM Trans. Inf. Syst.*, vol. 44, no. 1, art. no. 19, pp. 1–54, 2025.
6. C. Li, C. Guo, W. Ren, R. Cong, J. Hou and S. Kwong, "An Underwater Image Enhancement Benchmark Dataset and Beyond," *IEEE Transactions on Image Processing*, vol. 29, pp. 4376–4389, 2020.
7. L. Yang, J. Wu, H. Li, C. Liu and S. Wei, "Real-Time Runway Detection Using Dual-Modal Fusion of Visible and Infrared Data," *Remote Sensing*, vol. 17, no. 4, p. 669, 2025.

 doi:10.20944/preprints202603.2230.v1

8. J. Lei, H. Wang, Z. Lei, J. Li and S. Rong, "CNN–Transformer Hybrid Architecture for Underwater Sonar Image Segmentation," *Remote Sensing*, vol. 17, no. 4, p. 707, 2025.
9. K. He, X. Zhang, S. Ren and J. Sun, "Deep Residual Learning for Image Recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, 2016.
10. B. Zhou, A. Khosla, A. Lapedriza, A. Oliva and A. Torralba, "Learning Deep Features for Discriminative Localization," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2921–2929, 2016.
11. J. Chen, H. Zhang, M. Gong and Z. Gao, "Collaborative Compensative Transformer Network for Salient Object Detection," *Pattern Recognition*, vol. 154, art. no. 110600, Oct. 2024.
12. R. Azad, A. Kazerouni, B. Azad, E. Khodapanah Aghdam, Y. Velichko, U. Bagci and D. Merhof, "Laplacian-Former: Overcoming the Limitations of Vision Transformers in Local Texture Detection," in *Medical Image Computing and Computer Assisted Intervention (MICCAI)*, LNCS, vol. 14222, pp. 736–746, Oct. 2023.
13. X. Wang, L. Wan, D. Lin and W. Feng, "Phase-based fine-grained change detection," *Expert Systems with Applications*, vol. 227, pp. 120181, 2023.
14. F. Perazzi, P. Krähenbühl, Y. Pritch and A. Hornung, "Saliency filters: Contrast based filtering for salient region detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 733–740, 2012.
15. Q. Zhang, S. Wang, X. Wang, Z. Sun, S. Kwong and J. Jiang, "Geometry Auxiliary Salient Object Detection for Light Fields via Graph Neural Networks," *IEEE Transactions on Image Processing*, vol. 30, pp. 7578–7592, 2021.
16. M. Xu, Z. Sun, Y. Hu, H. Tang, Y. Hu, X. Song and L. Nie, "Superpixel Segmentation With Edge Guided Local-Global Attention Network," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 35, no. 12, pp. 11922–11934, 2025.
17. X. Yuan, B. Zhang, J. Zhou, C. Lian, Q. Zhang and J. Yue, "Gradient residual attention network for infrared image super-resolution," *Optics and Lasers in Engineering*, vol. 175, pp. 107998, 2024.
18. F. Wang, M. Jiang, C. Qian, S. Yang, "Residual Attention Network for Image Classification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6450–6458, 2017.
19. J. Bi, H. Wei, G. Zhang, K. Yang and Z. Song, "DyFusion: Cross-Attention 3D Object Detection with Dynamic Fusion," *IEEE Latin America Transactions*, vol. 22, no. 2, pp. 106–112, Feb. 2024.
20. R. Achanta, S. Hemami, F. Estrada and S. Susstrunk, "Frequency-tuned salient region detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1597–1604, 2009.
21. W. Luo, Y. Li, R. Urtasun and R. S. Zemel, "Understanding the Effective Receptive Field in Deep Convolutional Neural Networks," *Advances in Neural Information Processing Systems (NIPS)*, arXiv:1701.04128, 2016.
22. N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov and S. Zagoruyko, "End-to-End Object Detection with Transformers," in *Computer Vision*, Cham: Springer, vol. 12346, pp. 13, 2020.
23. Z. Liu, Y. Lin, Y. Cao, H. Hu, "Swin Transformer: Hierarchical Vision Transformer using Shifted Windows," in *IEEE International Conference on Computer Vision (ICCV)*, pp. 9992–10002, 2021.
24. C. Xie, C. Xia, M. Ma, Z. Zhao, X. Chen and J. Li, "Pyramid Grafting Network for One-Stage High Resolution Saliency Detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11707–11716, 2022.
25. K. Li, G. Wan, G. Cheng, L. Meng and J. Han, "Object detection in optical remote sensing images: A survey and a new benchmark," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 159, pp. 296–307, 2020.
26. O. Ronneberger, P. Fischer and T. Brox, "U-Net: Convolutional Networks for Biomedical Image Segmentation," in *Medical Image Computing and Computer-Assisted Intervention*, vol. 9351, pp. 28, 2015.
27. X. Zeng, M. Xu, Y. Hu, H. Tang, Y. Hu and L. Nie, "Adaptive Edge-Aware Semantic Interaction Network for Salient Object Detection in Optical Remote Sensing Images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 61, pp. 1–16, 2023.
28. L. Gao, B. Liu, P. Fu and M. Xu, "Adaptive Spatial Tokenization Transformer for Salient Object Detection in Optical Remote Sensing Images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 61, pp. 1–15, 2023.
29. B. Cheng, Z. Liu, H. Tang, Q. Wang, "Multimodal-Guided Transformer Architecture for Remote Sensing Salient Object Detection," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 22, pp. 1–5, 2025.
30. J. Li, C. Li, X. Zheng, X. Liu and C. Tang, "Global Context Relation-Guided Feature Aggregation Network for Salient Object Detection in Optical Remote Sensing Images," *Remote Sensing*, vol. 16, no. 16, p. 2978, 2024.

 doi:10.20944/preprints202603.2230.v1

31. Y. Liu, M. Xu, T. Xiao, H. Tang, Y. Hu and L. Nie, "Heterogeneous Feature Collaboration Network for Salient Object Detection in Optical Remote Sensing Images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 62, pp. 1–14, 2024.
32. F. Gao, M. Fu, J. Cao, J. Dong and Q. Du, "Adaptive Frequency Enhancement Network for Remote Sensing Image Semantic Segmentation," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 63, pp. 1–15, 2025.
33. M. Xu, C. Yu, Z. Li, H. Tang, Y. Hu and L. Nie, "HDNet: A Hybrid Domain Network With Multiscale High-Frequency Information Enhancement for Infrared Small-Target Detection," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 63, pp. 1–15, 2025.
34. P. Xiao, X. Feng, S. Zhao and J. She, "Segmentation of High-resolution Remotely Sensed Imagery Based on Phase Congruency," *ACTA GEODAETICA et CARTOGRAPHICA SINICA*, vol. 36, no. 2, pp. 146–151, May 2007.
35. Z. Zhang and M. R. Sabuncu, "Generalized cross entropy loss for training deep neural networks with noisy labels," in *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, pp. 8792–8802, 2018.
36. J. Yu, Y. Jiang, Z. Wang, Z. Cao, T. Huang, "Unitbox: An advanced object detection network," in *ACM Multimedia*, pp. 516–520, 2016.
37. C. Li, R. Cong, J. Hou, S. Zhang, Y. Qian and S. Kwong, "Nested Network With Two-Stream Pyramid for Salient Object Detection in Optical Remote Sensing Images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 57, no. 11, pp. 9156–9166, Nov. 2019.
38. Q. Zhang, R. Cong, C. Li, M. Cheng, Y. Fang, X. Cao, Y. Zhao and S. Kwong, "Dense Attention Fluid Network for Salient Object Detection in Optical Remote Sensing Images," *IEEE Transactions on Image Processing*, vol. 30, pp. 1305–1317, 2021.
39. Z. Tu, C. Wang, C. Li, M. Fan, H. Zhao, B. Luo, "ORSI Salient Object Detection via Multiscale Joint Region and Boundary Model," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–13, 2022.
40. D.-P. Fan, M.-M. Cheng, Y. Liu, T. Li and A. Borji, "Structure-Measure: A New Way to Evaluate Foreground Maps," in *IEEE International Conference on Computer Vision (ICCV)*, pp. 4558–4567, 2017.
41. D.-P. Fan, C. Gong, Y. Cao, B. Ren, M.-M. Cheng, A. Borji, "Enhanced-alignment measure for binary foreground map evaluation," in *Proc. Int. Joint Conf. Artif. Intell.*, pp. 698–704, 2018.
42. J.-G. Yu, J. Zhao, J. Tian, and Y. Tan, "Maximal entropy random walk for region-based visual saliency," *IEEE Transactions on Cybernetics*, vol. 44, no. 9, pp. 1661–1672, Sep. 2014.
43. Y. Yuan, C. Li, J. Kim, W. Cai, and D. D. Feng, "Reversion correction and regularized random walk ranking for saliency detection," *IEEE Transactions on Image Processing*, vol. 27, no. 3, pp. 1311–1322, Mar. 2018.
44. X. Zhou et al., "Edge-guided recurrent positioning network for salient object detection in optical remote sensing images," *IEEE Transactions on Cybernetics*, vol. 53, no. 1, pp. 539–552, Jan. 2023.
45. Y. Pang, X. Zhao, L. Zhang, and H. Lu, "Multi-scale interactive network for salient object detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9413–9422, 2020.
46. X. Zhao, Y. Pang, L. Zhang, H. Lu, and L. Zhang, "Suppress and balance: A simple gated network for salient object detection," in *Proc. Eur. Conf. Comput. Vis.*, pp. 35–51, 2020.
47. G. Li, Z. Liu, W. Lin, and H. Ling, "Multi-content complementation network for salient object detection in optical remote sensing images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, Art. no. 5614513, 2022.
48. G. Li, Z. Liu, Z. Bai, W. Lin, and H. Ling, "Lightweight salient object detection in optical remote sensing images via feature correlation," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–12, 2022.
49. Z. Tu, C. Wang, C. Li, M. Fan, H. Zhao, and B. Luo, "ORSI salient object detection via multiscale joint region and boundary model," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, Art. no. 5607913, 2022.
50. X. Zhou, K. Shen, Z. Liu, C. Gong, J. Zhang, and C. Yan, "Edge-aware multiscale feature integration network for salient object detection in optical remote sensing images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–15, 2022.
51. G. Li, Z. Liu, D. Zeng, W. Lin, and H. Ling, "Adjacent context coordination network for salient object detection in optical remote sensing images," *IEEE Transactions on Cybernetics*, vol. 53, no. 1, pp. 526–538, Jan. 2023.
52. J. Huang and K. Huang, "Dynamic Context Coordination for Salient Object Detection in Optical Remote Sensing Images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 22, pp. 1–5, 2025.

 doi:10.20944/preprints202603.2230.v1

53. S. Lee, S. Cho, C. Park, S. Park, J. Kim, S. Lee, "LSHNet: Leveraging Structure-Prior With Hierarchical Features Updates for Salient Object Detection in Optical Remote Sensing Images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 62, pp. 1–16, 2024.
54. K. Huang, N. Li, J. Huang and C. Tian, "Exploiting Memory-Based Cross-Image Contexts for Salient Object Detection in Optical Remote Sensing Images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 62, pp. 1–15, 2024.
55. Q. Wang, Y. Liu, Z. Xiong, and Y. Yuan, "Hybrid feature aligned network for salient object detection in optical remote sensing imagery," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, Art. no. 5624915, 2022.
56. J. Zhao, Y. Jia, L. Ma, and L. Yu, "Adaptive dual-stream sparse transformer network for salient object detection in optical remote sensing images," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 17, pp. 5173–5192, 2024.
57. M. Xu, S. Wang, Y. Hu, H. Tang, R. Cong and L. Nie, "Cross-Model Nested Fusion Network for Salient Object Detection in Optical Remote Sensing Images," *IEEE Transactions on Cybernetics*, vol. 55, no. 11, pp. 5332–5345, 2025.