

Integrating Vision Liquid Time-Constant and Wavelet for RGB-D Salient Object Detection

First Bei Cheng

Kunming University of Science and Technology

Second Danyang Yang

Kunming University of Science and Technology

Third Zhihong Zeng

azh@gpnu.edu.cn

Guangdong Polytechnic Normal University

Fourth Xu Lu

Guangdong Polytechnic Normal University

Fifth Haijun Liu

Chongqing University

Research Article

Keywords: RGB-D Salient Object Detection, State Space Model, single-backbone, Liquid Neural Network, Cross-modal Fusion, Wavelet transform

Posted Date: October 27th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-10809276/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: No competing interests reported.

Integrating Vision Liquid Time-Constant and Wavelet for RGB-D Salient Object Detection

First Bei Cheng¹, Second Danyang Yang^{1†},
Third Zhihong Zeng^{2*}, Fourth Xu Lu², Fifth Haijun Liu³

¹Faculty of Information Engineering and Automation, Kunming
University of Science and Technology, Kunming, 650500, China.

^{2*}Institute of Interdisciplinary Studies, Guangdong Polytechnic Normal
University, Guangzhou, 510665, China.

³School of Microelectronics and Communication Engineering,
Chongqing University, Chongqing, 400044, China.

*Corresponding author(s). E-mail(s): azh@gpnu.edu.cn;

Contributing authors: chengbei@kust.edu.cn;

yangdanyang@stu.kust.edu.cn; bruda@126.com; haijun_liu@cqu.edu.cn;

[†]These authors contributed equally to this work.

Abstract

RGB-D salient object detection (SOD) typically employs Convolutional Neural Networks (CNNs) or Transformers as backbones. However, CNNs suffer from limited receptive fields, while Transformers incur prohibitive computational costs. The emerging Mamba architecture addresses these limitations by enabling global contextual modeling with linear complexity, yet introduces new challenges: structural disruption from selective scanning causing local feature loss. To overcome this deficiency, we propose Liquid-Wavelet Network (LWNet), a novel framework integrating Vision Liquid Time-Constant (Vi-LTC) and Cross-Modal Wavelet Fusion (CMWF) module. First, we propose the Vi-LTC module, which are dynamic neural networks with input-dependent time constants for flexible temporal modeling. It leverages neuron-like adaptive dynamics to preserve structural integrity while focusing on salient regions, thereby enhancing the retention of local structure-aware features. Second, we propose the CMWF that employs multi-resolution decomposition to disentangle features into low- and high-frequency components, enabling frequency-aware cross-modal complementarity. Extensive experiments on five RGB-D SOD benchmarks demonstrate LWNet’s superiority over 18 cutting-edge methods. Additionally, it exhibits exceptional generalization on two remote sensing datasets, achieving competitive

performance against nine specialized models. This work bridges the gap between natural and remote sensing image analysis.

Keywords: RGB-D Salient Object Detection, State Space Model, single-backbone, Liquid Neural Network, Cross-modal Fusion, Wavelet transform.

1 Introduction

Salient object detection (SOD), as one of the core tasks in computer vision, aims to mimic the human visual system’s ability to perceive the most prominent regions or objects in an image, thereby rapidly localizing and segmenting the most attention-grabbing targets in a scene. With the exponential growth of digital image data, SOD has demonstrated broad value in multimedia processing field such as image segmentation[1, 2], object tracking[3, 4], visual compression[5, 6], and content-aware editing[7, 8]. However, as highlighted in prior studies [9, 10], conventional RGB-based SOD methods often struggle to achieve comprehensive segmentation of salient objects under challenging conditions such as varying illumination and low contrast. To address these limitations, depth maps have been introduced as supplementary inputs to enhance the understanding of spatial cues in images[9–11].

To date, numerous RGB-D feature fusion methods have been proposed for SOD. Cheng et al.[12] presented MobileSal, an efficient RGB-D SOD framework that leverages lightweight mobile backbones to achieve competitive accuracy with reduced computational cost. Zeng et al.[13] developed a lightweight RGB-D SOD model employing parallel attention shift convolution modules to capture long-range dependencies and local information for enhanced representation learning. CATNet[14] introduced a cascaded Transformer with adaptive enhancement and multi-head self-attention to refine RGB and depth features. This design improves global context modeling and multi-scale aggregation over CNN-based methods.

In recent years, Mamba [15] has demonstrated remarkable advantages in image processing[15–17], primarily due to its unique combination of efficient long-range dependency modeling and linear computational complexity. Traditional CNNs are limited by small receptive fields, while Transformers suffer from quadratic computational complexity. In contrast, Mamba employs trainable, time-varying parameterized state space equations to achieve efficient and adaptive sequence modeling. It achieves input-adaptive long-range interactions at linear computational complexity while preserving sequence modeling capabilities.

Motivation: Current Mamba-based architectures suffer from two core limitations in RGB-D SOD field. *First*, spatial continuity disruption occurs due to the flattening of 2D feature maps into 1D sequences for state space modeling (SSM). This operation severs inherent spatial correlations, causing adjacent pixels in the original image to be distributed non-contiguously across the sequence. Consequently, the causality-dependent structure of SSM is impaired, hindering accurate sequential prediction and degrading spatial fidelity in the output. *Second*, high-frequency feature degradation

arises from the inherent smoothing effect of SSMs. In Mamba, global modeling emphasizes contextual consistency by strengthening long-range dependencies, but this comes at the cost of blur local details such as object structures and edges. As a result, salient structural information and object boundaries encoded in high-frequency signals are weakened. This attenuation propagates through the decoder, leading to suboptimal localization and coarse segmentation results.

Therefore, we propose a novel framework integrating wavelet fusion and Liquid Time-Constant Networks (LTCs) [18] for RGB-D SOD. Specifically, to address spatial fragmentation, we propose the Vision Liquid Time-Constant (Vi-LTC) module. By leveraging 2D convolutional dynamics and an explicit differential solver, Vi-LTC models feature extraction as a continuous state evolution governed by input-adaptive time constants. This enables per-pixel adaptation to local complexity, effectively compensating for the local details lost by the Mamba backbone. Furthermore, to mitigate high-frequency loss, we introduce the Cross-Modal Wavelet Fusion (CMWF) module. It comprises a Wavelet Transform (WT) branch for edge enhancement via frequency decomposition, and an SSM branch for global semantic modeling. The synergy of these branches yields representations that preserve structural fidelity while capturing global context.

In summary, our key contributions are outlined below:

- We propose a novel Liquid-Wavelet Network, named LWNet, for RGB-D SOD, which effectively integrates Vi-LTC and wavelet fusion to achieve accurate detection. To the best of our knowledge, this is a pioneer work that introduces Vi-LTC into the field of SOD.
- We propose the Vi-LTC module, which first extends Liquid Time-Constant networks from 1D sequential domains to 2D visual modeling. This module adaptively adjusts the importance of different regions through dynamic modeling and gating mechanisms, enabling the network to better attend to salient regions.
- We propose the CMWF module, which leverages WT to decompose RGB and depth features into multiple frequency subbands and performs cross-modal fusion at both low- and high-frequency levels. Simultaneously, it enhances modality-specific representations through SSM, effectively achieving complementary feature integration across modalities.
- LWNet is compared with 18 existing methods on five natural image datasets and demonstrated favorable performance. It also shows competitive results on two remote sensing benchmarks, highlighting its robustness and generalization ability.

2 Related Work

2.1 Cross-modal feature fusion in RGB-D SOD

Cross-modal feature fusion is a key component in RGB-D salient object detection, as it aims to effectively exploit the complementary properties of RGB appearance and depth geometry. DPPNet [19] performs pixel-wise RGB-D fusion via static spatial potential modeling to enhance local saliency perception. CATNet [14] adopts a cascaded Transformer architecture to progressively aggregate RGB and depth features for long-range

dependency modeling. Fusion-Mamba [20] leverages Mamba-based state space models to achieve efficient cross-modal feature interaction with linear complexity. Unlike previous approaches, CMWF adopts a split-and-conquer fusion strategy: one branch utilizes the Wavelet Transform (WT) to decompose RGB-D features into frequency components for structurally consistent and frequency-aware fusion, while the other branch employs state space models (SSMs) to dynamically enhance modality-specific representations. By decoupling frequency-aware structural alignment from dynamic contextual modeling, CMWF effectively reduces redundant cross-modal interactions and achieves more precise salient object perception.

2.2 Wavelet transform in Computer Vision

Since the late 20th century, the wavelet transform (WT) [21] has been a powerful tool for signal processing, particularly for non-stationary signals like transients, localized features, and image edges. Unlike the Fourier transform, WT uses localized wavelet basis functions, enabling effective multi-scale feature extraction. In object detection, Huang et al. [22] integrated WT with CNNs, enhancing fire detection by combining multi-scale WT features with ResNet feature maps. Similar approaches have also been explored in recent work [23], demonstrating that wavelet-based downsampling can effectively preserve edge and structural information while maintaining efficiency. Yao et al. [24] used WT’s lossless downsampling property in combination with a Multi-scale Vision Transformer [25] to achieve reversible downsampling, while enhancing self-attention outputs with IWT to improve feature representation. Finder et al. [26] employed wavelet downsampling for larger receptive fields, reducing computational costs compared to conventional methods by scaling trainable parameters logarithmically with the receptive field size.

2.3 Liquid Time-Constant Networks

Neural ODEs [27] define hidden state evolution via continuous differential equations $\frac{dh(t)}{dt} = f(h(t), x(t), t)$. Building upon this, Liquid Time-Constant Networks (LTCs) [18] enhance stability through input-dependent time constants and bounded dynamics:

$$\frac{dh(t)}{dt} = - \left[\frac{1}{\tau} + f(x, h) \right] \cdot h(t) + f(x, h) \cdot A(x). \quad (1)$$

Here, τ regulates the "inertia" of the state (decay rate), while $f(\cdot)$ and $A(x)$ model the synaptic coupling and external drive, respectively. This mechanism allows the system to dynamically adapt its timescale to external stimuli. Despite their theoretical advantages, original LTCs are limited to vector inputs. We propose the Vision Liquid Time-Constant (Vi-LTC) module to extend this continuous dynamics to the spatial domain. By reformulating the ODE solvers with convolutional operators, Vi-LTC enables spatially adaptive feature evolution suitable for dense prediction tasks like SOD.

3 Methodology

3.1 Preliminaries

Wavelet Transform: Given an input image $\mathbf{I} \in \mathbb{R}^{C \times H \times W}$, we employ the 2D discrete wavelet transform (DWT) with a haar wavelet to decompose the image into one low-frequency subband and three high-frequency subbands:

$$\{\mathbf{I}_{LL}, \mathbf{I}_{LH}, \mathbf{I}_{HL}, \mathbf{I}_{HH}\} = \text{2D-DWT}(\mathbf{I}), \quad (2)$$

where $\mathbf{I}_{LL} \in \mathbb{R}^{C \times \frac{H}{2} \times \frac{W}{2}}$ denote the low-frequency component of image $\mathbf{I}$, and $\mathbf{I}_{LH}, \mathbf{I}_{HL}, \mathbf{I}_{HH} \in \mathbb{R}^{C \times \frac{H}{2} \times \frac{W}{2}}$ denote the high-frequency component. The low-frequency component $\mathbf{I}_{LL}$ retains the majority of the information in image $\mathbf{I}$. Specifically, the $\mathbf{I}_{LH}$ subband captures horizontal gradient variations and is sensitive to vertical edges. The $\mathbf{I}_{HL}$ subband captures vertical gradient variations and is sensitive to horizontal edges, while the $\mathbf{I}_{HH}$ subband is sensitive to diagonal edges. The original image $\mathbf{I}$ can be reconstructed through the inverse 2D discrete wavelet transform (2D-IDWT) using the four subbands ($\mathbf{I}_{LL}, \mathbf{I}_{LH}, \mathbf{I}_{HL}, \mathbf{I}_{HH}$) as follows:

$$\mathbf{I} = \text{2D-IDWT}(\mathbf{I}_{LL}, \mathbf{I}_{LH}, \mathbf{I}_{HL}, \mathbf{I}_{HH}). \quad (3)$$

3.2 Network Architecture

As illustrated in Fig. 1, the proposed framework consists of a Mamba-based backbone [16], the Vi-LTC module, the CMWF module, and a cascaded decoder. A single-stream backbone first extracts multi-level features from RGB and depth inputs. Subsequently, Vi-LTC reconstructs local details lost during global scanning. The CMWF then fuses cross-modal features by integrating global semantics and frequency-aware details. Finally, a multi-scale decoder generates the saliency maps.

Encoder: We employ a pre-trained VMamba [16] as a single-stream backbone to extract hierarchical features from RGB and depth images. This leverages the linear complexity of SSMs for efficient global context modeling compared to traditional CNNs or Transformers.

Vision Liquid Time-Constant Module: To mitigate local feature degradation caused by selective scanning, the Vi-LTC module processes the hierarchical RGB and depth features $\mathbf{R}_i$ and $\mathbf{D}_i$ ($i \in 1, 2, 3, 4, 5$) extracted by the backbone. By introducing input-dependent dynamics and continuous state evolution, Vi-LTC restores spatial continuity and enhances structural details. The process is formulated as $\bar{\mathbf{R}}_i = \text{Vi-LTC}(\mathbf{R}_i)$, $\bar{\mathbf{D}}_i = \text{Vi-LTC}(\mathbf{D}_i)$.

Cross-modal Wavelet Fusion: The CMWF module fuses $\bar{\mathbf{R}}_i$ and $\bar{\mathbf{D}}_i$ by exploiting complementary cues in both spatial and frequency domains. It operates via two parallel branches: an SSM branch that enhances global semantics ($\mathbf{F}_i^m$), and a Wavelet Transform (WT) branch that preserves high-frequency details via IWT reconstruction ($\mathbf{F}_i^w$). The final fused feature $\mathbf{F}_i^u$ combines these components to ensure comprehensive feature interaction.

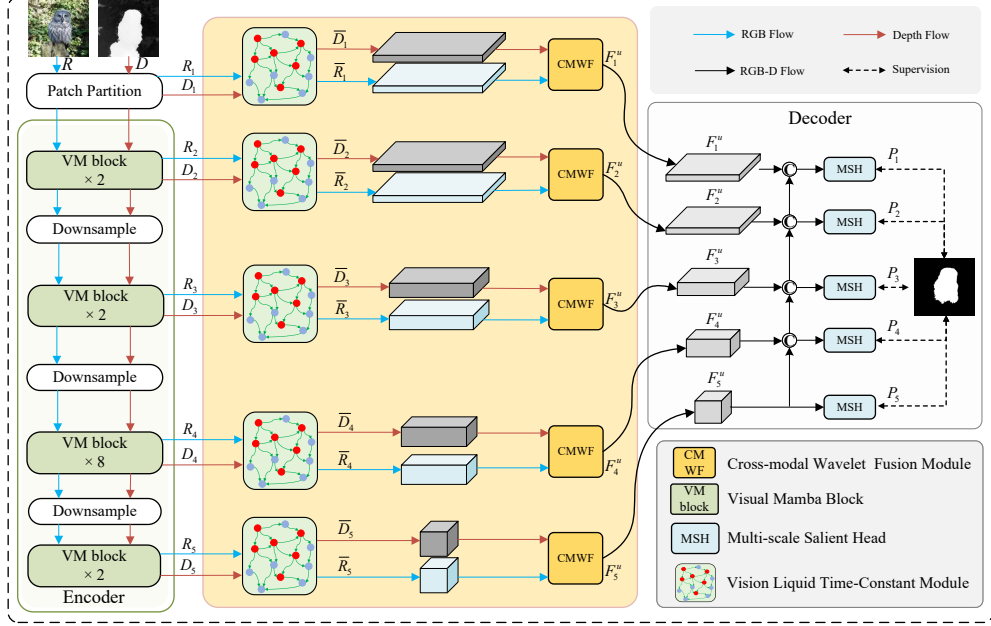


Fig. 1: The pipeline of our LWNNet. It includes the mamba-based encoder, the Vision Liquid Time-Constant module, the cross-modal wavelet fusion module, and multi-scale salient head.

Decoder: Following [28], we adopt a cascaded multi-scale decoder to generate predictions P_i from fused features. Deep supervision is applied at each scale to facilitate gradient flow and ensure the effective aggregation of high-level semantics and low-level details.

3.3 Visual Mamba Backbone

We employ the pre-trained VMamba [16] as the backbone to extract hierarchical features. Given the input RGB image R and depth map D , the backbone processes them through patch embedding and four stages of cascaded Visual State Space (VSS) blocks, generating multi-scale feature maps with strides of $\{4, 8, 16, 32\}$. The core component, the VSS block, leverages the 2D Selective Scan (SS2D) module to model global contextual dependencies with linear computational complexity, combined with a gating branch to modulate information flow. This architecture enables efficient long-range dependency modeling while avoiding the quadratic cost of Transformers.

3.4 Vision Liquid Time-Constant Network

To mitigate the loss of local structural details caused by global scanning mechanisms (e.g., SS2D) and to introduce dynamic context adaptability, we propose the Vi-LTC module. While traditional CNNs and Transformers treat feature extraction as a static mapping, Vi-LTC views it as a

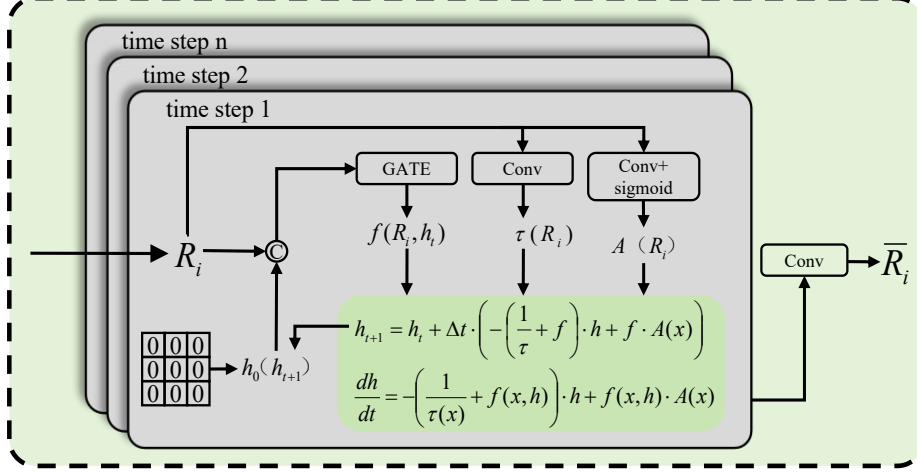


Fig. 2: Illustration of the Vi-LTC module.

continuous spatiotemporal evolution process[18]. Built upon the theory of Liquid Neural Networks, this module introduces a time-continuous hidden state $\mathbf{h}(t)$ that evolves based on input stimuli. The core motivation lies in the unique property of the liquid time-constant τ , which regulates the "inertia" of the system state. By learning input-dependent τ , the network can dynamically adjust its evolution speed across different spatial locations-maintaining stability in homogeneous regions while allowing rapid state transitions at object boundaries. This continuous dynamics constraint forces the model to balance global semantic consistency with fine-grained structural fidelity, making it intrinsically suitable for the precise localization demands of SOD.

Formally, the dynamics of a single neuron in the Vi-LTC are governed by the following ODE:

$$\frac{d\mathbf{h}(t)}{dt} = - \left[\frac{1}{\tau(\mathbf{x})} + f(\mathbf{x}, \mathbf{h}) \right] \cdot \mathbf{h}(t) + f(\mathbf{x}, \mathbf{h}) \cdot A(\mathbf{x}), \quad (4)$$

where $\mathbf{h}(t)$ represents the evolving hidden state, and $\mathbf{x}$ denotes the input features. $A(\mathbf{x})$ serves as the external driving force, while $f(\mathbf{x}, \mathbf{h})$ represents the synaptic coupling strength. Crucially, $\tau(\mathbf{x})$ is the input-dependent time constant. In our visual adaptation, we parameterize these terms using convolutional operators to handle 2D spatial dependencies.

The mechanism functions through the interplay between the state $\mathbf{h}$ and the time-constant τ .

Dynamic State Evolution ($\mathbf{h}$): The hidden state $\mathbf{h}(t)$ acts as a memory container that iteratively refines the saliency representation. The initial state $\mathbf{h}_0$ is initialized as an all-zero matrix with the same spatial dimensions as the input feature. Unlike standard residual connections, the update of $\mathbf{h}$ is constrained by the decay term $-[1/\tau + f]\mathbf{h}$, ensuring system stability and bounded outputs.

Spatially-Adaptive Regulation (τ): The time-constant $\tau(\mathbf{x})$ determines the "stiffness" of the state evolution. In regions with complex textures or edges (high-frequency), the network learns a smaller τ , reducing system inertia to allow $\mathbf{h}(t)$ to change rapidly and capture details. Conversely, in salient object interiors (low-frequency), a larger τ slows down the evolution, filtering out noise and enforcing spatial coherence.

Unlike standard gated networks that learn independent input and forget gates, Vi-LTC enforces a strict **physical coupling** derived from ODEs: the decay rate (determined by τ and f) and the input injection (determined by f) are mathematically locked. This constraint guarantees Bounded Input Bounded Output (BIBO) stability, preventing feature explosion in deep networks. Furthermore, the sensitivity of model performance to different ODE solvers shown as Table. 4, empirically confirms that Vi-LTC functions as a continuous dynamical system rather than a static discrete mapping.

To implement this continuous dynamics within a discrete deep learning framework, we employ an explicit Euler solver to approximate the solution of Eq. 4. This transforms the ODE into an iterative update rule, allowing for efficient end-to-end training:

$$\begin{aligned}
\mathbf{F} &= \text{CAT}(\mathbf{R}_i, \mathbf{h}_t), \\
\mathbf{x} &= A(\mathbf{R}_i) = \text{Conv}_{3 \times 3}(\mathbf{R}_i), \\
\mathbf{y} &= \tau(\mathbf{R}_i) = \sigma(\text{Conv}_{3 \times 3}(\mathbf{R}_i)), \\
\mathbf{z} &= f(\mathbf{R}_i, \mathbf{h}_t) = \sigma(\text{Conv}_{3 \times 3}(\text{ReLU}(\text{DWConv}(\mathbf{F})))), \\
\Delta \mathbf{h} &= - \left(\frac{1}{\mathbf{y}} + \mathbf{z} \right) \cdot \mathbf{h}_t + \mathbf{z} \cdot \mathbf{x}, \\
\mathbf{h}_{t+1} &= \mathbf{h}_t + \Delta t \cdot \Delta \mathbf{h}, \quad \bar{\mathbf{R}}_i = \text{Conv}_{1 \times 1}(\mathbf{h}_{t+1}),
\end{aligned} \tag{5}$$

where Δt is the integration step size. This formulation ensures that the output feature $\bar{\mathbf{R}}_i$ is not merely a gated mapping but the result of a physically grounded dynamical process.

Furthermore, to leverage cross-modal complementarity, we extend this framework to the Selective LTC (S-LTC) mechanism. S-LTC utilizes the evolved RGB state to modulate the dynamics of the depth branch. Specifically, we generate a saliency mask $\mathbf{M}$ from the fused features, which is then applied to the driving term of the depth stream. This effectively adjusts the "external force" acting on the depth state evolution, guiding the solver to focus computational updates on salient regions:

$$\begin{aligned}
\mathbf{F}_{fuse} &= \text{CAT}(\bar{\mathbf{R}}_i, \mathbf{D}_i), \\
\mathbf{M} &= \sigma(\text{Conv}_{1 \times 1}(\text{ReLU}(\text{Conv}_{3 \times 3}(\mathbf{F}_{fuse})))), \\
\bar{\mathbf{D}}_i &= \text{DWConv}(\text{Vi-LTC}(\text{DWConv}(\mathbf{M} \odot \mathbf{D}_i))).
\end{aligned} \tag{6}$$

Here, $\odot$ denotes element-wise multiplication. By coupling the driving force of the depth modality with RGB semantics, S-LTC ensures that the depth features evolve in alignment with the visual content.

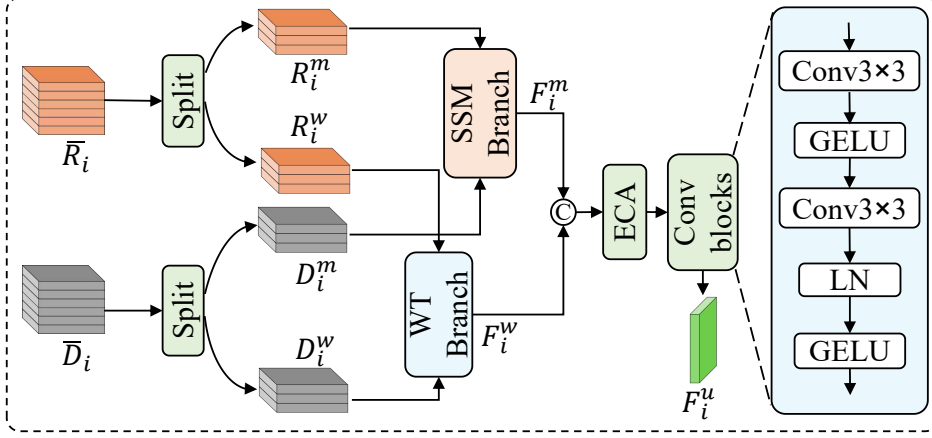


Fig. 3: Details of cross-modal wavelet fusion module.

3.5 Cross Modal Wavelet Fusion

In RGB-D SOD, RGB images provide color and texture information, while depth maps capture object contours and positional details. To enable complementary cross-modal interaction and preserve high-frequency edge information, we designed the CMWF module (Fig. 3). We first split the RGB and depth features into two separate streams:

$$\mathbf{R}_i^m, \mathbf{R}_i^w = \text{Split}(\bar{\mathbf{R}}_i); \quad \mathbf{D}_i^m, \mathbf{D}_i^w = \text{Split}(\bar{\mathbf{D}}_i). \quad (7)$$

The features $\mathbf{R}_i^m, \mathbf{D}_i^m$ are routed to the SSM branch for global semantic enhancement, while $\mathbf{R}_i^w, \mathbf{D}_i^w$ are directed to the WT branch for frequency-domain fusion.

WT Branch: As shown in Fig. 4, this branch employs Discrete Wavelet Transform (WT) to decompose features into low-frequency (LL) and high-frequency (H $\in$ {LH, HL, HH}) components:

$$\{\mathbf{R}_i^{\text{LL}}, \mathbf{R}_i^{\text{H}}\} = \text{WT}(\mathbf{R}_i^w); \quad \{\mathbf{D}_i^{\text{LL}}, \mathbf{D}_i^{\text{H}}\} = \text{WT}(\mathbf{D}_i^w). \quad (8)$$

Subsequently, we fuse these components using Inverse Wavelet Transform (IWT) and Efficient Channel Attention (ECA) to generate the enhanced feature:

$$\begin{aligned} \mathbf{F}_i^{\text{H}} &= \text{Convs}(\text{ECA}(\text{CAT}(\mathbf{R}_i^{\text{H}}, \mathbf{D}_i^{\text{H}}))), \\ \mathbf{F}_i^{\text{L}} &= \text{Convs}(\text{ECA}(\text{CAT}(\mathbf{R}_i^{\text{LL}}, \mathbf{D}_i^{\text{LL}}))), \\ \mathbf{F}_i^{\text{HL}} &= \text{IWT}(\mathbf{F}_i^{\text{L}} + \mathbf{R}_i^{\text{LL}} \odot \mathbf{D}_i^{\text{LL}}, \mathbf{F}_i^{\text{H}} + \mathbf{R}_i^{\text{H}} \odot \mathbf{D}_i^{\text{H}}), \\ \mathbf{F}_i^{\text{w}} &= \mathbf{F}_i^{\text{HL}} + \mathbf{R}_i^{\text{w}} + \mathbf{D}_i^{\text{w}}, \end{aligned} \quad (9)$$

where $\odot$ denotes element-wise multiplication. We use GELU activation in convolutional blocks to preserve negative values in frequency bands.

SSM Branch: This branch facilitates preliminary cross-modal fusion followed by

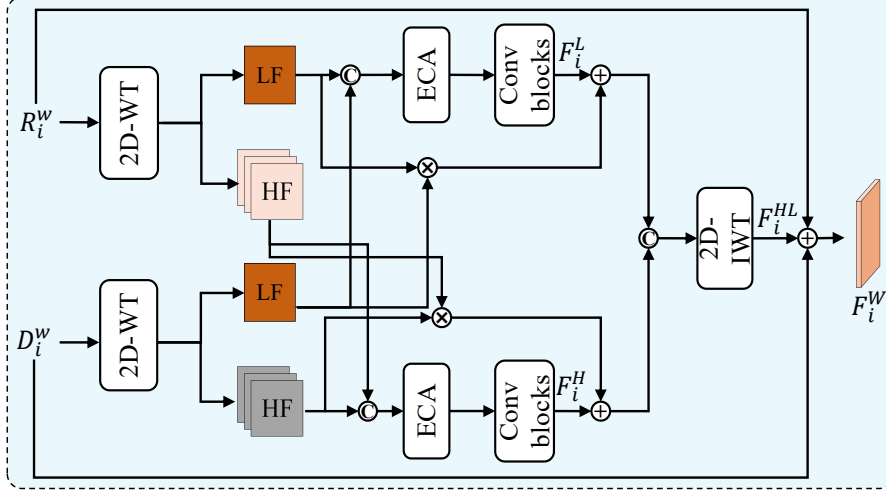


Fig. 4: Details of WT branch.

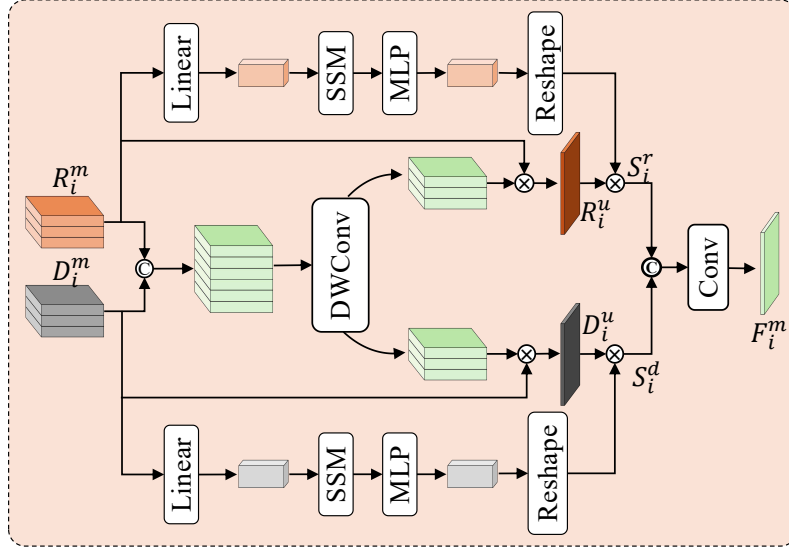


Fig. 5: Details of SSM branch.

self-modal enhancement using SSMs (Fig. 5). The RGB features are processed as:

$$\begin{aligned} R_i^u &= \text{DWConv}(\text{CAT}(R_i^m, D_i^m)) \odot R_i^m, \\ S_i^r &= \text{MLP}(\text{Linear}(\text{SSM}(R_i^m))) \odot R_i^u. \end{aligned} \quad (10)$$

The enhanced depth feature $\mathbf{S}_i^d$ is obtained identically. The branch outputs are then fused:

$$\mathbf{F}_i^m = \text{Conv-BN}(\text{CAT}(\mathbf{S}_i^r, \mathbf{S}_i^d)). \quad (11)$$

Finally, we aggregate the outputs from both branches to produce the unified fused feature:

$$\mathbf{F}_i^u = \text{Convs}(\text{ECA}(\text{CAT}(\mathbf{F}_i^m, \mathbf{F}_i^w))). \quad (12)$$

4 Experiments

4.1 Experiment Settings

Implementation setup: We train the LWNet on an GeForce RTX 4080 GPU. During the training and testing of the model, we resize the input images to 352×352 . To enhance the generalization ability and robustness of the model, we employ various data augmentation methods, including random flip, random crop, and random rotation. LWNet is trained using the Adam optimizer[29]. We train our model for 200 epochs with a batch size of 8 and an initial learning rate of $5e-5$. The learning rate is decayed by a factor of 10 every 60 epochs.

Dataset: Our experiments were conducted on five mainstream benchmark datasets: STERE[30], NLPR[31], NJU2K[32], SIP[33] and DUT[34].

Training and testing: Following prior studies [35, 36], we use 700 images from the NLPR dataset and 1,485 images from the NJU2K dataset as the training set, while the test set includes non-overlapping portions of the NJU2K and NLPR datasets, the entire STERE dataset, and the complete SIP dataset, ensuring comprehensive evaluation across diverse scenarios. In addition, we conduct separate training and testing on the DUT dataset.

Training Loss: To supervise the network, we employ a joint loss function $L_{total} = L_{wbce} + L_{wiou}$. This combined objective incorporates edge-aware weights to penalize errors in boundary regions more heavily than in smooth backgrounds, thereby refining the structural details of the predicted saliency maps. The weight generation process is mathematically formulated as:

$$weight = 1 + k \cdot \text{avgpool}(\text{GT}) - \text{GT}, \quad (13)$$

where avgpool denotes average pooling, and k is a weight coefficient used to control the importance of edges. The weight map $weight$ is then combined with the binary cross-entropy (BCE) to form the WBCE loss:

$$L_{WBCE} = \frac{\sum(\text{weight} \cdot L_{BCE})}{\sum \text{weight}}, \quad (14)$$

in the above equation, dividing by weight is for normalization. The calculation of WIOU comes next:

$$L_{WIOU} = 1 - \frac{\text{inter} + 1}{\text{union} - \text{inter} + 1}, \quad (15)$$

where *inter* is the weighted intersection, and *union* is the weighted union. They are computed as follows:

$$\begin{aligned} \text{inter} &= \sum \text{weit} \cdot \text{pred} \cdot \text{GT}, \\ \text{union} &= \sum [\text{weit} \cdot (\text{pred} + \text{GT})], \end{aligned} \quad (16)$$

where *pred* denotes predicted results in the decoder stage. The overall loss of the model is defined as:

$$L = \sum_{i=1}^5 [L_{W BCE}^i(P_i, GT) + L_{W IOU}^i(P_i, GT)], \quad (17)$$

where P_i is the i_{th} level prediction map.

Evaluation metrics: We evaluate our model using four widely recognized metrics in RGB-D SOD tasks: S-measure (S_α)[37], F-measure (F_β)[38], E-measure (E_ξ)[39], and Mean Absolute Error (MAE)[40].

4.2 Compare with State-of-the-arts Methods

Contenders: We perform experiments on our and another eighteen methods to make the comparison. To be fair, the compared results are produced using authorized codes or supplied by the authors if there is no code available.

Quantitative Evaluation: Table 1 presents the quantitative results of our LWNet compared with 18 approaches across five mainstream datasets. Overall, our method achieves competitive performance against numerous great baselines. Specifically, our method achieves best performance on five benchmark datasets: STERE[30], NJU2K[32], NLPR[31], DUTLF[34] and SIP[33] datasets. First, the Vi-LTC module possesses dynamic modeling capabilities, enabling the model to adaptively focus on salient object regions. As a result, it significantly improves the F_β of LWNet and effectively reduces the MAE. Second, benefiting from the frequency-specific fusion of the CMWF module and the self-modal enhancement of the SSM, LWNet is able to capture finer details and structural characteristics of salient objects. Consequently, it achieves competitive performance in terms of E_ξ and S_α .

Qualitative comparison: As illustrated in Fig. 6, we conduct a visual comparison with different methods across various challenging scenarios. In the (a) and (b) rows, the images contain objects with fine structural details. It is evident that LWNet exhibits superior capability in capturing object contours and internal structures, effectively preserving fine-grained spatial information that other methods tend to overlook. The (c) and (d) rows represent cases with low-quality depth maps, where cross-modal fusion becomes particularly crucial. These examples highlight the robustness of LWNet under modality degradation. In the (c) row, the presence of strong background interference-such as water reflections-poses challenges for existing models. In contrast, LWNet remains unaffected and produces a complete and precise detection of the ship’s structure, demonstrating its ability to selectively rely on more reliable modality cues. Finally, the (e) and (f) rows presents a large-scale target, which is typically difficult to detect due to weak visual saliency. While most compared methods fail to localize the entire object, LWNet accurately captures the full extent of the target.

Quantitative Evaluation on Remote Sensing Images: Although LWNet is

Table 1: Comparison with other RGB-D SOD methods. The **red** and **blue** values indicate the top two methods. $\uparrow$ ($\downarrow$) indicates that higher (lower) is better. PubYear denotes the year in which a reference was published.

Method		CR-Net	PICR-Net	LSNet	HINet	CATNet	DFGNet	AirSOD	DMGNet	MFUR-Net	DFormer	TPCL	SAINet	LAFB*	GL-DMNet	ALNet	SATNet	EATNet	HENet	Ours
Journal		AI	ACM MM	TIP	PR	TMM	TMM	TCSVT	NN	KBS	ICLR	TMM	OLE	TITS	TMM	JSJU	TIP	MS	TCSVT	
PubYear		23 [41]	23 [42]	23 [43]	23 [44]	24 [14]	24 [45]	24 [13]	24 [46]	24 [47]	24 [48]	24 [49]	24 [50]	25 [51]	25 [52]	25 [53]	25 [54]	25 [55]	25 [56]	
STERE	$F_{\beta}^{max} \uparrow$	0.917	0.927	0.854	0.883	0.902	0.896	0.900	0.910	0.919	0.921	0.916	0.901	0.917	0.918	0.908	0.907	0.919	0.919	0.929
	$E_{\xi}^{max} \uparrow$	0.944	0.957	0.913	0.933	0.935	-	0.939	0.934	-	0.945	0.956	0.943	0.936	0.947	0.942	0.953	0.958	0.944	0.955
	$S_{\alpha} \uparrow$	0.904	0.920	0.834	0.892	0.925	0.911	0.895	0.897	0.920	0.915	0.925	0.916	0.901	0.908	0.898	0.922	0.895	0.912	0.922
	MAE $\downarrow$	0.038	0.030	0.066	0.049	0.030	0.035	0.043	0.041	0.040	0.033	0.033	0.031	0.038	0.037	0.040	0.032	0.036	0.036	0.029
	Rank $\downarrow$	9	2	19	18	8	17	14	16	11	5	3	13	10	12	15	4	7	6	1
NLPR	$F_{\beta}^{max} \uparrow$	0.930	0.936	0.859	0.906	0.916	0.902	0.923	0.927	0.930	0.931	0.930	0.912	0.923	0.926	-	0.900	0.932	0.937	0.944
	$E_{\xi}^{max} \uparrow$	0.963	0.970	0.932	0.957	0.968	-	0.963	0.956	-	0.960	0.970	0.960	0.965	0.962	-	0.964	0.969	0.970	0.971
	$S_{\alpha} \uparrow$	0.930	0.935	0.873	0.922	0.939	0.928	0.924	0.924	0.931	0.930	0.935	0.921	0.929	0.927	-	0.932	0.938	0.938	0.942
	MAE $\downarrow$	0.021	0.018	0.038	0.026	0.018	0.021	0.023	0.022	0.022	0.021	0.017	0.022	0.017	0.022	-	0.019	0.019	0.019	0.014
	Rank $\downarrow$	7	3	18	17	6	15	13	14	9	8	5	16	10	11	-	12	4	2	1
NIJ2K	$F_{\beta}^{max} \uparrow$	0.932	0.935	0.921	0.914	0.929	0.914	0.918	0.927	0.937	0.934	0.930	0.909	0.932	0.930	0.913	0.917	0.941	0.938	0.945
	$E_{\xi}^{max} \uparrow$	0.954	0.957	0.949	0.945	0.933	-	0.944	0.930	-	0.953	0.959	0.943	0.951	0.950	0.935	0.952	0.964	0.959	0.963
	$S_{\alpha} \uparrow$	0.921	0.927	0.910	0.915	0.940	0.921	0.908	0.913	0.923	0.927	0.925	0.903	0.925	0.920	0.904	0.925	0.934	0.931	0.935
	MAE $\downarrow$	0.032	0.028	0.038	0.039	0.025	0.032	0.039	0.035	0.035	0.028	0.028	0.036	0.031	0.033	0.038	0.029	0.032	0.027	0.023
	Rank $\downarrow$	9	4	16	14	5	13	17	15	10	7	6	18	8	12	19	11	2	3	1
DUTLF	$F_{\beta}^{max} \uparrow$	-	0.956	0.916	-	0.951	0.925	0.920	0.853	0.948	-	0.946	0.878	0.938	0.947	0.937	-	-	0.946	0.963
	$E_{\xi}^{max} \uparrow$	-	0.969	0.940	-	0.971	-	0.946	0.888	-	-	0.967	0.922	0.951	0.962	0.957	-	-	0.968	0.972
	$S_{\alpha} \uparrow$	-	0.942	0.890	-	0.953	0.932	0.891	0.838	0.943	-	0.936	0.881	0.929	0.931	0.921	-	-	0.949	0.948
	MAE $\downarrow$	-	0.021	0.058	-	0.020	0.027	0.048	0.076	0.027	-	0.024	0.050	0.025	0.029	0.033	-	-	0.020	0.019
	Rank $\downarrow$	-	3	12	-	2	9	11	14	6	-	5	13	8	7	10	-	-	4	1
SIP	$F_{\beta}^{max} \uparrow$	0.911	0.897	0.788	0.889	0.918	0.879	0.887	0.904	0.910	0.922	0.917	0.890	0.926	0.915	0.895	0.896	0.901	0.914	0.935
	$E_{\xi}^{max} \uparrow$	0.924	0.924	0.869	0.901	0.944	-	0.903	0.924	-	0.935	0.943	0.918	0.940	0.936	0.923	0.926	0.924	0.936	0.944
	$S_{\alpha} \uparrow$	0.879	0.872	0.773	0.858	0.913	0.883	0.858	0.878	0.890	0.900	0.900	0.863	0.904	0.896	0.874	0.894	0.870	0.894	0.905
	MAE $\downarrow$	0.050	0.053	0.098	0.065	0.034	0.048	0.060	0.049	0.049	0.039	0.038	0.054	0.038	0.041	0.053	0.044	0.052	0.041	0.035
	Rank $\downarrow$	9	12	19	18	2	16	17	11	10	4	5	13	3	6	15	8	14	7	1
Params (M) $\downarrow$		49.2	112.0	5.3	98.9	262.6	42.1	2.36	-	-	5.9	129.4	-	-	-	-	5.2	222.3	10.4	40.9
FLOPs (G) $\downarrow$		37.6	27.0	1.2	389.7	341.8	74.8	0.91	-	-	4.5	212.0	-	-	-	-	1.5	116.5	10.7	64.5

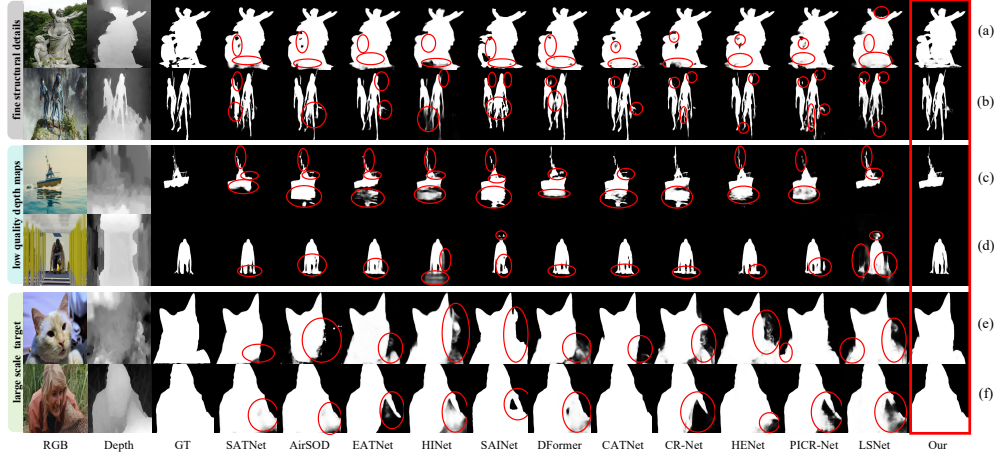


Fig. 6: Visual comparison with different methods. Including EATNet[55], HINet[44], SATNet[54], AirSOD[13], SAINet[50], DFormer[48], CATNet[14], CR-Net [41], HENet[56], PICR-Net[42], LSNet[43]. Our method outperforms other models in various scenes. Our LWNNet not only accurately detects salient objects, but also enables edges to be clearer.

Table 2:

QUANTITATIVE EXPERIMENTS ON TWO CHALLENGING REMOTE SENSING DATASETS.

Method	SeaNet	OMRA-Net	GCANet	GLSIN	GMA-Net	DHCont	LSHNet	SggNet	BCARNet		
Journal	TMM	AI	JISRS	GRSL	RS	TMM	TGRS	RS	PR		
PubYear	23 [57]	24 [58]	24 [59]	24 [60]	24 [61]	24 [62]	24 [63]	25 [64]	25 [65]	Ours	
ORSSD	$F_{\beta}^{max} \uparrow$	0.9227	0.867	0.9251	0.901	0.9096	0.9182	0.9200	0.9032	0.9082	0.9251
	$E_{\epsilon}^{max} \uparrow$	0.9767	0.958	0.9813	0.972	-	0.9804	0.9824	0.9759	0.9833	0.9725
	$S_{\alpha} \uparrow$	0.9503	0.902	0.9235	0.917	0.9268	0.9441	0.9491	0.9342	0.9466	0.9468
	MAE $\downarrow$	0.0070	0.015	0.0104	0.011	0.0184	0.0096	0.0075	0.0111	0.0071	0.0071
	Rank $\downarrow$	2	10	9	8	7	5	1	6	4	3
EORSSD	$F_{\beta}^{max} \uparrow$	0.8873	0.852	0.9036	0.866	0.8746	0.8833	0.8844	0.8770	0.8794	0.8971
	$E_{\epsilon}^{max} \uparrow$	0.9710	0.954	0.9703	0.965	-	0.9750	0.9761	0.9762	0.9761	0.9599
	$S_{\alpha} \uparrow$	0.9356	0.906	0.9207	0.918	0.9227	0.9349	0.9370	0.9279	0.9360	0.9300
	MAE $\downarrow$	0.0052	0.009	0.0079	0.007	0.0072	0.0061	0.0064	0.0068	0.0057	0.0057
	Rank $\downarrow$	3	10	9	8	7	4	1	6	2	5
Params(M) $\downarrow$		67.4	82.5	-	-	64.37	-	-	2.7	24	28.0
FLOPs(G) $\downarrow$		7.52	104.7	-	-	-	-	-	1.3	7	45.7

designed for RGB-D SOD on natural images, it also performs well on remote sensing datasets. As shown in Table 2, we compare LWNNet against nine methods specifically developed for remote sensing tasks. Evaluations are conducted on two widely used benchmarks: ORSSD [66] and EORSSD [67]. Since these datasets do not contain depth maps, we remove the depth branch from our model and retain only the wavelet component for a fair comparison. Despite this simplification, LWNNet maintains competitive performance.

Table 3:

ABLATION STUDY OF LWNET ON THE NJU2K [32] AND DUTLF [34] DATASETS.

Variant	Baseline	Vi-LTC	CMWF	NJU2K				DUTLF			
				$F_\beta^{\max} \uparrow$	$E_\xi^{\max} \uparrow$	$S_\alpha \uparrow$	MAE $\downarrow$	$F_\beta^{\max} \uparrow$	$E_\xi^{\max} \uparrow$	$S_\alpha \uparrow$	MAE $\downarrow$
(a)	✓			0.932	0.951	0.922	0.034	0.948	0.956	0.932	0.032
(b)	✓	✓		0.940	0.956	0.927	0.029	0.958	0.962	0.939	0.027
(c)	✓		✓	0.938	0.959	0.929	0.031	0.956	0.966	0.941	0.023
(d)	✓	✓	✓	0.945	0.963	0.935	0.023	0.963	0.972	0.948	0.019

Table 4:

PERFORMANCE OF VI-LTC WITH DIFFERENT ODE SOLVERS ON THE NJU2K [32] AND DUTLF [34] DATASETS.

Variant	Candidate	NJU2K				DUTLF				FLOPs (G) $\downarrow$
		$F_\beta^{\max} \uparrow$	$E_\xi^{\max} \uparrow$	$S_\alpha \uparrow$	MAE $\downarrow$	$F_\beta^{\max} \uparrow$	$E_\xi^{\max} \uparrow$	$S_\alpha \uparrow$	MAE $\downarrow$	
(a)	Euler	0.938	0.959	0.928	0.031	0.957	0.970	0.942	0.019	5.7
(b)	Heun2	0.939	0.961	0.930	0.026	0.959	0.967	0.940	0.022	8.1
(c)	RK4	0.942	0.961	0.932	0.028	0.960	0.966	0.943	0.018	14.3
(d)	S-LTC	0.945	0.963	0.935	0.023	0.963	0.972	0.948	0.019	7.3

4.3 Ablation Study

Several ablation experiments are conducted to assess the effectiveness of the proposed module on NJU2K[32], NLPR[31], STERE[30], SIP[33] and DUT[34] datasets.

4.3.1 The effectiveness of each module

To validate the effectiveness of each module, we conducted controlled ablation studies, as presented in Table 3. We begin with the baseline 3-(a) and incrementally integrate each module to construct the complete LWNet architecture 3-(e). The baseline consists of a single-stream VMamba backbone and element-wise addition for cross-modal fusion.

Progressing from 3-(a) to 3-(b), all evaluation metrics, particularly the F_β , demonstrate significant improvement. This validates that the Vi-LTC module successfully redirects model attention to salient objects. These advantages substantially enhance the model’s representational power. Subsequently, the transition to Table 3-(c) shows further gains across metrics, indicating effective inter-module interaction and successful cross-modal feature fusion through the CMWF module. Analyses in Table 3-(b) and 3-(c) respectively reveal the specialized contributions of these components: the Vi-LTC module elevates F_β by leveraging dynamic input processing and LTC equation-based neuronal modeling, enabling human visual mechanism emulation that refocuses attention on salient targets. Concurrently, the CMWF module improves E_ξ and S_α while reducing MAE. This is achieved through complementary cross-modal feature integration, where high-frequency wavelet components play a dual role: suppress background noise and enhance boundary sharpness in salient regions.

4.3.2 The effectiveness of Vi-LTC module

To demonstrate the effectiveness of Vi-LTC, we visualize the processed features using class activation maps, as shown in Fig. 7. The model often fails to accurately localize

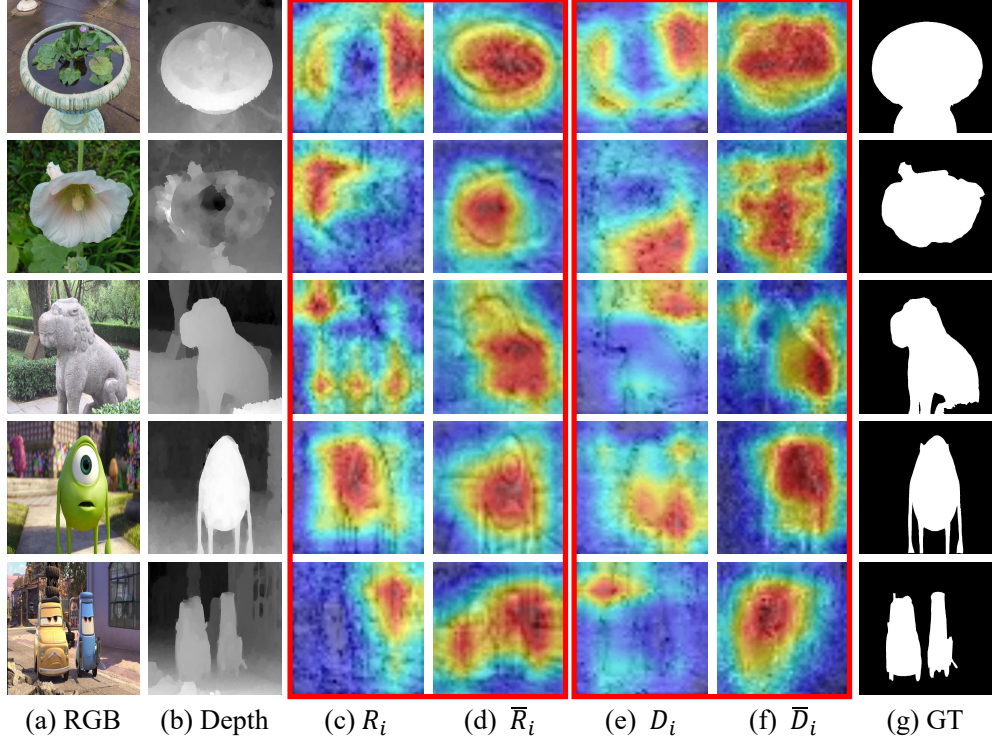


Fig. 7: Comparison of class activation map visualizations for the Vi-LTC module. Column (a) and (b) show the input RGB and depth image.

salient regions, or predicts shifted salient areas, preventing effective attention allocation to key regions, shown as columns (c) and (e). Compared with the backbone features, the ability to accurately identify salient regions is significantly improved after using the proposed Vi-LTC module, as shown in columns (d) and (f). These results clearly demonstrate that Vi-LTC effectively redirects the model’s attention toward salient object regions.

In addition, to validate the S-LTC mechanism, we designed substitution experiments. S-LTC mechanism integrates an Euler solver with a masking mechanism. First, the LTC equation serves as the computational core, providing memory capabilities and neuron-inspired responsiveness to inputs. When evaluating differential equation solvers with varying precision and computational cost, we observed consistent performance improvements with higher-precision methods: RK4 4-(a) achieved the highest accuracy, followed by Heun2 4-(b), while Euler solver 4-(c) delivered the most modest gains. Second, the S-LTC mechanism 4-(d) leverages RGB features to guide depth feature processing, concentrating computation on salient regions. Guided by the RGB modality, S-LTC employs a masking strategy in combination with the low computational complexity of the Euler solver.

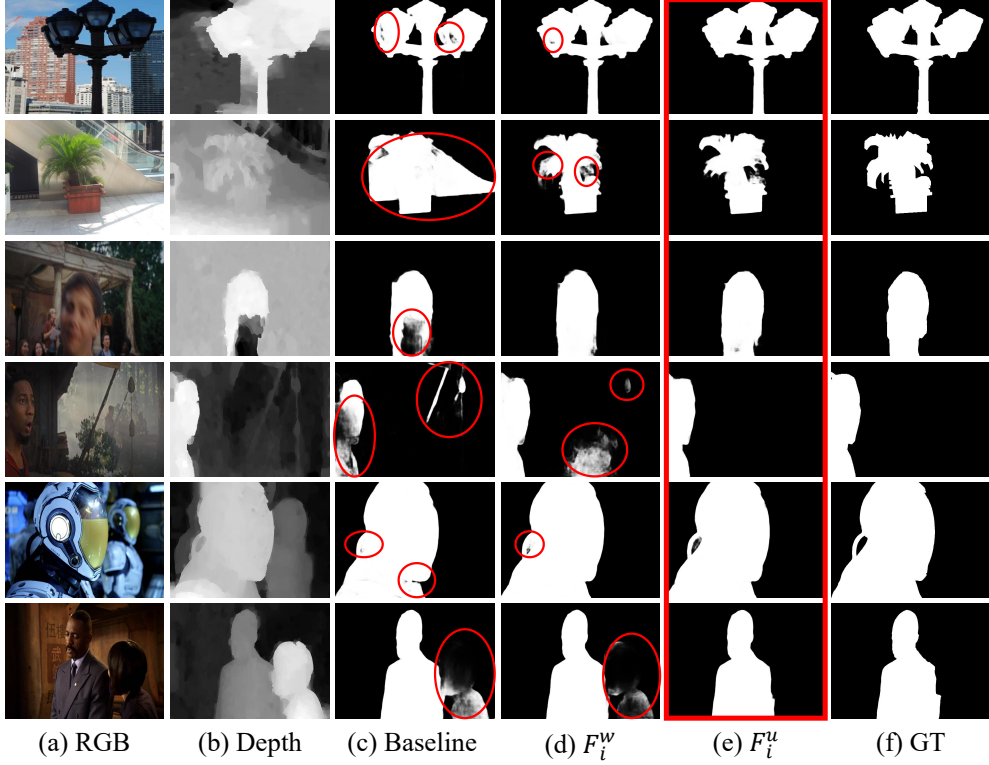


Fig. 8: Visual comparison illustrating the effectiveness of the CMWF.

Table 5:

COMPARISON OF DIFFERENT FUSION METHODS ON THE DUTLF [34] DATASET.

Method	Params (M) ↓	FLOPs (G) ↓	$F_\beta^{\max}$ ↑	$E_\xi^{\max}$ ↑	S_α ↑	MAE ↓
HINet	33.6	50.0	0.951	0.961	0.937	0.024
EATNet	25.3	30.9	0.950	0.962	0.938	0.025
Ours	40.9	64.5	0.963	0.972	0.948	0.019

4.3.3 The effectiveness of our CMWF module

Additionally, we conduct a qualitative visualization analysis of the CMWF module. As shown in Fig. 8, it can be clearly observed that the WT branch significantly reduces false positives and false negatives, while the full CMWF module produces more refined and accurate predictions.

To further evaluate the effectiveness of our CMWF, we replaced it with the cross-modal fusion strategies used in HINet [44] and EATNet [55], respectively, as shown in Table 5. The results demonstrate that, compared with the fusion methods of HINet and EATNet, our CMWF achieves consistent improvements across all evaluation metrics.

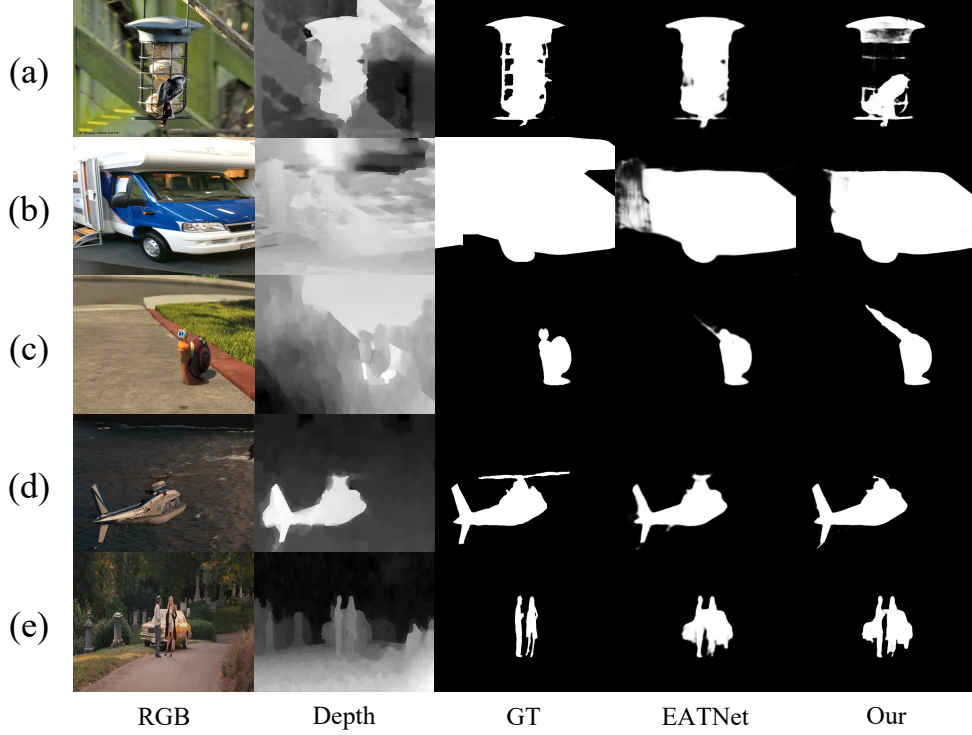


Fig. 9: Visual examples of failure cases.

Table 6:

ABLATION STUDY OF CMWF ON THE NJU2K [32] AND DUTLF [34] DATASETS.

Variant	Candidate	NJU2K				DUTLF			
		$F_{\beta}^{\max} \uparrow$	$E_{\xi}^{\max} \uparrow$	$S_{\alpha} \uparrow$	MAE $\downarrow$	$F_{\beta}^{\max} \uparrow$	$E_{\xi}^{\max} \uparrow$	$S_{\alpha} \uparrow$	MAE $\downarrow$
(a)	w/o WT	0.941	0.960	0.931	0.028	0.953	0.963	0.935	0.025
(b)	w/o SSM	0.940	0.961	0.932	0.027	0.956	0.964	0.937	0.025
(c)	Ours	0.945	0.963	0.935	0.023	0.963	0.972	0.948	0.019

In the key component ablation experiments, systematic removal of the WT branch 6-(a) and SSM branch 6-(b) consistently degraded performance across all metrics.

5 Failure Cases

As illustrated in Fig. 9, two particularly challenging cases are observed. First, when the RGB image exhibits high intra-object contrast while the depth map is of poor quality (as in rows (a) and (b)). In this case, the model tends to treat different parts of the salient object as separate entities due to the high contrast, and the degraded depth map fails to provide holistic structural cues. Consequently, neither modality can supply reliable structural information.

Second, when the salient object in the RGB image has low contrast with the background and the depth map is also of poor quality (as in rows (b), (c), and (d)). Under such circumstances, the RGB modality cannot clearly delineate the salient object from the background, while the low-quality depth map offers little compensation, resulting in inaccurate predictions.

6 CONCLUSION

In this paper, we propose LWNet, a novel method for accurate RGB-D SOD, which employs a shared VMamba backbone to efficiently extract cross-modal features. Concretely, we propose the Vi-LTC module, leveraging LTCs equations and dynamic modeling to enable the model to focus on salient objects within images, thereby facilitating superior learning of local features. To the best of our knowledge, we are the first to apply LTCs to the SOD task. Furthermore, we propose the CMWF module to perform cross-modal feature fusion across different frequency bands, achieving effectively complementary features fusion. Our experiments demonstrate that LWNet achieves excellent performance on five benchmark datasets and delivers competitive results on two remote sensing datasets. These results demonstrate the strong robustness of our approach and its potential to bridge the gap between natural and remote sensing images.

Declarations

- Funding: This work was supported in part by the Yunnan Province Fundamental Research Projects under Grant 202401AU070154, and the Key-Area Research and Development Program of Guangdong Province 2023B0303020001.
- Conflict of interest: Conflict of interest The authors declare no competing interests.
- Data availability: The data can be obtained upon request from the authors.
- Code availability: The source code is publicly available at <https://github.com/1243731970/LWNet>.

References

- [1] Minaee, S., Boykov, Y., Porikli, F., Plaza, A., Kehtarnavaz, N., Terzopoulos, D.: Image segmentation using deep learning: A survey. *IEEE transactions on pattern analysis and machine intelligence* **44**(7), 3523–3542 (2021)
- [2] Huang, X., Deng, Z., Li, D., Yuan, X., Fu, Y.: Missformer: An effective transformer for 2d medical image segmentation. *IEEE transactions on medical imaging* **42**(5), 1484–1494 (2022)
- [3] Meinhardt, T., Kirillov, A., Leal-Taixe, L., Feichtenhofer, C.: Trackformer: Multi-object tracking with transformers. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8844–8854 (2022)

- [4] Ma, F., Shou, M.Z., Zhu, L., Fan, H., Xu, Y., Yang, Y., Yan, Z.: Unified transformer tracker for object tracking. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 8781–8790 (2022)
- [5] Yu, S., Chen, T., Shen, J., Yuan, H., Tan, J., Yang, S., Liu, J., Wang, Z.: Unified visual transformer compression. arXiv preprint arXiv:2203.08243 (2022)
- [6] Ye, X., Gan, Y., Huang, X., Ge, Y., Tang, Y.: Voco-llama: Towards vision compression with large language models. arXiv preprint arXiv:2406.12275 (2024)
- [7] Tu, S., Dai, Q., Cheng, Z.-Q., Hu, H., Han, X., Wu, Z., Jiang, Y.-G.: Motioned-itor: Editing video motion via content-aware diffusion. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 7882–7891 (2024)
- [8] Ren, Y., Shi, J., Zhang, Z., Fan, Y., Lin, Z., He, B., Shrivastava, A.: Content-aware image color editing with auxiliary color restoration tasks. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, pp. 5192–5201 (2024)
- [9] Zhai, Y., Fan, D.-P., Yang, J., Borji, A., Shao, L., Han, J., Wang, L.: Bifurcated backbone strategy for rgb-d salient object detection. *IEEE Transactions on Image Processing* **30**, 8727–8742 (2021)
- [10] Chen, Z., Cong, R., Xu, Q., Huang, Q.: Dpanet: Depth potentiality-aware gated attention network for rgb-d salient object detection. *IEEE Transactions on Image Processing* **30**, 7012–7024 (2020)
- [11] Huang, R., Xing, Y., Wang, Z.: Rgb-d salient object detection by a cnn with multiple layers fusion. *IEEE Signal Processing Letters* **26**(4), 552–556 (2019)
- [12] Wu, Y.-H., Liu, Y., Xu, J., Bian, J.-W., Gu, Y.-C., Cheng, M.-M.: Mobilesal: Extremely efficient rgb-d salient object detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **44**(12), 10261–10269 (2021)
- [13] Zeng, Z., Liu, H., Chen, F., Tan, X.: Airsod: A lightweight network for rgb-d salient object detection. *IEEE Transactions on Circuits and Systems for Video Technology* **34**(3), 1656–1669 (2023)
- [14] Sun, F., Ren, P., Yin, B., Wang, F., Li, H.: Catnet: A cascaded and aggregated transformer network for rgb-d salient object detection. *IEEE Transactions on Multimedia* **26**, 2249–2262 (2024) <https://doi.org/10.1109/TMM.2023.3294003>
- [15] Gu, A., Dao, T.: Mamba: Linear-time sequence modeling with selective state spaces. arXiv preprint arXiv:2312.00752 (2023)
- [16] Liu, Y., Tian, Y., Zhao, Y., Yu, H., Xie, L., Wang, Y., Ye, Q., Jiao, J., Liu, Y.:

- Vmamba: Visual state space model. *Advances in neural information processing systems* **37**, 103031–103063 (2024)
- [17] Zhan, Y., Zeng, Z., Liu, H., Tan, X., Tian, Y.: Mambasod: Dual mamba-driven cross-modal fusion network for rgb-d salient object detection. *Neurocomputing* **631**, 129718 (2025)
 - [18] Hasani, R., Lechner, M., Amini, A., Rus, D., Grosu, R.: Liquid time-constant networks. In: *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, pp. 7657–7666 (2021)
 - [19] Yuan, J., Wang, Y., Wang, Z., Xu, Q., Veeravalli, B., Yang, X.: Dppnet: A depth pixel-wise potential-aware network for rgb-d salient object detection. *IEEE Transactions on Multimedia* **27**, 4256–4268 (2025) <https://doi.org/10.1109/TMM.2025.3535386>
 - [20] Dong, W., Zhu, H., Lin, S., Luo, X., Shen, Y., Guo, G., Zhang, B.: Fusion-mamba for cross-modality object detection. *IEEE Transactions on Multimedia* **27**, 7392–7406 (2025) <https://doi.org/10.1109/TMM.2025.3599020>
 - [21] Daubechies, I.: *Ten Lectures on Wavelets*. SIAM, ??? (1992)
 - [22] Huang, L., Liu, G., Wang, Y., Yuan, H., Chen, T.: Fire detection in video surveillances using convolutional neural networks and wavelet transform. *Engineering Applications of Artificial Intelligence* **110**, 104737 (2022) <https://doi.org/10.1016/j.engappai.2022.104737>
 - [23] Zhang, W., Zhou, L., Zhuang, P., Li, G., Pan, X., Zhao, W., Li, C.: Underwater image enhancement via weighted wavelet visual perception fusion. *IEEE Transactions on Circuits and Systems for Video Technology* **34**(4), 2469–2483 (2023)
 - [24] Yao, T., Pan, Y., Li, Y., Ngo, C.-W., Mei, T.: Wave-vit: Unifying wavelet and transformers for visual representation learning. In: *European Conference on Computer Vision*, pp. 328–345 (2022). Springer
 - [25] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929* (2020)
 - [26] Finder, S.E., Amoyal, R., Treister, E., Freifeld, O.: Wavelet convolutions for large receptive fields. In: *European Conference on Computer Vision*, pp. 363–380 (2024). Springer
 - [27] Chen, R.T., Rubanova, Y., Bettencourt, J., Duvenaud, D.K.: Neural ordinary differential equations. *Advances in neural information processing systems* **31**

(2018)

- [28] Cheng, B., Liu, Z., Wang, Q., Shen, T., Fu, C., Tian, A.: Lightweight progressive multi-level feature collaborative network for remote sensing image salient object detection. *IEEE Transactions on Geoscience and Remote Sensing* (2024)
- [29] Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014)
- [30] Niu, Y., Geng, Y., Li, X., Liu, F.: Leveraging stereopsis for saliency analysis. In: 2012 IEEE Conference on Computer Vision and Pattern Recognition, pp. 454–461 (2012). IEEE
- [31] Peng, H., Li, B., Xiong, W., Hu, W., Ji, R.: Rgbd salient object detection: A benchmark and algorithms. In: *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6–12, 2014, Proceedings, Part III 13*, pp. 92–109 (2014). Springer
- [32] Ju, R., Ge, L., Geng, W., Ren, T., Wu, G.: Depth saliency based on anisotropic center-surround difference. In: 2014 IEEE International Conference on Image Processing (ICIP), pp. 1115–1119 (2014). IEEE
- [33] Fan, D.-P., Lin, Z., Zhang, Z., Zhu, M., Cheng, M.-M.: Rethinking rgb-d salient object detection: Models, data sets, and large-scale benchmarks. *IEEE Transactions on neural networks and learning systems* **32**(5), 2075–2089 (2020)
- [34] Piao, Y., Ji, W., Li, J., Zhang, M., Lu, H.: Depth-induced multi-scale recurrent attention network for saliency detection. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 7254–7263 (2019)
- [35] Gao, S., Zhang, P., Yan, T., Lu, H.: Multi-scale and detail-enhanced segment anything model for salient object detection. In: *Proceedings of the 32nd ACM International Conference on Multimedia*, pp. 9894–9903 (2024)
- [36] Zhang, Q., Qin, Q., Yang, Y., Jiao, Q., Han, J.: Feature calibrating and fusing network for rgb-d salient object detection. *IEEE Transactions on Circuits and Systems for Video Technology* **34**(3), 1493–1507 (2023)
- [37] Fan, D.-P., Cheng, M.-M., Liu, Y., Li, T., Borji, A.: Structure-measure: A new way to evaluate foreground maps. In: *Proceedings of the IEEE International Conference on Computer Vision*, pp. 4548–4557 (2017)
- [38] Achanta, R., Hemami, S., Estrada, F., Susstrunk, S.: Frequency-tuned salient region detection. In: 2009 IEEE Conference on Computer Vision and Pattern Recognition, pp. 1597–1604 (2009). IEEE

- [39] Fan, D.-P., Gong, C., Cao, Y., Ren, B., Cheng, M.-M., Borji, A.: Enhanced-alignment measure for binary foreground map evaluation. In: Proceedings of the 27th International Joint Conference on Artificial Intelligence, pp. 698–704 (2018)
- [40] Perazzi, F., Krähenbühl, P., Pritch, Y., Hornung, A.: Saliency filters: Contrast based filtering for salient region detection. In: 2012 IEEE Conference on Computer Vision and Pattern Recognition, pp. 733–740 (2012). IEEE
- [41] Zong, G., Wei, L., Guo, S., Wang, Y.: A cascaded refined rgb-d salient object detection network based on the attention mechanism. *Applied Intelligence* **53**(11), 13527–13548 (2023)
- [42] Cong, R., Liu, H., Zhang, C., Zhang, W., Zheng, F., Song, R., Kwong, S.: Point-aware interaction and cnn-induced refinement network for rgb-d salient object detection. In: Proceedings of the 31st ACM International Conference on Multimedia, pp. 406–416 (2023)
- [43] Zhou, W., Zhu, Y., Lei, J., Yang, R., Yu, L.: Lsnet: Lightweight spatial boosting network for detecting salient objects in rgb-thermal images. *IEEE Transactions on Image Processing* **32**, 1329–1340 (2023) <https://doi.org/10.1109/TIP.2023.3242775>
- [44] Bi, H., Wu, R., Liu, Z., Zhu, H., Zhang, C., Xiang, T.-Z.: Cross-modal hierarchical interaction network for rgb-d salient object detection. *Pattern Recognition* **136**, 109194 (2023) <https://doi.org/10.1016/j.patcog.2022.109194>
- [45] Xiao, F., Pu, Z., Chen, J., Gao, X.: Dgfnet: Depth-guided cross-modality fusion network for rgb-d salient object detection. *IEEE Transactions on Multimedia* **26**, 2648–2658 (2023)
- [46] Tang, Y., Li, M.: Dmgnet: Depth mask guiding network for rgb-d salient object detection. *Neural Networks* **180**, 106751 (2024) <https://doi.org/10.1016/j.neunet.2024.106751>
- [47] Feng, Z., Wang, W., Li, W., Li, G., Li, M., Zhou, M.: Mfur-net: Multimodal feature fusion and unimodal feature refinement for rgb-d salient object detection. *Knowledge-Based Systems* **299**, 112022 (2024) <https://doi.org/10.1016/j.knosys.2024.112022>
- [48] Yin, B., Zhang, X., Li, Z., Liu, L., Cheng, M.-M., Hou, Q.: Dformer: Rethinking rgb-d representation learning for semantic segmentation. *arXiv preprint arXiv:2309.09668* (2023)
- [49] Wu, J., Hao, F., Liang, W., Xu, J.: Transformer fusion and pixel-level contrastive learning for rgb-d salient object detection. *IEEE Transactions on Multimedia* **26**, 1011–1026 (2023)

- [50] Yan, Y., Jia, X., Song, K., Cui, W., Zhao, Y., Liu, C., Guo, J.: Specificity auto-correlation integration network for surface defect detection of no-service rail. *Optics and Lasers in Engineering* **172**, 107862 (2024) <https://doi.org/10.1016/j.optlaseng.2023.107862>
- [51] Chen, C., Song, M., Pang, S., Peng, C.: Adapting generic rgb-d salient object detection for specific traffic scenarios. *IEEE Transactions on Intelligent Transportation Systems* (2025)
- [52] Yi, K., Tang, H., Li, Y., Xu, J., Zhang, J.: Dual mutual learning network with global-local awareness for rgb-d salient object detection. *arXiv preprint arXiv:2501.01648* (2025)
- [53] Wang, Y., Zheng, W., Xia, Y., Wang, S.: Asymmetric dual-stream networks for lightweight rgb-d salient object detection. *Journal of Shanghai Jiaotong University (Science)*, 1–11 (2025)
- [54] Duan, S., Yang, X., Wang, N., Gao, X.: Lightweight rgb-d salient object detection from a speed-accuracy tradeoff perspective. *IEEE Transactions on Image Processing* (2025)
- [55] Liu, X., Liu, C., Zhou, X., Fan, G.: Eatnet: edge-aware and transformer-based network for rgb-d salient object detection. *Multimedia Systems* **31**(2), 128 (2025)
- [56] Gao, H., Wang, F., Wang, M., Sun, F., Li, H.: Highly efficient rgb-d salient object detection with adaptive fusion and attention regulation. *IEEE Transactions on Circuits and Systems for Video Technology* (2024)
- [57] Li, G., Liu, Z., Zhang, X., Lin, W.: Lightweight salient object detection in optical remote-sensing images via semantic matching and edge alignment. *IEEE Transactions on Geoscience and Remote Sensing* **61**, 1–11 (2023)
- [58] Yan, R., Yan, L., Cao, Y., Geng, G., Zhou, P.: One-stop multiscale reconciliation attention network with scribble supervision for salient object detection in optical remote sensing images. *Applied Intelligence* **54**(5), 3737–3755 (2024)
- [59] Yan, L., Yan, R., Geng, G., Zhou, M., Chen, R.: Salient object detection in optical remote sensing images based on global context mixed attention. *Journal of the Indian Society of Remote Sensing* **52**(7), 1489–1499 (2024)
- [60] Yan, R., Yan, L., Cao, Y., Geng, G., Zhou, P., Meng, Y., Wang, T.: Global–local semantic interaction network for salient object detection in optical remote sensing images with scribble supervision. *IEEE Geoscience and Remote Sensing Letters* **21**, 1–5 (2024)
- [61] Huo, L., Hou, J., Feng, J., Wang, W., Liu, J.: Global and multiscale aggregate network for saliency object detection in optical remote sensing images. *Remote*

- [62] Qiu, Y., Sun, Y., Mei, J., Xu, J.: Deeply hybrid contrastive learning based on semantic pseudo-label for salient object detection in optical remote sensing images. *IEEE Transactions on Multimedia* (2024)
- [63] Lee, S., Cho, S., Park, C., Park, S., Kim, J., Lee, S.: Lshnet: Leveraging structure-prior with hierarchical features updates for salient object detection in optical remote sensing images. *IEEE Transactions on Geoscience and Remote Sensing* (2024)
- [64] Liu, J., He, J., Chen, H., Yang, R., Huang, Y.: A lightweight semantic-and graph-guided network for advanced optical remote sensing image salient object detection. *Remote Sensing* **17**(5), 861 (2025)
- [65] Gu, Y., Chen, S., Sun, X., Ji, J., Zhou, Y., Ji, R.: Optical remote sensing image salient object detection via bidirectional cross-attention and attention restoration. *Pattern Recognition* **164**, 111478 (2025)
- [66] Li, C., Cong, R., Hou, J., Zhang, S., Qian, Y., Kwong, S.: Nested network with two-stream pyramid for salient object detection in optical remote sensing images. *IEEE Transactions on Geoscience and Remote Sensing* **57**(11), 9156–9166 (2019)
- [67] Zhang, Q., Cong, R., Li, C., Cheng, M.-M., Fang, Y., Cao, X., Zhao, Y., Kwong, S.: Dense attention fluid network for salient object detection in optical remote sensing images. *IEEE Transactions on Image Processing* **30**, 1305–1317 (2020)