# RA-SOD: Reliability-Aware RGB-T Salient Object Detection under Modality Degradation

Hongbo Gao[1,2]†, Zhengyu Li[1]†, Xueru Nie[1], Dihao Zhu[2], Lijun Zhao[1], Yunke Wang[2]✉, and Chang Xu[2]✉

[1] State Key Laboratory of Robotics and Systems, Harbin Institute of Technology, No. 92 West Dazhi Street, Nangang District, Harbin, Heilongjiang 150006, China
[2] School of Computer Science, University of Sydney, Camperdown, Sydney, NSW 2006, Australia
{23b908032,24s108484}@stu.hit.edu.cn, {yunke.wang,c.xu}@sydney.edu.au
†Equal contribution. ✉Corresponding authors.

**Abstract.** RGB–Thermal (RGB-T) salient object detection leverages complementary cues from visible and thermal modalities to improve robustness in challenging environments. However, in real-world scenarios, the reliability of each modality is inherently unstable: RGB images degrade under low illumination, motion blur, and noise, while thermal imagery often suffers from contrast compression and sensor artifacts. Such degradation introduces unreliable perceptual evidence that can mislead cross-modal fusion and significantly deteriorate detection performance. To address this challenge, we propose RA-SOD, a reliability-aware RGB-T salient object detection framework that explicitly models modality reliability and integrates it into feature learning and cross-modal fusion. First, we introduce a reliability-conditioned representation that adaptively compensates degraded modality features while preserving structural cues. Second, an uncertainty-guided dual-stream refinement strategy progressively corrects cross-modal representations while suppressing unreliable evidence. Finally, we propose a pixel-wise modality competition mechanism that dynamically selects modality cues according to spatial reliability for fine-grained fusion. Extensive experiments on four benchmarks (VT821, VT1000, VT5000, and VT-IMAG) demonstrate that RA-SOD achieves state-of-the-art performance and exhibits strong robustness under severe modality degradation. Code and models are available at `https://github.com/zaoxienian/RA-SOD`.

**Keywords:** Salient Object Detection · Multi-Modality · Trustworthy

## 1 Introduction

RGB–Thermal (RGB-T) salient object detection aims to identify visually prominent regions by jointly leveraging complementary information from visible and thermal infrared modalities [38, 43]. Compared with single-modality approaches,

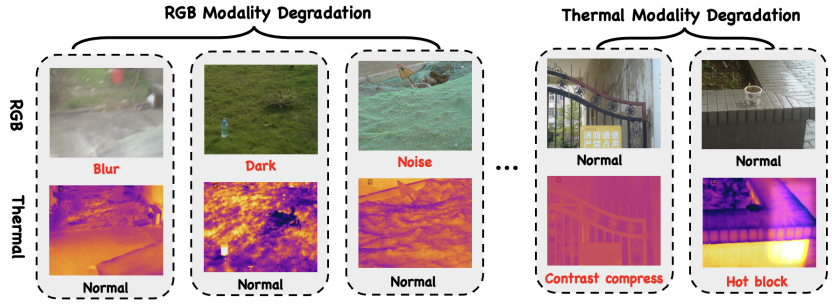


**Fig. 1:** Examples of **modality degradation** in RGB–Thermal scenes. The RGB modality may suffer from blur, low illumination, and noise, while the thermal modality can exhibit contrast compression and block artifacts. Such degradations distort local responses and obscure object boundaries, posing significant challenges for reliable cross-modal fusion.

RGB-T methods benefit from cross-modal complementarity, achieving improved robustness under low illumination and complex background conditions [5, 16, 17, 44]. However, in real-world environments, the quality of each modality is inherently unstable. The RGB modality may degrade under low-light conditions, motion blur, or heavy noise, while the thermal modality can suffer from contrast compression, suppressed temperature gradients, or block-like artifacts that distort local responses and obscure object boundaries. We refer to these phenomena as **modality degradation** [12, 15, 36, 50], as shown in Fig. 1. Under such conditions, a modality may provide unreliable or even misleading perceptual evidence, which can adversely affect cross-modal fusion.

Most existing RGB-T salient object detection methods emphasize elaborate fusion strategies, such as two-stream architectures [2, 35, 53] or attention-based modality weighting mechanisms [26, 60]. These approaches implicitly assume that modality features extracted during encoding remain sufficiently reliable and cross-modal inconsistencies mainly arise at later fusion stages. However, in practice, modality degradation often emerges early and spreads through the network hierarchy [11, 37, 39]. When degradation occurs, standard backbones apply fixed feature transformations that cannot adapt to input-dependent distributional shifts, causing noisy or distorted signals to be progressively propagated and even amplified. Consequently, subsequent fusion modules operate on already corrupted representations, making it difficult to disentangle structural information from modality-specific noise.

To address this issue, we propose a reliability-aware RGB-T salient object detection framework called RA-SOD. Instead of treating reliability as an auxiliary prior or a late-stage weighting cue, RA-SOD models reliability as a core principle and progressively propagates it across representation, refinement, and fusion stages. **First**, at the encoding stage, we construct reliability-conditioned backbone representations through a parallel residual modulation branch with feature-conditioned routing. This design enables the network to adaptively compensate

degraded modality features while preserving stable structural priors from the backbone. **Second**, during decoding, we introduce an uncertainty-guided dual-stream recursive refinement scheme. Explicitly predicted uncertainty maps act as structured reliability signals to regulate bidirectional cross-modal correction, preventing unreliable features from being reinforced during hierarchical propagation. **Finally**, at the fusion stage, we propose a hierarchical pixel-wise modality competition mechanism that dynamically selects RGB or thermal cues according to spatial reliability, enabling fine-grained cross-modal integration. By modeling and propagating reliability across representation, refinement, and fusion stages, RA-SOD establishes a unified hierarchical reliability framework.

Extensive experiments on four RGB-T benchmarks (VT821, VT1000, VT5000, and VT-IMAG) demonstrate state-of-the-art performance and strong robustness under challenging modality degradation scenarios.

## 2 Related Work

### 2.1 RGB Salient Object Detection

Early RGB SOD methods rely on handcrafted features and heuristic priors (e.g., contrast [3, 20] and center prior [19]), but show limited robustness in complex scenes. Deep learning, particularly fully convolutional networks (FCNs) [21], later enables end-to-end pixel-wise prediction, leading to representative models such as Amulet [63] and UCF [64], with related efforts on weakly supervised object localization [61]. Subsequent works further enhance representation through multi-scale aggregation [51], attention mechanisms [23, 68], and boundary guidance [27, 52, 67]. More recently, generalist and state-space architectures have been explored, including VSCode [25] and Mamba-based Samba [13]. However, unimodal RGB methods remain vulnerable to adverse conditions such as low illumination, motion blur, and sensor noise.

### 2.2 RGB-X Salient Object Detection

Multi-modal SOD enhances robustness via auxiliary data. Specifically, RGB-D methods leverage geometric priors from depth maps to distinguish foreground from background [2, 8, 57], optimizing detection results via multi-modal fusion networks (e.g., JL-DCF [9], DFormer [62], [54]). However, depth information is often compromised by low illumination or complex reflections. Conversely, RGB-T SOD exploits thermal data to detect intrinsic radiation cues, benefiting from their robustness against lighting changes [5,70]. Existing works have transitioned from input-level concatenation to feature fusion mechanisms, involving feature interaction [48, 49], graph learning (e.g., MGAI [33], M3S-NIR [39], [40]), and boundary enhancement(e.g., LSNet [71]), which maximize cross-modal complementarity, and define the state-of-the-art for SOD.

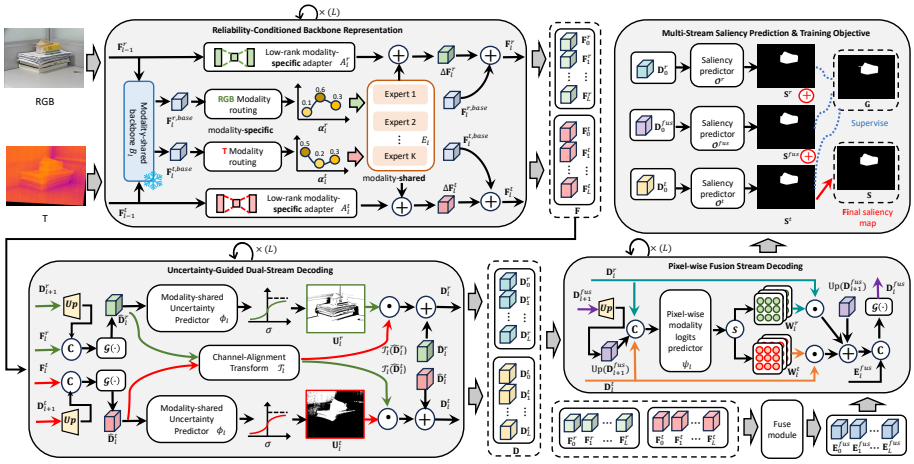


**Fig. 2:** Overall architecture of RA-SOD. It consists of three key components: (1) **Reliability-Conditioned Backbone Representation (RCBR)**: Extracts adaptive features by integrating structural priors from a frozen backbone with residual modulation via modality-specific adapters and a shared expert bank. (2) **Uncertainty-Guided Dual-Stream Decoding**: Performs recursive cross-modal refinement regulated by predicted uncertainty maps to mitigate localized degradation. (3) **Pixel-wise Fusion Stream Decoding**: Implements a competition mechanism to adaptively select modality-dependent cues at each spatial location. The final saliency map is obtained by aggregating predictions from the RGB, Thermal, and Fusion streams.

### 2.3 Multi-modal Learning in RGB-T SOD

Existing RGB-T SOD methods typically formulate multi-modal learning as feature fusion [24, 41, 66], designing multi-scale modules to progressively integrate cross-modal information [10, 45, 65]. Later works move beyond simple fusion to end-to-end representation and interaction modeling [26, 46, 60]. For example, Tu et al. [37] propose a multi-interactive dual-decoder for cross-modal interaction, while ConTriNet [35] introduces a divide-and-conquer triple-flow architecture, and SMR-Net [59] explores semantic-guided mutual reinforcement. Despite these advances, modality-degradation-aware representation learning remains less explored. Fixed feature transformations may propagate unreliable cues when one modality is degraded. Recent mixture-of-experts (MoE) models improve input-adaptive representation learning by routing features to specialized experts through learned gates [18, 31, 32]. Based on this input-adaptive routing mechanism, RA-SOD employs a lightweight shared expert bank to adapt RGB-T representations under modality degradation.

## 3 Methodology

In this section, we introduce the framework of RA-SOD for RGB-T salient object detection under modality degradation. The key idea is to model modality relia-

bility and propagate it across representation, refinement, and fusion stages in the network. As illustrated in Fig. 2, RA-SOD consists of three main components: (1) a **Reliability-Conditioned Backbone Representation (RCBR)** that adaptively modulates modality features during encoding, (2) an **Uncertainty-Guided Dual-Stream Decoding** mechanism that progressively refines RGB and thermal representations under reliability awareness, and (3) a **Pixel-wise Fusion Stream Decoding** module that performs spatially adaptive modality competition for fine-grained cross-modal integration. These components jointly form a hierarchical reliability modeling framework that mitigates the propagation of unreliable cues caused by modality degradation.

### 3.1 Overall Architecture

RA-SOD adopts a reliability-conditioned backbone followed by a tri-stream hierarchical decoding architecture for RGB-Thermal salient object detection. Given aligned RGB and thermal inputs $I^r$ and $I^t$, the network predicts a saliency map $\mathbf{S} \in [0,1]^{H \times W}$.

Using a shared frozen backbone, we first extract stage-wise base features $\{\mathbf{F}_l^{r,\text{base}}\}_{l=0}^{L}$ and $\{\mathbf{F}_l^{t,\text{base}}\}_{l=0}^{L}$. Instead of directly using these features, RA-SOD constructs reliability-conditioned representations $\{\mathbf{F}_l^{r}\}_{l=0}^{L}$ and $\{\mathbf{F}_l^{t}\}_{l=0}^{L}$ by introducing a parallel residual modulation branch at each stage. This design preserves a stable structural prior from the frozen backbone while enabling adaptive modulation under modality-dependent reliability variations (Sec. 3.2).

Based on the resulting representations, the network decodes three hierarchical streams in parallel: an RGB stream, a Thermal stream, and a Fusion stream. The RGB and Thermal streams perform uncertainty-guided dual-stream decoding through recursive refinement, while the Fusion stream conducts pixel-wise modality competition within a top-down decoding hierarchy. Each stream generates a saliency prediction via a lightweight head with identical architecture but independent parameters, and the three predictions are linearly aggregated to produce the final saliency map.

### 3.2 Reliability-Conditioned Backbone Representation

Reliable feature representation is critical under modality degradation. However, standard backbones apply fixed transformations to all inputs, limiting their ability to adapt to modality-dependent reliability variations. To address this issue, we construct a Reliability-Conditioned Backbone Representation (RCBR), which augments the frozen backbone with a lightweight residual modulation branch. This design preserves stable structural priors from the pretrained backbone while allowing adaptive feature correction for degraded modalities.

Let $\mathbf{F}_{l-1}^{m}$ denote the input feature of modality $m \in \{r, t\}$ at stage $l$. The frozen backbone block produces a base feature

$$\mathbf{F}_l^{m,\text{base}} = B_l(\mathbf{F}_{l-1}^{m}), \tag{1}$$

where $B_l(\cdot)$ denotes the shared frozen backbone transformation.

**Feature-conditioned modality routing.** To adaptively regulate residual modulation, we first predict expert routing weights from the base feature:

$$\boldsymbol{\alpha}_l^m = \mathrm{Softmax}\Big(g_l^m(\mathbf{F}_l^{m,\mathrm{base}})\Big), \qquad \sum_{k=1}^{K} \alpha_{l,k}^m = 1, \tag{2}$$

where $g_l^m(\cdot)$ is a lightweight gating function. Since modality degradation often induces distributional shifts in the backbone feature space, conditioning routing decisions on $\mathbf{F}_l^{m,\mathrm{base}}$ allows the model to implicitly capture modality-dependent reliability.

**MoE-guided residual representation adaptation.** To enhance the expressive capacity of residual modulation, we introduce a mixture-of-experts (MoE) mechanism that provides multiple transformation subspaces. Each expert captures a distinct feature adaptation pattern, while routing weights dynamically determine their contributions according to modality reliability.

In parallel to the frozen backbone transformation, a residual branch operates directly on $\mathbf{F}_{l-1}^m$:

$$\Delta\mathbf{F}_l^m = \sum_{k=1}^{K} \alpha_{l,k}^m E_{l,k}(\mathbf{F}_{l-1}^m) + A_l^m(\mathbf{F}_{l-1}^m), \tag{3}$$

where $E_{l,k}(\cdot)$ are shared expert transformations, and $A_l^m(\cdot)$ is a modality-specific low-rank adapter.

The final reliability-conditioned representation is

$$\mathbf{F}_l^m = \mathbf{F}_l^{m,\mathrm{base}} + \Delta\mathbf{F}_l^m. \tag{4}$$

This parallel design preserves the stable structural prior of the frozen backbone while enabling adaptive modulation of unreliable feature components. In the subsequent decoding stage (Sec. 3.3), reliability is further modeled explicitly via uncertainty-guided cross-modal refinement.

### 3.3 Uncertainty-Guided Dual-Stream Decoding

During decoding, modality degradation may lead to spatially localized ambiguity, where certain regions contain unreliable cues. Directly propagating such features can amplify errors during hierarchical decoding. To alleviate this issue, we introduce an Uncertainty-Guided Dual-Stream Decoding mechanism, which explicitly estimates uncertainty maps to regulate cross-modal feature correction.

Specifically, RGB and Thermal modalities are decoded through two parallel top-down streams. At stage $l$, a preliminary decoding feature $\hat{\mathbf{D}}_l^m$ is computed from $\mathbf{F}_l^m$, upon which bidirectional uncertainty-guided refinement is applied. The refined state $\mathbf{D}_l^m$ is propagated to the next finer stage, forming a hierarchical recursion.

**Top-down recursion.** Given the reliability-conditioned encoder representations, we first construct preliminary decoding features through a top-down hierarchical decoding process.

At the coarsest stage $L$,

$$\hat{\mathbf{D}}_L^r = \mathcal{G}_L^r(\mathbf{F}_L^r), \quad \hat{\mathbf{D}}_L^t = \mathcal{G}_L^t(\mathbf{F}_L^t). \tag{5}$$

For $l = L-1, \ldots, 0$,

$$\hat{\mathbf{D}}_l^m = \mathcal{G}_l^m\left([\mathbf{F}_l^m; \mathrm{Up}(\mathbf{D}_{l+1}^m)]\right), \quad m \in \{r, t\}, \tag{6}$$

where $[\cdot;\cdot]$ denotes channel concatenation and $\mathrm{Up}(\cdot)$ represents bilinear upsampling.

**Uncertainty-guided cross-modal refinement.** Although the preliminary decoding features provide modality-specific predictions, their reliability may vary across spatial locations under modality degradation. To explicitly model such reliability variations, we introduce an uncertainty-guided refinement mechanism.

An uncertainty map is predicted as

$$\mathbf{U}_l^m = \sigma\Big(\phi_l(\hat{\mathbf{D}}_l^m)\Big), \tag{7}$$

where $\phi_l(\cdot)$ is a modality-shared lightweight uncertainty predictor.

The uncertainty map regulates cross-modal correction

$$\mathbf{D}_l^r = \hat{\mathbf{D}}_l^r + \mathbf{U}_l^r \odot \mathcal{T}_l(\hat{\mathbf{D}}_l^t), \quad \mathbf{D}_l^t = \hat{\mathbf{D}}_l^t + \mathbf{U}_l^t \odot \mathcal{T}_l(\hat{\mathbf{D}}_l^r), \tag{8}$$

where $\mathcal{T}_l(\cdot)$ denotes a channel-alignment transform.

The refined states are recursively propagated to form decoding pyramids $\{\mathbf{D}_l^r\}$ and $\{\mathbf{D}_l^t\}$.

### 3.4 Pixel-wise Fusion Stream Decoding

Besides modality-specific decoding pyramids, a third fusion stream $\{\mathbf{D}_l^{fus}\}_{l=0}^L$ is decoded hierarchically. Encoder features are first fused through the MFM module [35], producing modality-integrated encoder representations $\{\mathbf{E}_l^{fus}\}_{l=0}^L$, which serve as the input to the fusion decoding process.

**Fusion recursion.** Starting from the coarsest level, the fusion stream progressively aggregates encoder features and modality-guided cues in a top-down manner:

$$\mathbf{D}_l^{fus} = \mathcal{G}_l^{fus}\Big([\mathbf{E}_l^{fus}; \mathbf{H}_l]\Big), \quad l = L, \ldots, 0, \tag{9}$$

with $\mathbf{H}_L = \mathbf{0}$.

**Pixel-wise modality competition.** Although the dual-stream decoding stage produces refined RGB and thermal representations, their reliability may still vary across spatial locations under modality degradation. Therefore, instead of using fixed fusion weights, we introduce a pixel-wise modality competition mechanism to dynamically determine which modality provides more reliable cues at each location.

For $l = L-1, \ldots, 0$,

$$[\mathbf{W}_l^r, \mathbf{W}_l^t] = \mathrm{Softmax}\Big(\psi_l([\mathrm{Up}(\mathbf{D}_{l+1}^{fus}); \mathbf{D}_l^r; \mathbf{D}_l^t])\Big), \tag{10}$$

where $\psi_l(\cdot)$ predicts modality logits at each spatial location.

The fusion guidance feature is

$$\mathbf{H}_l = \mathbf{W}_l^r \odot \mathbf{D}_l^r + \mathbf{W}_l^t \odot \mathbf{D}_l^t + \mathrm{Up}(\mathbf{D}_{l+1}^{fus}). \tag{11}$$

This design allows the model to adaptively emphasize the more reliable modality while suppressing unreliable responses caused by local degradation.

### 3.5 Multi-Stream Saliency Prediction

After hierarchical decoding, each stream produces a modality-specific representation that captures complementary saliency cues. Therefore, we attach an independent prediction head to each stream to generate saliency maps. Each stream predicts saliency independently:

$$\mathbf{S}^r = \mathcal{O}^r(\mathbf{D}_0^r), \quad \mathbf{S}^t = \mathcal{O}^t(\mathbf{D}_0^t), \quad \mathbf{S}^{fus} = \mathcal{O}^{fus}(\mathbf{D}_0^{fus}), \tag{12}$$

where $\mathcal{O}^r$, $\mathcal{O}^t$, and $\mathcal{O}^{fus}$ denote lightweight prediction heads. They share the same architecture but have independent parameters. The final saliency map is obtained via linear aggregation:

$$\mathbf{S} = \mathbf{S}^r + \mathbf{S}^t + \mathbf{S}^{fus}. \tag{13}$$

This multi-stream prediction strategy enables the model to exploit complementary cues from modality-specific and fused representations.

### 3.6 Training Objective

To encourage reliability-aware learning while avoiding modality dominance, we impose deep supervision on the RGB, Thermal, and Fusion predictions. The three streams are optimized jointly, and the overall objective is

$$\mathcal{L} = \sum_{m \in \{r,t,fus\}} \left(\lambda_1 \mathcal{L}_{\mathrm{BCE}}(\mathbf{S}^m, \mathbf{G}) + \lambda_2 \mathcal{L}_{\mathrm{IoU}}(\mathbf{S}^m, \mathbf{G})\right), \tag{14}$$

where $\mathbf{G}$ is the ground-truth map. This multi-stream supervision stabilizes optimization and yields complementary saliency cues that improve the final aggregated prediction.

**Table 1:** Comparison with recent state-of-the-art CNN-based methods in RGB-D/RGB-T SOD on VT821, VT1000 and VT5000 benchmarks. The best and second-best results are highlighted in red and blue, respectively.

| Method | VT821 | | | | | VT1000 | | | | | VT5000 | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | $S_m \uparrow$ | $F_\beta \uparrow$ | $F_\beta^w \uparrow$ | $E_m \uparrow$ | $\mathcal{M} \downarrow$ | $S_m \uparrow$ | $F_\beta \uparrow$ | $F_\beta^w \uparrow$ | $E_m \uparrow$ | $\mathcal{M} \downarrow$ | $S_m \uparrow$ | $F_\beta \uparrow$ | $F_\beta^w \uparrow$ | $E_m \uparrow$ | $\mathcal{M} \downarrow$ |
| DCMF [42] | 0.856 | 0.834 | 0.740 | 0.866 | 0.055 | 0.917 | 0.834 | 0.841 | 0.915 | 0.028 | 0.857 | 0.753 | 0.728 | 0.873 | 0.052 |
| CIR-Net [4] | 0.861 | 0.777 | 0.724 | 0.884 | 0.045 | 0.914 | 0.861 | 0.823 | 0.924 | 0.028 | 0.871 | 0.791 | 0.744 | 0.899 | 0.041 |
| RAFNet [55] | 0.883 | 0.826 | 0.811 | 0.905 | 0.038 | 0.932 | 0.893 | 0.898 | 0.941 | 0.020 | 0.891 | 0.840 | 0.829 | 0.921 | 0.033 |
| PopNet [56] | 0.861 | 0.795 | 0.774 | 0.887 | 0.043 | 0.929 | 0.893 | 0.895 | 0.939 | 0.020 | 0.887 | 0.839 | 0.827 | 0.918 | 0.034 |
| HiDAnet [57] | 0.890 | 0.842 | 0.832 | 0.920 | 0.031 | 0.927 | 0.894 | 0.897 | 0.942 | 0.020 | 0.890 | 0.846 | 0.836 | 0.923 | 0.032 |
| MIA [22] | 0.844 | 0.740 | 0.720 | 0.850 | 0.070 | 0.924 | 0.868 | 0.864 | 0.926 | 0.025 | 0.878 | 0.793 | 0.780 | 0.893 | 0.040 |
| ECFFNet [69] | 0.877 | 0.810 | 0.801 | 0.902 | 0.034 | 0.923 | 0.876 | 0.885 | 0.930 | 0.021 | 0.874 | 0.806 | 0.801 | 0.906 | 0.038 |
| OSRNet [17] | 0.875 | 0.813 | 0.801 | 0.896 | 0.043 | 0.926 | 0.892 | 0.891 | 0.935 | 0.022 | 0.875 | 0.823 | 0.807 | 0.908 | 0.040 |
| LSNet [71] | 0.878 | 0.825 | 0.809 | 0.911 | 0.033 | 0.925 | 0.885 | 0.887 | 0.935 | 0.023 | 0.877 | 0.825 | 0.806 | 0.915 | 0.037 |
| CAVER [29] | 0.891 | 0.839 | 0.835 | 0.919 | 0.033 | 0.936 | 0.903 | 0.909 | 0.945 | 0.017 | 0.892 | 0.842 | 0.835 | 0.924 | 0.032 |
| LAFB [49] | 0.887 | 0.843 | 0.817 | 0.915 | 0.027 | 0.927 | 0.905 | 0.905 | 0.945 | 0.018 | 0.890 | 0.857 | 0.841 | 0.931 | 0.030 |
| SMR-Net [59] | 0.889 | 0.846 | 0.835 | 0.921 | 0.031 | 0.928 | 0.902 | 0.908 | 0.947 | 0.019 | 0.893 | 0.859 | 0.845 | 0.935 | 0.030 |
| ConTriNet [35] | 0.883 | 0.846 | 0.836 | 0.920 | 0.033 | 0.927 | 0.903 | 0.903 | 0.948 | 0.019 | 0.894 | 0.860 | 0.846 | 0.934 | 0.030 |
| **RA-SOD (Ours)** | 0.892 | 0.857 | 0.846 | 0.925 | 0.031 | 0.932 | 0.915 | 0.913 | 0.951 | 0.018 | 0.897 | 0.870 | 0.855 | 0.937 | 0.029 |

# 4 Experiments

## 4.1 Experimental Setup

**Implementation Details.** Experiments are conducted on a single NVIDIA RTX 3090 GPU. Following ConTriNet [35], we adopt an ImageNet-pretrained Res2Net-50 as the shared backbone. RGB and thermal images are resized to $352 \times 352$ with random flipping, rotation, and border cropping for augmentation. The network is optimized with Adam using an initial learning rate of $5 \times 10^{-5}$ and a cosine decay schedule for 100 epochs. During inference, saliency maps are resized to the original resolution and the final prediction is obtained by linearly aggregating the three outputs. Performance is evaluated using the Saliency-Evaluation-Toolbox.

**Datasets.** We evaluate our method on three RGB-T benchmarks: VT821 [43], VT1000 [40], and VT5000 [38]. To further assess robustness under severe degradations, we additionally adopt the challenging VT-IMAG [35]. Following ConTriNet [35], 2,500 pairs from VT5000 are used for training, and the remaining 2,500 pairs together with the full VT821, VT1000, and VT-IMAG datasets are used for evaluation.

**Evaluation Metrics.** Specifically, we adopt five standard metrics: Structure-measure ($S_m$) [6], Mean F-measure ($F_\beta$) [1], Weighted F-measure ($F_\beta^w$) [28], Mean Enhanced-alignment measure ($E_m$) [7], and Mean Absolute Error ($\mathcal{M}$) [30]. Among them, $S_m$ evaluates structural similarity; $F_\beta$ and $F_\beta^w$ measure the precision–recall trade-off; $E_m$ considers both pixel-level alignment and global statistics; and $\mathcal{M}$ computes the average pixel-wise error. We report mean $F_\beta$ and $E_m$

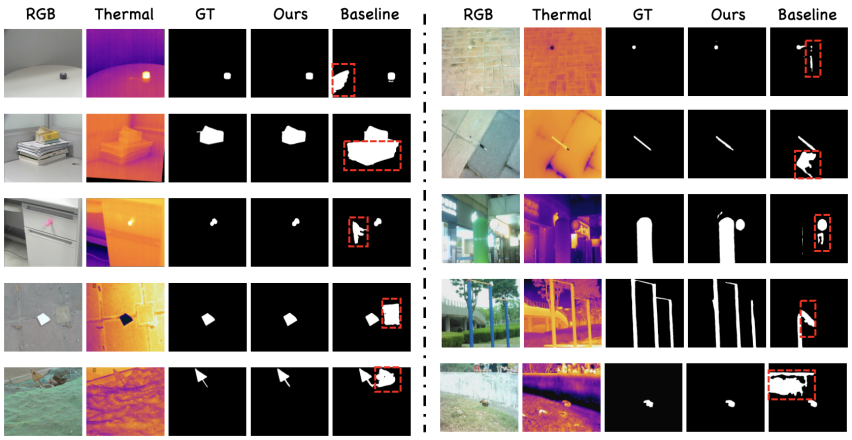


**Fig. 3:** Qualitative comparison between RA-SOD (ours) and ConTriNet (baseline) on standard RGB-T SOD benchmarks (VT821, VT1000, and VT5000). Red dashed boxes highlight typical failure cases of the baseline.

scores averaged over all thresholds. Higher values indicate better performance for all metrics except $\mathcal{M}$.

### 4.2 Comparison on Standard Benchmarks

We compare RA-SOD with several recent state-of-the-art methods, including five RGB-D SOD models (DCMF [42], CIR-Net [4], RAFNet [55], PopNet [56], and HiDAnet [57]) and eight RGB-T SOD models (MIA [22], ECFFNet [69], OSRNet [17], LSNet [71], CAVER [29], LAFB [49], SMR-Net [59], and ConTriNet [35]). For fair comparison, all RGB-T methods are retrained on the RGB-T dataset using their official implementations and consistent experimental settings, while RGB-D models follow their default hyperparameters. For SMR-Net [59], we replace the original ResNet-30 backbone with Res2Net-50 [14] to ensure a fair evaluation.

**Quantitative Results.** Table 1 reports quantitative comparisons under five evaluation metrics. The results show that RA-SOD achieves the best or second-best performance for all five metrics on all three benchmarks. On the large-scale VT5000 benchmark, RA-SOD ranks first on every metric. Compared with ConTriNet [35], RA-SOD improves $F_\beta$ by 0.010 and $F_\beta^w$ by 0.009, and also brings a 0.003 gain in $E_m$, while reducing $\mathcal{M}$ by 0.001. On the smaller VT821 and VT1000 benchmarks, RA-SOD remains consistently within the top two. On VT1000, RA-SOD achieves the best $F_\beta$, $F_\beta^w$, and $E_m$, and it stays highly competitive in structure measure and $\mathcal{M}$ compared with CAVER [29]. Overall, these results demonstrate strong generalization from small and medium-scale benchmarks to the larger and more diverse VT5000.

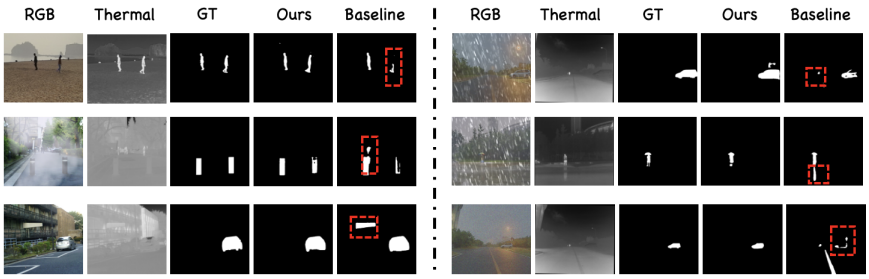


**Fig. 4:** Qualitative comparison between RA-SOD (ours) and ConTriNet (baseline) on the challenging VT-IMAG dataset.

**Qualitative Results and Visualization.** The visual comparisons in Fig. 3 reveal the limitations of ConTriNet [35] (Baseline). Without explicit reliability modeling, it struggles to distinguish salient targets from background noise, leading to severe false detections (red dashed boxes). In contrast, our reliability-aware framework effectively suppresses unreliable cues, enabling accurate localization with complete structures and sharp boundaries even under extremely low signal-to-noise conditions.

### 4.3 Robustness Analysis under Degradations

**Synthetic Degradation Settings.** To verify the effectiveness of reliability awareness under modality degradation, we construct synthetic degraded test sets derived from the VT1000 dataset [40] by degrading only one modality at a time while keeping the other modality and the ground-truth masks unchanged. We focus on five representative degradations, including low-light, additive noise, and blur on the RGB modality, as well as local hot-block artifacts and global contrast compression on the thermal modality. The degradation severity is randomly sampled per image from preset ranges with a fixed seed to ensure reproducibility.

**Table 2:** Average comparison under synthetic degradations on VT1000.

| Method | $S_m \uparrow$ | $F_\beta \uparrow$ | $F_\beta^w \uparrow$ | $E_m \uparrow$ | $\mathcal{M} \downarrow$ |
|---|---|---|---|---|---|
| LAFB [49] | 0.919 | 0.887 | 0.887 | 0.938 | 0.022 |
| SMR-Net [59] | 0.917 | 0.895 | 0.893 | 0.945 | 0.021 |
| ConTriNet [35] | 0.921 | 0.893 | 0.895 | 0.945 | 0.021 |
| **RA-SOD** | **0.927** | **0.906** | **0.907** | **0.949** | **0.020** |

**Comparison under Synthetic Degradations.** Table 2 compares RA-SOD with representative RGB-T SOD methods under synthetic degradations. Existing methods suffer noticeable performance drops in corrupted scenarios due to limited mechanisms for handling unreliable modality features. In contrast, RA-SOD consistently improves performance by explicitly modeling modality reliability. Compared with the strongest competing results, RA-SOD improves $S_m$ from 0.921 to 0.927, $F_\beta$ from 0.895 to 0.906, and $F_\beta^w$ from 0.895 to 0.907, while reducing $\mathcal{M}$ to 0.020. These results demonstrate the effectiveness of reliability-aware modeling for robust RGB-T saliency detection under modality degradation.

**Table 3:** Quantitative comparison with state-of-the-art methods on the VT-IMAG benchmark. Red and Blue indicate the best and second-best performance, respectively.

| Method | $S_m \uparrow$ | $F_\beta \uparrow$ | $F_\beta^w \uparrow$ | $E_m \uparrow$ | $\mathcal{M} \downarrow$ |
|---|---|---|---|---|---|
| CGFNet [46] | 0.795 | 0.696 | 0.664 | 0.861 | 0.042 |
| OSRNet [17] | 0.782 | 0.665 | 0.626 | 0.839 | 0.051 |
| LSNet [71] | 0.779 | 0.712 | 0.636 | 0.896 | 0.038 |
| CAVER [29] | 0.815 | 0.724 | 0.693 | 0.885 | 0.032 |
| ConTriNet [35] | 0.828 | 0.745 | 0.727 | 0.902 | 0.029 |
| **RA-SOD (Ours)** | 0.829 | 0.765 | 0.735 | 0.908 | 0.031 |

**Performance on the VT-IMAG Challenge Set.** Table 3 reports the quantitative comparison on the challenging VT-IMAG benchmark [35], which is designed to evaluate robustness under severe sensor failures and environmental degradations (e.g., thermal crossover, bad weather, and strong noise). RA-SOD achieves superior performance on four out of five metrics among state-of-the-art CNN-based methods. Notably, it significantly outperforms the strong baseline ConTriNet [35] on its own benchmark, improving $F_\beta$ by 0.020 and $F_\beta^w$ by 0.008. These results indicate that while the baseline struggles to preserve structural integrity under heavy modality corruption, RA-SOD effectively suppresses unreliable cues via reliability-aware modeling, achieving the highest $S_m$ of 0.829 and $E_m$ of 0.908. The qualitative comparisons in Fig. 4 further demonstrate that RA-SOD produces more reliable and complete saliency predictions under severe degradation.

### 4.4 Analysis of RA-SOD

**Model Complexity and Efficiency.** Table 4 compares the model complexity of different methods. For fair evaluation, all compared methods are implemented with a Res2Net-50 backbone. ConTriNet [35] is the most lightweight model with 34.78M parameters and 55.42G FLOPs, while RA-SOD ranks second with 46.72M parameters and 60.13G FLOPs. RA-SOD also runs faster than SMR-Net and LAFB, showing a good balance between complexity and performance.

**Table 4:** Comparison of model complexity and inference speed.

| Method | Backbone | Params. | FLOPs | FPS |
|---|---|---|---|---|
| ConTriNet [35] | R2N-50 | 34.78 | 55.42 | 13.53 |
| LAFB [49] | R2N-50 | 453.03 | 139.70 | 4.62 |
| SMR-Net [59] | R2N-50 | 47.70 | 62.80 | 10.15 |
| **RA-SOD** | R2N-50 | **46.72** | **60.13** | **11.24** |

**Backbone Generalization.** Table 5 reports the comparison with Transformer-based RGB-T SOD methods on VT5000. With Swin-B, RA-SOD$^\dagger$ achieves the best results on all five metrics and clearly outperforms ConTriNet. These results show that the proposed reliability-aware design can generalize from Res2Net-50 to Transformer-based backbones.

**Table 5:** Transformer-backbone comparison on VT5000. RA-SOD† uses Swin-B. "–" denotes metrics not reported.

| Method | Backbone | $S_m \uparrow$ | $F_\beta \uparrow$ | $F_\beta^w \uparrow$ | $E_m \uparrow$ | $\mathcal{M} \downarrow$ |
|---|---|---|---|---|---|---|
| HRTransNet [34] | HRFormer | 0.912 | 0.871 | 0.870 | 0.945 | 0.025 |
| XMSNet [58] | Transformer | 0.907 | 0.871 | 0.865 | 0.939 | 0.028 |
| SACNet [48] | Swin-B | 0.917 | – | 0.888 | 0.957 | 0.021 |
| PCNet [47] | Swin-B | 0.920 | 0.899 | – | 0.956 | – |
| ConTriNet [35] | Swin-B | 0.923 | 0.898 | 0.895 | 0.956 | 0.020 |
| **RA-SOD†** | Swin-B | **0.930** | **0.904** | **0.899** | **0.963** | **0.017** |

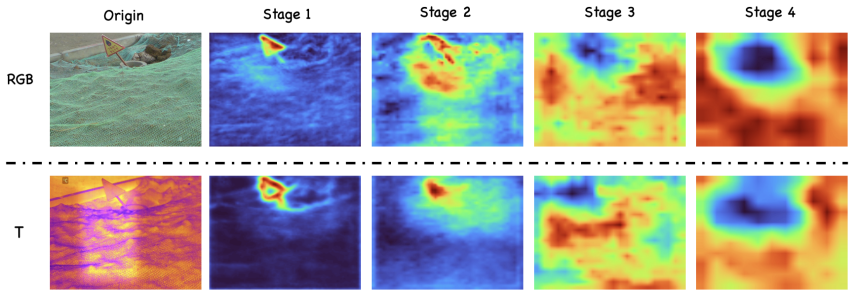


**Fig. 5:** Visualization of the predicted uncertainty maps across four decoding stages for the RGB (top) and Thermal (bottom) branches. Brighter regions indicate higher predicted uncertainty. The evolution of these maps demonstrates the model's progressive refinement from boundary-level ambiguity to region-level reliability.

**Uncertainty Visualization.** Fig. 5 visualizes the predicted uncertainty maps at different decoding stages, where brighter regions indicate higher uncertainty. At early stages, uncertainty mainly appears around object boundaries due to limited semantic capacity and foreground–background ambiguity. As decoding progresses, uncertainty within object interiors gradually decreases, reflecting more stable object-level predictions from deeper features. Meanwhile, RGB and thermal streams exhibit distinct uncertainty patterns, suggesting that the learned uncertainty captures modality-dependent reliability rather than random noise.

### 4.5 Ablation Studies

**Effectiveness of Uncertainty-Guided Dual-Stream Decoding.** Table 7 evaluates UGDD under both standard and degraded settings using $F_\beta^w$. Removing UGDD decreases performance from 0.913 to 0.908 on VT1000 and from 0.735 to 0.728 on VT-IMAG, while No-Unc. (dual-stream decoding without uncertainty guidance) also drops to 0.903 under degradations. Table 8 further shows that $U_l^m$ responds to corrupted modalities, e.g., $U_l^r$ rises to 0.57 under RGB low-light, while $U_l^t$ rises to 0.55 under thermal contrast. These results verify the effectiveness of uncertainty-guided correction.

**Table 6:** Ablation study of the Reliability-Conditioned Backbone Representation (RCBR) on three benchmarks. "A" denotes the low-rank modality-specific adapter, and "E" represents the shared expert bank (MoE). B0 is the base network without residual modulation. The best results are highlighted in **bold**.

| Variant | Module | | VT821 | | | | | VT1000 | | | | | VT5000 | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | A | E | $S_m\uparrow$ | $F_\beta\uparrow$ | $F_\beta^w\uparrow$ | $E_m\uparrow$ | $M\downarrow$ | $S_m\uparrow$ | $F_\beta\uparrow$ | $F_\beta^w\uparrow$ | $E_m\uparrow$ | $M\downarrow$ | $S_m\uparrow$ | $F_\beta\uparrow$ | $F_\beta^w\uparrow$ | $E_m\uparrow$ | $M\downarrow$ |
| Frozen Backbone (B0) | × | × | 0.886 | 0.852 | 0.840 | 0.924 | 0.032 | 0.929 | 0.906 | 0.906 | 0.948 | 0.019 | 0.896 | 0.865 | 0.851 | 0.935 | 0.030 |
| +A only (B1) | ✓ | × | 0.888 | 0.856 | 0.843 | **0.925** | 0.032 | **0.933** | **0.916** | **0.914** | **0.951** | **0.018** | 0.896 | 0.868 | 0.853 | 0.936 | 0.031 |
| +E only (B2) | × | ✓ | 0.888 | **0.858** | 0.842 | 0.924 | 0.032 | 0.931 | 0.910 | 0.910 | 0.949 | 0.020 | 0.896 | 0.866 | 0.851 | **0.937** | **0.029** |
| **Ours (A+E) (B3)** | ✓ | ✓ | **0.892** | 0.857 | **0.846** | **0.925** | **0.031** | 0.932 | 0.915 | 0.913 | **0.951** | **0.018** | **0.897** | **0.870** | **0.855** | **0.937** | **0.029** |

**Table 7:** Ablation study of UGDD using $F_\beta^w$. "Deg." averages five RGB/T degradation settings on VT1000.

| Method | VT1000 | VT5000 | VT-IMAG | Deg. |
|---|---|---|---|---|
| w/o UGDD | 0.908 | 0.849 | 0.728 | 0.897 |
| No-Unc. | 0.911 | 0.852 | 0.731 | 0.903 |
| **Full (Ours)** | **0.913** | **0.855** | **0.735** | **0.907** |

**Table 8:** Mean spatial $U_l^m$ on VT1000 under synthetic corruptions. The gate rises only for the degraded modality.

| Cond. | Clean | Low-light | Noise | Blur | Contrast | Hot-block |
|---|---|---|---|---|---|---|
| $U_l^r$ | 0.32 | **0.57** | **0.48** | **0.51** | 0.33 | 0.32 |
| $U_l^t$ | 0.30 | 0.31 | 0.30 | 0.31 | **0.55** | **0.53** |

**Effectiveness of Reliability-Conditioned Backbone Representation.** Table 6 reports the ablation results of the Reliability-Conditioned Backbone Representation (RCBR). Introducing modality-specific adapters (B1) and the shared expert bank (B2) consistently improves performance over the frozen backbone (B0). The full model (B3) achieves the best results, particularly on VT5000, improving $F_\beta^w$ from 0.851 to 0.855 and reducing $M$ to 0.029, demonstrating the effectiveness of routing-aware residual modulation for handling modality reliability variations.

**Impact of the Number of Experts.** Table 9 analyzes the effect of the number of experts $K$. The best trade-off between accuracy and robustness is achieved at $K = 3$, which yields the highest $S_m$ (0.897) and $F_\beta^w$ (0.855) and the lowest $\mathcal{M}$ (0.029) on VT5000. Smaller $K$ limits transformation diversity, while larger $K$ leads to performance saturation and slight degradation, likely due to redundancy and increased optimization difficulty. Therefore, we set $K = 3$ as the default configuration.

**Effectiveness of Pixel-wise Modality Competition.** We compare the proposed Pixel-wise Modality Competition (PMC) with the MDAM module used in

**Table 9:** Ablation on the number of experts $K$. Best results are highlighted in **bold**.

| $K$ | VT821 | | | | | VT1000 | | | | | VT5000 | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | $S_m\uparrow$ | $F_\beta\uparrow$ | $F_\beta^w\uparrow$ | $E_m\uparrow$ | $\mathcal{M}\downarrow$ | $S_m\uparrow$ | $F_\beta\uparrow$ | $F_\beta^w\uparrow$ | $E_m\uparrow$ | $\mathcal{M}\downarrow$ | $S_m\uparrow$ | $F_\beta\uparrow$ | $F_\beta^w\uparrow$ | $E_m\uparrow$ | $\mathcal{M}\downarrow$ |
| 1 | 0.890 | 0.856 | 0.844 | 0.925 | 0.032 | **0.933** | 0.914 | 0.912 | 0.950 | 0.019 | 0.896 | 0.868 | 0.853 | 0.936 | 0.030 |
| 2 | 0.891 | 0.856 | 0.844 | 0.924 | **0.031** | 0.932 | 0.913 | 0.911 | 0.949 | 0.019 | 0.896 | 0.858 | 0.854 | 0.933 | 0.030 |
| 3 | **0.892** | 0.857 | **0.846** | **0.925** | **0.031** | **0.932** | **0.915** | **0.913** | 0.951 | **0.018** | **0.897** | 0.870 | **0.855** | **0.937** | **0.029** |
| 4 | 0.890 | **0.859** | 0.844 | 0.923 | 0.032 | 0.930 | 0.912 | 0.911 | 0.950 | 0.019 | 0.895 | **0.871** | 0.852 | 0.935 | 0.030 |
| 5 | 0.889 | 0.855 | 0.842 | 0.921 | 0.032 | 0.929 | 0.909 | 0.909 | **0.952** | 0.020 | 0.894 | 0.868 | 0.850 | 0.933 | 0.031 |

ConTriNet [35]. As shown in Table 10, replacing MDAM with PMC consistently improves performance across all benchmarks. On VT5000, PMC increases $F_\beta$ from 0.865 to 0.870 and $F_\beta^w$ from 0.850 to 0.855, with similar gains observed on VT821 ($E_m$: 0.919 $\rightarrow$ 0.925). These results suggest that pixel-wise competition better models spatially varying reliability, enabling adaptive modality selection and more accurate saliency predictions.

**Table 10:** Ablation study of the proposed PMC compared with the baseline MDAM [35]. Best results are highlighted in **bold**.

| Variant | VT821 | | | | | VT1000 | | | | | VT5000 | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | $S_m\uparrow$ | $F_\beta\uparrow$ | $F_\beta^w\uparrow$ | $E_m\uparrow$ | $M\downarrow$ | $S_m\uparrow$ | $F_\beta\uparrow$ | $F_\beta^w\uparrow$ | $E_m\uparrow$ | $M\downarrow$ | $S_m\uparrow$ | $F_\beta\uparrow$ | $F_\beta^w\uparrow$ | $E_m\uparrow$ | $M\downarrow$ |
| MDAM [35] | 0.890 | 0.853 | 0.840 | 0.919 | **0.030** | **0.932** | 0.913 | 0.911 | **0.951** | **0.018** | 0.896 | 0.865 | 0.850 | 0.933 | **0.029** |
| PMC(Ours) | **0.892** | **0.857** | **0.846** | **0.925** | 0.031 | **0.932** | **0.915** | **0.913** | **0.951** | **0.018** | **0.897** | **0.870** | **0.855** | **0.937** | **0.029** |

## 5 Conclusion

In this work, we present RA-SOD, a reliability-aware RGB-T salient object detection framework designed to address modality degradation in real-world environments. Unlike existing approaches that mainly focus on fusion strategies, RA-SOD treats reliability as a fundamental principle and models it across representation learning, decoding refinement, and fusion. Specifically, a reliability-conditioned backbone adaptively compensates degraded features while preserving structural priors, an uncertainty-guided dual-stream decoder prevents unreliable cues from being propagated, and a pixel-wise modality competition strategy enables spatially adaptive cross-modal integration. Extensive experiments on multiple benchmarks demonstrate that RA-SOD achieves state-of-the-art performance and strong robustness under severe degradation conditions, highlighting the importance of reliability modeling for robust cross-modal perception.

## References


1. Achanta, R., Hemami, S., Estrada, F., Susstrunk, S.: Frequency-tuned salient region detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 1597–1604 (2009)
2. Chen, H., Li, Y.: Three-stream attention-aware network for rgb-d salient object detection. IEEE Transactions on Image Processing **28**(6), 2825–2835 (2019). https://doi.org/10.1109/TIP.2019.2891104
3. Cheng, M.M., Mitra, N.J., Huang, X., Torr, P.H.S., Hu, S.M.: Global contrast based salient region detection. IEEE Transactions on Pattern Analysis and Machine Intelligence **37**(3), 569–582 (2015). https://doi.org/10.1109/TPAMI.2014.2345401
4. Cong, R., Lin, Q., Zhang, C., Li, C., Zhao, Y., Kwong, S., Hoi, S.C.: Cir-net: Cross-modality interaction and refinement for rgb-d salient object detection. IEEE Transactions on Image Processing **31**, 6800–6815 (2022)
5. Cong, R., Zhang, K., Zhang, C., Zheng, F., Zhao, Y., Huang, Q., Kwong, S.: Does thermal really always matter for rgb-t salient object detection? IEEE Transactions on Multimedia **25**, 6971–6982 (2022)
6. 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 (ICCV). pp. 4548–4557 (2017)
7. 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 International Joint Conference on Artificial Intelligence (IJCAI). pp. 698–704 (2018)
8. Feng, D., Barnes, N., You, S., McCarthy, C.: Local background enclosure for rgb-d salient object detection. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 2343–2350 (2016). https://doi.org/10.1109/CVPR.2016.257
9. Fu, K., Fan, D.P., Ji, G.P., Zhao, Q.: Jl-dcf: Joint learning and densely-cooperative fusion framework for rgb-d salient object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 3052–3062 (2020)
10. Gao, W., Liao, G., Ma, S., Li, G., Liang, Y., Lin, W.: Unified information fusion network for multi-modal rgb-d and rgb-t salient object detection. IEEE Transactions on Circuits and Systems for Video Technology **32**(4), 2091–2106 (2022)
11. Han, Z., Zhang, C., Fu, H., Zhou, J.T.: Trusted multi-view classification with dynamic evidential fusion. IEEE transactions on pattern analysis and machine intelligence **45**(2), 2551–2566 (2022)
12. Hao, S., Zhong, C., Tang, H.: Cola: Conditional dropout and language-driven robust dual-modal salient object detection. In: European Conference on Computer Vision. pp. 354–371. Springer (2024)
13. He, J., Fu, K., Liu, X., Zhao, Q.: Samba: A unified mamba-based framework for general salient object detection. In: 2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 25314–25324 (2025). https://doi.org/10.1109/CVPR52734.2025.02357
14. He, T., Zhang, Z., Zhang, H., Zhang, Z., Xie, J., Li, M.: Bag of tricks for image classification with convolutional neural networks. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 558–567 (2019)
15. Hu, G., Wang, Z., Li, C., Yuan, D., He, B., Tang, J.: Missingness-aware prompting for modality-missing rgbt tracking. Journal of King Saud University Computer and Information Sciences **37**(6), 128 (2025)

16. Huo, F., Zhu, X., Zhang, L., Liu, Q., Shu, Y.: Efficient context-guided stacked refinement network for rgb-t salient object detection. IEEE Transactions on Circuits and Systems for Video Technology **32**(5), 3111–3124 (2021)
17. Huo, F., Zhu, X., Zhang, Q., Liu, Z., Yu, W.: Real-time one-stream semantic-guided refinement network for rgb-thermal salient object detection. IEEE Transactions on Instrumentation and Measurement **71**, 1–12 (2022)
18. Jacobs, R.A., Jordan, M.I., Nowlan, S.J., Hinton, G.E.: Adaptive mixtures of local experts. Neural Computation **3**(1), 79–87 (1991)
19. Jiang, Z., Davis, L.S.: Submodular salient region detection. In: 2013 IEEE Conference on Computer Vision and Pattern Recognition. pp. 2043–2050 (2013). https://doi.org/10.1109/CVPR.2013.266
20. Kim, J., Han, D., Tai, Y.W., Kim, J.: Salient region detection via high-dimensional color transform and local spatial support. IEEE Transactions on Image Processing **25**(1), 9–23 (2016). https://doi.org/10.1109/TIP.2015.2495122
21. Li, X., Zhao, L., Wei, L., Yang, M.H., Wu, F., Zhuang, Y., Ling, H., Wang, J.: Deepsaliency: Multi-task deep neural network model for salient object detection. IEEE Transactions on Image Processing **25**(8), 3919–3930 (2016). https://doi.org/10.1109/TIP.2016.2579306
22. 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)
23. Liu, N., Han, J., Yang, M.H.: Picanet: Pixel-wise contextual attention learning for accurate saliency detection. IEEE Transactions on Image Processing **29**, 6438–6451 (2020). https://doi.org/10.1109/TIP.2020.2988568
24. Liu, Z., Huang, X., Zhang, G., Fang, X., Wang, L., Tang, B.: Scribble-supervised rgb-t salient object detection. In: 2023 IEEE International Conference on Multimedia and Expo (ICME). pp. 2369–2374. IEEE (2023)
25. Luo, Z., Liu, N., Zhao, W., Yang, X., Zhang, D., Fan, D.P., Khan, F., Han, J.: Vscode: General visual salient and camouflaged object detection with 2d prompt learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 17169–17180 (2024)
26. Lv, C., Zhou, X., Wan, B., Wang, S., Sun, Y., Zhang, J., Yan, C.: Transformer-based cross-modal integration network for rgb-t salient object detection. IEEE Transactions on Consumer Electronics **70**(2), 4741–4755 (2024). https://doi.org/10.1109/TCE.2024.3390841
27. Ma, M., Xia, C., Xie, C., Chen, X., Li, J.: Boosting broader receptive fields for salient object detection. IEEE Transactions on Image Processing **32**, 1026–1038 (2023). https://doi.org/10.1109/TIP.2022.3232209
28. Margolin, R., Zelnik-Manor, L., Tal, A.: How to evaluate foreground maps? In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 248–255 (2014)
29. Pang, Y., Zhao, X., Zhang, L., Lu, H.: Caver: Cross-modal view-mixed transformer for bi-modal salient object detection. IEEE Transactions on Image Processing **32**, 892–904 (2023)
30. Perazzi, F., Krähenbühl, P., Pritch, Y., Hornung, A.: Saliency filters: Contrast based filtering for salient region detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 733–740 (2012)
31. Riquelme, C., Puigcerver, J., Mustafa, B., Neumann, M., Jenatton, R., Pinto, A.S., Keysers, D., Houlsby, N.: Scaling vision with sparse mixture of experts. In: Advances in Neural Information Processing Systems. vol. 34, pp. 8583–8595 (2021)

32. Shazeer, N., Mirhoseini, A., Maziarz, K., Davis, A., Le, Q., Hinton, G., Dean, J.: Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. arXiv preprint arXiv:1701.06538 (2017)
33. Song, K., Huang, L., Gong, A., Yan, Y.: Multiple graph affinity interactive network and a variable illumination dataset for rgbt image salient object detection. IEEE Transactions on Circuits and Systems for Video Technology **33**(7), 3104–3118 (2023). https://doi.org/10.1109/TCSVT.2022.3233131
34. Tang, B., Liu, Z., Tan, Y., He, Q.: Hrtransnet: Hrformer-driven two-modality salient object detection. IEEE Transactions on Circuits and Systems for Video Technology **33**(2), 728–742 (2022)
35. Tang, H., Li, Z., Zhang, D., He, S., Tang, J.: Divide-and-conquer: Confluent triple-flow network for rgb-t salient object detection. IEEE Transactions on Pattern Analysis and Machine Intelligence **47**(3), 1958–1974 (2025). https://doi.org/10.1109/TPAMI.2024.3511621
36. Tian, C., Yang, C., Zhu, G., Wang, Q., He, Z.: Learning a robust rgb-thermal detector for extreme modality imbalance. Pattern Recognition Letters **196**, 1–8 (2025)
37. Tu, Z., Li, Z., Li, C., Lang, Y., Tang, J.: Multi-interactive dual-decoder for rgb-thermal salient object detection. IEEE Transactions on Image Processing **30**, 5678–5691 (2021)
38. Tu, Z., Ma, Y., Li, Z., Li, C., Xu, J., Liu, Y.: Rgbt salient object detection: A large-scale dataset and benchmark. IEEE Transactions on Multimedia (2022), also arXiv:2007.03262
39. Tu, Z., Xia, T., Li, C., Lu, Y., Tang, J.: M3s-nir: Multi-modal multi-scale noise-insensitive ranking for rgb-t saliency detection. In: IEEE Conference on Multimedia Information Processing and Retrieval (MIPR). pp. 141–146 (2019)
40. Tu, Z., Xia, T., Li, C., Wang, X., Ma, Y., Tang, J.: Rgb-t image saliency detection via collaborative graph learning. IEEE Transactions on Multimedia **22**(1), 160–173 (2020)
41. Wan, B., Zhou, X., Sun, Y., Wang, T., Lv, C., Wang, S., Yin, H., Yan, C.: Mffnet: Multi-modal feature fusion network for v-d-t salient object detection. IEEE Transactions on Multimedia **26**, 2069–2081 (2024)
42. Wang, F., Pan, J., Xu, S., Tang, J.: Learning discriminative cross-modality features for rgb-d saliency detection. IEEE Transactions on Image Processing **31**, 1285–1297 (2022)
43. Wang, G., Li, C., Ma, Y., Zheng, A., Tang, J., Luo, B.: Rgb-t saliency detection benchmark: Dataset, baselines, analysis and a novel approach. In: China Conference on Image and Graphics Technologies and Applications. pp. 359–369. Springer (2018)
44. Wang, H., Song, K., Huang, L., Wen, H., Yan, Y.: Thermal images-aware guided early fusion network for cross-illumination rgb-t salient object detection. Engineering Applications of Artificial Intelligence **118**, 105640 (2023)
45. Wang, H., Guo, H., Chai, X., Mu, B., Shao, F.: Cognition-inspired dynamic feature integration network for rgb-d and rgb-t salient object detection. IEEE Transactions on Instrumentation and Measurement **74**, 1–16 (2025). https://doi.org/10.1109/TIM.2025.3600718
46. Wang, J., Song, K., Bao, Y., Huang, L., Yan, Y.: Cgfnet: Cross-guided fusion network for rgb-t salient object detection. IEEE Transactions on Circuits and Systems for Video Technology **32**(5), 2949–2961 (2022)

47. Wang, K., Chen, K., Li, C., Tu, Z., Luo, B.: Alignment-free rgb-t salient object detection: A large-scale dataset and progressive correlation network. In: Proceedings of the AAAI conference on artificial intelligence. vol. 39, pp. 7780–7788 (2025)
48. Wang, K., Lin, D., Li, C., Tu, Z., Luo, B.: Alignment-free rgbt salient object detection: Semantics-guided asymmetric correlation network and a unified benchmark. IEEE Transactions on Multimedia **26**, 10692–10707 (2024)
49. Wang, K., Tu, Z., Li, C., Zhang, C., Luo, B.: Learning adaptive fusion bank for multi-modal salient object detection. IEEE Transactions on Circuits and Systems for Video Technology **34**(8), 7344–7358 (2024)
50. Wang, Q., Sun, Y., Chi, Y., Shen, T.: Rgb-t object detection with failure scenarios. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing **18**, 3000–3010 (2024)
51. Wang, T., Borji, A., Zhang, L., Zhang, P., Lu, H.: A stagewise refinement model for detecting salient objects in images. In: 2017 IEEE International Conference on Computer Vision (ICCV). pp. 4039–4048 (2017). https://doi.org/10.1109/ICCV.2017.433
52. Wang, T., Zhang, L., Wang, S., Lu, H., Yang, G., Ruan, X., Borji, A.: Detect globally, refine locally: A novel approach to saliency detection. In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3127–3135 (2018). https://doi.org/10.1109/CVPR.2018.00330
53. Wang, Y., Wei, S., Xu, S., Qin, Y., Zhao, Y.: Confidence-driven unimodal interference removal for enhanced multimodal object detection. IEEE Transactions on Circuits and Systems for Video Technology (2025)
54. Wang, Y., Zhang, L., Zhang, P., Zhuge, Y., Wu, J., Yu, H., Lu, H.: Learning local-global representation for scribble-based rgb-d salient object detection via transformer. IEEE Transactions on Circuits and Systems for Video Technology **34**(11), 11592–11604 (2024). https://doi.org/10.1109/TCSVT.2024.3424651
55. Wu, Z., Gobichettipalayam, S., Tamadazte, B., Allibert, G., Paudel, D.P., Demonceaux, C.: Robust rgb-d fusion for saliency detection. In: International Conference on 3D Vision (3DV). pp. 403–413. IEEE (2022)
56. Wu, Z., Allibert, G., Meriaudeau, F., Demonceaux, C.: Source-free depth for object pop-out. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 1032–1042 (2023)
57. Wu, Z., Allibert, G., Mériaudeau, F., Ma, C., Demonceaux, C.: Hidanet: Rgb-d salient object detection via hierarchical depth awareness. IEEE Transactions on Image Processing **32**, 2160–2173 (2023)
58. Wu, Z., Wang, J., Zhou, Z., An, Z., Jiang, Q., Demonceaux, C., Sun, G., Timofte, R.: Object segmentation by mining cross-modal semantics. In: Proceedings of the 31st ACM International Conference on Multimedia. pp. 3455–3464 (2023)
59. Xiao, G., Liu, X., Lin, Z., Ming, R.: Smr-net: Semantic-guided mutually reinforcing network for cross-modal image fusion and salient object detection. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 39, pp. 8637–8645 (2025)
60. Xie, Z., Shao, F., Chen, G., Chen, H., Jiang, Q., Meng, X., Ho, Y.S.: Cross-modality double bidirectional interaction and fusion network for rgb-t salient object detection. IEEE Transactions on Circuits and Systems for Video Technology **33**(8), 4149–4163 (2023). https://doi.org/10.1109/TCSVT.2023.3241196
61. Xu, R., Luo, Y., Hu, H., Du, B., Shen, J., Wen, Y.: Rethinking the localization in weakly supervised object localization. In: ACM MM. pp. 5484–5494 (2023)
62. Yin, B.W., Cao, J.L., Cheng, M.M., Hou, Q.: Dformerv2: Geometry self-attention for rgbd semantic segmentation. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 19345–19355 (2025)

63. Zhang, P., Wang, D., Lu, H., Wang, H., Ruan, X.: Amulet: Aggregating multi-level convolutional features for salient object detection. In: Proceedings of the IEEE international conference on computer vision. pp. 202–211 (2017)
64. Zhang, P., Wang, D., Lu, H., Wang, H., Yin, B.: Learning uncertain convolutional features for accurate saliency detection. In: Proceedings of the IEEE International Conference on computer vision. pp. 212–221 (2017)
65. Zhang, Q., Huang, N., Yao, L., Zhang, D., Shan, C., Han, J.: Rgb-t salient object detection via fusing multi-level cnn features. IEEE Transactions on Image Processing **29**, 3321–3335 (2020). https://doi.org/10.1109/TIP.2019.2959253
66. Zhang, Q., Xiao, T., Huang, N., Zhang, D., Han, J.: Revisiting feature fusion for rgb-t salient object detection. IEEE Transactions on Circuits and Systems for Video Technology **31**(5), 1804–1818 (2021). https://doi.org/10.1109/TCSVT.2020.3014663
67. Zhao, J.X., Liu, J.J., Fan, D.P., Cao, Y., Yang, J., Cheng, M.M.: Egnet: Edge guidance network for salient object detection. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 8779–8788 (2019)
68. Zhao, T., Wu, X.: Pyramid feature attention network for saliency detection. In: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 3080–3089 (2019). https://doi.org/10.1109/CVPR.2019.00320
69. Zhou, W., Guo, Q., Lei, J., Yu, L., Hwang, J.N.: Ecffnet: Effective and consistent feature fusion network for rgb-t salient object detection. IEEE Transactions on Circuits and Systems for Video Technology **32**(3), 1224–1235 (2022)
70. Zhou, W., Sun, F., Jiang, Q., Cong, R., Hwang, J.N.: Wavenet: Wavelet network with knowledge distillation for rgb-t salient object detection. IEEE Transactions on Image Processing **32**, 3027–3039 (2023). https://doi.org/10.1109/TIP.2023.3275538
71. 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)