G-SalAlignMamba: Geometry-Aware Vision Mamba for Dual-Modal Salient Object Detection

Haixiao Gao[1], **Yimin Zheng**[1], **Mengke Song**[2*], **Linyou Xiao**[1], **Tian-Tian Zhang**[3], **Zhi-Ri Tang**[1*]
[1]School of Intelligent Systems Science and Engineering, Jinan University, China
[2]China University of Petroleum (East China), China
[3]Department of Electronic Engineering and Computer Science, Hong Kong Metropolitan University, Hong Kong
{ghx2004, zymmm}@stu2023.jnu.edu.cn, songsook@163.com, pingfkl@stu2023.jnu.edu.cn, ttzhang@hkmu.edu.hk, tangzhiri@jnu.edu.cn

Abstract

Recently, Visual State Space Models offer powerful global modeling for Dual-modal Salient Object Detection (SOD). However, they are still constrained by three inherent limitations: first, Mamba's strict reliance on sequential ordering makes it sensitive to cross-modal geometric misalignment, where spatial shifts disrupt token correspondence; second, general indiscriminate scanning treating all tokens equally may lead to signal dilution, where sparse foreground features are overwhelmed by background noise; third, conventional decoders rely on implicit upsampling, causing boundary degradation during resolution recovery. To address these challenges, we propose G-SalAlignMamba, a geometry-aware framework tailored for dual-modal SOD. We introduce Geometry-Aware Encoding with explicit alignment to correct spatial shifts, Semantics-Informed Refinement to prevent signal dilution by prioritizing foregrounds, and Structure-Preserving Decoding that integrates explicit alignment with unsupervised boundary refinement. Extensive experiments show that G-SalAlignMamba achieves state-of-the-art performance on RGB-D and RGB-T benchmarks with favorable efficiency (30.41 FPS, 83.80M parameters). The code is available at https://github.com/PC1-99/G-SalAlignMamb.git.

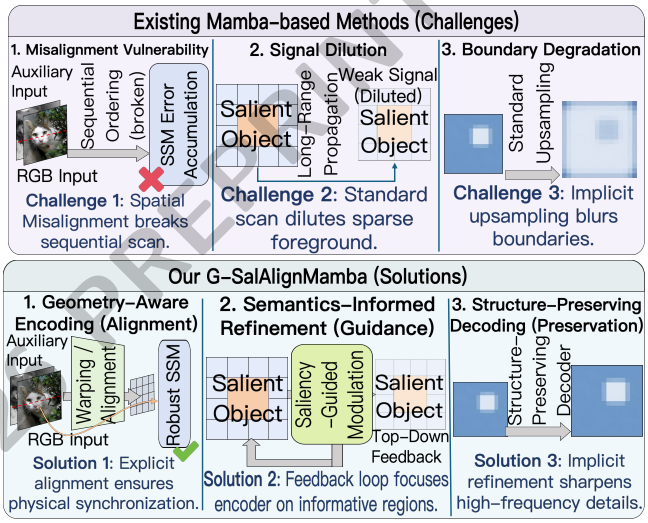

Figure 1: Top row illustrates three critical limitations in existing Mamba-based SOD: (a) Misalignment Vulnerability disrupts sequential state transitions; (b) Signal Dilution weakens sparse foreground signals during long-range propagation; (c) Boundary Degradation results from implicit upsampling. Bottom row presents our proposed G-SalAlignMamba solutions: (d) Geometry-Aware Encoding enforces physical synchronization; (e) Semantics-Informed Refinement prioritizes informative regions via feedback; (f) Structure-Preserving Decoding achieves sharp boundaries via implicit, unsupervised refinement.

1 Introduction

Dual-modal Salient Object Detection (SOD) aims to robustly identify visually attractive regions by integrating RGB images with auxiliary modalities (e.g., Depth or Thermal). This field has primarily relied on Convolutional Neural Networks (CNNs) [Fan *et al.*, 2020a] or Transformer architectures [Liu *et al.*, 2021]. Although these models have achieved significant progress, CNNs are limited by local receptive fields that restrict their ability to model complex global contexts. Transformers possess global receptive fields, but face computational costs that grow quadratically with resolution. To better balance computational efficiency and global modeling, Visual State Space Models (SSMs), represented by VMamba [Liu *et al.*, 2024b], have recently garnered widespread attention for their ability to model long-range dependencies with linear computational complexity $\mathcal{O}(N)$ [Gu and Dao, 2024]. While promising, we argue that the direct application of VMamba to dual-modal SOD is constrained by three critical limitations rooted in current methodologies, as shown in Figure 1:

1. Misalignment Vulnerability in Mamba Architecture. Most pioneering Mamba-based fusion frameworks follow a simple "concatenate-then-scan" paradigm [He *et al.*, 2025], while leaving cross-modal geometric misalignment insufficiently addressed. Since Mamba relies on a continuous se-

*Corresponding authors.

quential scanning order to update hidden states, physical spatial shifts caused by parallax or calibration errors can disrupt token correspondence within the scanning sequence. This Misalignment Vulnerability makes simple fusion strategies sensitive to spatial mismatches between modalities, causing the SSM model to construct inaccurate dependencies and accumulate errors.

2. Signal Dilution Caused by General Scanning Mechanisms. General Vision Mamba backbones (e.g., ViM [Zhu *et al.*, 2024], VMamba [Liu *et al.*, 2024b]) are largely designed for image classification tasks, where feature distributions are assumed to be relatively uniform. However, in SOD tasks, salient objects typically occupy only a sparse fraction of the image. General raster scanning mechanisms treat all tokens indiscriminately, lacking a discriminative bias towards target regions. In the recursive propagation mechanism of SSMs, the current state depends on the accumulation of all historical information. This implies that sparse foreground signals must traverse extremely long spatial sequences. During this long-range propagation, effective information is prone to being "diluted" or overwhelmed by the vast majority of background noise. Without task-specific modulation, the Mamba encoder easily fails to maintain the prominence of salient features in deep layers.

3. Boundary Degradation During Sequence Reconstruction. Existing Vision Mamba variants also face a "coupling gap" between the SSMs' sequence representation and the integrity of the image's 2D physical structure during resolution recovery. Current decoders in general Vision Mamba models still adhere to traditional practices in vision tasks, employing spatially-agnostic implicit upsampling operators (e.g., bilinear interpolation) [Zhu *et al.*, 2024; Liu *et al.*, 2024b]. These operations lack awareness of Mamba's scanning characteristics and fail to utilize the geometric alignment cues captured during the encoding stage. This dimensional projection distortion during the "sequence-to-space" reconstruction not only triggers pixel-level Correspondence Confusion, but also leads to blurred edges and structural deterioration of salient targets, forming a structural bottleneck for accurate SOD [Qin *et al.*, 2019; Zhao *et al.*, 2019].

To address these interconnected issues, we propose **G-SalAlignMamba**, a unified geometry-aware framework tailored for dual-modal SOD. Guided by an *Alignment-Guidance-Preservation* philosophy, we first introduce a **Geometry-Aware Encoding** scheme with coarse-to-fine explicit alignment to enforce physical synchronization, resolving the misalignment vulnerability. Second, a **Semantics-Informed Refinement** strategy is proposed to modulate the scanning process via a top-down feedback loop, preventing signal dilution. Finally, a **Structure-Preserving Decoding** scheme is presented for extending geometric constraints to the upsampling stage, ensuring sharp boundary recovery without explicit supervision. The code is available at https://github.com/PC1-99/G-SalAlignMamb.git.

Our main contributions are summarized as follows:

- We first identify the intrinsic Misalignment Vulnerability of Mamba's sequential modeling. To address this, we propose **G-SalAlignMamba**, pioneering a **Geometry-Aware Encoding** scheme that enforces physical synchronization to construct a robust foundation for multi-modal SSMs.
- We propose a **Semantics-Informed Refinement** strategy to counter the Signal Dilution inherent in generic scanning. By constructing a top-down semantic feedback loop, we empower the encoder to prioritize sparse salient signals, preventing them from being overwhelmed by background noise.
- We design a **Structure-Preserving Decoding** scheme to bridge the geometric gap in conventional upsampling. This module enforces geometry-consistent reconstruction with an unsupervised boundary refinement mechanism, effectively solving Boundary Degradation without explicit supervision.
- Extensive experiments on RGB-D and RGB-T benchmarks demonstrate that **G-SalAlignMamba** achieves state-of-the-art performance, validating the necessity of geometric constraints in Mamba-based fusion.

2 Related Work

2.1 Visual Mamba

The success of SSMs has driven their extension to the visual domain, where Vision Mamba (ViM) [Zhu *et al.*, 2024] and VMamba [Liu *et al.*, 2024b] achieve efficient linear-complexity modeling for fundamental recognition tasks. This advantage has rapidly sparked widespread adoption across dense prediction fields [Xing *et al.*, 2024]. In the context of SOD, recent pioneers like Samba [He *et al.*, 2025] have also validated the potential of Mamba for maintaining spatial continuity. However, current works primarily focus on single-modal representations or simple fusion strategies, usually overlooking the inherent cross-modal spatial and geometric disparities in dual-modal settings. This limitation inspires us to propose a geometry-aware framework that explicitly incorporates alignment mechanisms into the VMamba architecture to suppress misalignment noise.

2.2 RGB-Depth Salient Object Detection (RGB-D SOD)

To address RGB-only SOD limitations, many approaches incorporate depth information, named RGB-D SOD, which aims to enhance the model's geometric and 3D structural understanding. Depth helps detect salient objects when color/texture contrasts are weak. A representative work [Fu *et al.*, 2021] used Siamese networks with joint learning and densely-cooperative fusion, improving robustness. Another method [Ji *et al.*, 2021] tackled depth noise by using depth calibration followed by cross-modal fusion, boosting saliency detection. However, existing methods still struggle with low-quality depth maps and cross-modal spatial misalignment.

2.3 RGB-Thermal Salient Object Detection (RGB-T SOD)

RGB-T SOD fuses RGB and thermal imagery, particularly useful in low-light, nocturnal, or high-thermal-contrast environments. RGB-T methods utilize multi-stream networks, attention-based fusion, and multi-scale decoding to integrate

RGB and thermal information [Pang *et al.*, 2023]. For example, [Zhou *et al.*, 2023] has shown that modeling pixel-wise positional relations and using Transformer-based encoders and decoders leads to more accurate saliency masks in challenging thermal and visible scenarios. Thermal data significantly enhances performance when RGB data alone is insufficient. Nevertheless, they often fail to handle severe parallax and thermal noise interference in complex scenes [Tu *et al.*, 2022a].

3 Method

To systematically resolve the issues of spatial misalignment and feature degradation in Mamba-based SOD, given paired RGB and auxiliary inputs $I_R, I_A \in \mathbb{R}^{3\times H\times W}$, we propose a geometry-aware dual-modal SOD framework to address cross-modal misalignment and signal dilution, as illustrated in Figure 2. The encoder employs a dual-branch Visual State-Space (VSS) backbone with lightweight Early Interaction to stabilize intermediate features, followed by a Coarse-to-Fine Explicit Alignment module that corrects spatial shifts via multi-scale displacement fields and a Complementary Gated Fusion block to produce a deeply fused representation F_{fused}. The refinement module applies Semantics-Informed Refinement to shallow features, modulating them via top-down Saliency Priors derived from F_{fused} to prioritize foreground signals. Finally, the Structure-Preserving Decoding scheme progressively reconstructs the saliency map by integrating Context-Aware Alignment Upsampling with unsupervised Multi-Scale Boundary Refinement, ensuring precise structural recovery and sharp edges essential for high-performance SOD.

3.1 Geometry-Aware Encoding: Cross-Modal Alignment and Fusion

Hierarchical Feature Extraction with Early Interaction

To extract robust representations from RGB image $\mathbf{I}_R$ and auxiliary input $\mathbf{I}_A$, we employ a dual-branch encoder based on VSS architecture. Different from CNNs, VSS captures long-range dependencies with linear complexity $\mathcal{O}(N)$.

We share parameters θ_{VSS} between branches to learn modality-agnostic geometric representations and reduce parameter overhead as:

$$\mathbf{F}_R^{(i)}, \mathbf{F}_A^{(i)} = \text{VSS}(\mathbf{I}_R, \mathbf{I}_A; \theta_{\text{VSS}}), \quad i \in \{1,2,3,4\}. \tag{1}$$

To prevent semantic divergence in deep layers, we introduce a lightweight Early Interaction mechanism. At stage $i = 2$, intermediate features are stabilized via window-based cross-attention as:

$$\mathbf{F}_R'^{(2)} = \mathbf{F}_R^{(2)} + g_{\text{early}} \cdot \text{WindowCrossAttn}(\mathbf{F}_R^{(2)}, \mathbf{F}_A^{(2)}), \tag{2}$$

where $g_{\text{early}} = \sigma(\gamma_{\text{early}})$ is a learnable gate with γ_{early} (initialized to weakly inject cross-modal information to prevent early degradation). This ensures that the subsequent alignment module operates on features with compatible semantic levels.

Coarse-to-Fine Explicit Alignment

Implicit alignment in SOD often struggles with parallax-induced displacements, leading to blurred boundaries. Hence, we propose a **Coarse-to-Fine Explicit Alignment** strategy that predicts a pixel-wise displacement field $\Delta\mathbf{p}$ through multi-scale feature extraction and cross-attention fusion.

To capture misalignment at varying magnitudes, we construct a feature pyramid at three resolutions (original, $1/2$, and $1/4$) by projecting the stage-4 representations via scale-specific convolutions. For each scale $s \in \{0,1,2\}$, these multi-scale features are subsequently enhanced using cross-attention to yield the fused representation $\mathbf{F}_{\text{fused}}^{(s)}$. The final displacement field $\Delta\mathbf{p}$ is an adaptive aggregation of multi-scale predictions as:

$$\Delta\mathbf{p} = \sum_{s=0}^{2} w_s \cdot \text{Predictor}_s(\mathbf{F}_{\text{fused}}^{(s)}), \quad \text{s.t.} \sum w_s = 1, \tag{3}$$

where w_s are learnable weights derived from Softmax. The auxiliary features are then explicitly aligned to the RGB coordinate system via differentiable warping $\mathcal{W}$:

$$\mathbf{F}_A^{\text{align}} = \mathcal{W}(\mathbf{F}_A^{(4)}, \Delta\mathbf{p}), \tag{4}$$

where $\Delta\mathbf{p}$ is bounded by a learnable maximum offset $\alpha_{\max}$ through $\tanh$ normalization. The explicit rectification eliminates geometric discrepancies, ensuring the spatial consistency critical for preserving sharp boundaries.

Complementary Gated Fusion

To address the failure of naive fusion in dual-modal SOD caused by indiscriminately treating noisy inputs, we propose a **Complementary Gated Fusion** block to balance shared and complementary cues:

1. **Commonality Path:** We integrate shared semantics using concatenation and Selective Scanning via the Concat Mamba Fusion Block:

$$\mathbf{F}_{\text{common}} = \text{ConcatMamba}(\mathbf{F}_R^{(4)}, \mathbf{F}_A^{\text{align}}). \tag{5}$$

This path concatenates RGB and aligned auxiliary features, then sequentially processes them through linear projection, reshape, depthwise convolution, Saliency-Guided SS2D (SG-SS2D) [He *et al.*, 2025], layer normalization (LN), and a final reshape to enhance feature representation.

2. **Difference-Aware Path:** To explicitly model complementary signals, we compute feature differences between the RGB and aligned auxiliary modalities across multiple scales. The difference maps are encoded, gated via adaptive thresholds, and processed by lightweight SS2D blocks, before being upsampled and fused into the final difference representation $\mathbf{F}_{\text{diff}}$. A learnable gating mechanism dynamically balances these paths as:

$$\mathbf{F}_{\text{fused}} = (1-g)\cdot\mathbf{F}_{\text{common}} + g\cdot\mathbf{F}_{\text{diff}} + \mathbf{F}_R^{(4)}, \tag{6}$$

where $g = \sigma(\gamma_{\text{gate}})$ with γ_{gate} (initialized to bias towards the commonality path). The residual connection with $\mathbf{F}_R$ ensures gradient flow and feature preservation. Consequently, this mechanism adaptively exploits complementarity with suppressing noise, which could ensure robust SOD performance even with low-quality inputs.

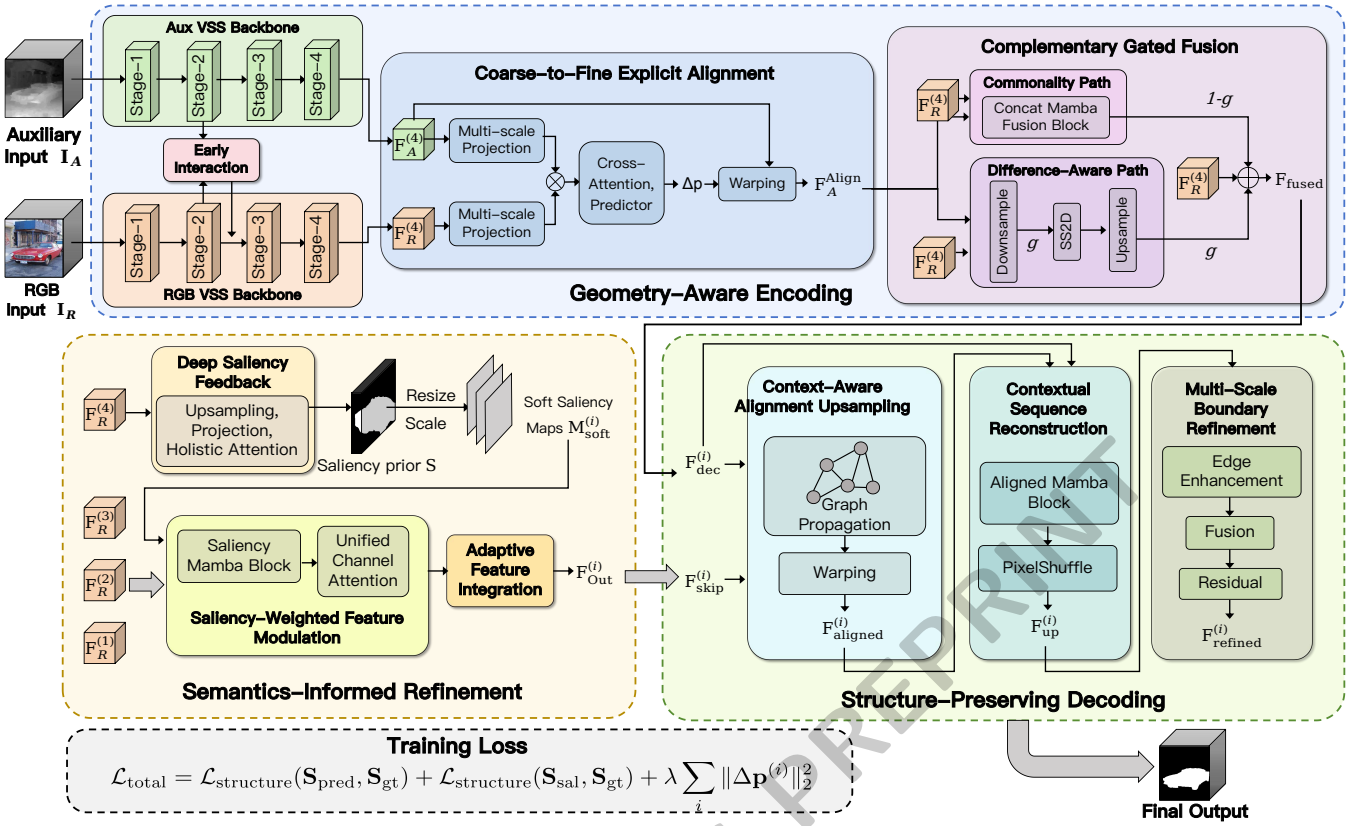

Figure 2: An overview of our proposed G-SalAlignMamba, including Geometry-Aware Encoding for dual-modal features alignment and fusion, Semantics-Informed Refinement to prevent signal dilution, and Structure-Preserving Decoding for geometry-consistent reconstruction.

3.2 Semantics-Informed Refinement: Mitigating Signal Dilution

In general Vision Mamba architectures, the sequential scanning mechanism treats all spatial tokens equally. However, salient objects often occupy a small fraction of the image in SOD tasks. Hence, general methods may lead to *signal dilution*, where strong foreground signals are overwhelmed by background tokens during state transitions. We address this issue via a feedback-driven refinement scheme.

Deep Saliency Feedback Mechanism

We first introduce a top-down feedback loop inspired by cognitive attention. We utilize the deeply fused representation $\mathbf{F}_R^{(4)}$ (from the final encoder stage) to generate a coarse **Saliency Prior S** to guide shallow and high-resolution features.

Formally, we generate a raw prior $\mathbf{S}_{\text{raw}}$ by projecting and upsampling the deep semantic features, followed by spatial smoothing via a Holistic Attention (HA) module to ensure coherence, ultimately yielding the final coarse saliency prior $\mathbf{S}$. The prior is then resized and transformed into a soft modulation mask via temperature scaling to guide the refinement:

$$\mathbf{S} = \text{HA}(\sigma(\mathbf{S}_{\text{raw}})) \in [0,1]^{H \times W}, \tag{7}$$

where $\text{HA}(\cdot)$ denotes the Holistic Attention module that applies Gaussian smoothing for spatial coherence. This method leverages high-level semantic context to modulate low-level details, ensuring the enhancement module focuses computational resources on potential foreground regions before features propagate to the decoder.

Saliency-Weighted Feature Modulation

General state-space models' equal token treatment causes *feature dilution* in SOD tasks, where background often overwhelms sparse foregrounds. We propose a **Saliency-Weighted Modulation** strategy that utilizes a saliency-derived mask to modulate scanning, preserving foreground signals without altering sequence length.

For an input feature $\mathbf{F}_R^{(i)}$ at stage $i \in \{1,2,3\}$, the saliency prior $\mathbf{S}$ is resized to match the specific spatial resolution (H_i, W_i), yielding the mask $\mathbf{M}^{(i)}$. The mask is shaped using temperature scaling as:

$$\mathbf{M}_{\text{soft}}^{(i)} = \sigma\left(\frac{\mathbf{M}^{(i)}}{\tau}\right)^{\gamma}, \tag{8}$$

where τ is the temperature parameter and γ is the power scaling factor. Guided by the soft mask $\mathbf{M}_{\text{soft}}^{(i)}$, the input features $\mathbf{F}_R^{(i)}$ are processed by the Saliency Mamba Block—comprising linear projection, reshape, DWConv, SG-SS2D [He *et al.*, 2025], LN, and a final reshape to yield the refined representations.

The soft modulation ensures that the hidden state of the SSMs is primarily influenced by informative foreground signals, thereby preserving the integrity of salient features with suppressing background noise to yield the prior $\mathbf{F}^{(i)}_{\text{refined}}$. The saliency-weighted modulation adds negligible computational overhead, maintaining inference efficiency comparable to general scanning approaches.

Additionally, to enhance channel specificity, we integrate a unified channel attention module $\mathcal{A}_{\text{channel}}$ (combining SE [Hu *et al.*, 2018] and ECA mechanisms [Wang *et al.*, 2020]) within the Saliency Mamba Block as:

$$\mathbf{F}^{(i)}_{\text{refined}} = \mathbf{F}^{(i)}_{\text{refined}} + \alpha \cdot \mathcal{A}_{\text{channel}}(\mathbf{F}^{(i)}_{\text{refined}}), \tag{9}$$

where α is a learnable balancing parameter, enhancing the signal-to-noise ratio for precise and complete SOD results.

Adaptive Feature Integration

Finally, to regulate the information flow to the decoder, we employ an Adaptive Feature Integration mechanism. The refined features $\mathbf{F}^{(i)}_{\text{refined}}$ are modulated by a learnable gating factor $\sigma(\gamma_i)$. This method allows the network to dynamically determine the importance of the refined features, selectively accepting informative signals with suppressing any residual artifacts or noise before decoding, ultimately producing the integrated feature representation $\mathbf{F}^{(i)}_{\text{out}}$.

3.3 Structure-Preserving Decoding: Precise Upsampling

General decoders often suffer from feature blurring during upsampling, which is particularly detrimental to SOD, where precise boundary delineation is critical for high evaluation metrics. To address this issue, we propose a structure-preserving scheme that enforces explicit alignment. The decoder restores the resolution of the deepest encoder feature $\mathbf{F}_{\text{fused}}$ to the original scale.

Context-Aware Alignment Upsampling

In SOD tasks, spatial misalignment in skip connections often leads to boundary ghosting. The decoding process operates over three stages $i \in \{1, 2, 3\}$. At each stage, the module simultaneously processes the current decoder feature $\mathbf{F}^{(i)}_{\text{dec}}$ and the skip-connected feature $\mathbf{F}^{(i)}_{\text{skip}}$ from the encoder's shallow layers. Specifically, $\mathbf{F}^{(i)}_{\text{dec}} \in \mathbb{R}^{B\times C_i\times H_i\times W_i}$ denotes the feature from the previous decoding stage (initialized as $\mathbf{F}_{\text{fused}}$ for $i = 1$). Correspondingly, $\mathbf{F}^{(i)}_{\text{skip}} \in \mathbb{R}^{B\times C^{(i)}_{\text{skip}}\times H_i\times W_i}$ represents the skip feature derived from encoder stage $(4-i)$ (i.e., $\mathbf{F}^{(3)}_{\text{out}}$, $\mathbf{F}^{(2)}_{\text{out}}$, or $\mathbf{F}^{(1)}_{\text{out}}$ for $i = 1, 2, 3$ respectively).

To maintain feature correspondence with recovering details, we introduce a **Context-Aware Alignment Upsampling** mechanism. We first project the concatenated decoder and skip features into node embeddings $\mathbf{N}^{(i)}$ and refine them via anisotropic graph propagation to capture boundary cues:

$$\mathbf{N}^{(i)} = \text{NodeEmbed}(\text{Concat}(\mathbf{F}^{(i)}_{\text{dec}}, \mathbf{F}^{(i)}_{\text{skip}})), \tag{10}$$

$$\mathbf{N}^{(i)} \leftarrow \mathbf{N}^{(i)} + \beta \cdot \sum_{d\in\mathcal{D}} \mathbf{w}_d \odot (\mathbf{N}_d - \mathbf{N}^{(i)}), \tag{11}$$

where β is a learnable diffusion strength, and $\mathcal{D} = \{\text{up}, \text{down}, \text{left}, \text{right}\}$ denotes the set of four cardinal neighbor directions used for anisotropic propagation. The refined nodes guide the prediction of a dense offset field $\Delta\mathbf{p}^{(i)}$, which aligns the skip features to the decoder's coordinate system via differentiable warping $\mathcal{W}$:

$$\Delta\mathbf{p}^{(i)} = \tanh(\mathcal{R}_\Delta(\text{Concat}(\mathbf{F}^{(i)}_{\text{concat}}, \mathbf{N}^{(i)}_{\text{prop}}))) \cdot \alpha_{\max}, \tag{12}$$

$$\mathbf{F}^{(i)}_{\text{aligned}} = \mathcal{W}(\mathbf{F}^{(i)}_{\text{skip}}, \text{Grid} + \Delta\mathbf{p}^{(i)}). \tag{13}$$

The explicit spatial shift ensures feature correspondence and preserves object structure during upsampling. Besides, this rectification guarantees the spatial precision required for high-quality SOD.

Contextual Sequence Reconstruction

In the decoding phase, the Contextual Sequence Reconstruction block (AlignedMambaBlock) plays a vital role in recovering spatial details. It takes the upsampled decoder feature $\mathbf{F}^{(i)}_{\text{dec}}$ and the explicitly aligned skip-connection feature $\mathbf{F}^{(i)}_{\text{aligned}}$ as inputs. Under the guidance of the resized saliency prior $\mathbf{P}_{\text{sal}}$, these features are fused to model global context and dependencies. The reconstructed sequence is then expanded and upsampled via a PixelShuffle operation to generate the high-resolution feature $\mathbf{F}^{(i)}_{\text{up}}$ for the subsequent stage, aiming to preserve the semantic completeness in SOD tasks.

Multi-Scale Boundary Refinement

Preserving high-frequency details is critical for distinguishing salient foregrounds in SOD tasks. To avoid boundary degradation, we employ an unsupervised Multi-Scale Boundary Refinement module that implicitly sharpens high-frequency details without requiring explicit edge supervision. The module operates by extracting local boundary cues from $\mathbf{F}^{(i)}_{\text{up}}$ using parallel convolutional branches at three distinct scales ($k \in \{1, 2, 4\}$). These multi-scale boundary features are subsequently upsampled, concatenated, and fused back into the main feature stream via a residual connection. This process yields the final refined output $\mathbf{F}^{(i)}_{\text{refined}}$, characterized by sharper edges and clearer structural delineations.

Loss Functions

To stabilize training, we apply L2 regularization on the offset fields. The total loss combines the structure loss $\mathcal{L}_{\text{structure}}$ for both main and auxiliary outputs with the offset regularization:

$$\begin{aligned}\mathcal{L}_{\text{total}} &= \mathcal{L}_{\text{structure}}(\mathbf{S}_{\text{pred}}, \mathbf{S}_{\text{gt}}) \\ &\quad + \mathcal{L}_{\text{structure}}(\mathbf{S}_{\text{sal}}, \mathbf{S}_{\text{gt}}) + \lambda \sum_i \|\Delta\mathbf{p}^{(i)}\|_2^2,\end{aligned} \tag{14}$$

where no explicit boundary supervision is required. This composite loss drives structural accuracy, ensuring robust convergence for SOD tasks.

4 Experiment

4.1 Datasets and Metrics

For RGB-D SOD, we use five benchmark datasets: NJUD [Ju *et al.*, 2014], NLPR [Peng *et al.*, 2014], SIP [Fan *et al.*,

Method	Type	NJUD			NLPR			SIP			STERE			DUTLF-D			Params (M)	FPS
		$S_m\uparrow$	$F_m\uparrow$	$E_m\uparrow$	$S_m\uparrow$	$F_m\uparrow$	$E_m\uparrow$	$S_m\uparrow$	$F_m\uparrow$	$E_m\uparrow$	$S_m\uparrow$	$F_m\uparrow$	$E_m\uparrow$	$S_m\uparrow$	$F_m\uparrow$	$E_m\uparrow$		
BBSNet [Fan *et al.*, 2020b]	CNN	0.921	0.919	0.949	0.931	0.918	0.961	0.879	0.884	0.922	0.908	0.903	0.942	0.882	0.870	0.912	49.77	45.21
JL-DCF [Fu *et al.*, 2020]		0.877	0.892	0.941	0.931	0.918	0.965	0.885	0.894	0.931	0.900	0.895	0.942	0.894	0.891	0.927	143.52	22.15
SP-Net [Zhou *et al.*, 2021]		0.925	0.928	0.957	0.927	0.919	0.962	0.894	0.904	0.933	0.907	0.906	0.949	0.895	0.899	0.933	67.88	47.38
DCF [Ji *et al.*, 2021]		0.904	0.905	0.943	0.922	0.910	0.957	0.874	0.886	0.922	0.906	0.904	0.948	0.925	0.930	0.956	53.92	32.64
SPSN [Lee *et al.*, 2022]		0.918	0.921	0.952	0.923	0.912	0.960	0.892	0.900	0.936	0.907	0.902	0.945	-	-	-	-	-
SwinNet-B [Liu *et al.*, 2021]	Trans.	0.920	0.924	0.956	0.941	0.936	0.974	0.911	0.927	0.950	0.919	0.918	0.956	0.918	0.920	0.949	199.18	11.42
CATNet [Sun *et al.*, 2023]		0.932	0.937	0.960	0.938	0.934	0.971	0.910	0.928	0.951	0.920	0.922	0.958	0.952	0.958	0.975	262.73	23.09
VST-S++ [Liu *et al.*, 2024a]		0.928	0.928	0.957	0.935	0.925	0.964	0.904	0.918	0.946	0.921	0.916	0.954	0.945	0.950	0.969	143.15	27.85
CPNet [Hu *et al.*, 2024]		0.935	0.941	0.963	0.940	0.936	0.971	0.907	0.927	0.946	0.920	0.922	0.960	0.951	0.959	0.974	216.50	20.17
VSCode-S [Luo *et al.*, 2024]		0.944	0.949	0.970	0.941	0.932	0.968	0.924	0.942	0.958	0.931	0.928	0.958	0.960	0.967	0.980	74.72	30.56
DiMSOD [Zhang *et al.*, 2025]	Diff.	0.947	0.947	0.969	-	-	-	-	-	-	-	-	-	0.957	0.967	0.951	-	-
Samba [He *et al.*, 2025]	Mamba	0.945	0.938	0.969	0.939	0.926	0.976	0.925	0.929	0.946	0.925	0.929	0.951	0.958	0.941	0.968	62.86	37.24
G-SalAlignMamba (Ours)		0.956	0.948	0.983	0.947	0.932	0.978	0.929	0.941	0.955	0.935	0.932	0.958	0.977	0.954	0.981	83.80	30.41

Table 1: Quantitative comparison of our method against other recent advanced RGB-D SOD methods on five benchmark datasets. "-" indicates the result is not available. "↑" denotes that the larger value is better. The best two results are stressed in red and blue.

Method	Type	VT5000			VT1000			VT821			un-VT5000			un-VT1000			un-VT821			UVT2000			Params (M)	FPS
		$S_m\uparrow$	$F_m\uparrow$	$E_m\uparrow$	$S_m\uparrow$	$F_m\uparrow$	$E_m\uparrow$	$S_m\uparrow$	$F_m\uparrow$	$E_m\uparrow$	$S_m\uparrow$	$F_m\uparrow$	$E_m\uparrow$	$S_m\uparrow$	$F_m\uparrow$	$E_m\uparrow$	$S_m\uparrow$	$F_m\uparrow$	$E_m\uparrow$	$S_m\uparrow$	$F_m\uparrow$	$E_m\uparrow$		
CSRNet [Huo *et al.*, 2021]	CNN	0.868	0.811	0.905	0.918	0.877	0.925	0.884	0.831	0.909	0.642	0.475	0.713	0.705	0.582	0.732	0.737	0.619	0.783	0.655	0.420	0.658	1.03	40.51
CGFNet [Wang *et al.*, 2021]		0.883	0.851	0.922	0.923	0.906	0.944	0.881	0.845	0.912	0.833	0.757	0.868	0.877	0.817	0.878	0.837	0.776	0.874	0.764	0.539	0.704	18.37	29.73
DCNet [Zhu *et al.*, 2025]		0.895	0.847	0.920	0.929	0.911	0.948	0.899	0.842	0.912	0.858	0.803	0.879	0.858	0.850	0.880	0.817	0.793	0.869	0.767	0.632	0.808	24.15	14.77
CAVER [Pang *et al.*, 2023]		0.892	0.841	0.924	0.936	0.903	0.945	0.891	0.839	0.924	0.850	0.805	0.893	0.873	0.838	0.881	0.818	0.780	0.856	0.786	0.616	0.782	93.84	16.32
SPNet [Zhang *et al.*, 2023]		0.914	0.880	0.948	0.941	0.925	0.954	0.913	0.873	0.936	0.900	0.848	0.929	0.931	0.902	0.938	0.894	0.833	0.910	0.765	0.558	0.733	113.59	29.93
SwinNet [Liu *et al.*, 2021]	Trans.	0.912	0.865	0.942	0.938	0.896	0.947	0.904	0.847	0.926	0.837	0.767	0.901	0.853	0.802	0.871	0.854	0.783	0.903	0.790	0.592	0.780	198.74	26.19
LSNet [Liu *et al.*, 2022]		0.877	0.825	0.915	0.925	0.885	0.935	0.878	0.825	0.911	0.847	0.767	0.892	0.868	0.797	0.875	0.842	0.754	0.875	0.763	0.527	0.711	4.64	88.83
SACNet [Wang *et al.*, 2024]		0.917	0.901	0.957	0.942	0.923	0.958	0.906	0.868	0.932	0.872	0.780	0.899	0.852	0.803	0.868	0.876	0.812	0.916	0.795	0.601	0.792	530.90	19.00
PCNet [Wang *et al.*, 2025]		0.920	0.899	0.956	0.943	0.924	0.958	0.915	0.879	0.941	0.879	0.861	0.936	0.922	0.904	0.947	0.893	0.869	0.936	0.819	0.686	0.851	-	-
DiMSOD [Zhang *et al.*, 2025]	Diff.	0.921	0.898	0.959	0.953	0.935	0.955	0.923	0.917	0.949	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Samba [He *et al.*, 2025]	Mamba	0.924	0.904	0.949	0.936	0.936	0.959	0.909	0.918	0.935	0.853	0.834	0.907	0.913	0.891	0.904	0.858	0.880	0.918	0.832	0.689	0.763	62.86	37.24
G-SalAlignMamba (Ours)		0.939	0.913	0.956	0.941	0.939	0.961	0.921	0.919	0.940	0.926	0.849	0.931	0.935	0.910	0.913	0.871	0.909	0.930	0.905	0.704	0.788	83.80	30.41

Table 2: Quantitative comparison of our method against other recent advanced RGB-T SOD methods on seven benchmark datasets. "-" indicates the result is not available. "↑" denotes that the larger value is better. The best two results are stressed in red and blue.

2020a], STERE [Niu *et al.*, 2012] and DUTLF-D [Piao *et al.*, 2019]. For RGB-T SOD, we adopt seven datasets: VT5000 [Tu *et al.*, 2022b], VT1000 [Tu *et al.*, 2019], VT821 [Wang *et al.*, 2018], un-VT5000 [Tu *et al.*, 2022a], un-VT1000 [Tu *et al.*, 2022a], un-VT821 [Tu *et al.*, 2022a] and UVT2000 [Wang *et al.*, 2024]. For RGB-D and RGB-T SOD, we employ three standard metrics: **E-measure** [Fan *et al.*, 2018] (E_m) jointly assesses pixel-level accuracy and image-level consistency; **S-measure** [Fan *et al.*, 2017] (S_m) evaluates structural similarity; and **F-measure** [Achanta *et al.*, 2009] (F_m) balances precision and recall.

4.2 Implementation Details

Our method is implemented in PyTorch and trained on an NVIDIA GeForce RTX 4090 GPU. We train separate models for the RGB-D SOD and RGB-T SOD tasks, respectively. Following previous works, we arrange the training sets for each task as follows: the training sets of NJUD, NLPR, and DUTLF-D for RGB-D SOD; the training sets of VT5000 and UVT2000 for RGB-T SOD. For the specific hyperparameter settings, we initialize the learnable gating factors γ_{early} and γ_{gate} to -2.0, while the adaptive integration factor γ_i is initialized to 2.0. The maximum offset for explicit alignment is bounded by $\alpha_{\max} = 3.0$. Both the temperature parameter τ and the power scaling factor γ are set to 1.0. The weight λ in the loss function is set to 0.01. In the training process, we adopt the AdamW optimizer with an initial learning rate of 1×10^{-4} and a batch size of 2. All input images are uniformly resized to 448×448 for training and testing. The model converges after 30 training epochs.

4.3 Model Comparison

Quantitative Evaluation. To demonstrate the universality of our unified framework in dual-modal SOD, we compare it with 12 RGB-D and 11 RGB-T state-of-the-art methods, including the representative Mamba-based Samba baseline, in Table 1 and Table 2. Overall, G-SalAlignMamba delivers consistently competitive or superior performance across all three metrics.

For RGB-D (Table 1), our model gives the best S_m scores on all five datasets and remains highly competitive in F_m and E_m; compared with Samba, it improves NJUD from $[0.945, 0.938, 0.969]$ to $[0.956, 0.948, 0.983]$ across (S_m, F_m, E_m), showing the benefit of explicit cross-modal alignment over "concatenate-then-scan" fusion. For RGB-T (Table 2), it also achieves state-of-the-art results, improving VT5000 over Samba from $[0.924, 0.904, 0.949]$ to $[0.939, 0.913, 0.956]$ and yielding clear gains on misaligned

Idx	Encoder			Refinement			Decoder			NJUD			NLPR			VT5000			UVT2000		
	EI	CFEA	CGF	DSFM	SWM	AFI	CAAU	CSR	MBR	$S_m \uparrow$	$F_m \uparrow$	$E_m \uparrow$	$S_m \uparrow$	$F_m \uparrow$	$E_m \uparrow$	$S_m \uparrow$	$F_m \uparrow$	$E_m \uparrow$	$S_m \uparrow$	$F_m \uparrow$	$E_m \uparrow$
Base										0.852	0.903	0.911	0.842	0.892	0.909	0.850	0.835	0.910	0.747	0.589	0.663
E1	✓									0.864	0.913	0.915	0.858	0.902	0.911	0.856	0.847	0.911	0.762	0.601	0.681
E2		✓								0.869	0.917	0.917	0.866	0.903	0.912	0.860	0.852	0.914	0.775	0.612	0.694
E3			✓							0.879	0.921	0.921	0.888	0.905	0.916	0.877	0.859	0.916	0.789	0.626	0.707
E4	✓	✓								0.883	0.922	0.922	0.906	0.910	0.919	0.879	0.860	0.915	0.801	0.638	0.721
E5	✓	✓	✓							0.903	0.923	0.931	0.909	0.917	0.930	0.889	0.864	0.917	0.818	0.652	0.735
H1	✓	✓	✓	✓						0.905	0.924	0.933	0.909	0.917	0.931	0.895	0.870	0.918	0.824	0.660	0.742
H2	✓	✓	✓		✓					0.907	0.925	0.936	0.910	0.918	0.933	0.901	0.879	0.920	0.833	0.668	0.750
H3	✓	✓	✓			✓				0.911	0.922	0.932	0.910	0.920	0.930	0.911	0.901	0.921	0.838	0.672	0.747
H4	✓	✓	✓	✓	✓					0.910	0.927	0.940	0.919	0.919	0.935	0.912	0.901	0.922	0.846	0.681	0.758
H5	✓	✓	✓	✓	✓	✓				0.921	0.931	0.950	0.927	0.929	0.943	0.914	0.909	0.941	0.861	0.690	0.769
D1	✓	✓	✓	✓	✓	✓	✓			0.935	0.937	0.958	0.930	0.922	0.940	0.921	0.903	0.940	0.873	0.692	0.771
D2	✓	✓	✓	✓	✓	✓		✓		0.936	0.935	0.958	0.933	0.924	0.945	0.926	0.906	0.945	0.884	0.696	0.776
D3	✓	✓	✓	✓	✓	✓			✓	0.938	0.934	0.957	0.935	0.927	0.950	0.931	0.909	0.951	0.891	0.698	0.779
D4	✓	✓	✓	✓	✓	✓	✓	✓		0.939	0.936	0.959	0.938	0.929	0.951	0.936	0.912	0.955	0.899	0.701	0.784
D5	✓	✓	✓	✓	✓	✓	✓	✓	✓	**0.956**	**0.948**	**0.983**	**0.947**	**0.932**	**0.978**	**0.939**	**0.913**	**0.956**	**0.905**	**0.704**	**0.788**

Table 3: Ablation study of the proposed framework. The best results are highlighted in **bold**.

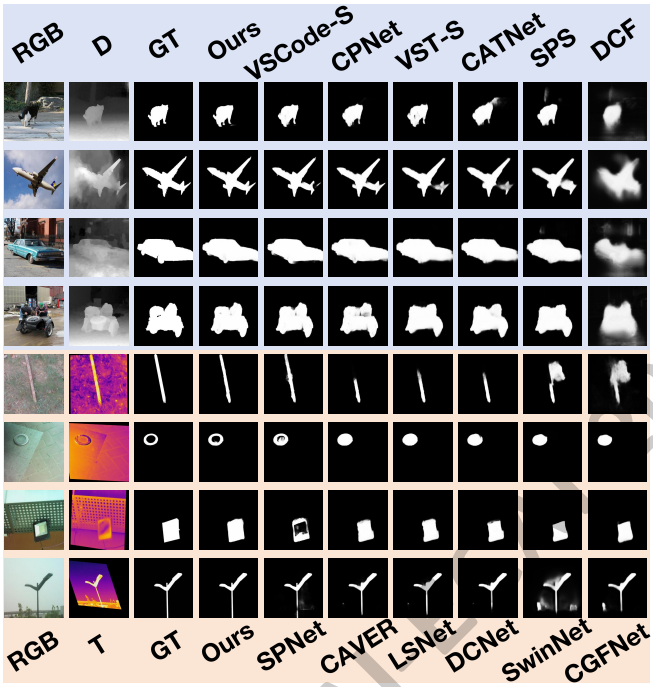

Figure 3: Visual comparison against other RGB-D SOD and RGB-T SOD methods on the NJUD and UVT2000 dataset.

UVT2000 (+7.3% S_m, +1.5% F_m, +2.5% E_m). Benefiting from Mamba's linear complexity $\mathcal{O}(N)$, G-SalAlignMamba balances accuracy and cost with 83.80M parameters, about 61%–68% fewer than CPNet and CATNet, while maintaining real-time inference at 30.41 FPS and outperforming heavy Transformer baselines such as SwinNet-B (11.42 FPS) in speed.

Qualitative Evaluation. Figure 3 compares our method with state-of-the-art approaches on RGB-D (top) and RGB-T (bottom) benchmarks. G-SalAlignMamba yields sharper boundaries and clearer object delineation under complex RGB-D topologies (e.g., 2^{nd} row) or noisy depth (e.g., 3^{rd} row), while suppressing artifacts and preserving fine-grained structures for thin instances (5^{th}, 8^{th} rows) and small objects (6^{th} row) in RGB-T scenes with thermal degradation.

4.4 Ablation Study

We conduct ablations on NJUD, NLPR, VT5000, and UVT2000 to quantify the individual and combined contributions of all modules in G-SalAlignMamba.

Geometry-Aware Encoding. We use a strong Base with a pre-trained VMamba backbone, channel-concatenation fusion, and bilinear upsampling, excluding our geometry-aware and refinement modules. Table 3 then examines *Early Interaction* (EI), *Coarse-to-Fine Explicit Alignment* (CFEA), and *Complementary Gated Fusion* (CGF), testing them independently (E1–E3), in combinations, and finally as the full encoder (E5).

Semantics-Informed Refinement. For the refinement stage, we decouple *Deep Saliency Feedback Mechanism* (DSFM), *Saliency-Weighted Modulation* (SWM), and *Adaptive Feature Integration* (AFI), reporting their individual effects (H1–H3), feedback-modulation synergy (H4), and the complete refinement scheme (H5).

Structure-Preserving Decoding. Finally, we progressively add *Context-Aware Alignment Upsampling* (CAAU), *Contextual Sequence Reconstruction* (CSR), and *Multi-Scale Boundary Refinement* (MBR), assessing isolated components (D1–D3) and stepwise integration (D4–D5). The full decoder achieves the best spatial alignment and boundary quality compared with general methods.

5 Conclusion

In this paper, we proposed G-SalAlignMamba, a geometry-aware framework for dual-modal SOD that targets three key issues: *Misalignment Vulnerability*, *Signal Dilution*, and *Boundary Degradation*. Despite the promising performance on RGB-D and RGB-T benchmarks, our proposed G-SalAlignMamba may still encounter challenges in scenarios with extreme occlusion or highly cluttered backgrounds, where the target features are severely obstructed in both modalities. Future work will focus on incorporating stronger semantic priors to improve robustness in these edge cases.

Acknowledgments

This work was partially supported by the National Natural Science Foundation of China (Grant No. 62401226) and the Hong Kong Metropolitan University Research Grant (Project Reference No. RD/2025/2.6).

Contribution Statement

Haixiao Gao and Yimin Zheng contributed equally to this work. Haixiao Gao and Yimin Zheng developed the main methodology and conducted the experiments. Mengke Song and Zhi-Ri Tang supervised the research and revised the manuscript. Linyou Xiao and Tian-Tian Zhang contributed to the analysis and manuscript preparation.

References

[Achanta *et al.*, 2009] Radhakrishna Achanta, Sheila Hemami, Francisco Estrada, and Sabine Susstrunk. Frequency-tuned salient region detection. In *2009 IEEE conference on computer vision and pattern recognition*, pages 1597–1604. IEEE, 2009.

[Fan *et al.*, 2017] Deng-Ping Fan, Ming-Ming Cheng, Yun Liu, Tao Li, and Ali Borji. Structure-measure: A new way to evaluate foreground maps. In *Proceedings of the IEEE international conference on computer vision*, pages 4548–4557, 2017.

[Fan *et al.*, 2018] Deng-Ping Fan, Cheng Gong, Yang Cao, Bo Ren, Ming-Ming Cheng, and Ali Borji. Enhanced-alignment measure for binary foreground map evaluation. *arXiv preprint arXiv:1805.10421*, 2018.

[Fan *et al.*, 2020a] Deng-Ping Fan, Zheng Lin, Zhao Zhang, Menglong Zhu, and Ming-Ming Cheng. Rethinking rgb-d salient object detection: Models, data sets, and large-scale benchmarks. *IEEE Transactions on neural networks and learning systems*, 32(5):2075–2089, 2020.

[Fan *et al.*, 2020b] Deng-Ping Fan, Yingjie Zhai, Ali Borji, Jufeng Yang, and Ling Shao. Bbs-net: Rgb-d salient object detection with a bifurcated backbone strategy network. In *European conference on computer vision*, pages 275–292. Springer, 2020.

[Fu *et al.*, 2020] Keren Fu, Deng-Ping Fan, Ge-Peng Ji, and Qijun Zhao. 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*, pages 3052–3062, 2020.

[Fu *et al.*, 2021] Keren Fu, Deng-Ping Fan, Ge-Peng Ji, Qijun Zhao, Jianbing Shen, and Ce Zhu. Siamese network for rgb-d salient object detection and beyond. *IEEE transactions on pattern analysis and machine intelligence*, 44(9):5541–5559, 2021.

[Gu and Dao, 2024] Albert Gu and Tri Dao. Mamba: Linear-time sequence modeling with selective state spaces. In *First conference on language modeling*, 2024.

[He *et al.*, 2025] Jiahao He, Keren Fu, Xiaohong Liu, and Qijun Zhao. Samba: A unified mamba-based framework for general salient object detection. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 25314–25324, 2025.

[Hu *et al.*, 2018] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7132–7141, 2018.

[Hu *et al.*, 2024] Xihang Hu, Fuming Sun, Jing Sun, Fasheng Wang, and Haojie Li. Cross-modal fusion and progressive decoding network for rgb-d salient object detection. *International Journal of Computer Vision*, 132(8):3067–3085, 2024.

[Huo *et al.*, 2021] Fushuo Huo, Xuegui Zhu, Lei Zhang, Qifeng Liu, and Yu Shu. 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.

[Ji *et al.*, 2021] Wei Ji, Jingjing Li, Shuang Yu, Miao Zhang, Yongri Piao, Shunyu Yao, Qi Bi, Kai Ma, Yefeng Zheng, Huchuan Lu, et al. Calibrated rgb-d salient object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9471–9481, 2021.

[Ju *et al.*, 2014] Ran Ju, Ling Ge, Wenjing Geng, Tongwei Ren, and Gangshan Wu. Depth saliency based on anisotropic center-surround difference. In *2014 IEEE international conference on image processing (ICIP)*, pages 1115–1119. IEEE, 2014.

[Lee *et al.*, 2022] Minhyeok Lee, Chaewon Park, Suhwan Cho, and Sangyoun Lee. Spsn: Superpixel prototype sampling network for rgb-d salient object detection. In *European conference on computer vision*, pages 630–647. Springer, 2022.

[Liu *et al.*, 2021] Zhengyi Liu, Yacheng Tan, Qian He, and Yun Xiao. Swinnet: Swin transformer drives edge-aware rgb-d and rgb-t salient object detection. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(7):4486–4497, 2021.

[Liu *et al.*, 2022] Biyuan Liu, Huaixin Chen, Zhixi Wang, Wenqiang Xie, and LingYu Shuai. Lsnet: Extremely light-weight siamese network for change detection of remote sensing image. In *IGARSS 2022-2022 IEEE International Geoscience and Remote Sensing Symposium*, pages 2358–2361. IEEE, 2022.

[Liu *et al.*, 2024a] Nian Liu, Ziyang Luo, Ni Zhang, and Junwei Han. Vst++: Efficient and stronger visual saliency transformer. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(11):7300–7316, 2024.

[Liu *et al.*, 2024b] Yue Liu, Yunjie Tian, Yuzhong Zhao, Hongtian Yu, Lingxi Xie, Yaowei Wang, Qixiang Ye, Jianbin Jiao, and Yunfan Liu. Vmamba: Visual state space model. *Advances in neural information processing systems*, 37:103031–103063, 2024.

[Luo *et al.*, 2024] Ziyang Luo, Nian Liu, Wangbo Zhao, Xuguang Yang, Dingwen Zhang, Deng-Ping Fan, Fahad Khan, and Junwei Han. 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*, pages 17169–17180, 2024.

[Niu *et al.*, 2012] Yuzhen Niu, Yujie Geng, Xueqing Li, and Feng Liu. Leveraging stereopsis for saliency analysis. In *2012 IEEE conference on computer vision and pattern recognition*, pages 454–461. IEEE, 2012.

[Pang *et al.*, 2023] Youwei Pang, Xiaoqi Zhao, Lihe Zhang, and Huchuan Lu. Caver: Cross-modal view-mixed transformer for bi-modal salient object detection. *IEEE Transactions on Image Processing*, 32:892–904, 2023.

[Peng *et al.*, 2014] Houwen Peng, Bing Li, Weihua Xiong, Weiming Hu, and Rongrong Ji. Rgbd salient object detection: A benchmark and algorithms. In *European conference on computer vision*, pages 92–109. Springer, 2014.

[Piao *et al.*, 2019] Yongri Piao, Wei Ji, Jingjing Li, Miao Zhang, and Huchuan Lu. Depth-induced multi-scale recurrent attention network for saliency detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 7254–7263, 2019.

[Qin *et al.*, 2019] Xuebin Qin, Zichen Zhang, Chenyang Huang, Chao Gao, Masood Dehghan, and Martin Jagersand. Basnet: Boundary-aware salient object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7479–7489, 2019.

[Sun *et al.*, 2023] Fuming Sun, Peng Ren, Bowen Yin, Fasheng Wang, and Haojie Li. Catnet: A cascaded and aggregated transformer network for rgb-d salient object detection. *IEEE Transactions on Multimedia*, 26:2249–2262, 2023.

[Tu *et al.*, 2019] Zhengzheng Tu, Tian Xia, Chenglong Li, Xiaoxiao Wang, Yan Ma, and Jin Tang. Rgb-t image saliency detection via collaborative graph learning. *IEEE Transactions on Multimedia*, 22(1):160–173, 2019.

[Tu *et al.*, 2022a] Zhengzheng Tu, Zhun Li, Chenglong Li, and Jin Tang. Weakly alignment-free rgbt salient object detection with deep correlation network. *IEEE Transactions on Image Processing*, 31:3752–3764, 2022.

[Tu *et al.*, 2022b] Zhengzheng Tu, Yan Ma, Zhun Li, Chenglong Li, Jieming Xu, and Yongtao Liu. Rgbt salient object detection: A large-scale dataset and benchmark. *IEEE Transactions on Multimedia*, 25:4163–4176, 2022.

[Wang *et al.*, 2018] Guizhao Wang, Chenglong Li, Yunpeng Ma, Aihua Zheng, Jin Tang, and Bin Luo. Rgb-t saliency detection benchmark: Dataset, baselines, analysis and a novel approach. In *Chinese Conference on Image and Graphics Technologies*, pages 359–369. Springer, 2018.

[Wang *et al.*, 2020] Qilong Wang, Banggu Wu, Pengfei Zhu, Peihua Li, Wangmeng Zuo, and Qinghua Hu. Eca-net: Efficient channel attention for deep convolutional neural networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11534–11542, 2020.

[Wang *et al.*, 2021] Jie Wang, Kechen Song, Yanqi Bao, Liming Huang, and Yunhui Yan. Cgfnet: Cross-guided fusion network for rgb-t salient object detection. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(5):2949–2961, 2021.

[Wang *et al.*, 2024] Kunpeng Wang, Danying Lin, Chenglong Li, Zhengzheng Tu, and Bin Luo. Alignment-free rgbt salient object detection: Semantics-guided asymmetric correlation network and a unified benchmark. *IEEE Transactions on Multimedia*, 26:10692–10707, 2024.

[Wang *et al.*, 2025] Kunpeng Wang, Keke Chen, Chenglong Li, Zhengzheng Tu, and Bin Luo. Alignment-free rgb-t salient object detection: A large-scale dataset and progressive correlation network. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 7780–7788, 2025.

[Xing *et al.*, 2024] Zhaohu Xing, Tian Ye, Yijun Yang, Guang Liu, and Lei Zhu. Segmamba: Long-range sequential modeling mamba for 3d medical image segmentation. In *International conference on medical image computing and computer-assisted intervention*, pages 578–588. Springer, 2024.

[Zhang *et al.*, 2023] Zihao Zhang, Jie Wang, and Yahong Han. Saliency prototype for rgb-d and rgb-t salient object detection. In *Proceedings of the 31st ACM international conference on multimedia*, pages 3696–3705, 2023.

[Zhang *et al.*, 2025] Shuo Zhang, Jiaming Huang, Wenbing Tang, Yan Wu, Terrence Hu, Xiaogang Xu, and Jing Liu. Dimsod: A diffusion-based framework for multimodal salient object detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 10103–10111, 2025.

[Zhao *et al.*, 2019] Jia-Xing Zhao, Jiang-Jiang Liu, Deng-Ping Fan, Yang Cao, Jufeng Yang, and Ming-Ming Cheng. Egnet: Edge guidance network for salient object detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 8779–8788, 2019.

[Zhou *et al.*, 2021] Tao Zhou, Huazhu Fu, Geng Chen, Yi Zhou, Deng-Ping Fan, and Ling Shao. Specificity-preserving rgb-d saliency detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4681–4691, 2021.

[Zhou *et al.*, 2023] Heng Zhou, Chunna Tian, Zhenxi Zhang, Chengyang Li, Yuxuan Ding, Yongqiang Xie, and Zhongbo Li. Position-aware relation learning for rgb-thermal salient object detection. *IEEE Transactions on Image Processing*, 32:2593–2607, 2023.

[Zhu *et al.*, 2024] Lianghui Zhu, Bencheng Liao, Qian Zhang, Xinlong Wang, Wenyu Liu, and Xinggang Wang. Vision mamba: Efficient visual representation learning with bidirectional state space model. *arXiv preprint arXiv:2401.09417*, 2024.

[Zhu *et al.*, 2025] Jiayi Zhu, Xuebin Qin, and Abdulmotaleb Elsaddik. Dc-net: Divide-and-conquer for salient object detection. *Pattern Recognition*, 157:110903, 2025.