Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

RGBA-Net: Reliability-Gated Asymmetric Fusion with Boundary Awareness for Lightweight RGB-D Salient Object Detection

Tianlun Yuan
Shanghai University of Electric Power
Mengchun Yu
Shanghai University of Electric Power
Yongfeng Jin
Shanghai University of Electric Power
Qinglin Huang
Shanghai University of Electric Power
Jing Li
24208168@mail.shiep.edu.cn
Shanghai University of Electric Power

Research Article

Keywords: RGB-D salient object detection, Lightweight, Cross-modality interaction, Depth reliability calibration

Posted Date: July 24th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10403910/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

RGBA-Net: Reliability-Gated Asymmetric Fusion with Boundary Awareness for Lightweight RGB-D Salient Object Detection

Tianlun Yuan[1], Mengchun Yu[1], Yongfeng Jin[1], Qinglin Huang[1], Jing Li[1*]

[1*]Faculty of Artificial Intelligence, Shanghai University of Electric Power, No. 1851 Hucheng Ring Road, Pudong New Area, Shanghai, 201306, Shanghai, China.

*Corresponding author(s). E-mail(s): y24208168@mail.shiep.edu.cn;
Contributing authors: j24suep@outlook.com; mengcy@mail.shiep.edu.cn;
lamaaaa@mail.shiep.edu.cn; qinglin.h@foxmail.com;

Abstract

RGB-D salient object detection leverages RGB images and depth maps to identify visually conspicuous regions. Despite progress, two challenges persist: depth maps often contain noise or missing data, and high-precision models suffer from excessive parameters and complexity. This paper presents RGBA-Net, a lightweight network with an asymmetric dual-stream encoder using EdgeNeXt and MobileNetV3 for RGB and depth features. A Depth Reliability Gate mitigates low-quality depth inputs via gradient-guided calibration. Cross-modality interaction is achieved through a decoupled dual-branch structure, while a Multi-scale Boundary Enhancement Fusion Module refines boundaries via adaptive fusion and progressive decoding. Experiments on seven datasets show that RGBA-Net outperforms twelve mainstream methods with only 3.49 million parameters. It sets new state-of-the-art benchmarks for lightweight models, confirming its balance between efficiency and accuracy.

Keywords: RGB-D salient object detection, Lightweight, Cross-modality interaction, Depth reliability calibration

1 Introduction

Salient Object Detection (SOD) aims to emulate the human visual attention mechanism to automatically locate and segment the most perceptually salient regions within an image. It has been extensively deployed in various computer vision tasks, including but not limited to image segmentation [1], object tracking [2], and video analysis [3]. Early SOD methodologies predominantly relied on RGB images and handcrafted low-level features [4, 5]. However, the single RGB modality struggles to provide sufficient discriminative cues in challenging scenarios, such as fluctuating illumination, complex backgrounds, or low contrast between objects and their surroundings.

With the widespread proliferation of depth sensors, the acquisition cost of depth maps has been significantly reduced. Depth maps provide abundant spatial geometric cues, which effectively complement the textural and chromatic limitations of RGB images, thereby substantially enhancing model robustness in complex environments [6]. Consequently, RGB-D SOD has emerged as a prominent research focal point.

In recent years, a multitude of deep learning-based RGB-D SOD frameworks have been developed. Initial efforts largely adopted symmetric dual-stream architectures, striving to optimize detection accuracy through complex cross-modal interaction modules [7, 8]. For instance, Zhao [6] introduced a contrast prior and fluid pyramid integration module to improve precision. Piao [9] designed a multi-scale recurrent attention network to bolster adaptability to complex scenes. CIRNet, proposed by Cong [10], utilizes progressive attention-guided units to facilitate deep cross-modal interaction. Additionally, Bi [11] presented HINet, which employs a feedback mechanism for enhanced information exchange, while CATNet by Sun [12] adopts a cascaded and aggregated Transformer structure to strengthen multi-scale feature representation.

Although the aforementioned methods attain high precision, their prohibitive parameter counts and computational complexity preclude real-time deployment on resource-constrained mobile platforms. To reconcile the trade-off between accuracy and efficiency, lightweight RGB-D SOD methodologies have emerged. Examples include MobileSal [13], which introduced implicit depth restoration; MoADNet [14], utilizing an asymmetric dual-stream structure for faster inference; LSNet [15], enhancing boundary quality via lightweight spatial boosting; MAGNet [16], combining multi-scale awareness with global fusion; BTNet [17], adopting a bio-inspired two-stage optimization; and HENet [18], reducing overhead through adaptive fusion. Nevertheless, existing lightweight methods confront two critical bottlenecks: first, suboptimal adaptability to low-quality depth maps, which fail to effectively suppress noise interference during feature fusion [19, 20]; and second, limited preservation of fine-grained boundary details, often leading to blurred edges or loss of detail in the prediction maps [21, 22].

To address the trade-off between lightweight model design and high detection performance, we propose RGBA-Net, a computationally efficient and robust network tailored for RGB-D SOD. Specifically, we rethink the feature extraction process and adopt an asymmetric dual-stream architecture by selecting the lightweight EdgeNeXt [23] and MobileNetV3 [24] as the backbones for RGB and depth images, respectively. These networks are inherently optimized for resource-constrained edge devices, maintaining competitive accuracy while significantly reducing computational burden. More

importantly, this asymmetric design perfectly aligns with the heterogeneous nature of the two modalities: EdgeNeXt efficiently captures rich semantic textures from RGB inputs, while MobileNetV3 extracts essential spatial geometric priors from depth maps without redundant parameters. Building upon this efficient foundation, RGBA-Net further incorporates specialized modules to tackle depth noise and boundary degradation.

Building upon this efficient foundation, RGBA-Net further incorporates specialized modules to tackle depth noise and boundary degradation. The primary contributions are three-fold:

- **Depth Reliability Gate (DRG)**: We introduce a DRG module that dynamically modulates depth feature weights through a gradient-guided soft-isolation mechanism, effectively suppressing noise in low-quality depth data.
- **Cross-Modality Synergy (CMS)**: We design a CMS module with a decoupled dual-branch architecture to explore both shared semantics and complementary information, thereby bolstering the representational capacity of fused features.
- **Multi-scale Boundary Enhancement Fusion Module (MBEFM)**: We construct an MBEFM that integrates boundary awareness and adaptive fusion to markedly sharpen the localization precision of salient object edges.

2 Related Work

Early RGB-D SOD methodologies primarily relied on handcrafted features, such as depth contrast [4] and anisotropic center-surround difference [5], to facilitate salient region localization. However, these approaches struggle to capture high-level semantic abstractions, leading to constrained generalization capabilities. With the rapid evolution of deep convolutional neural networks (CNNs), researchers have transitioned toward constructing end-to-end RGB-D SOD models. Notable examples include the fluid pyramid integration by Zhao [6], multi-scale recurrent attention networks by Piao [9], and the progressive attention-guided units in CIRNet [10]. Other advancements include HINet's feedback mechanism [11], CATNet's aggregated Transformer structure [12], and SPSN's superpixel prototype sampling [25]. Furthermore, TPCL [26] combined pixel-level contrastive learning with Transformer architectures, while STANet [27] employed a triple-attention framework to handle low-quality depth maps. Despite their superior precision, these methods predominantly rely on parameter-heavy backbones like VGG and ResNets, characterized by prohibitive parameter counts and computational intensity, which hinder their deployment in real-time or mobile scenarios.

To mitigate model complexity, researchers have increasingly explored lightweight architectures driven by the evolution of efficient image classification networks. Traditional heavy backbones are being supplanted by optimized alternatives such as MobileNetV3 and EdgeNeXt. MobileNetV3 leverages hardware-aware architecture search to deliver high-accuracy, low-latency performance suited for mobile devices. Conversely, EdgeNeXt efficiently bridges CNNs and Transformers via a split-depth transposed attention encoder, which encodes multi-scale features while implicitly

expanding the receptive field with minimal overhead. Inspired by these advancements, our proposed model adopts an asymmetric strategy, employing EdgeNeXt and MobileNetV3 as the backbones for RGB and depth modalities, respectively, to forge a highly efficient RGB-D SOD framework. In the broader context of lightweight SOD, Zhang [19] proposed a feature enhancement module driven by depth quality awareness. MoADNet [14] similarly adopted an asymmetric dual-stream structure to significantly accelerate inference speed. MobileSal [13] introduced implicit depth restoration and compact pyramid refinement, establishing itself as a representative framework for lightweight models. Subsequent works, such as LSNet [15], utilize lightweight spatial boosting, while MAGNet [16] integrates multi-scale awareness, and BTNet [17] employs bio-inspired two-stage optimization. Additionally, HENet [18] and AirSOD [28] focus on adaptive fusion and parallel attention shifts to reduce computational costs. Nevertheless, these methods still exhibit suboptimal robustness against low-quality depth maps and show limited proficiency in preserving fine-grained salient object boundaries—critical issues that the present study aims to address.

Cross-modal fusion represents a core challenge in RGB-D SOD. Early strategies predominantly resorted to elementary operations, such as concatenation or summation [7], which fail to capture intricate inter-modal correlations. Recently, attention mechanisms have been extensively leveraged to facilitate cross-modal interaction, including the middle-level feature fusion (MFF) structure [29], cross-modal attention modules [30], and parallel Transformers for global interaction [26]. Other efforts include heterogeneous fusion with integrity learning [31] and cross-level attention interaction networks [32]. Simultaneously, boundary refinement has emerged as a pivotal research trajectory for enhancing detection precision. Zhou [15] and Zhang [20] proposed boundary enhancement and edge-guided fusion to improve localization. Ren [17] introduced a bio-inspired mask-guided decoder for coarse-to-fine boundary refinement. The proposed MBEFM integrates multi-scale boundary awareness with adaptive feature selection, facilitating high-fidelity modeling of salient object contours.

3 Proposed Model

As illustrated in Fig. 1, the proposed RGBA-Net utilizes an asymmetric encoder-decoder architecture, which primarily consists of three stages: feature extraction (encoding), feature fusion and optimization, and progressive decoding.

During the encoding phase, we employ lightweight EdgeNeXt and MobileNetV3 as the backbone networks for the RGB and depth streams, respectively, yielding multi-scale features denoted as $\{r_1, r_2, r_3\}$ and $\{d_1, d_2, d_3\}$. To mitigate interference from low-quality depth data, depth features are initially processed by the Depth Reliability Gate (DRG) module for quality assessment and filtering, thereby extracting robust geometric cues.

In the feature fusion and optimization stage, RGB features at the same hierarchy and the DRG-refined depth features are fed into the Cross-Modality Synergy (CMS) module to facilitate profound cross-modal interaction via a dual-branch structure, generating fused features $\{fuse_1, fuse_2, fuse_3\}$ that enter the optimization pipeline. These features are first processed by the Boundary Refinement (BR) module

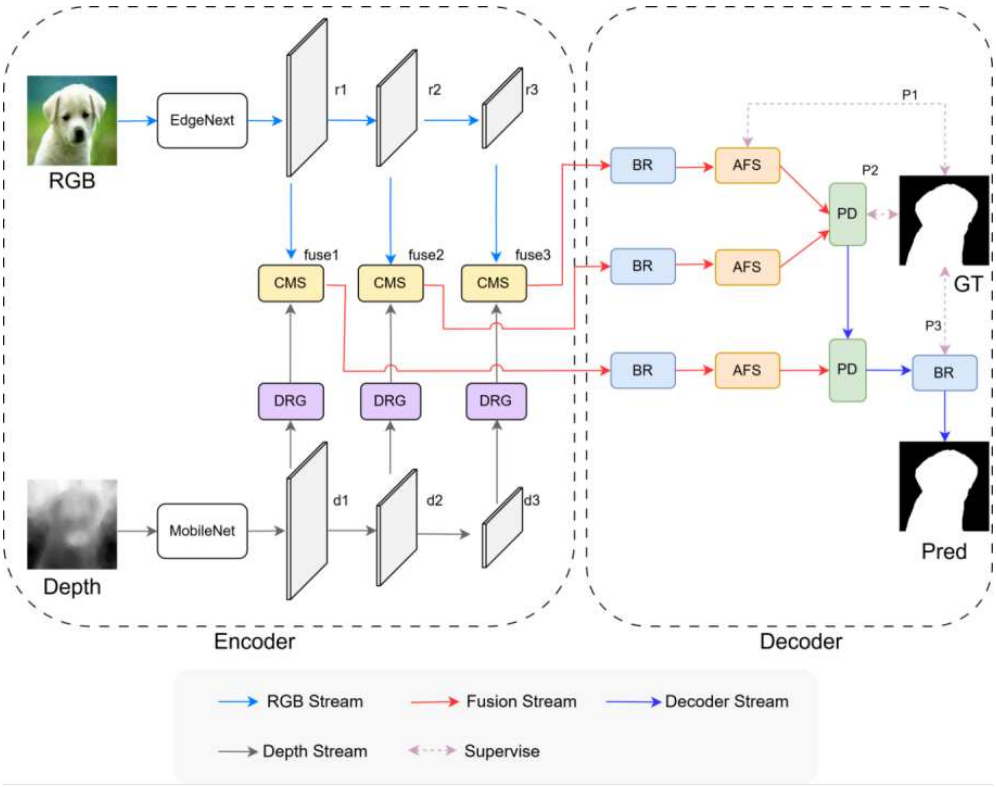


Fig. 1 Overall architecture of the proposed RGBA-Net. It consists of three key components: a Depth Reliability Gate (DRG) for depth quality filtering, a Cross-Modality Synergy (CMS) module for feature interaction, and a Multi-scale Boundary Enhancement Fusion Module (MBEFM) for boundary refinement

to enhance edge responses, followed by the Adaptive Feature Selector (AFS), which dynamically reconciles boundary details with semantic consistency to select the most discriminative feature representations.

In the decoding phase, the network utilizes a Progressive Decoder (PD) to incrementally aggregate the optimized multi-level features in a top-down manner. To ensure the fidelity of feature learning, a multi-level supervision mechanism (P_1, P_2, P_3) is introduced during the decoding process, employing Ground Truth (GT) to provide auxiliary supervision for outputs at various hierarchies. Finally, the aggregated features are processed by a terminal Boundary Refinement (BR) module to further sharpen the predicted boundaries, generating a high-quality final saliency map ($Pred$).

3.1 Depth Reliability Gate

In RGB-D Salient Object Detection (SOD) tasks, depth map quality typically exhibits spatial non-uniformity, particularly at object boundaries, occluded regions, and areas characterized by substantial sensor noise. Unlike traditional strategies that assume

uniform depth reliability, actual confidence levels vary spatially. Depth measurements in high-gradient regions (e.g., object boundaries) frequently exhibit pronounced uncertainty due to the intrinsic physical limitations of depth sensors, whereas depth information in flat regions remains relatively robust.

To address this issue, we propose the Depth Reliability Gate (DRG) module, which adaptively modulates depth feature importance weights based on local gradient characteristics. By quantifying spatial credibility, the DRG module attenuates the contribution of noisy depth information during fusion while amplifying reliable features, thereby bolstering the network's robustness against low-quality depth inputs. Unlike most existing attention mechanisms that indiscriminately fuse depth maps, our DRG module innovatively models the high uncertainty of depth sensors in high-gradient areas as a negative-correlation gating problem. Utilizing fixed-weight edge awareness and a gating relaxation strategy, the DRG achieves "soft-isolation" of depth noise with negligible computational overhead, providing a purified information source for subsequent multi-modal synergy.

Specifically, the DRG commences by evaluating the local uncertainty of depth features using a grouped convolutional layer as a gradient operator, with weights fixed as a Sobel-like operator to extract the local gradient magnitude $G = C_{grad}(D)$. The response map G is averaged across the channel dimension and spatially smoothed via a 3×3 average pooling layer to generate the final reliability assessment map S:

$$S = \text{AvgPool}_{3\times3}\left(\frac{1}{C}\sum_{c=1}^{C} G_c\right) \tag{1}$$

Within S, regions with higher numerical values correspond to high-gradient areas (such as edges or noise), implying diminished reliability.

To transform the assessment map into feature selection weights, we devise a negative-correlation mapping mechanism. A 1×1 convolutional layer transforms S, followed by a Sigmoid function σ to generate the gating weights:

$$R = \sigma\left(\text{Conv}_{1\times1}(-S)\right) \tag{2}$$

The incorporation of a negative sign ensures that high-gradient regions yield weights approaching zero, thereby facilitating effective noise suppression. Furthermore, to prevent the over-suppression of legitimate object boundaries, a gating relaxation mechanism is introduced. We establish a minimum gating threshold γ (set to 0.15 in this study) to ensure that depth information is not entirely discarded even in regions identified with the lowest reliability. The finalized gating weight R is calculated as:

$$R = R \times (1.0 - \gamma) + \gamma \tag{3}$$

This effectively maps the gating range from $(0, 1)$ to $[\gamma, 1]$. Finally, the resultant gating weight map R is integrated with the original input depth features D via element-wise multiplication:

$$D_{out} = D \odot R \tag{4}$$

To visually demonstrate the effectiveness of the proposed DRG module, we present the learned spatial attention map in Fig. 2.

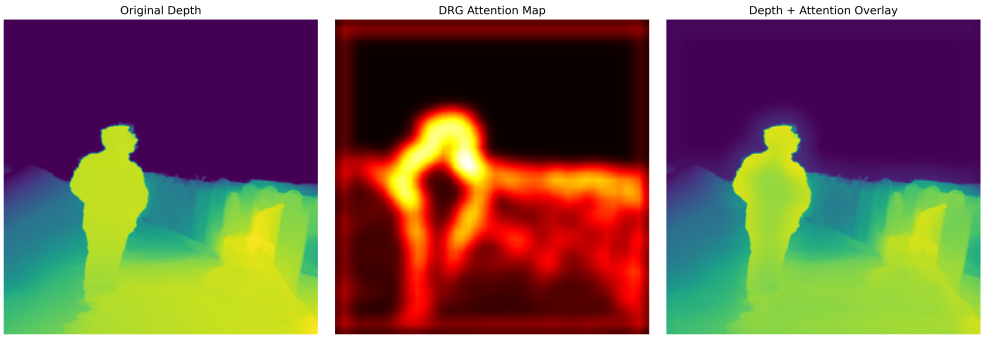


Fig. 2 Visualization of the Depth Reliability Gate (DRG) attention mechanism. From left to right: original input depth map, learned DRG attention map, and overlay of the attention map on the original depth map

As shown in Fig. 2, the leftmost panel shows an imperfect input depth map with inherent sensor noise and sharp depth discontinuities at object boundaries. The middle panel depicts the spatial confidence weights automatically learned by the DRG module, where darker regions correspond to lower reliability. It can be clearly observed that the DRG module assigns significantly lower weights to regions with abrupt depth changes and noise, particularly around object contours. The rightmost overlay further confirms that the DRG effectively suppresses unreliable depth noise and guides the network to focus on high-quality, smooth depth regions, thereby enhancing the robustness of subsequent cross-modal fusion.

Through the DRG module, the network dynamically prioritizes flat, reliable depth regions while mitigating the influence of noisy edges, significantly enhancing the robustness of multi-modal fusion. This mechanism ensures that the subsequent stages of the network receive a "purified" depth representation, effectively preventing the propagation of sensor-induced uncertainties into the high-level semantic fusion process.

3.2 Cross-Modality Synergy Module

Following the separate extraction of RGB features and the reliability-based "purification" of depth features via the DRG, the pivotal challenge lies in achieving efficient and profound fusion of these heterogeneous modalities. Traditional linear fusion often fails to capture non-linear correlations between RGB and depth cues. Such rudimentary operations constrain the model's ability to learn discriminative features, particularly in challenging scenarios where foreground and background exhibit feature similarities in a single modality.

To break the trade-off between feature consistency and informational complementarity, we propose the Cross-Modality Synergy (CMS) mechanism. The CMS module innovatively adopts a decoupled parallel architecture: one branch specializes in excavating shared semantic consensus through deep multiplicative interaction to suppress background noise, while the other preserves modality-specific cues via concatenation to

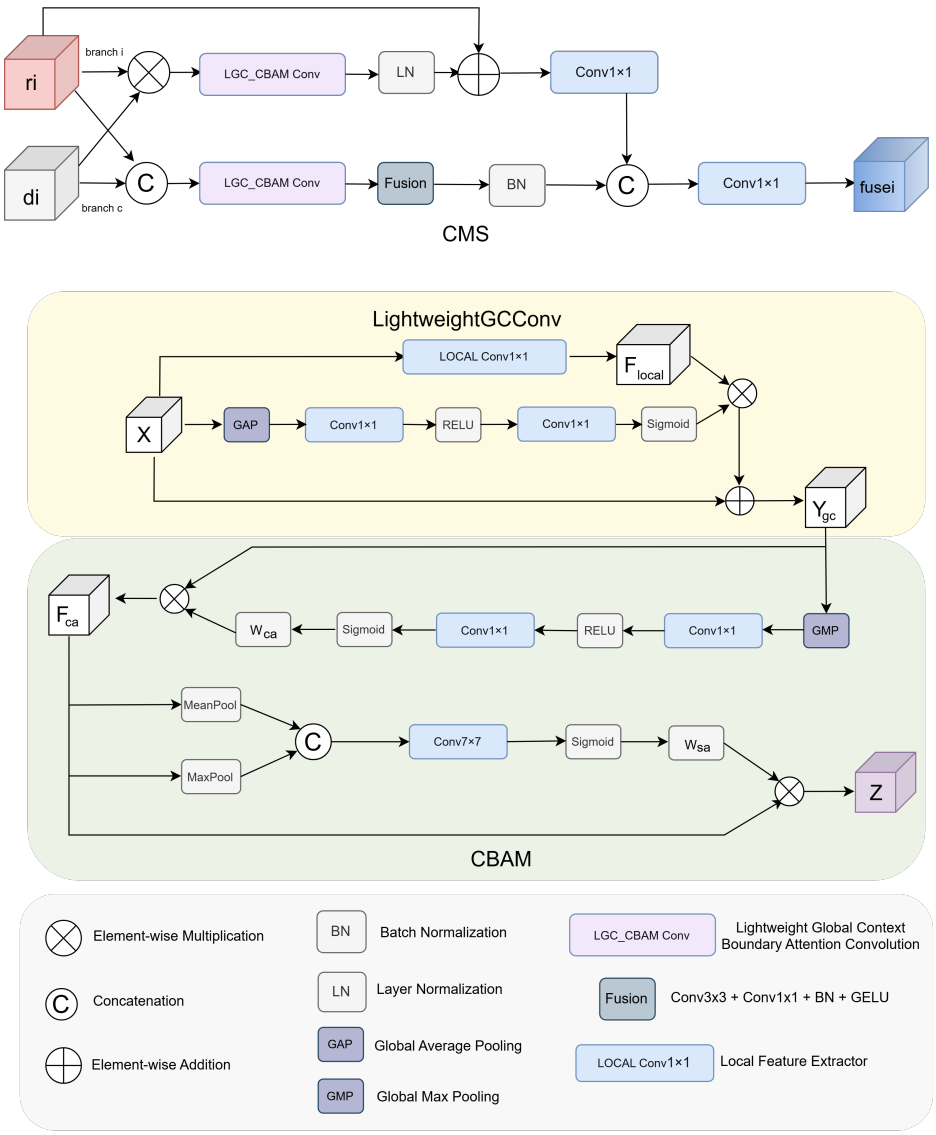


Fig. 3 Detailed structure of the proposed Cross-Modality Synergy (CMS) module and the lightweight LGC-CBAM Conv unit

ensure informational integrity. This dual-branch structure integrates shared semantics and complementary cues for superior representation.

As illustrated in Fig. 3, the CMS module receives RGB features R and DRG-refined depth features D_{out} as inputs. Its structure comprises an Interaction Branch (Branch i) and a Complementary Branch (Branch c).

1. **Interaction Branch (Branch i)**: This branch is dedicated to deeply mining the commonalities and correlations between RGB and depth features. Initially, the

input RGB features r_i and depth features d_i undergo element-wise multiplication to produce a preliminary interaction feature $X = r_i \odot d_i$. This operation effectively highlights spatially co-activated regions across both modalities and is regarded as a potent mechanism for feature selection and enhancement.

Subsequently, the interacted features are fed into our core LGC-CBAM Conv unit (detailed in the lower portion of Fig. 3) for profound refinement. Through LGC-CBAM Conv processing, the interaction features retain local details while acquiring global semantic information, simultaneously bolstering the response of critical channels. The output of the interaction branch F_{sync} can be expressed as:

$$F_{sync} = \text{Conv}_{1\times1}\left(\mathcal{L}_{LGC-CBAM}(r_i \odot d_i) + r_i\right) \tag{5}$$

2. **Complementary Branch (Branch c)**: This branch aims to preserve and integrate raw information from both modalities as a necessary adjunct to the shared semantics of the interaction branch. We concatenate r_i and d_i along the channel dimension to form an aggregated feature containing complete raw information. Next, LGC-CBAM Conv is utilized for feature interaction, followed by channel attention-based fusion via the Fusion Branch and batch normalization. The output of this branch F_{comp} is defined as:

$$F_{comp} = \text{BN}\left(\mathcal{F}_{branch}\left(\mathcal{L}_{LGC-CBAM}([r_i, d_i])\right)\right) \tag{6}$$

Finally, the deep synergistic features F_{sync} and raw complementary features F_{comp} are concatenated and fused via a 1×1 convolution for dimensional reduction, yielding the final CMS output F_{fuse}:

$$F_{fuse} = \text{Conv}_{1\times1}\left([F_{sync}, F_{comp}]\right) \tag{7}$$

3.2.1 LGC-CBAM Conv Unit

LGC-CBAM Conv is a lightweight, attention-augmented convolutional unit tailored for the CMS module, as depicted in the lower portion of Fig. 3. It integrates lightweight global context modeling with channel-spatial attention mechanisms, aiming to capture long-range dependencies at minimal computational cost while adaptively enhancing informative feature channels. The module comprises two primary components:

1. **Lightweight Global Context Convolution (Lightweight GCConv)**: We devised an efficient convolutional strategy to mitigate computational complexity. The structure adaptively adjusts based on kernel size: standard 1×1 convolutions are used when $k = 1$, whereas depthwise separable convolutions are employed when $k > 1$. This design substantially reduces computational complexity from $C_{in} \times C_{out} \times k^2$ to $C_{in} \times k^2 + C_{in} \times C_{out}$, significantly curtailing parameter counts while preserving feature extraction capabilities.

   Building upon this, we introduce global context modeling. Initially, local spatial features are extracted via convolution, while channel-level statistics are captured using Global Average Pooling (GAP). Two 1×1 convolutions then generate channel attention weights to achieve adaptive feature calibration:

$$Y_{gc} = F_{local} \cdot \sigma\left(\text{Conv}_{1\times1}\left(\text{ReLU}\left(\text{Conv}_{1\times1}\left(\text{GAP}(X)\right)\right)\right)\right) + X \tag{8}$$

2. **Lightweight Convolutional Block Attention Module (CBAM)**: To further strengthen channel discriminability, we introduce a lightweight CBAM. This module first calibrates features via channel attention: diverging from traditional methods that utilize both average and max pooling, we found that Global Max Pooling (GMP) alone effectively captures salient feature statistics, thereby further reducing computational overhead. Channel attention weights W_{ca} are generated via a two-layer MLP for secondary feature weighting:

$$W_{ca} = \sigma\left(\text{Conv}_{1\times1}\left(\text{ReLU}\left(\text{Conv}_{1\times1}\left(\text{GMP}(Y_{gc})\right)\right)\right)\right) \tag{9}$$

Subsequently, the spatial attention mechanism generates spatial context via average and max pooling along the channel dimension, followed by a 7×7 convolution to produce spatial weights W_{sa} for adaptive spatial enhancement. The final output feature Z is calculated as:

$$W_{sa} = \sigma\left(\text{Conv}_{7\times7}\left(\text{Concat}\left(\text{AvgPool}(Y_{gc}), \text{MaxPool}(Y_{gc})\right)\right)\right) \tag{10}$$

$$Z = Y_{gc} \otimes W_{ca} \otimes W_{sa} \tag{11}$$

Where $\otimes$ denotes element-wise multiplication.

Through this serial design, LGC-CBAM Conv first utilizes GCConv to extract and calibrate base features, followed by fine-grained dual-attention enhancement via CBAM, achieving efficient feature distillation and reinforcement.

Overall, the CMS module ingeniously decouples deep interaction and information preservation into two parallel branches, which are subsequently integrated efficiently via concatenation and dimensional reduction. This enables the model to learn complex cross-modal associations without sacrificing modality-specific cues—which are equally vital for saliency judgment—thereby generating high-fidelity fused features.

3.3 Multi-scale Boundary Enhancement Fusion Module

Following successful cross-modal feature fusion, the effective integration of multi-scale information and the precise capture of salient object boundary structures emerge as critical challenges for enhancing detection performance. Traditional multi-scale fusion methodologies, such as the Feature Pyramid Network (FPN), are capable of integrating semantic information across various levels but frequently neglect the preservation and enhancement of boundary details. This oversight results in blurred edges in the generated saliency maps, particularly yielding suboptimal performance in complex backgrounds or low-contrast scenarios.

Furthermore, features at different hierarchies possess distinct receptive fields and semantic granularities, making rudimentary upsampling and concatenation insufficient for achieving adaptive inter-feature fusion. As network depth increases, high-level features attain robust semantics at the cost of severe dilution of spatial boundary details. Existing FPN variants tend to treat all hierarchies indiscriminately or employ static fusion strategies. In practice, however, low-level features are better suited for boundary

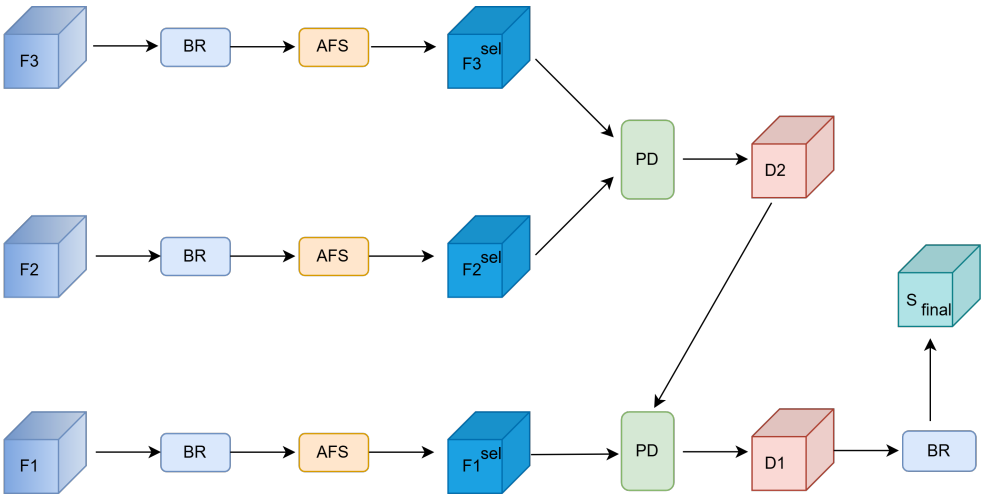


Fig. 4 Overall architecture of the proposed Multi-scale Boundary Enhancement Fusion Module (MBEFM)

delineation, whereas high-level features excel at capturing global semantics. Moreover, the degree of reliance on boundary versus semantic cues varies significantly across different input images, necessitating a more dynamic and awareness-driven fusion mechanism.

To address the gradual attrition of boundary information as network depth increases in conventional FPNs, we propose the Multi-scale Boundary Enhancement Fusion Module (MBEFM). MBEFM introduces a dual mechanism comprising explicit boundary modeling and Adaptive Feature Selection (AFS). Through a hierarchical processing workflow, the module sequentially executes boundary-aware enhancement, adaptive feature fusion, and feature refinement, ultimately yielding high-fidelity saliency features via progressive decoding. The overall architecture of MBEFM is illustrated in Fig. 4, with its core design tenets as follows: 1) explicitly modeling and augmenting multi-scale boundary information; 2) facilitating adaptive feature fusion via attention mechanisms; 3) reconciling boundary details with semantic consistency during feature selection; and 4) employing a lightweight decoding strategy for efficient feature integration.

3.3.1 Boundary Refinement Module

To explicitly model and bolster boundary information across multi-scale features, we incorporate a Boundary Refinement (BR) module at each level of the feature pyramid. This module employs multi-scale edge detection operators to extract boundary responses at varying granularities from the input features, while utilizing an adaptive weight-learning mechanism to reinforce edge structures relevant to the salient objects.

Specifically, given the i-th level fused feature $F_i \in \mathbb{R}^{C_i \times H_i \times W_i}$ from the CMS module, we apply multi-scale edge detection operators—specifically, fixed 3×3 and 5×5 Sobel kernels—to extract local and global edge responses, denoted as $E_{i,1}$ and

$E_{i,2}$, respectively. Subsequently, these multi-scale edge features are concatenated and fused through a lightweight 1×1 convolutional layer to generate a unified edge map E_i:

$$E_i = \text{Conv}_{1\times1}\left(\text{Concat}(E_{i,1}, E_{i,2})\right) \tag{12}$$

To adaptively enhance boundary regions within the original feature maps, we developed a boundary importance gating mechanism that generates a spatial attention weight map W_{ib} based on the edge feature E_i. Consequently, the boundary-enhanced feature F_{ib} is derived via a weighted residual connection as follows:

$$W_{ib} = \sigma\left(\text{Conv}_{1\times1}\left(\text{GELU}\left(\text{Conv}_{1\times1}(E_i)\right)\right)\right) \tag{13}$$

$$F_{ib} = F_i \odot (1 + W_{ib}) + F_i \tag{14}$$

This design empowers the model to focus on boundary regions while simultaneously preserving underlying semantic information, ensuring that edge localization is sharpened without compromising global contextual consistency.

3.3.2 Adaptive Feature Selection and Progressive Decoding

Conventional Feature Pyramid Networks (FPNs) typically employ simplistic linear summation for feature fusion, thereby neglecting the complementary relationships between features across different scales. To address this, we leverage a dual-attention mechanism—incorporating both channel and spatial attention—to dynamically modulate feature weights.

To refine the boundary-enhanced features F_{ib}, standard channel and spatial attention mechanisms (as described in Sec. 3.2.1) are applied to dynamically modulate feature weights. This dual-attention approach re-calibrates F_{ib} to emphasize salient regions while suppressing background noise.

The resulting feature is integrated via element-wise multiplication with the dual-attention weights. To further augment the representational capacity of the features, we employ depthwise separable convolutions and channel shuffle operations to bolster inter-feature information exchange, resulting in the enhanced multi-scale feature F_{enh}:

$$F_{enh} = \text{ChannelShuffle}\left(\text{DWConv}\left(\text{PWConv}(F_{fus})\right)\right) \tag{15}$$

Considering the diverse requirements for boundary precision and semantic consistency across various datasets, we propose the Adaptive Feature Selection (AFS) module, whose internal architecture is depicted in Fig. 5. The AFS module initially extracts global statistical descriptors from the enhanced feature F^{enh} by employing parallel Global Average Pooling (GAP) and Global Max Pooling (GMP) layers. These descriptors are concatenated and sequentially processed through a CBG unit (comprising 1×1 Conv, Batch Normalization, and GELU activation) and a 1×1 convolutional layer. Subsequently, a SoftMax function is applied to dynamically generate two complementary weighting coefficients, α and $1 - \alpha$.

The module incorporates two specialized parallel processing streams to handle different image characteristics: the upper branch utilizes a 3×3 Lightweight Global

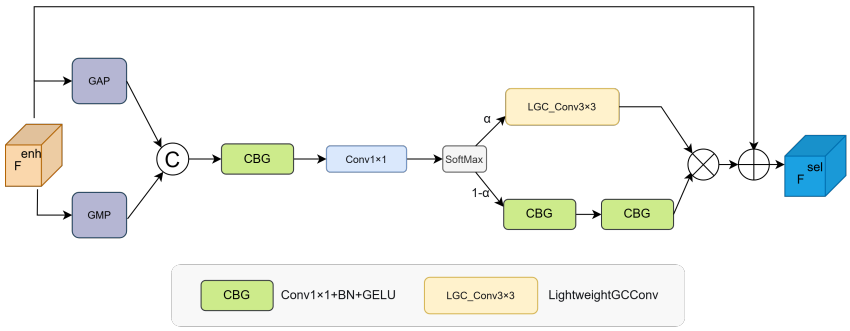


Fig. 5 Detailed internal structure of the proposed Adaptive Feature Selection (AFS) module

Context Convolution (LGC_Conv3 × 3) to emphasize high-frequency boundary details, while the lower branch comprises two sequential CBG blocks focused on smoothing and maintaining global semantic consistency. The outputs of these two branches are weighted by their respective coefficients and integrated with the original input feature F^{enh} through a residual connection. This formulation is expressed as:

$$F^{sel} = \alpha \cdot \mathrm{LGC}(F^{enh}) + (1-\alpha) \cdot \mathrm{CBG}(\mathrm{CBG}(F^{enh})) + F^{enh} \tag{16}$$

Upon obtaining the multi-scale refined features $\{F_1^{sel}, F_2^{sel}, F_3^{sel}\}$, where lower indices indicate higher spatial resolutions, a Progressive Decoder (PD) is employed to integrate these features in a top-down manner. To maintain a lightweight footprint, each PD unit utilizes our proposed Lightweight GCConv in place of conventional convolutions, facilitating large receptive field context aggregation while significantly reducing computational complexity. The decoding sequence is formulated as:

$$D_2 = \mathrm{PD}\left(F_2^{sel}, F_3^{sel}\right) \tag{17}$$

$$D_1 = \mathrm{PD}\left(F_1^{sel}, D_2\right) \tag{18}$$

Each decoder unit incorporates feature upsampling, lightweight convolution, channel attention, and residual connections to ensure the simultaneous preservation of fine structural details and high-level semantics. Finally, to further sharpen the boundaries of the final prediction, we re-employ the Boundary Refinement (BR) module at the highest-resolution output layer. This "dual-boundary optimization strategy," which spans the entire pipeline from initial feature extraction to final prediction output, markedly enhances the localization precision for complex object contours, yielding an edge-sharp and internally consistent saliency map S_{final}:

$$S_{final} = D_1 \odot (1 + W_b) + D_1 \tag{19}$$

3.4 Loss Function

To guide the model in generating high-quality saliency maps characterized by sharp boundaries and internal consistency, we designed a hybrid loss function that integrates pixel-level classification, region overlap metrics, and boundary-aware constraints.

Furthermore, to facilitate effective multi-scale feature learning, we employ a deep supervision strategy by imposing supervisory signals on the outputs of all three decoding stages. The total loss function $\mathcal{L}_{total}$ is formulated as the weighted summation of the losses across these scales:

$$\mathcal{L}_{total} = \sum_{k=1}^{3} \mathcal{L}_{stage}^{(k)} \tag{20}$$

where $k \in \{1, 2, 3\}$ denotes the hierarchical index of the decoder, with outputs corresponding to a range from coarse-to-fine spatial resolutions.

Each stage-wise loss $\mathcal{L}_{stage}$ comprises four components: weighted Binary Cross-Entropy (BCE) loss $\mathcal{L}_{BCE}$, Intersection-over-Union (IoU) loss $\mathcal{L}_{IoU}$, Dice loss $\mathcal{L}_{Dice}$, and boundary-aware loss $\mathcal{L}_{Edge}$:

$$\mathcal{L}_{stage} = \mathcal{L}_{BCE} + \mathcal{L}_{IoU} + \mathcal{L}_{Dice} + \lambda \mathcal{L}_{Edge} \tag{21}$$

Here, λ serves as the balancing coefficient for the boundary loss, which is empirically set to 0.2 in our experiments. The individual loss components are detailed as follows:

3.4.1 Pixel-level Classification Loss

We employ the Binary Cross-Entropy (BCE) loss to quantify the pixel-wise discrepancy between the predicted saliency map P and the ground truth G, providing a fundamental supervisory signal:

$$\mathcal{L}_{BCE} = -\frac{1}{N} \sum_{(x,y)} [G(x,y) \log(\sigma(P(x,y))) + (1 - G(x,y)) \log(1 - \sigma(P(x,y)))] \tag{22}$$

where (x, y) denotes the pixel coordinates, N is the total number of pixels, and $\sigma(\cdot)$ represents the Sigmoid activation function.

3.4.2 Region Consistency Losses

Relying solely on the BCE loss may cause the model to converge to local optima and neglect the global structural integrity when processing large-scale salient objects. To address this, we incorporate Intersection-over-Union (IoU) and Dice losses. The IoU loss directly optimizes the degree of overlap between the predicted and ground-truth regions:

$$\mathcal{L}_{IoU} = 1 - \frac{\sum_{(x,y)} P(x,y) \cdot G(x,y)}{\sum_{(x,y)} [P(x,y) + G(x,y) - P(x,y) \cdot G(x,y)]} \tag{23}$$

Conversely, the Dice loss mitigates the foreground-background class imbalance by utilizing an F1-score formulation:

$$\mathcal{L}_{Dice} = 1 - \frac{2 \sum_{(x,y)} P(x,y) \cdot G(x,y) + \epsilon}{\sum_{(x,y)} P(x,y) + \sum_{(x,y)} G(x,y) + \epsilon} \tag{24}$$

where ϵ is a small constant (typically 1×10^{-6}) added to avoid division by zero.

3.4.3 Boundary-Aware Loss

To further sharpen salient object boundaries and alleviate the pervasive issue of edge blurring inherent in traditional methods, we introduce an explicit boundary constraint. This loss term initially utilizes Sobel operators to extract edge response maps, denoted as P_e and G_e, from the prediction and ground truth, respectively:

$$P_e = \left|\mathrm{Conv}_{sobel}(P)\right|, \quad G_e = \left|\mathrm{Conv}_{sobel}(G)\right| \tag{25}$$

Subsequently, these edge response maps are normalized, and their pixel-wise BCE loss is computed:

$$\mathcal{L}_{Edge} = \mathrm{BCE}\left(P_{edge}^{norm}, G_{edge}^{norm}\right) \tag{26}$$

This design compels the model to prioritize high-frequency boundary information during training, thereby yielding saliency predictions with more precise contours.

4 Experimental Settings

4.1 Datasets

We evaluate the performance of RGBA-Net across seven benchmark RGB-D datasets. DUT-RGBD comprises 1,200 image pairs collected from real-world scenes; LFSD contains 100 pairs with a spatial resolution of 360×360; SIP includes 929 pairs featuring multiple individuals in outdoor environments; STERE consists of 1,000 stereoscopic images with corresponding saliency masks; SSD covers 80 pairs spanning both indoor and outdoor settings; NJU2K is a large-scale dataset with 1,985 pairs; and NLPR provides 1,000 pairs showcasing various salient objects.

Following the experimental protocol established in MoADNet [14], we partition these datasets by selecting 800, 700, and 1,485 pairs from DUT-RGBD, NLPR, and NJU2K, respectively, to construct a comprehensive training set, while all remaining images across the seven datasets are reserved for testing.

4.2 Evaluation Metrics

We employ five widely recognized metrics to quantitatively assess the detection performance of various models: Structure-measure (S_α), Precision-Recall (PR) curves, mean E-measure (E_ξ^{mean}), Mean Absolute Error (M), and mean F-measure (F_β^{mean}). In addition to these accuracy-oriented metrics, we also evaluate efficiency-related parameters, including the total parameter count (Params), computational complexity in terms of floating-point operations (FLOPs), and inference speed measured by frames per second (FPS), to provide a holistic comparison of model efficiency.

4.3 Implementation Details

RGBA-Net is implemented in PyTorch and trained on a single NVIDIA RTX 4090 GPU. During both the training and inference phases, input RGB and depth images are

resized to 384×384 pixels. To bolster model generalization and mitigate overfitting, we adopt a comprehensive data augmentation strategy, including random flipping, cropping, rotation, and color jittering.

The model is trained for 300 epochs using AdamW (weight decay 1×10^{-4}, batch size 32): the initial learning rate is set to 5×10^{-4} for the first 50 epochs, decaying to 5×10^{-5} for the next 150 epochs, and finally to 5×10^{-6} for the remaining 100 epochs. A cosine annealing warm-restart scheduler is utilized for dynamic learning rate adjustment. The entire training procedure requires approximately 7 hours. For supervision, we utilize the previously described hybrid loss function to monitor multi-level predictions.

4.4 Comparison with State-of-the-Art Methods

We evaluate the proposed RGBA-Net against 12 state-of-the-art (SOTA) models across seven benchmark public datasets. For a fair comparison, all models were trained on a unified dataset comprising 2,985 image pairs from DUT-RGBD, NLPR, and NJU2K, following the training protocols specified by the respective authors. To measure inference speed, we standardized the evaluation environment using a batch size of 1, FP32 precision, an input resolution of 384×384, and the PyTorch framework. All speed benchmarks were conducted on a single NVIDIA GeForce RTX 4090 GPU.

The quantitative comparison result of the 12 most recent RGB-D SOD models with RGBA-Net on 7 commonly used datasets is presented in Table 1.

Table 1 Quantitative comparison with state-of-the-art methods. **Red bold**: best, **Blue bold**: second best, **Black bold**: third best

Datasets	Metric	Large Models						Lightweight Models						
		SPSN [25] ECCV22	CIRNet [10] TIP22	CAVER [33] TIP23	TPCL [26] TMM24	STANet [27] IJON24	BCNet [31] PR25	MobileSal [13] TPAMI22	MoADNet [14] TCSVT22	MFF [29] TIP22	LSNet [15] TIP23	MAGNet [16] KBS24	BTNet [17] NN25	**Ours** -
Params (M) ↓		37.0	103.2	55.5	112.7	85.0	–	6.5	5.0	**3.9**	4.6	5.2	**3.8**	**3.49**
Speed (FPS) ↑		7.0	42.4	26.6	23.8	10.0	–	**64.6**	52.5	49.6	**71.4**	56.6	63.3	**66.7**
NLPR	E_ξ ↑	0.950	0.955	**0.961**	**0.963**	0.957	**0.966**	0.961	0.952	0.949	0.952	0.954	0.956	**0.964**
	S_α ↑	0.919	**0.934**	**0.929**	0.925	0.924	0.932	0.920	0.916	0.908	0.921	0.912	**0.926**	**0.929**
	F_β ↑	0.884	0.884	**0.921**	0.908	0.918	**0.923**	0.908	0.900	0.895	0.881	0.879	0.915	**0.919**
	M ↓	0.024	0.023	**0.022**	0.023	**0.021**	**0.021**	0.025	0.027	0.031	0.024	0.026	0.022	**0.021**
NJU2K	E_ξ ↑	0.910	0.926	0.949	0.928	0.947	**0.958**	0.919	0.942	0.935	0.926	0.921	0.944	**0.951**
	S_α ↑	0.916	**0.925**	0.920	0.916	0.919	0.921	0.906	0.904	0.901	0.912	0.904	**0.925**	**0.917**
	F_β ↑	0.874	0.906	**0.923**	0.918	0.920	0.918	0.904	0.905	0.910	0.900	0.893	0.913	**0.920**
	M ↓	0.038	**0.034**	0.035	0.038	**0.034**	**0.032**	0.042	0.041	0.039	0.036	0.039	0.035	**0.034**
SIP	E_ξ ↑	0.912	0.917	0.943	**0.945**	0.929	0.937	0.910	0.916	0.905	0.928	0.917	0.923	**0.937**
	S_α ↑	0.828	0.888	**0.904**	0.900	0.891	0.901	0.867	0.877	0.864	0.889	0.873	0.893	**0.896**
	F_β ↑	0.826	0.874	**0.915**	0.903	0.906	0.911	0.889	0.898	0.874	0.887	0.878	0.893	**0.906**
	M ↓	0.062	0.053	**0.040**	**0.039**	0.044	0.041	0.051	0.048	0.054	0.046	0.052	0.046	**0.042**
LFSD	E_ξ ↑	0.839	0.902	0.914	0.897	0.896	–	0.856	0.867	0.844	0.891	0.899	0.903	**0.925**
	S_α ↑	0.825	0.875	0.876	**0.885**	0.861	–	0.830	0.845	0.831	0.858	0.865	0.878	**0.888**
	F_β ↑	0.782	0.879	0.880	0.877	0.868	–	0.804	0.821	0.811	0.860	0.867	**0.882**	**0.883**
	M ↓	0.085	0.068	0.061	0.059	0.070	–	0.079	0.072	0.079	0.068	**0.063**	**0.063**	**0.054**
STERE	E_ξ ↑	0.923	0.929	0.931	0.928	**0.952**	0.944	0.940	0.942	0.912	0.924	0.921	0.938	**0.953**
	S_α ↑	0.902	0.916	0.914	0.916	0.916	0.904	0.903	0.899	0.878	0.893	0.895	**0.920**	**0.913**
	F_β ↑	0.867	0.890	0.897	0.896	**0.916**	0.905	0.895	0.892	0.874	0.878	0.867	0.894	**0.914**
	M ↓	0.038	0.039	**0.033**	0.034	**0.033**	0.039	0.041	0.042	0.049	0.043	0.045	0.038	**0.034**
DUT-D	E_ξ ↑	0.944	0.951	**0.963**	0.960	–	0.956	0.939	0.958	0.943	0.947	0.939	0.957	**0.967**
	S_α ↑	0.910	0.932	0.927	**0.936**	–	0.927	0.905	0.927	0.902	0.921	0.912	0.929	**0.937**
	F_β ↑	0.898	0.922	**0.936**	0.935	–	0.934	0.895	0.933	0.879	0.916	0.906	0.931	**0.945**
	M ↓	0.036	0.031	**0.026**	**0.025**	–	0.031	0.040	0.031	0.043	0.032	0.035	0.028	**0.024**
SSD	E_ξ ↑	0.876	0.897	0.901	0.899	0.906	–	0.904	0.902	0.883	0.910	0.913	0.921	**0.925**
	S_α ↑	0.832	0.861	0.880	0.878	0.880	–	0.865	0.854	0.851	0.868	0.858	**0.879**	**0.882**
	F_β ↑	0.793	0.816	0.863	**0.867**	0.870	–	0.848	0.840	0.834	0.840	0.832	0.852	**0.871**
	M ↓	0.063	0.054	0.059	**0.053**	**0.049**	–	0.052	0.057	0.061	0.049	0.050	0.042	**0.046**

Comparison with Large-scale Models: As summarized in Table 1, although RGBA-Net is optimized for lightweight design, its performance is comparable to, and in several cases surpasses, existing large-scale models. Specifically, compared to TPCL [26], our

model possesses only 3.1% of the parameter count while achieving superior detection precision. Notably, on the challenging DUT-RGBD dataset, RGBA-Net attains the best performance across all four evaluation metrics, significantly outperforming TPCL. In comparison with CIRNet [10], RGBA-Net curtails the parameter count by approximately 96.6% and boosts inference speed by 57.3%. Furthermore, compared to the recent BCNet [31], although results are competitive, RGBA-Net attains a higher E_ξ and a lower M on the STERE dataset, demonstrating its enhanced robustness in complex scenarios. Overall, RGBA-Net achieves comparable or superior performance to large models with significantly lower computational overhead, validating its high efficiency.

Comparison with Lightweight Models: Within the lightweight category, RGBA-Net exhibits a distinct advantage, maintaining the minimal parameter footprint (3.49M) among all compared methods while offering a highly competitive inference speed of 66.7 FPS. Compared to the state-of-the-art BTNet [17], RGBA-Net further reduces the parameter count by 8.2% and achieves superior performance across multiple datasets. For instance, on the NLPR dataset, our E_ξ and F_β scores improve by 0.8% and 0.4%, respectively. Compared to MobileSal [13], RGBA-Net reduces the parameter count by 46.3% while improving the F_β on the SIP dataset from 0.889 to 0.906. Even when compared to LSNet [15], which features the highest speed (71.4 FPS), RGBA-Net maintains a slightly lower speed but offers an overwhelming advantage in accuracy, particularly on DUT-RGBD where our F_β reaches 0.945, far exceeding LSNet's 0.916.

In conclusion, RGBA-Net sets new SOTA benchmarks for lightweight models on the DUT-RGBD and STERE datasets. As illustrated in Fig. 7, our Precision-Recall (PR) curves consistently outperform most advanced methods, further verifying the superiority of our approach. Additionally, visual comparisons in Fig. 6 demonstrate that while other models may fail in noisy scenarios or complex depth conditions, our method effectively extracts salient objects with high structural integrity.

4.5 Ablation Studies

To verify the individual contribution and effectiveness of each module within the proposed RGBA-Net, we conducted extensive ablation studies on the SIP and LFSD datasets. Furthermore, we evaluated efficiency-related metrics, including parameter counts (Params), floating-point operations (FLOPs), and inference speed (FPS). Detailed quantitative results of these ablation experiments are summarized in Table 2.

(1) Effectiveness of DRG: To validate the role of the DRG module in filtering low-quality depth information, we conducted a targeted ablation study. Specifically, we removed the DRG module such that depth features were directly fed into subsequent layers without being subjected to reliability-based screening or weighting. In the absence of DRG, the model lacks the capability to dynamically suppress low-quality or noisy depth regions based on local gradient characteristics. As indicated by the quantitative analysis in Table 2, the removal of DRG results in a conspicuous performance degradation across all evaluation metrics on the SIP and LFSD datasets. These results confirm DRG's effectiveness in mitigating low-quality depth interference. Furthermore, qualitative analysis in Fig. 8 reveals that models lacking DRG are prone to false detections when confronted with significant depth noise, whereas the integration

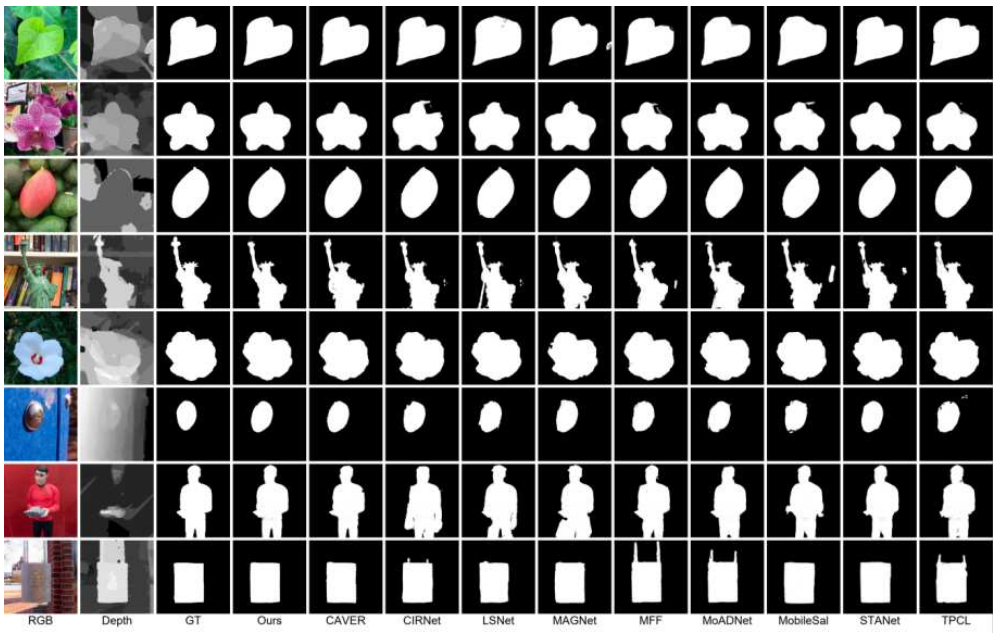


Fig. 6 Qualitative comparison with state-of-the-art methods on challenging samples

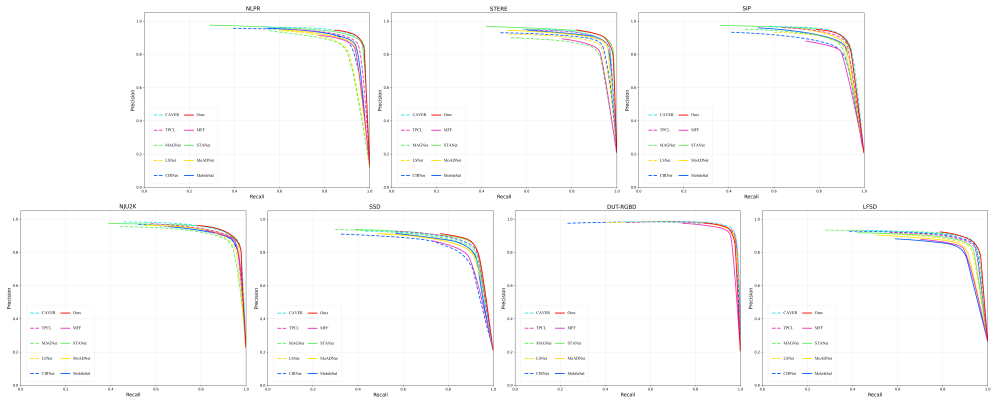


Fig. 7 PR curves comparison of the proposed model (bold red line) against other models across 7 datasets

of DRG enables the model to focus on salient objects, markedly bolstering prediction robustness.

(2) Effectiveness of CMS: To evaluate the contribution of the CMS module to multi-modal feature fusion, we compared the full RGBA-Net with a variant that utilizes a simple concatenation (Concat) operation. In the variant without CMS, RGB and depth features are linearly combined via channel-wise concatenation, thereby precluding

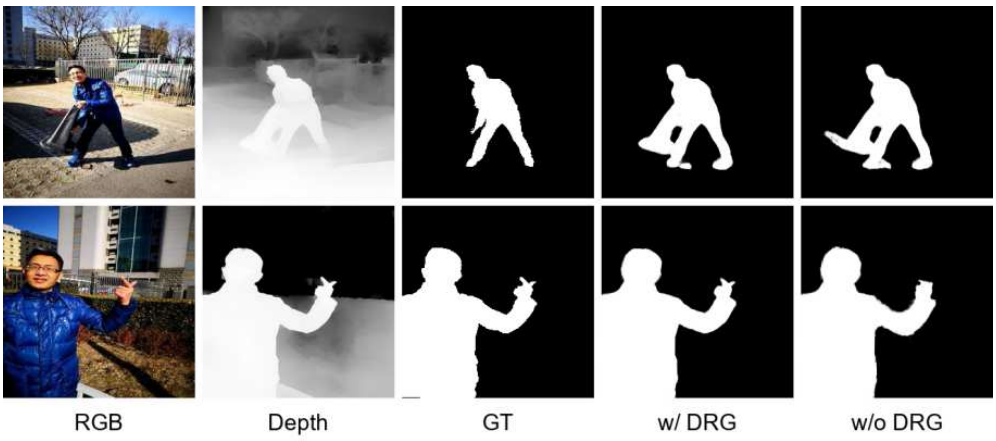


Fig. 8 Qualitative comparison of the DRG module ablation

deep cross-modal interaction and adaptive complementary enhancement. Specifically, CMS facilitates nonlinear synergy through depthwise separable convolutions and dual-attention mechanisms. Quantitative comparisons in Table 2 demonstrate that omitting the CMS module leads to a substantial performance decline in complex scenarios, such as those with low contrast or fluctuating illumination; notably, the Mean Absolute Error (M) increases significantly on the LFSD dataset. Visualizations in Fig. 9 further confirm that models lacking CMS often fail to fully exploit depth cues to compensate for deficiencies in the RGB modality, resulting in incomplete object detection. In contrast, the integration of CMS achieves successful inter-modal synergy, yielding more consistent and prominent saliency maps.

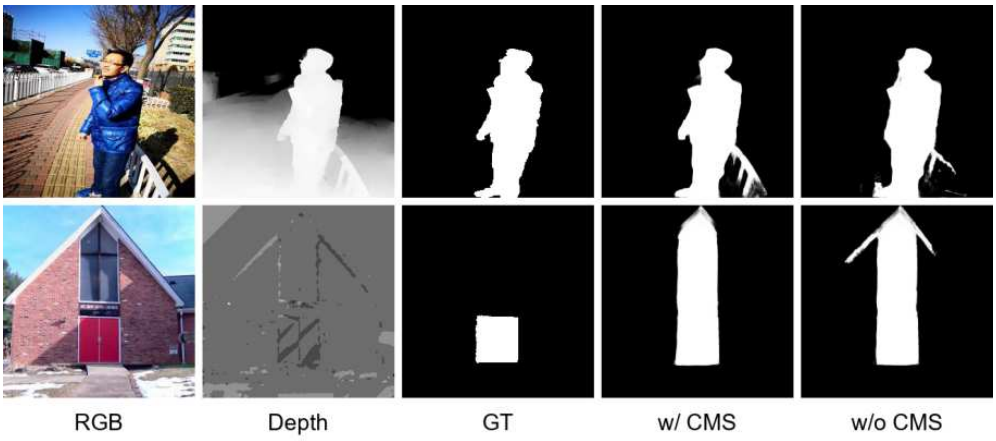


Fig. 9 Qualitative comparison of the CMS module ablation

(3) Effectiveness of MBEFM: To investigate the impact of the MBEFM module on boundary localization precision, we compared the full network with a baseline version devoid of this module. Without MBEFM, the network no longer leverages high-frequency edge information extracted by multi-scale Sobel operators to refine the fused features, nor does it utilize adaptive boundary gating to enhance object contours. Instead, feature maps are processed through standard convolutional layers. Quantitative results in Table 2 indicate that the absence of MBEFM precipitates a significant drop in performance metrics, particularly in scenes with intricate object boundaries. Qualitative comparisons in Fig. 10 clearly illustrate that saliency maps generated without MBEFM exhibit blurred and coarse boundaries. In contrast, the integration of MBEFM yields sharp predictions that align closely with the actual object contours, validating the module's core value in fine-grained edge preservation.

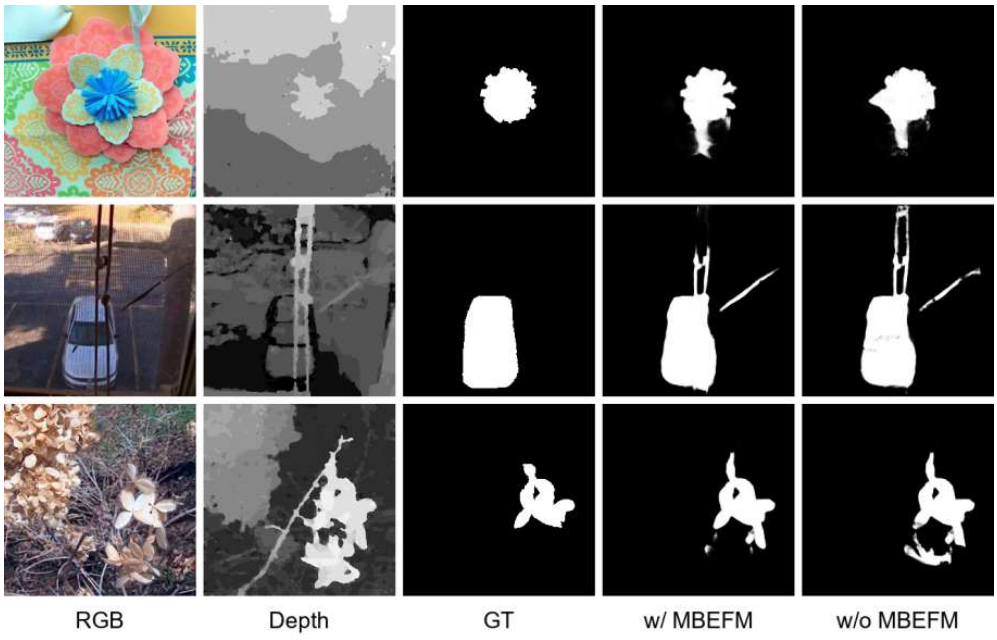


Fig. 10 Qualitative comparison of the MBEFM module ablation

(4) Synergistic Effect of Modules: Finally, we evaluated the combined impact of the DRG, CMS, and MBEFM modules. Experimental results demonstrate that simultaneously enabling all three modules yields the superior performance on the LFSD and SIP datasets, outperforming any individual module or partial combination. This performance gain is attributed to the mutual reinforcement between components: DRG first ensures the purity of input depth features, providing a high-quality complementary source for CMS; CMS subsequently achieves deep inter-modal fusion to construct robust semantic representations; and finally, MBEFM builds upon these features to sharpen boundaries and ensure structural precision. The organic integration of these

three modules enables RGBA-Net to achieve comprehensive and precise salient object detection while maintaining a highly efficient, lightweight architecture.

Table 2 Ablation study of the proposed modules on the LFSD and SIP datasets.

NO.	DRG	CMS	MBEFM	Params (M) ↓	FLOPs (G) ↓	FPS ↑	LFSD				SIP			
							$E_\xi \uparrow$	$S_\alpha \uparrow$	$F_\beta \uparrow$	$M \downarrow$	$E_\xi \uparrow$	$S_\alpha \uparrow$	$F_\beta \uparrow$	$M \downarrow$
1	–	✓	✓	3.49	4.30	68.8	0.916	0.875	0.871	0.058	0.930	0.885	0.894	0.046
2	✓	–	✓	3.18	3.95	78.5	0.898	0.862	0.855	0.061	0.908	0.872	0.879	0.048
3	✓	✓	–	3.32	4.15	73.0	0.905	0.870	0.868	0.059	0.915	0.878	0.886	0.049
4	✓	✓	✓	3.49	4.32	66.7	**0.925**	**0.888**	**0.883**	**0.054**	**0.937**	**0.896**	**0.906**	**0.042**

4.6 Failure Analysis and Future Work

Although RGBA-Net achieves superior detection performance across most benchmark datasets, it still encounters challenges when processing exceptionally complex scenarios. Figure 11 illustrates two representative failure cases, revealing the inherent limitations of the model under specific conditions.

(1) Challenges in low-contrast scenarios: As shown in the first row, detecting salient objects in texture-heavy backgrounds remains difficult when depth maps lack sufficient discriminative geometric contrast. The depth sensor fails to provide reliable structural cues in these cases, causing the lightweight encoder to struggle with spatial guidance and producing blurred boundaries.

(2) Slender structure preservation: The second row demonstrates a limitation in capturing thin objects (e.g., a traffic cone) when their depth values merge with the background. Although the human torso is correctly identified, the thin structure's depth information is severely diluted, leading to its omission in the final prediction.

Summary and Outlook: These observations indicate that the efficacy of RGBA-Net remains partially dependent on depth map quality. Future research will focus on integrating frequency-domain enhancement (e.g., Fast Fourier Transform) to restore high-frequency boundary details and exploring depth restoration modules for enhanced geometric guidance.

5 Conclusion

In this paper, we propose RGBA-Net, a lightweight and efficient boundary-aware reliability-gated network for RGB-D salient object detection. Possessing only 3.49M parameters and a computational complexity of 4.32 GFLOPs, the model achieves a real-time inference speed of 66.7 FPS. Specifically, the network leverages an asymmetric dual-stream encoder to extract multi-modal features and introduces the Depth Reliability Gate (DRG) to adaptively evaluate depth map quality, effectively mitigating the impact of depth noise. Furthermore, the Cross-Modality Synergy (CMS) module facilitates profound feature fusion via LGC-CBAM convolutions, while the Multi-scale Boundary Enhancement Fusion Module (MBEFM) explicitly reinforces object contours. Finally, the Progressive Decoder (PD) and Adaptive Feature Selection (AFS) module dynamically balance boundary details with semantic consistency. Extensive evaluations on seven benchmark datasets demonstrate that RGBA-Net

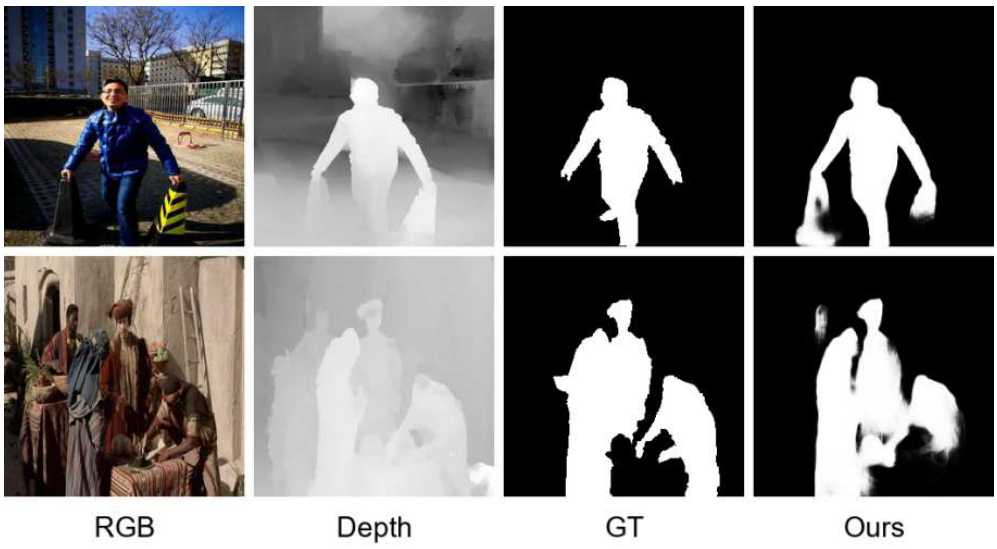


Fig. 11 Two representative failure cases of the proposed RGBA-Net

outperforms existing state-of-the-art methods while maintaining a minimal computational footprint. Future work will focus on developing more sophisticated depth-quality assessment strategies and exploring further model compression to support seamless deployment on resource-constrained mobile platforms.

Declarations

Funding This work was supported by the National Natural Science Foundation of China under Grant No. U1936213, Program of Shanghai Academic Research Leader No. 21XD1421500, and Shanghai Science and Technology Commission Project No. 20020500600.

Competing interests The authors declare no competing interests.

Author contributions Tianlun Yuan: Methodology, Software, Validation, Formal analysis, Writing-original draft, Visualization. Mengchun Yu: Formal analysis, Investigation, Writing-review & editing, Visualization. Yongfeng Jin: Software, Validation, Data curation, Investigation. Qinglin Huang: Writing-review & editing, Formal analysis, Resources. Jing Li: Conceptualization, Supervision, Writing-review & editing, Funding acquisition, Project administration, Resources.

Data availability All datasets used in this study are publicly available benchmark datasets, which can be accessed from the corresponding original references.

References

[1] Liu, Y., Wu, Y.-H., Wen, P., Shi, Y., Qiu, Y., Cheng, M.-M.: Leveraging instance-,

image- and dataset-level information for weakly supervised instance segmentation. IEEE Trans. Pattern Anal. Mach. Intell. **44**(3), 1415–1428 (2020) https://doi.org/10.1109/TPAMI.2020.3023152

[2] Yang, P., Wang, Q., Dou, J., Dou, L.: Sdcs-cf: Saliency-driven localization and cascade scale estimation for visual tracking. J. Vis. Commun. Image Represent. **98**(C) (2024) https://doi.org/10.1016/j.jvcir.2023.104040

[3] Wang, J., Yu, S., Zhang, B., Zhao, X., García-Fernández, Á.F., Lim, E.G., Xiao, J.: Cross-frame feature-saliency mutual reinforcing for weakly supervised video salient object detection. Pattern Recognit. **150**, 110302 (2024) https://doi.org/10.1016/j.patcog.2024.110302

[4] Ju, R., Ge, L., Geng, W., Ren, T., Wu, G.: Depth saliency based on anisotropic center-surround difference. In: Proc. IEEE Int. Conf. Image Process. (ICIP), pp. 1115–1119. IEEE, Paris, France (2014). https://doi.org/10.1109/ICIP.2014.7025222

[5] Cong, R., Lei, J., Zhang, C., Huang, Q., Cao, X., Hou, C.: Saliency detection for stereoscopic images based on depth confidence analysis and multiple cues fusion. IEEE Signal Proc. Lett. **23**(6), 819–823 (2016) https://doi.org/10.1109/LSP.2016.2557347 . Published online 21 April 2016; print issue June 2016

[6] Zhao, J.-X., Cao, Y., Fan, D.-P., Cheng, M.-M., Li, X.-Y., Zhang, L.: Contrast prior and fluid pyramid integration for RGBD salient object detection. In: Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recogn. (CVPR), pp. 3927–3936. IEEE, ??? (2019). https://doi.org/10.1109/CVPR.2019.00405

[7] Zhong, M., Sun, J., Wang, F., Sun, F.: LESOD: Lightweight and efficient network for RGB-D salient object detection. Pattern Recogn. **171**(Part A), 112103 (2026) https://doi.org/10.1016/j.patcog.2025.112103

[8] Liu, Z., Wang, Y., Tu, Z., Xiao, Y., Tang, B.: TriTransNet: RGB-D salient object detection with a triplet transformer embedding network. In: Proceedings of the 29th ACM International Conference on Multimedia (MM '21), pp. 4481–4490 (2021). https://doi.org/10.1145/3474085.3475601

[9] Piao, Y., Ji, W., Li, J., Zhang, M., Lu, H.: Depth-induced multi-scale recurrent attention network for saliency detection. In: Proc. IEEE/ICCV Int. Conf. Comput. Vis. (ICCV), Seoul, Korea (South), pp. 7254–7263 (2019). https://doi.org/10.1109/ICCV.2019.00735

[10] Cong, R., Lin, Q., Zhang, C., Li, C., Cao, X., Huang, Q.: CIR-Net: Cross-modality interaction and refinement for RGB-D salient object detection. IEEE Trans. Image Proc. **31**, 6800–6815 (2022) https://doi.org/10.1109/TIP.2022.3216198

[11] 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 Recogn. **136**, 109194 (2023) https://doi.org/10.1016/j.patcog.2022.109194

[12] Sun, F., Ren, P., Yin, B., Wang, F., Li, H.: CATNet: A cascaded and aggregated transformer network for RGB-D salient object detection. IEEE Trans. Multimedia **26**, 2249–2262 (2023) https://doi.org/10.1109/TMM.2023.3294003

[13] 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 Trans. Pattern Anal. Mach. Intell. **44**(12), 10261–10269 (2022) https://doi.org/10.1109/TPAMI.2021.3134684

[14] Jin, X., Yi, K., Xu, J.: MoADNet: Mobile asymmetric dual-stream networks for real-time and lightweight RGB-D salient object detection. IEEE Trans. Circuits Syst. Video Technol. **32**(11), 7632–7645 (2022) https://doi.org/10.1109/TCSVT.2022.3180274

[15] Zhou, W., Zhu, Y., Lei, J., Yang, R., Yu, L.: LSNet: Lightweight spatial boosting network for detecting salient objects in RGB-Thermal images. IEEE Trans. Image Proc. **32**, 1329–1340 (2023) https://doi.org/10.1109/TIP.2023.3242775

[16] Zhong, M., Sun, J., Ren, P., Wang, F., Sun, F.: MAGNet: Multi-scale awareness and global fusion network for RGB-D salient object detection. Knowl.-Based Syst. **299**, 112126 (2024) https://doi.org/10.1016/j.knosys.2024.112126

[17] Ren, P., Bai, T., Sun, F.: Bio-inspired two-stage network for efficient RGB-D salient object detection. Neural Netw. **185**, 107244 (2025) https://doi.org/10.1016/j.neunet.2025.107244

[18] Gao, H., Wang, F., Wang, M., Sun, F., Li, H.: Highly efficient RGB-D salient object detection with adaptive fusion and attention regulation. IEEE Trans. Circuits Syst. Video Technol. **35**(4), 3104–3118 (2025) https://doi.org/10.1109/TCSVT.2024.3502244

[19] Zhang, W., Ji, G.-P., Wang, Z., Fu, K., Zhao, Q.: Depth quality-inspired feature manipulation for efficient RGB-D salient object detection. In: Proc. ACM Int. Conf. Multimedia (ACM MM), pp. 731–740 (2021). https://doi.org/10.1145/3474085.3475240

[20] Zhang, Q., Qin, Q., Yang, Y., Jiao, Q., Han, J.: Feature calibrating and fusing network for RGB-D salient object detection. IEEE Trans. Circuits Syst. Video Technol. **34**(3), 1493–1507 (2024) https://doi.org/10.1109/TCSVT.2023.3296581

[21] Liang, Y., Qin, G., Sun, M., Qin, J., Yan, J., Zhang, Z.: Multi-modal interactive attention and dual progressive decoding network for rgb-d/t salient object detection. Neurocomputing **490**, 132–145 (2022) https://doi.org/10.1016/j.neucom.2022.03.029

[22] Zhang, C., Cong, R., Lin, Q., Ma, L., Li, F., Zhao, Y., Kwong, S.: Cross-modality discrepant interaction network for RGB-D salient object detection. In: Proc. ACM Int. Conf. Multimedia (ACM MM), pp. 2094–2102 (2021). https://doi.org/10.1145/3474085.3475364

[23] Maaz, M., Shaker, A., Cholakkal, H., Khan, S., Zamir, S.W., Anwer, R.M., Khan, F.S.: EdgeNeXt: Efficiently amalgamated CNN-Transformer architecture for mobile vision applications. In: Proc. Eur. Conf. Comput. Vis. (ECCV). Lecture Notes in Computer Science, vol. 13664, pp. 3–20. Springer, ??? (2022). https://doi.org/10.1007/978-3-031-20053-3_1

[24] Howard, A., Sandler, M., Chen, B., Wang, W., Chen, L.-C., Tan, M.: Searching for MobileNetV3. In: Proc. IEEE/ICCV Int. Conf. Comput. Vis. (ICCV), pp. 1314–1324 (2019). https://doi.org/10.1109/ICCV.2019.00140

[25] Lee, M., Park, C., Cho, S., Lee, S.: SPSN: Superpixel prototype sampling network for RGB-D salient object detection. In: Proc. Eur. Conf. Comput. Vis. (ECCV), vol. 13689, pp. 630–647. Springer, ??? (2022). https://doi.org/10.1007/978-3-031-19818-2_36

[26] Wu, J., Hao, F., Liang, W., Xu, J.: Transformer fusion and pixel-level contrastive learning for RGB-D salient object detection. IEEE Trans. Multimedia **26**, 1011–1026 (2024) https://doi.org/10.1109/TMM.2023.3275308

[27] Song, P., Li, W., Zhong, P., Zhang, J., Koniusz, P., Duan, F., Barnes, N.: Synergizing triple attention with depth quality for RGB-D salient object detection. Neurocomputing **589**, 127672 (2024) https://doi.org/10.1016/j.neucom.2024.127672

[28] Zeng, Z., Liu, H., Chen, F., Tan, X.: AirSOD: A lightweight network for RGB-D salient object detection. IEEE Trans. Circuits Syst. Video Technol. **34**(3), 1656–1669 (2024) https://doi.org/10.1109/TCSVT.2023.3295588

[29] Huang, N., Jiao, Q., Zhang, Q., Han, J.: Middle-level feature fusion for lightweight RGB-D salient object detection. IEEE Trans. Image Proc. **31**, 6621–6634 (2022) https://doi.org/10.1109/TIP.2022.3214092

[30] Wang, F., Su, Y., Wang, R., Sun, J., Sun, F., Li, H.: Cross-modal and cross-level attention interaction network for salient object detection. IEEE Trans. Artif. Intell. **5**(6), 2907–2920 (2024) https://doi.org/10.1109/TAI.2023.3333827

[31] Yang, Y., Huang, N., Zhang, Q., Han, J.: Mitigating fusion bias for RGB-D salient object detection. Pattern Recogn. **171**, 112135 (2026) https://doi.org/10.1016/j.patcog.2025.112135

[32] 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

Trans. Neural Netw. Learn. Syst. **32**(5), 2075–2089 (2021) https://doi.org/10.1109/TNNLS.2020.2996406

[33] Pang, Y., Zhao, X., Zhang, L., Lu, H.: CAVER: Cross-modal view-mixed transformer for bi-modal salient object detection. IEEE Trans. Image Proc. **32**, 892–904 (2023) https://doi.org/10.1109/TIP.2023.3234702