Mamba4SOD: RGB-T Salient Object Detection Using Mamba-Based Fusion Module

Yi Xu[1] | Ruichao Hou[1] | Ziheng Qi[2] | Tongwei Ren[1]

[1]State Key Laboratory for Novel Software Technology, Nanjing University, Nanjing, China | [2]Department of Electrical and Computer Engineering, University of Illinois Urbana-Champaign, Champaign, Illinois, USA

Correspondence: Ruichao Hou (rchou@nju.edu.cn)

Received: 5 December 2024 | **Revised:** 4 March 2025 | **Accepted:** 26 March 2025

Handling Editor: Henghui Ding

Funding: This work was supported by the National Natural Science Foundation of China (62072232), the Key R&D Project of Jiangsu Province (BE2022138), the Fundamental Research Funds for the Central Universities (0217-14380026), the Innovation Project of State Key Laboratory for Novel Software Technology Nanjing University (ZZKT-2024B20) and the Collaborative Innovation Center of Novel Software Technology and Industrialization.

Keywords: Mamba | RGB and thermal | salient object detection | state space mode | Swin Transformer

ABSTRACT

RGB and thermal salient object detection (RGB-T SOD) aims to accurately locate and segment salient objects in aligned visible and thermal image pairs. However, existing methods often struggle to produce complete masks and sharp boundaries in challenging scenarios due to insufficient exploration of complementary features from the dual modalities. In this paper, we propose a novel mamba-based fusion network for RGB-T SOD task, named Mamba4SOD, which integrates the strengths of Swin Transformer and Mamba to construct robust multi-modal representations, effectively reducing pixel misclassification. Specifically, we leverage Swin Transformer V2 to establish long-range contextual dependencies and thoroughly analyse the impact of features at various levels on detection performance. Additionally, we develop a novel Mamba-based fusion module with linear complexity, boosting multi-modal enhancement and fusion. Experimental results on VT5000, VT1000 and VT821 datasets demonstrate that our method outperforms the state-of-the-art RGB-T SOD methods.

1 | Introduction

Salient object detection (SOD) aims to locate and segment the most salient object(s) in a natural image, which is widely used in image segmentation [1], action recognition [2], image retrieval [3], and object tracking [4]. Solely RGB-based SOD methods' [5–7] performance is often hindered in extreme scenarios, such as low light or adverse weather conditions, due to the inherent limitations of visible sensors. Hence, thermal modality is used as an effective supplement to RGB modality in SOD, because it captures radiation emitted by objects above absolute zero, providing critical information that is invariant to ambient lighting conditions. RGB and thermal (RGB-T) SOD has emerged as a promising solution, leveraging complementary cues to boost detection robustness [8–10].

Traditional RGB-T SOD methods [11, 12] primarily rely on handcrafted features or heuristic priors to predict salient regions. Their effectiveness in complex scenarios is limited by the lack of rich semantic information and the well-designed fusion strategy. Recently, the success of Transformers in the visual domain [13, 14] has been widely demonstrated, where Transformer-based backbones like PVT [15] and Swin Transformer [16] have been widely applied and achieved remarkable results by capturing richer contextual information and facilitating multi-scale feature integration.

However, as illustrated in Figure 1, existing methods still face several challenges, such as blurred boundaries and background misclassification. Specifically, in the first two rows, both ADNet [17] and WGOFNet [18], two state-of-the-art RGB-T SOD methods, encounter challenges with blurred boundaries. ADNet fails to segment the legs, while WGOFNet produces a blurry mass. Specifically, in the second row, the blurred boundaries between the car body and shadow lead to incorrect segmentation of the rear wheels and shadow. The last two rows show the challenge of background misclassification. Both methods mistakenly treat the pole as part of the target and fail to distinguish the hollow parts of the headboard from the background. The underlying reason for these challenges lies in their insufficient exploration of feature extraction and fusion mechanisms, particularly under low-contrast conditions and scenarios with background clutter. These limitations become more severe when one modality undergoes degradation. Moreover, their progressive fusion of multi-level features introduces redundant information that ignores the hierarchical feature selection, ultimately hindering overall performance.

To address these issues, we propose a mamba-based fusion network for RGB-T SOD, named Mamba4SOD, which absorbs the strengths of Swin Transformer V2 [19] and Mamba [20] architectures to exploit the semantic complementary information and suppress the modality bias. Specifically, Mamba4SOD employs Swin Transformer V2 as the backbone to enhance feature extraction, capturing both fine-grained local details and global contextual information, effectively mitigating issues like blurred boundaries and enhancing overall detection performance. Swin Transformer is a sophisticated network architecture, and feature selection plays a crucial role in determining its performance. Although previous works [9, 21] typically employ the outputs from the four stages for multi-level feature fusion, they primarily focus on analysing shallow features while neglecting the optimal combination of deeper features. This indiscriminate utilisation of all stages often results in information redundancy, leading to inefficiencies and increased computational complexity. To overcome this issue, we perform a comprehensive analysis of multi-scale features across different stages of the Swin Transformer. By identifying and selecting the most informative stages, we develop an optimised feature selection strategy, ensuring a balanced trade-off between fine-grained details and semantic context, while simultaneously reducing computational complexity.

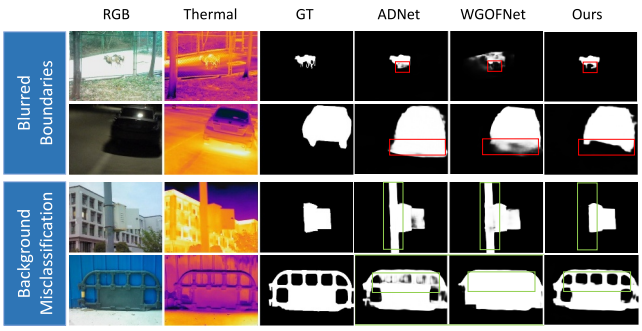

FIGURE 1 | Visualization in complex scenes. Challenges such as blurred boundaries and background classification are highlighted in red and green, respectively. The results show that our method effectively addresses both challenges.

To facilitate cross-modal fusion, we develop a novel Mamba-based fusion module (MFM), leveraging the advantages of Mamba in modelling long-range dependencies to achieve adaptive feature recalibration between RGB and thermal modalities. It dynamically adjusts the contribution of each modality and mitigates issues arising from poor-quality RGB images and misleading thermal information. MFM combines complementary features from both modalities, preserving critical details and enabling differentiation of visually similar but semantically different objects. To the best of our knowledge, this is the first work to explore and reveal the potential of the Mamba in the RGB-T SOD task.

Our main contributions can be summarised as follows:

- We propose a novel RGB-T SOD network that integrates a hybrid architecture of Swin Transformer and Mamba, which produce accurate masks and outperforms state-of-the-art methods.
- We analyse the impact of multi-scale features at different stages of the Swin Transformer on detection performance, optimising the feature selection process for ensuring a balanced trade-off between fine-grained details and semantic context.
- We propose a Mamba-based fusion module to fuse RGB and thermal modalities effectively for constructing robust multi-modal representations.

2 | Related Work

2.1 | RGB-T SOD

RGB-T SOD focuses on accurately locating and segmenting common salient objects in aligned visible and thermal infrared image pairs. It relies on the alignment and fusion of different modalities to enhance object-level saliency detection.

Existing RGB-T SOD methods can be categorised into three types: early traditional methods, CNN-based methods and Transformer-based methods. Early methods mostly relied on the manifold ranking algorithm. For instance, Wang et al. [12] proposed a graph-based multi-task manifold ranking algorithm and created the RGB-T SOD benchmark VT821.

With the advance of deep learning, CNN-based methods became the mainstream for RGB-T SOD. For example, Tu et al. [8] advanced the field by combining multi-level feature extraction with attention mechanisms. They contributed two critical datasets: VT1000 [22] and VT5000 [8], with VT5000 as a large dataset designed to support deep learning-based RGB-T SOD models. CSRNet [23] utilised a lightweight backbone and context-guided cross-modality fusion modules to improve feature integration and reduce computational costs. OSRNet [24] utilised a lightweight decoder for efficient feature refinement and real-time performance, outperforming competitors with real-time speed on a single GPU.

Recently, Transformer-based methods have emerged as a powerful alternative, addressing the limitations of CNNs by leveraging self-attention mechanisms for global context modelling. For instance, CAVER [25] employed a transformer-based model for multi-modal SOD, allowing global alignment and a novel attention mechanism. ACMANet [26] introduced an asymmetric cross-modal activation network that fuses diverse features and uses self-attention for precise salient object detection. SwinNet [21] integrated Swin Transformer with CNN to extract hierarchical features, aligning and recalibrating them across modalities to achieve sharp contours and well-defined boundaries. CWFNet [27] introduced a global illumination learning module to emphasise reliable saliency cues. WGOFNet [18] further optimised cross-modal fusion by adaptively weighting different modalities.

Unlike existing RGB-T SOD methods, our method revisits and optimises the feature selection in the Swin Transformer while exploring the potential of Mamba to facilitate multi-modal feature fusion.

2.2 | Mamba

Mamba is an advanced state-space model (SSM) that offers an efficient approach to long-range dependency modelling with linear computational complexity. SSMs have emerged as a promising alternative to Transformers in sequence modelling, particularly in scenarios where computational efficiency is critical. The structured state-space sequence model [28] improved computational efficiency by introducing a new parameterisation to the SSM. Then, the simplified state space layers for sequence modelling [29] extended the structured state-space sequence model by introducing multiple input multiple output SSM and efficient parallel scanning. Mamba [20] further advanced the SSM framework by introducing a data-dependent SSM layer, significantly enhancing sequence modelling capabilities.

Building on these advancements, Vision Mamba [30] incorporates bidirectional selective state space models and position embeddings to improve vision tasks. VMamba [31] incorporated a Cross-Scan Module, enabling 1D selective scanning in 2D image space. Mamba-based architectures have demonstrated exceptional performance in various vision applications, such as image classification [30, 32], object detection [33, 34] and image segmentation [35, 36]. Specifically, UMamba [35] integrated CNN and SSM blocks to effectively capture both local features and long-range dependencies, improving performance in various segmentation tasks. LocalMamba [37] designed a novel local scanning strategy that preserves 2D dependencies and dynamically selects optimal scan patterns, enhancing modelling capability.

The ability to capture long-range dependency modelling with linear computational complexity makes Mamba an attractive choice for multi-modal fusion. Motivated by these advantages, we explore the potential of Mamba to effectively integrate RGB-T features.

3 | Method

3.1 | Overview

The framework of the proposed Mamba4SOD is shown in Figure 2. It consists of three main components: a symmetric two-stream Swin Transformer V2 [19] backbone, a Mamba-based

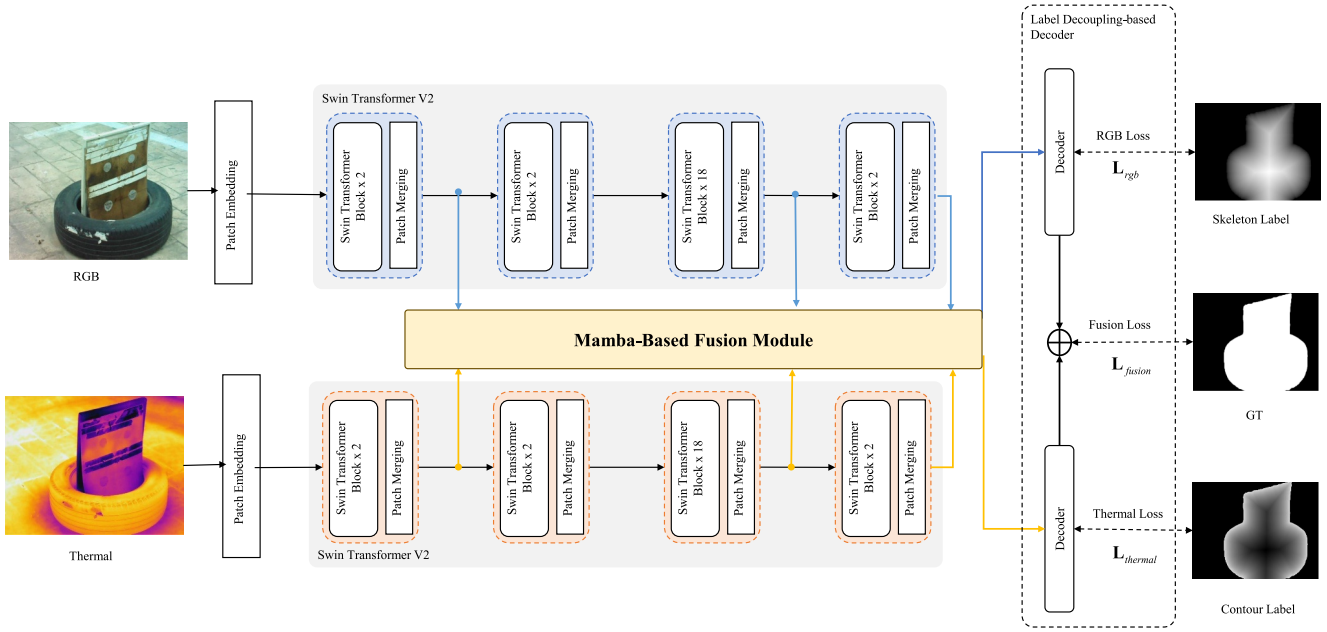

FIGURE 2 | The framework of Mamba4SOD, which is divided into three components, that is, Swin Transformer V2 backbone, Mamba-based fusion module and label decoupling-based decoder.

fusion module and a label decoupling-based decoder. Specifically, the Swin Transformer V2 backbone extracts multi-scale features from selected stages to provide robust and efficient feature representation. MFM effectively aligns and integrates the RGB and thermal features, using adaptive weighting to facilitate cross-modal fusion. Finally, the label decoupling-based decoder predicts accurate saliency maps under the supervision of skeleton and contour labels, enhancing boundary localisation and structural consistency.

3.2 | Swin Transformer V2 Backbone

Recently, numerous multi-modal transformers have demonstrated strong capabilities in multi-modal feature fusion. For example, the vision-language transformer (VLT) for referring segmentation [38] interprets natural language queries to identify specific objects in images, relying heavily on the interaction between vision and language. However, in RGB-T SOD, which lacks language inputs, language-driven mechanisms like VLT are not directly applicable. Instead, the primary focus is on robust feature fusion and noise suppression within visual modalities.

Swin Transformer has proven to be a powerful backbone in state-of-the-art RGB-T SOD methods [9, 21]. Hierarchical features for the two modalities are extracted using two independent Swin Transformer V2 backbones, an enhanced version of the original Swin Transformer, selected for the robust feature extraction capabilities that efficiently capture both local and global dependencies within multi-modal data. Furthermore, specific feature layers are selected based on their effectiveness and are subsequently processed for fusion and enhancement.

Previous works on Transformer-based RGB-T SOD have adopted varying stage selection strategies. For instance, SwinNet [21] uses features from stages 1–4, while MCNet [9] and WGOFNet [18] select stages 2–5. These methods typically employ outputs from four stages for multi-level feature fusion without thoroughly considering the optimal combination of features, resulting in potential information redundancy and unnecessary computational costs. Moreover, studies [39] have shown that deeper features generally contribute more to performance than shallower ones in deep aggregation methods. Therefore, we have retained the last stage.

To achieve efficient feature extraction, we perform a comprehensive analysis of multi-scale features across different stages of the Swin Transformer. Based on the results, we select features from three representative layers: one from the shallow levels (stages 1 and 2) and two from the deeper levels (stages 3–5). This carefully crafted selection strategy enables the model to capture both fine-grained details and high-level semantic information. In Mamba4SOD, we retain features $F^i_{\text{rgb}} = \{F^i_{\text{rgb}} | i = 2, 4, 5\}$ and $F^i_{\text{t}} = \{F^i_{\text{t}} | i = 2, 4, 5\}$, which are resized to 64 channels through a 1×1 convolution. These features are then fed into the MFM for effective multi-modal feature integration.

3.3 | Mamba-Based Fusion Module

ALGORITHM 1 | Mamba process.

Input: token sequence $\mathbf{F}_{l-1}$: (B, L, C)
Output: token sequence $\mathbf{F}_l$: (B, L, C)
$\mathbf{F}'_{l-1}$: (B, L, C) $\leftarrow \text{Norm}(\mathbf{F}_{l-1})$
$\mathbf{x}$: (B, L, E) $\leftarrow Linear^{x}(\mathbf{F}'_{l-1})$
$\mathbf{z}$: (B, L, E) $\leftarrow Linear^{x}(\mathbf{F}'_{l-1})$
/* Space State Model (SSM) process with different directions */
for o in {forward, backward} do
  x'_o : (B, L, E) $\leftarrow \text{SiLU}(\text{Conv1d}_o(\text{x}))$
  $\mathbf{B}_o$: (B, L, N) $\leftarrow \text{Linear}^{\text{B}}_o(\text{x}'_o)$
  $\mathbf{C}_o$: (B, L, N) $\leftarrow \text{Linear}^{\text{C}}_o(\text{x}'_o)$
  Δ_o : (B, L, E) $\leftarrow \log(1 + \exp(\text{Linear}^{\mathbf{\Delta}}_o(\mathbf{x}'_o) + \text{Parameter}^{\Delta}_o))$
  $\overline{\mathbf{A}}_o$: (B, L, E, N) $\leftarrow \Delta_o \otimes \text{Parameter}^{\mathbf{A}}_o$
  $\overline{\mathbf{B}}_o$: (B, L, E, N) $\leftarrow \Delta_o \otimes \mathbf{B}_o$
  $\mathbf{y}_o$: (B, L, E) $\leftarrow \mathbf{SSM}(\overline{\mathbf{A}}_o, \overline{\mathbf{B}}_o, \mathbf{C}_o)(\text{x}'_o)$
end for
$\mathbf{y}'$: (B, L, E) $\leftarrow \left(\mathbf{y}_{\text{forward}} + \mathbf{y}_{\text{backward}}\right) \odot \text{SiLU}(\mathbf{z})$
$\mathbf{F}_l$: (B, L, C) $\leftarrow \text{Linear}^{\mathbf{F}}(\mathbf{y}') + \mathbf{F}_{l-1}$
Return $\mathbf{F}_l$

Traditional fusion modules in the RGB-T SOD task rely on CNNs and the lack of long-range dependencies leads to insufficient modality alignment and fusion performance. Mamba leverages state-space models to efficiently model long-range dependencies [20]. Unlike static fusion approaches, it dynamically prioritises task-relevant features while suppressing noise. Furthermore, its directional scanning mechanism preserves spatial coherence, facilitating robust multi-modal or cross-scale alignment [40].

Inspired by this, we propose a Mamba-based fusion module designed to enhance and align features from the RGB and thermal modalities. The fusion process is further optimised with adaptive weights, which enable more effective and dynamic fusion. The detailed structure of MFM is shown in Figure 3. First, the multi-modal features undergo enhancement through the Mamba. The Mamba architecture closely resembles

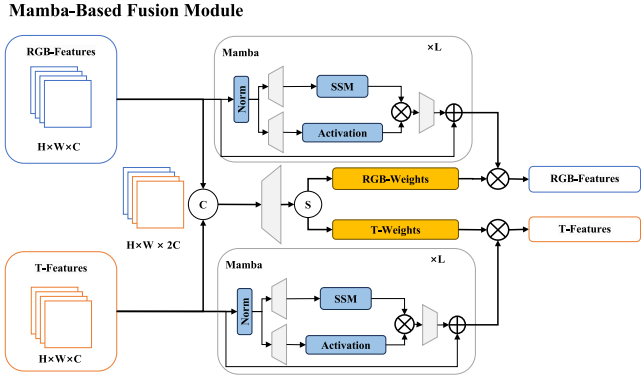

FIGURE 3 | The detailed design of the Mamba-based fusion module.

the Vim block of Vision Mamba [30] as illustrated in Algorithm 1. The Mamba processes visual features in the form of token sequences, similar to the patch-based method employed by Vision Transformers [41]. Given an input feature map F^i of shape (B, C, H, W), where B is the batch size, C is channel dimension, H and W are the spatial dimensions, it is reshaped into (B, L, C), where L is the number of tokens, and C is the channel dimension. These tokens are then fed into the Mamba for enhanced feature alignment and fusion. By leveraging the bidirectional state-space modelling capability, the Mamba effectively captures both local and long-range dependencies across modalities. The enhanced features are calculated as follows:

$$\tilde{F}^i_{\text{rgb}} = \mathcal{M}(F^i_{\text{rgb}}), \tag{1}$$

$$\tilde{F}^i_{\text{t}} = \mathcal{M}(F^i_{\text{t}}), \tag{2}$$

where $\tilde{F}^i_{\text{rgb}}$ and $\tilde{F}^i_{\text{t}}$ represent the enhanced features of the RGB and thermal modalities, respectively. F^i_{rgb} and F^i_{t} represent the output feature maps. $\mathcal{M}$ refers to the Mamba.

To determine the weights assigned to each modality, MFM concatenates the corresponding RGB and thermal features for each stage output feature map of the symmetric two-stream Swin Transformer V2 backbone:

$$F^i_{\text{c}} = \text{Concat}(F^i_{\text{rgb}}, F^i_{\text{t}}), \tag{3}$$

where F^i_{c} represents the concatenated feature, which is then fed into a fully connected (FC) layer to generate modality-specific weights W^i:

$$W^i = \text{FC}(F^i_{\text{c}}). \tag{4}$$

The softmax function is employed to determine the adaptive weights for both RGB and thermal features:

$$W^i_{\text{rgb}} = \text{softmax}(W^i), \tag{5}$$

$$W^i_{\text{t}} = 1 - W^i_{\text{rgb}}, \tag{6}$$

where W^i_{rgb} and W^i_{t} denote the adaptive weights for the RGB and thermal features, respectively. The enhanced modality features $\tilde{F}^i_{\text{rgb}}$ and $\tilde{F}^i_{\text{t}}$ are multiplied by their respective modality weights to fuse RGB and thermal features:

$$\text{A}^i_{\text{rgb}} = \tilde{F}^i_{\text{rgb}} \times W^i_{\text{rgb}}, \tag{7}$$

$$\text{A}^i_{\text{t}} = \tilde{F}^i_{\text{t}} \times W^i_{\text{t}}, \tag{8}$$

where A^i_{rgb} and A^i_{t} represent the attention maps of the MFM for the RGB and thermal modalities based on the outputs from backbone stage i, respectively.

3.4 | Label Decoupling-Based Decoder

Inspired by Jiang et al. and Wei et al. [9, 42], our decoder consists of two branches with the same structure but nonshared parameters. Each branch processes RGB features A^i_{rgb} and thermal image features A^i_{t} from MFM. Each branch is composed of three residual blocks, where each block includes a 3×3 convolutional layer and a BatchNorm layer. The decoding branches upsample the output of each residual block to match the input size of the subsequent block using bilinear interpolation. These branch-specific outputs are then concatenated to generate the final fused saliency map, which combines complementary information from both modalities to achieve precise and robust detection results.

Moreover, during the training stage, we employ skeleton label supervision for the RGB branch to emphasise the structural details of salient objects, leveraging the rich colour and texture information inherent in RGB images. For the thermal branch, we adopt contour label supervision, as thermal images are particularly effective at capturing boundary information, especially under challenging conditions such as low light or adverse weather. This label decoupling supervision strategy allows each branch to capitalise on its respective strengths, enhancing the effectiveness of cross-modal feature extraction and fusion.

3.5 | Loss Function

Following [9, 42], the loss function consists of three parts to measure the gap between the predicted results and the SOD ground truths, which is calculated as follows:

$$\mathcal{L} = \mathcal{L}_{\text{rgb}} + \mathcal{L}_{\text{thermal}} + \mathcal{L}_{\text{fusion}}, \tag{9}$$

$$\mathcal{L}_{\text{rgb}} = \mathcal{L}_{\text{BCE}} + \mathcal{L}_{\text{SSIM}}, \tag{10}$$

$$\mathcal{L}_{\text{thermal}} = \mathcal{L}_{\text{BCE}} + \mathcal{L}_{\text{SSIM}}, \tag{11}$$

$$\mathcal{L}_{\text{fusion}} = \mathcal{L}_{\text{BCE}} + \mathcal{L}_{\text{SSIM}} + \mathcal{L}_{\text{IoU}}, \tag{12}$$

where $\mathcal{L}_{\text{rgb}}$, $\mathcal{L}_{\text{thermal}}$ and $\mathcal{L}_{\text{fusion}}$ represent the RGB loss, thermal loss and fusion loss, respectively. $\mathcal{L}_{\text{BCE}}$ represents the binary cross-entropy loss, $\mathcal{L}_{\text{SSIM}}$ denotes the structural similarity index measure and $\mathcal{L}_{\text{IoU}}$ is the Intersection-over-Union loss.

4 | Experiments

4.1 | Experimental Setting

4.1.1 | Datasets

We evaluate our Mamba4SOD on three publicly available RGB-T SOD datasets: VT821, VT1000 and VT5000. We leverage 2500 pairs of aligned dual-modality images from VT5000 as the training set. The remaining 2500 pairs from VT5000, along with the 821 pairs from VT821 and 1000 pairs from VT1000, are used as the test set.

4.1.2 | Evaluation Metrics

We evaluate our method on seven widely-used metrics [43–47], precision-recall (PR) curve, mean F-measure (F_{avg}), maximum F-measure ($F_{\max}$), the weighted F-measure (F_ω), mean absolute error (MAE), E-measure (E_m) and S-measure (S_m).

The PR curve is generated by binarizing the saliency map at various probability thresholds (ranging from 0 to 1) and comparing the resulting binary maps with the ground truth.

F_m computes the weighted harmonic mean of the threshold precision and recall, defined as follows:

$$F_m = \frac{\left(1+\beta^2\right)\cdot \text{Precision}\cdot \text{Recall}}{\beta^2\cdot \text{Precision} + \text{Recall}}, \tag{13}$$

where β^2 is typically set to 0.3, giving equal weight to precision and recall. The mean F-measure F_{avg} is the average F-measure computed over all thresholds. We also report the maximum value of the F-measure as $F_{\max}$. Furthermore, following [44], we calculate the weighted F-measure, denoted as F_ω, to provide a more comprehensive evaluation.

MAE measures the average absolute difference between the predicted saliency map and the ground truth:

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|P_i - G_i|, \tag{14}$$

where N is the number of pixels, P_i is the predicted saliency value, and G_i is the ground truth.

E_m combines local pixel values with the overall image-level information:

$$E_m = \frac{1}{N}\sum_{i=1}^{N}\frac{2\cdot P_i\cdot G_i}{P_i^2 + G_i^2}. \tag{15}$$

S_m evaluates the structural similarity between the predicted saliency map and the ground truth:

$$S_m = \alpha\cdot S_o + (1-\alpha)\cdot S_r, \tag{16}$$

where S_o measures the object similarity, S_r measures the region similarity and α is a weight parameter that balances these two components.

4.1.3 | Implementation Details

We utilise dual NVIDIA RTX 4090 GPUs to train the proposed Mamba4SOD method. We choose SwinV2-B as the backbone, pre-trained on the ImageNet-22K dataset. RGB and thermal images are both resized to the shape of 384 × 384 pixels. During the training phase, we set the batch size to 6 and apply various data augmentation techniques, including horizontal flipping, random cropping, and multi-scale input strategies, following best practices from previous works [9, 42]. The entire network is trained end-to-end using stochastic gradient descent (SGD), with a momentum of 0.9 and a weight decay of 0.0005. The maximum learning rate is set to 0.005. The training process runs for a maximum of 70 epochs to achieve optimal convergence.

4.2 | Quantitative Evaluation

To thoroughly validate the effectiveness of our method, we conduct extensive comparisons with 13 state-of-the-art RGB-T SOD methods, including ADF [8], CSRNet [23], OSRNET [24], UMINet [48], TNet [49], MCFNet [50], CGFNet [51], CAVER [25], ADNet [17], ACMANet [26], SwinNet [21], CMDBIF-Net [52] and WGOFNet [18].

Table 1 provides a comprehensive quantitative comparison between our Mamba4SOD and 13 comparison methods, validating the superior performance of our method. In particular, we select the recently proposed ADNet, which achieves the second-best results, for detailed comparison. On the VT5000 dataset, our method improves F_{avg} and F_ω by 0.9% and 1.5%, respectively. E_m and S_m also show significant increases of 0.5% and 0.8%. On the VT1000 dataset, our method still maintains noticeably stronger performance, with F_{avg} and F_ω improving by 0.7% and 0.9%, respectively. On the VT821 dataset, our results show a substantial improvement over ADNet, with F_{avg} increasing by 0.9%, F_ω by 1.6%, and E_m and S_m each increasing by 0.9% and 1.0%. These performance gains can be attributed to the ability of MFM to selectively emphasise informative regions while suppressing noise and redundancy, particularly in cases where one modality is degraded. The improvements in E_m and S_m further validate the effectiveness of Mamba4SOD in preserving structural details and enhancing boundary accuracy.

Besides these key metrics, other metrics also exhibit significant advantages or strong competitiveness. As illustrated in Figure 4, the PR and F-measure curves of the compared methods, indicate that our proposed method surpasses other RGB-T SOD methods in terms of precision-recall balance and overall F-measure performance. In summary, our proposed method consistently achieves superior performance on VT5000, VT1000 and VT821 datasets.

4.3 | Qualitative Evaluation

We present a qualitative comparison with 13 state-of-the-art methods, as illustrated in Figure 5.

In Figure 5a–c, our method successfully eliminates interference from saliency objects leveraging robust semantic feature extraction, where other state-of-the-art methods struggle to differentiate the target from its surroundings. Figure 5d–f shows the hollow region misclassification challenges. In Figure 5e, despite noisy RGB inputs, our method precisely distinguishes hollow regions from complex regions, demonstrating its robustness against image degradation. In Figure 5g,h where small targets are present, other models fail to accurately identify salient objects and delineate their boundaries. Our method has superior performance in these cases can be attributed to its

TABLE 1 | Performance comparison with state-of-the-art methods on three RGB-T SOD datasets.

Methods	VT5000						VT1000						VT821					
	$F_{avg}\uparrow$	$F_{max}\uparrow$	$F_{\omega}\uparrow$	$MAE\downarrow$	$E_m\uparrow$	$S_m\uparrow$	$F_{avg}\uparrow$	$F_{max}\uparrow$	$F_{\omega}\uparrow$	$MAE\downarrow$	$E_m\uparrow$	$S_m\uparrow$	$F_{avg}\uparrow$	$F_{max}\uparrow$	$F_{\omega}\uparrow$	$MAE\downarrow$	$E_m\uparrow$	$S_m\uparrow$
CGFNet [51]	0.851	0.887	0.831	0.035	0.922	0.883	0.906	0.936	0.900	0.023	0.944	0.923	0.845	0.885	0.829	0.038	0.912	0.881
CSRNet [23]	0.811	0.857	0.796	0.042	0.905	0.868	0.877	0.918	0.878	0.024	0.925	0.918	0.831	0.880	0.821	0.038	0.909	0.885
SwinNet [21]	0.865	0.915	0.846	0.026	0.942	0.912	0.896	0.948	0.894	0.018	0.947	0.938	0.847	0.903	0.818	0.030	0.926	0.904
ADF [8]	0.778	0.863	0.722	0.048	0.891	0.864	0.847	0.923	0.804	0.034	0.921	0.910	0.717	0.804	0.627	0.077	0.843	0.810
OSRNet [24]	0.823	0.866	0.807	0.040	0.908	0.875	0.892	0.929	0.891	0.022	0.935	0.926	0.814	0.862	0.801	0.043	0.896	0.875
MCFNet [50]	0.848	0.886	0.836	0.033	0.924	0.887	0.902	0.939	0.906	0.019	0.944	0.932	0.844	0.889	0.835	0.029	0.918	0.891
TNet [49]	0.846	0.895	0.840	0.033	0.927	0.895	0.889	0.937	0.895	0.021	0.937	0.929	0.842	0.904	0.841	0.030	0.919	0.899
ACMANet [26]	0.858	0.890	0.823	0.033	0.932	0.887	0.904	0.933	0.889	0.021	0.945	0.927	0.837	0.873	0.807	0.035	0.914	0.883
CAVER [25]	0.856	0.897	0.849	0.028	0.935	0.899	0.906	0.945	0.912	0.016	0.949	0.938	0.854	0.897	0.846	0.026	0.928	0.897
CMDBIF-Net [52]	0.868	0.892	0.846	0.032	0.933	0.886	0.914	0.931	0.909	0.019	0.952	0.927	0.856	0.887	0.837	0.032	0.923	0.882
UMINet [48]	0.831	0.877	0.820	0.035	0.919	0.882	0.892	0.935	0.896	0.021	0.941	0.926	0.791	0.849	0.782	0.054	0.879	0.858
ADNet [17]	0.893	0.924	0.884	0.022	0.953	0.922	0.916	0.952	0.920	0.015	0.952	0.944	0.869	0.915	0.860	0.024	0.930	0.915
WGOFNet [18]	0.883	0.912	0.873	0.025	0.945	0.911	0.919	0.946	0.922	0.016	0.951	0.940	0.875	0.911	0.868	0.025	0.934	0.908
Ours	0.902	0.932	0.899	0.019	0.958	0.930	0.926	0.958	0.931	0.013	0.956	0.949	0.886	0.924	0.884	0.021	0.943	0.925

Note: The best and second-best results are highlighted in red and blue, respectively.

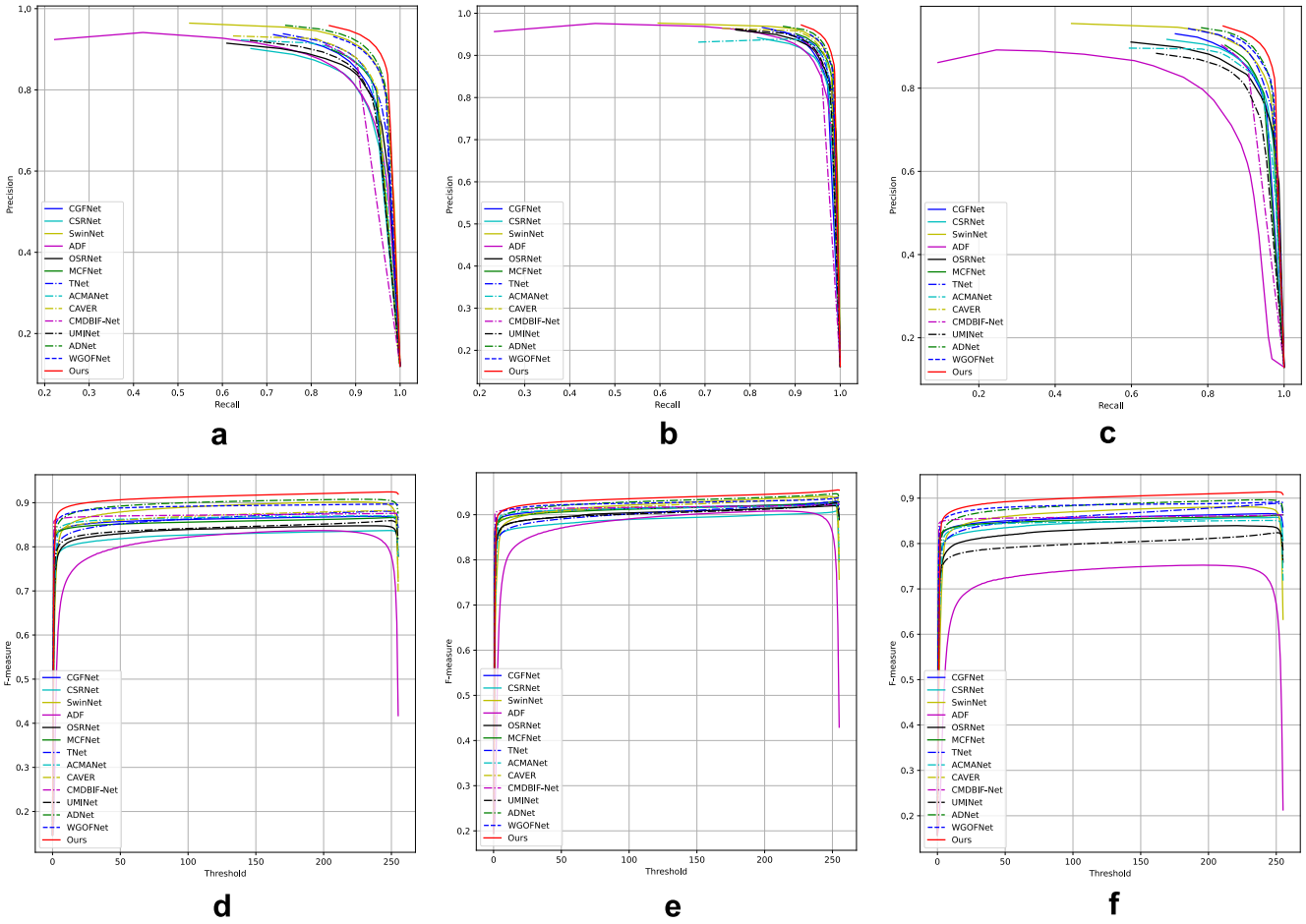

FIGURE 4 | Quantitative comparisons of Mamba4SOD with state-of-the-art RGB-T SOD methods on VT5000, VT1000 and VT821. The first row presents PR curves, while the second row shows *F*-measure curves under different thresholds. Specifically, (a) shows the PR curve comparison on VT5000, (b) shows the PR curve comparison on VT1000, (c) shows the PR curve comparison on VT821, (d) shows the *F*-measure curve comparison on VT5000, (e) shows the *F*-measure curve comparison on VT1000, and (f) shows the *F*-measure curve comparison on VT821.

ability to enhance fine-grained features, enabling the detection and segmentation of small but critical details. Figure 5i illustrates a scenario with blurred RGB images and thermal reflections. Although other methods are misled by these distortions, our method successfully aligns RGB and thermal data, achieving precise segmentation through cross-modal feature fusion. Figure 5j–m depicts scenarios involving blurred boundaries, where other models suffer significant degradation. Figure 5j involves multiple salient targets, Figure 5k features small targets, Figure 5l shows similar foreground and background and Figure 5m involves shadows blending with vehicle contours under lighting. In all these cases, our method successfully learns complete structural information, sharpens boundaries and maintains superior performance.

Overall, the superior performance of Mamba4SOD can be attributed to the powerful feature fusion and enhancement capabilities of the Mamba, combined with the multi-scale feature extraction potential of Swin Transformer V2. This architecture allows the model to achieve a well-balanced integration of global contextual information and fine-grained local details. The Mamba-based fusion module, through adaptive weighting, effectively facilitates cross-modal fusion by dynamically leveraging complementary features from both modalities. As a result, Mamba4SOD has robust adaptability and outperforms other methods across a diverse range of challenging scenarios.

4.4 | Ablation Analysis

4.4.1 | Effectiveness of Different Backbone Stages

To evaluate the impact of selecting different stages of the Swin Transformer V2, we conduct a series of experiments under computational constraints. We systematically explore various combinations of high-level features from different stages, and the results for the VT5000, VT1000 and VT821 datasets are summarised in Table 2, with the best results highlighted in red and the second-best in blue. The experimental results show that selecting features from the second, fourth and fifth stages yields superior performance. This can be attributed to the fact that RGB-T SOD tasks require a delicate balance between fine-grained edge details and high-level semantic understanding. Features from the second stage provide a richer representation of medium-grained local structures compared to the first stage, which primarily captures low-level details such as edges and textures. Although these low-level features are useful for detail recovery, they can introduce

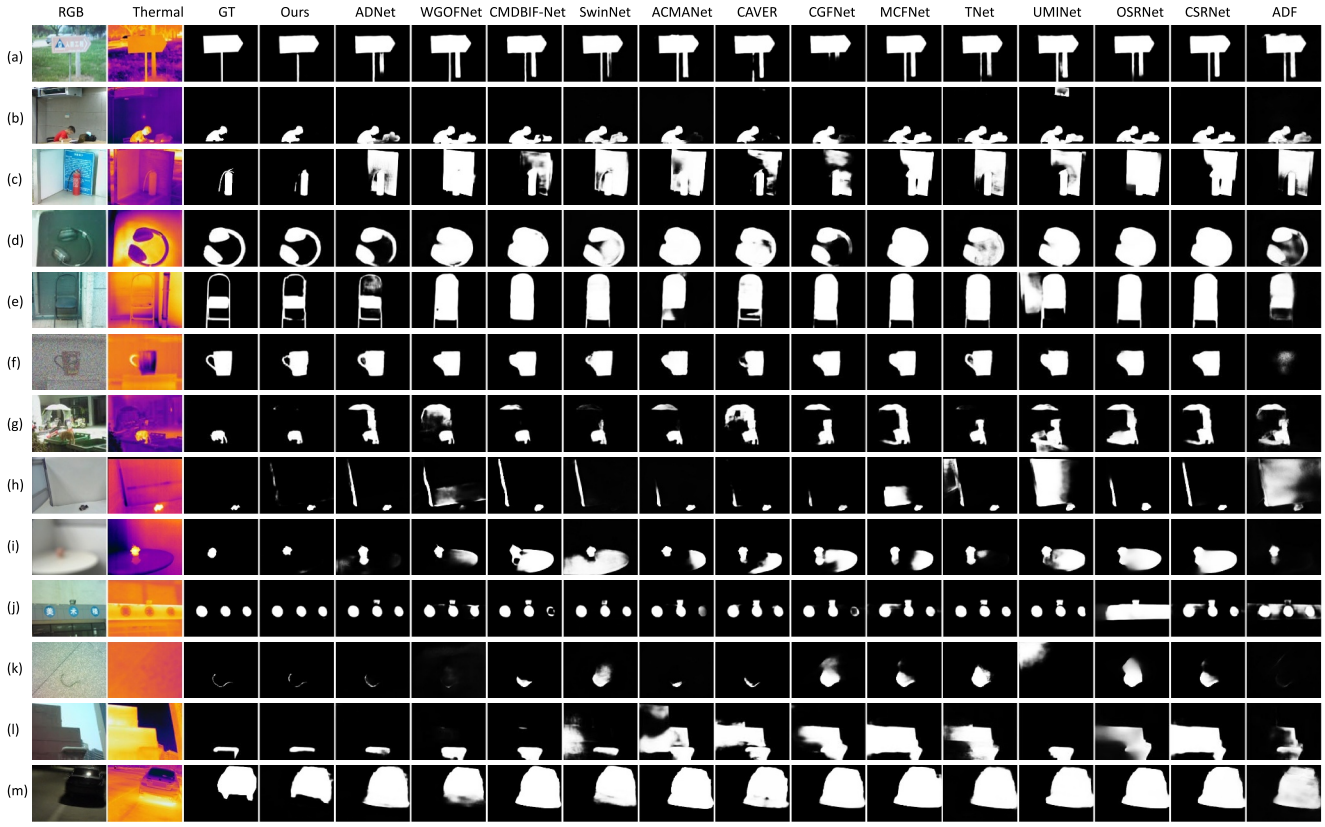

FIGURE 5 | Qualitative comparisons with 13 state-of-the-art methods. We select 13 RGB-T image pairs with diverse challenges to compare the quality of the saliency maps. From the left to right columns are RGB image, thermal image, ground-truth, and the results of 13 methods, respectively. Each row (a–m) depicts a different scene in the dataset.

noise in complex scenarios, making the second-stage features more robust and effective. Similarly, the fourth stage offers better high-level semantic information than the third stage, enabling the model to better understand the global context and relationships within the scene. By retaining features from the second, fourth, and fifth stages, the Mamba4SOD achieves an optimal balance between local details and global semantic comprehension. The combination not only enhances the model's ability to accurately locate and segment salient objects but also reduces computational costs.

4.4.2 | Effectiveness of Different Backbones

We compare the effectiveness of mainstream backbones, including ResNet50, ResNet-101 and Swin Transformer. The quantitative comparison results are presented in the first three rows of Table 3, showing the performance of our method with different backbones on the VT5000 dataset. While ResNet50 and ResNet-101 are widely used in computer vision tasks, they exhibit several limitations in the context of RGB-T SOD. Due to their inherently local convolutional operations, ResNet-based models struggle to capture long-range dependencies, which are essential for effectively integrating multi-modal information in complex scenes. Among the evaluated backbones, Swin Transformer V2 shows a notable improvement in detection performance, thanks to its global context and fine-grained details extraction capabilities.

4.4.3 | Effectiveness of the Mamba-Based Fusion Module

To validate the effectiveness of MFM in integrating RGB and thermal modalities, we conduct an ablation study of the architecture on VT5000 and VT821 datasets.

First, we train the proposed model using only the RGB modality (R + R) and only the thermal modality (T + T) and then conducted a quantitative comparison. The results are shown in the fourth and fifth rows of Tables 3 and 4, which indicate the significant advantages of dual-modality fusion. On the VT5000 dataset, the dual-modality model outperforms the single-modality models, achieving a 1.7% higher F_{avg} score than the RGB-only model and a 5.7% improvement over the thermal-only model. The F_ω values also increase by 1.7% and 5.9%, respectively. VT821 is a more challenging dataset with many low-quality inputs, so the importance of modality fusion becomes even more apparent. The dual-modality model shows a larger improvement, with F_{avg} increasing by 2% over the RGB-only model and 8.4% over the thermal-only model. Similarly, F_ω improves by 2.7% compared to RGB-only and 9.3% compared to thermal-only. These results validate that dual-modality training effectively combines complementary features from both modalities, leading to enhanced performance.

To further evaluate the contribution of the MFM, we conduct additional experiments by removing Mamba (w/o Mamba),

TABLE 2 | Ablation study of different stages in the proposed method.

Stages	VT5000						VT1000						VT821					
	$F_{avg}\uparrow$	$F_{max}\uparrow$	$F_{\omega}\uparrow$	$MAE\downarrow$	$E_m\uparrow$	$S_m\uparrow$	$F_{avg}\uparrow$	$F_{max}\uparrow$	$F_{\omega}\uparrow$	$MAE\downarrow$	$E_m\uparrow$	$S_m\uparrow$	$F_{avg}\uparrow$	$F_{max}\uparrow$	$F_{\omega}\uparrow$	$MAE\downarrow$	$E_m\uparrow$	$S_m\uparrow$
{5}	0.824	0.878	0.809	0.031	0.926	0.896	0.880	0.929	0.875	0.021	0.938	0.932	0.828	0.889	0.812	0.031	0.916	0.902
{1, 5}	0.880	0.916	0.877	0.023	0.949	0.918	0.914	0.949	0.921	0.015	0.951	0.944	0.862	0.911	0.862	0.026	0.930	0.912
{2, 5}	0.898	0.927	0.892	0.021	0.956	0.925	0.926	0.955	0.927	0.014	0.958	0.948	0.882	0.918	0.866	0.025	0.938	0.916
{3, 5}	0.870	0.918	0.877	0.021	0.947	0.919	0.904	0.951	0.917	0.015	0.948	0.943	0.857	0.918	0.864	0.023	0.933	0.915
{4, 5}	0.872	0.915	0.872	0.023	0.948	0.918	0.908	0.951	0.918	0.015	0.952	0.945	0.860	0.913	0.859	0.024	0.932	0.919
{1, 2, 5}	0.880	0.921	0.881	0.022	0.950	0.920	0.912	0.952	0.922	0.015	0.951	0.944	0.865	0.916	0.867	0.024	0.932	0.916
{1, 3, 5}	0.883	0.921	0.883	0.022	0.950	0.920	0.916	0.954	0.924	0.014	0.952	0.944	0.872	0.922	0.876	0.022	0.939	0.919
{1, 4, 5}	0.894	0.927	0.891	0.021	0.954	0.923	0.924	0.957	0.930	0.014	0.957	0.947	0.882	0.924	0.881	0.023	0.942	0.920
{2, 4, 5}	0.895	0.931	0.896	0.020	0.956	0.929	0.919	0.957	0.929	0.014	0.954	0.949	0.876	0.925	0.879	0.022	0.939	0.924
{3, 4, 5}	0.886	0.923	0.886	0.021	0.950	0.923	0.917	0.954	0.926	0.014	0.954	0.947	0.866	0.913	0.867	0.026	0.929	0.916
{2, 3, 5}	0.881	0.919	0.874	0.023	0.950	0.919	0.915	0.953	0.920	0.015	0.950	0.946	0.853	0.909	0.851	0.027	0.926	0.911
{1, 2, 3, 5}	0.885	0.923	0.886	0.022	0.950	0.922	0.917	0.954	0.926	0.014	0.952	0.945	0.870	0.920	0.874	0.023	0.937	0.917
{1, 2, 4, 5}	0.888	0.929	0.893	0.020	0.954	0.926	0.915	0.956	0.926	0.014	0.952	0.946	0.873	0.926	0.880	0.022	0.938	0.922
{1, 3, 4, 5}	0.891	0.926	0.890	0.021	0.954	0.924	0.921	0.958	0.930	0.013	0.954	0.948	0.850	0.900	0.842	0.031	0.923	0.898
{2, 3, 4, 5}	0.894	0.929	0.894	0.020	0.955	0.927	0.920	0.958	0.929	0.014	0.955	0.948	0.876	0.923	0.877	0.022	0.937	0.919
{1, 2, 3, 4, 5}	0.888	0.926	0.890	0.021	0.953	0.924	0.918	0.956	0.928	0.014	0.955	0.948	0.877	0.926	0.883	0.021	0.941	0.924

Note: The best and second best results are highlighted in red and blue respectively.

TABLE 3 | Ablation study on backbone and architecture on VT5000.

Type	Setting	VT5000					
		$F_{avg}\uparrow$	$F_{max}\uparrow$	$F_{\omega}\uparrow$	$MAE\downarrow$	$E_m\uparrow$	$S_m\uparrow$
Backbone	ResNet-50	0.852	0.898	0.842	0.030	0.927	0.896
	ResNet-101	0.858	0.900	0.845	0.029	0.933	0.897
	Swin-B	0.895	0.926	0.890	0.021	0.955	0.924
Architecture	RGB + RGB	0.885	0.921	0.882	0.023	0.950	0.919
	T + T	0.845	0.888	0.840	0.029	0.934	0.893
	W/o Mamba	0.894	0.931	0.896	0.020	0.956	0.928
	W/o weights	0.899	0.931	0.897	0.020	0.957	0.927
	Cross attention	0.883	0.927	0.886	0.021	0.952	0.925
	Ours	0.902	0.932	0.899	0.019	0.958	0.930

Note: The best results are highlighted in red.

TABLE 4 | Ablation study on backbone and architecture on VT821.

Type	Setting	VT821					
		$F_{avg}\uparrow$	$F_{max}\uparrow$	$F_{\omega}\uparrow$	$MAE\downarrow$	$E_m\uparrow$	$S_m\uparrow$
Backbone	ResNet-50	0.848	0.904	0.843	0.028	0.920	0.898
	ResNet-101	0.837	0.899	0.833	0.029	0.917	0.897
	Swin-B	0.876	0.919	0.873	0.022	0.941	0.918
Architecture	RGB + RGB	0.866	0.903	0.857	0.026	0.929	0.909
	T + T	0.802	0.855	0.791	0.038	0.909	0.862
	W/o Mamba	0.872	0.919	0.874	0.023	0.938	0.919
	W/o weights	0.878	0.921	0.875	0.024	0.939	0.917
	Cross attention	0.866	0.913	0.863	0.024	0.933	0.915
	Ours	0.886	0.924	0.884	0.021	0.943	0.925

Note: The best results are highlighted in red.

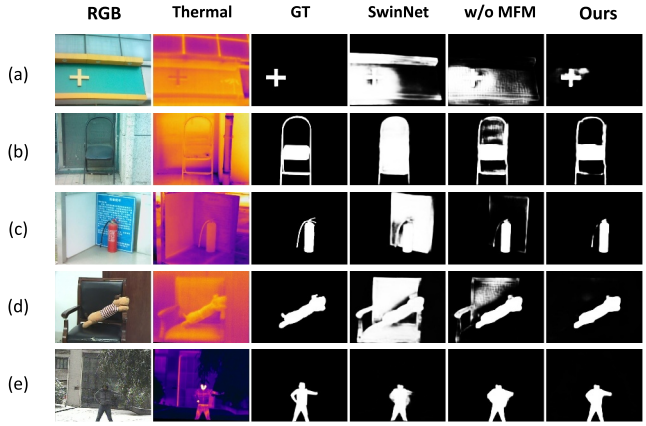

FIGURE 6 | Qualitative comparisons with SwinNet [21] and w/o MFM. Each row (a–e) depicts a different scene in the dataset.

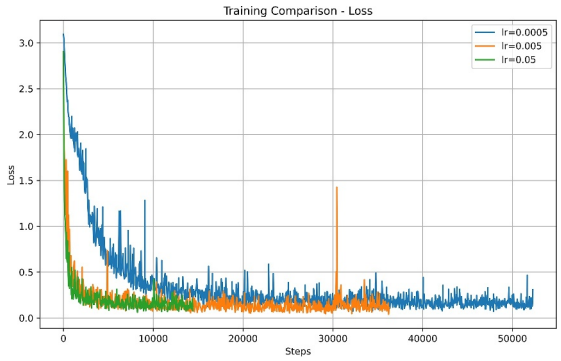

FIGURE 7 | Loss convergence curves for different learning rates (0.05, 0.005 and 0.0005).

excluding the adaptive weights (w/o weights) and replacing the MFM with cross-attention for feature fusion. The results are presented in Tables 3 and 4, highlighting the performance improvements brought by the proposed components. While cross-attention is a widely used feature fusion mechanism, it does not lead to significant improvements in RGB-T SOD. These results demonstrate that the Mamba with adaptive weights significantly enhances performance by dynamically adjusting feature contributions, thereby improving the model's ability to capture relevant information while suppressing noises.

Moreover, the visual comparison results are shown in Figure 6, which indicates that including the MFM enables better distinction between the target and background, resulting in more accurate boundary and complete objects.

TABLE 5 | Performance comparison with state-of-the-art methods on RGB-D datasets.

Datasets	Metric	D3Net	ICNet	DCMF	DRLF	SSF	UC-Net	JL-DCF	CoNet	DANet	EBFSP	CDNet	HAINet	RD3D	DSA2F	MMNet	SwinNet	Ours
NLPR	$S_m\uparrow$	0.912	0.923	0.900	0.903	0.914	0.920	0.925	0.908	0.920	0.915	0.902	0.924	0.930	0.918	0.925	0.941	0.953
	$F_\omega\uparrow$	0.861	0.870	0.839	0.843	0.875	0.890	0.878	0.846	0.875	0.897	0.848	0.897	0.892	0.892	0.889	0.908	0.915
	$E_m\uparrow$	0.944	0.944	0.933	0.936	0.949	0.953	0.953	0.934	0.951	0.952	0.935	0.957	0.958	0.950	0.950	0.967	0.962
	$MAE\downarrow$	0.030	0.028	0.035	0.032	0.026	0.025	0.022	0.031	0.027	0.026	0.032	0.024	0.022	0.024	0.024	0.018	0.022
NJU2K	$S_m\uparrow$	0.901	0.894	0.889	0.886	0.899	0.897	0.902	0.895	0.899	0.903	0.885	0.912	0.916	0.904	0.911	0.935	0.937
	$F_\omega\uparrow$	0.865	0.868	0.859	0.849	0.886	0.889	0.885	0.872	0.871	0.894	0.866	0.900	0.901	0.898	0.900	0.922	0.929
	$E_m\uparrow$	0.914	0.905	0.897	0.901	0.913	0.903	0.913	0.912	0.908	0.907	0.911	0.922	0.918	0.922	0.919	0.934	0.933
	$MAE\downarrow$	0.046	0.052	0.052	0.055	0.043	0.043	0.041	0.046	0.045	0.039	0.048	0.038	0.036	0.039	0.038	0.027	0.025
STERE	$S_m\uparrow$	0.899	0.903	0.883	0.888	0.887	0.903	0.903	0.905	0.901	0.900	0.896	0.907	0.911	0.897	0.891	0.919	0.924
	$F_\omega\uparrow$	0.859	0.865	0.841	0.845	0.867	0.885	0.869	0.884	0.868	0.870	0.873	0.885	0.886	0.893	0.880	0.893	0.902
	$E_m\uparrow$	0.920	0.915	0.904	0.915	0.921	0.922	0.919	0.927	0.921	0.912	0.922	0.925	0.927	0.927	0.924	0.929	0.929
	$MAE\downarrow$	0.046	0.045	0.054	0.050	0.046	0.039	0.040	0.037	0.043	0.045	0.042	0.040	0.037	0.039	0.045	0.033	0.031
DES	$S_m\uparrow$	0.898	0.920	0.877	0.895	0.905	0.933	0.931	0.911	0.924	0.937	0.875	0.935	0.935	0.916	0.830	0.945	0.945
	$F_\omega\uparrow$	0.870	0.889	0.820	0.868	0.876	0.917	0.900	0.861	0.899	0.913	0.839	0.924	0.917	0.901	0.746	0.926	0.936
	$E_m\uparrow$	0.951	0.959	0.923	0.954	0.948	0.974	0.969	0.945	0.968	0.974	0.921	0.974	0.975	0.955	0.893	0.980	0.980
	$MAE\downarrow$	0.031	0.027	0.040	0.030	0.025	0.018	0.020	0.027	0.023	0.018	0.034	0.018	0.019	0.023	0.058	0.016	0.015
SIP	$S_m\uparrow$	0.860	0.854	0.859	0.850	0.868	0.875	0.880	0.858	0.875	0.885	0.823	0.880	0.885	0.862	0.836	0.911	0.912
	$F_\omega\uparrow$	0.835	0.836	0.819	0.813	0.851	0.868	0.873	0.842	0.855	0.869	0.805	0.875	0.874	0.865	0.839	0.912	0.920
	$E_m\uparrow$	0.902	0.899	0.898	0.891	0.911	0.913	0.921	0.909	0.914	0.917	0.880	0.919	0.920	0.908	0.882	0.943	0.945
	$MAE\downarrow$	0.063	0.069	0.068	0.071	0.056	0.051	0.049	0.063	0.054	0.049	0.076	0.053	0.048	0.057	0.075	0.035	0.033

Note: The best results are highlighted in red.

4.4.4 | Hyper-Parameters Analysis

To further verify the effect of hyper-parameters, we conduct experiments comparing different learning rates (lr = 0.05, 0.005 and 0.0005), with the corresponding loss convergence curves shown in Figure 7. The results indicate that convergence speed increases with higher learning rates. Specifically, while lr = 0.05 leads to faster initial convergence, it causes less stable optimisation and slightly lower final performance. In contrast, lr = 0.005 offers a balanced trade-off, significantly accelerating convergence compared to lr = 0.0005, while maintaining stability and achieving the best final performance.

4.5 | Application to RGB-D SOD Task

To evaluate the performance of Mamba4SOD on another multi-modal task, we train our method under the same experimental conditions on the RGB-D SOD DUT dataset [53], and test it on several other RGB-D SOD datasets, including NLPR [54], NJU2K [55], STERE [56], DES [57] and SIP [58]. We compare our method with several state-of-the-art RGB-D SOD methods, including D3Net [58], ICNet [59], DCMF [60], DRLF [61], SSF [62], UC-Net [63], JL-DCF [64], CoNet [65], DANet [66], EBFSP [67], CDNet [68], HAINet [69], RD3D [70], DSA2F [71], MMNet [72] and SwinNet [21]. The saliency maps for evaluation are either provided by the published papers or generated by running the corresponding source codes to ensure a fair comparison. As shown in Table 5, our method achieves the best performance across multiple datasets, confirming its ability to handle both RGB-T and RGB-D modalities efficiently.

4.6 | Complexity Analysis

Table 6 shows the number of parameters and floating point operations (FLOPs) of different methods. We select representative F_ω for comparison due to its well-balanced trade-off between precision and recall. Compared with the most recent method CGFNet [51], *Ours_Res-Net50* achieves higher quantitative results with fewer parameters and FLOPs, demonstrating the effectiveness of the MFM. While ResNet50 is a widely used backbone, we chose not to rely solely on it due to its limited capability to capture long-range dependencies and multi-scale features efficiently.

TABLE 6 | Comparison of complexity and performance (weighted *F*-measure metrics on VT5000).

Backbone	Algorithm	params↓	*FLOPs*↓	F_ω↑
CNN	CAVER [25]	93.77 M	63.91 G	0.849
	ACMANet [26]	124.39 M	51.76 G	0.823
	CGFNet [51]	66.38 M	139.97 G	0.831
	Ours_ResNet50	49.16 M	32.06 G	0.890
Transformer	SwinNet [21]	199.18 M	124.14 G	0.846
	WGOFNet [18]	61.76 M	48.47 G	0.873
	ADNet [17]	93.32 M	56.68 G	0.884
	Ours_SwinV2-B	132.90 M	74.01 G	0.899

Note: The best and second best results are highlighted in red and blue, respectively.

By integrating the Mamba with the Transformer, *Ours_SwinV2-B* demonstrates a significant performance enhancement compared to all other methods. In comparison to SwinNet [21], which shares a similar backbone as our proposed method, we observe an improvement of over 5% in F_ω performance with fewer parameters and FLOPs, achieving state-of-the-art results. Furthermore, Mamba4SOD also exhibits notable advantages over WGOFNet [18] utilising PVT as the backbone and ADNet [17] using asymmetric structure. Specifically, based on the different scales of feature input, the parameter counts for the Mamba corresponding to the three selected stages in our method are 0.68, 0.71 and 1.26 M, respectively. Additionally, the average running time for the MFM is 23.74 ms under the same experimental condition, and our method achieves a real-time performance of over 18 frames per second (FPS).

4.7 | Failure Cases and Analysis

While our comprehensive evaluation metrics demonstrate the effectiveness of the proposed method, it still encounters challenges in certain complex scenarios. As shown in the first two rows of Figure 8, when both visible light and thermal modalities have interference, such as occlusions or similar foreground and background, our method struggles to accurately localise and fully segment the instances. Additionally, as shown in the last two rows of Figure 8, in highly challenging scenes where RGB images are significantly affected by noise and thermal images provide limited information, our method fails to produce satisfactory segmentation results. Despite these limitations, Mamba4SOD still outperforms existing comparison methods in these scenarios.

5 | Conclusion

In this paper, we proposed a novel RGB-T SOD network named Mamba4SOD, which integrates Swin Transformer V2 and Mamba architectures. The proposed method investigates optimal strategies for selecting and fusing different stages of Swin Transformer V2, which reduces model complexity and enhances performance. Furthermore, we proposed a Mamba-based fusion module to enhance the alignment and fusion of

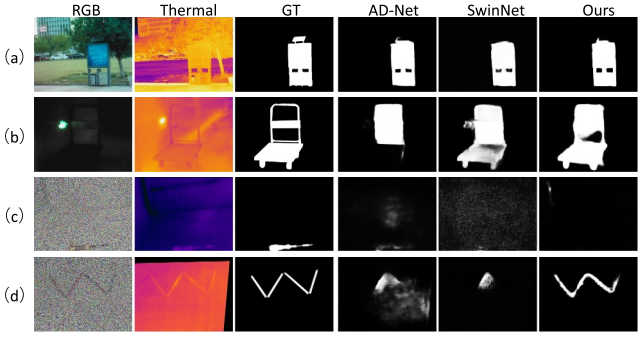

FIGURE 8 | Some failure cases of the proposed method. Each row (a–d) depicts a different scene in the dataset.

multi-modal features, significantly improving the integration of complementary information. Experimental results demonstrate that the proposed method outperforms existing state-of-the-art methods, particularly in scenarios involving blurred boundaries and background misclassification, accurately detecting salient objects under challenging conditions.

Additionally, recognising that perfect alignment of RGB-T data may not always be feasible in real-world scenarios, we plan to explore methods for addressing misaligned or partially aligned RGB-T data to boost the model's performance under such situations.

Author Contributions

Yi Xu: conceptualization, methodology, software, writing – original draft. **Ruichao Hou:** project administration, writing – original draft. **Ziheng Qi:** investigation, writing – original draft. **Tongwei Ren:** funding acquisition, resources, writing – review and editing.

Acknowledgements

This work was supported by the National Natural Science Foundation of China (62072232), the Key R&D Project of Jiangsu Province (BE2022138), the Fundamental Research Funds for the Central Universities (0217-14380026), the Innovation Project of State Key Laboratory for Novel Software Technology Nanjing University (ZZKT-2024B20) and the Collaborative Innovation Center of Novel Software Technology and Industrialization.

Conflicts of Interest

The authors declare no conflicts of interest.

Data Availability Statement

The data that support the findings of this study are available from the corresponding author upon reasonable request.

References

1. P. Zhang, T. Yan, Y. Liu, and H. Lu, "Fantastic Animals and Where to Find Them: Segment Any Marine Animal With Dual SAM," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2024), 2578–2587.

2. C. Lin, C. Xu, D. Luo, et al., "Learning Salient Boundary Feature for Anchor-Free Temporal Action Localization," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2021), 3320–3329.

3. S. Sowmyayani and P. A. J. Rani, "Content Based Video Retrieval System Using Two Stream Convolutional Neural Network," *Multimedia Tools and Applications* 82, no. 16 (2023): 24465–24483, https://doi.org/10.1007/s11042-023-14784-5.

4. X. Li, T. Zhang, Z. Liu, et al., "Saliency Guided Siamese Attention Network for Infrared Ship Target Tracking," *IEEE Transactions on Intelligent Vehicles* (2024): 1–18, https://doi.org/10.1109/tiv.2024.3370233.

5. J.-J. Liu, Q. Hou, Z.-A. Liu, and M.-M. Cheng, "PoolNet+: Exploring the Potential of Pooling for Salient Object Detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence* 45, no. 1 (2023): 887–904, https://doi.org/10.1109/tpami.2021.3140168.

6. Y.-H. Wu, Y. Liu, L. Zhang, M.-M. Cheng, and B. Ren, "EDN: Salient Object Detection via Extremely-Downsampled Network," *IEEE Transactions on Image Processing* 31 (2022): 3125–3136, https://doi.org/10.1109/tip.2022.3164550.

7. N. Liu, Z. Luo, N. Zhang, and J. Han, "VST++: Efficient and Stronger Visual Saliency Transformer," *IEEE Transactions on Pattern Analysis and Machine Intelligence* 46, no. 11 (2024): 7300–7316, https://doi.org/10.1109/tpami.2024.3388153.

8. Z. Tu, Y. Ma, Z. Li, C. Li, J. Xu, and Y. Liu, "RGBT Salient Object Detection: A Large-Scale Dataset and Benchmark," *IEEE Transactions on Multimedia* 25 (2022): 4163–4176, https://doi.org/10.1109/tmm.2022.3171688.

9. X. Jiang, Y. Hou, H. Tian, and L. Zhu, "Mirror Complementary Transformer Network for RGB-Thermal Salient Object Detection," *IET Computer Vision* 18, no. 1 (2024): 15–32, https://doi.org/10.1049/cvi2.12221.

10. K. Wang, D. Lin, C. Li, Z. Tu, and B. Luo, "Alignment-Free RGBT Salient Object Detection: Semantics-Guided Asymmetric Correlation Network and a Unified Benchmark," *IEEE Transactions on Multimedia* 26 (2024): 10692–10707, https://doi.org/10.1109/tmm.2024.3410542.

11. Q. Wang, P. Yan, Y. Yuan, and X. Li, "Multi-Spectral Saliency Detection," *Pattern Recognition Letters* 34, no. 1 (2013): 34–41, https://doi.org/10.1016/j.patrec.2012.06.002.

12. G. Wang, C. Li, Y. Ma, A. Zheng, J. Tang, and B. Luo, RGB-T Saliency Detection Benchmark: Dataset, Baselines, Analysis and a Novel Approach, IGTA 2018, Beijing, China, April 8–10, 2018, Revised Selected Papers, 13 (2018), 359–369, https://doi.org/10.1007/978-981-13-1702-6_36.

13. X. Li, H. Ding, H. Yuan, et al., "Transformer-Based Visual Segmentation: A Survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence* 46, no. 12 (2024): 10138–10163, https://doi.org/10.1109/tpami.2024.3434373.

14. H. Ding, C. Liu, S. He, X. Jiang, P. H. Torr, and S. Bai, "MOSE: A New Dataset for Video Object Segmentation in Complex Scenes," in *Proceedings of the IEEE/CVF International Conference on Computer Vision* (2023), 20224–20234.

15. W. Wang, E. Xie, X. Li, et al., "Pyramid Vision Transformer: A Versatile Backbone for Dense Prediction Without Convolutions," in *Proceedings of the IEEE International Conference on Computer Vision* (2021), 568–578.

16. Z. Liu, Y. Lin, Y. Cao, et al., "Swin Transformer: Hierarchical Vision Transformer Using Shifted Windows," in *Proceedings of the IEEE International Conference on Computer Vision* (2021), 10012–10022.

17. Y. Fang, R. Hou, J. Bei, T. Ren, and G. Wu, "ADNet: An Asymmetric Dual-Stream Network for RGB-T Salient Object Detection," in *Proceedings of ACM International Conference on Multimedia in Asia* (2023), 1–7.

18. J. Wang, G. Li, J. Shi, and J. Xi, "Weighted Guided Optional Fusion Network for RGB-T Salient Object Detection," *ACM Transactions on Multimedia Computing, Communications, and Applications* 20, no. 5 (2024): 1–20, https://doi.org/10.1145/3624984.

19. Z. Liu, H. Hu, Y. Lin, et al., "Swin Transformer V2: Scaling Up Capacity and Resolution," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2022), 12009–12019.

20. A. Gu and T. Dao, "Mamba: Linear-Time Sequence Modeling With Selective State Spaces," preprint, arXiv:2312.00752 (2024).

21. Z. Liu, Y. Tan, Q. He, and Y. 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, no. 7 (2022): 4486–4497, https://doi.org/10.1109/tcsvt.2021.3127149.

22. Z. Tu, T. Xia, C. Li, X. Wang, Y. Ma, and J. Tang, "RGB-T Image Saliency Detection via Collaborative Graph Learning," *IEEE Transactions on Multimedia* 22, no. 1 (2019): 160–173, https://doi.org/10.1109/tmm.2019.2924578.

23. F. Huo, X. Zhu, L. Zhang, Q. Liu, and Y. Shu, "Efficient Context-Guided Stacked Refinement Network for RGB-T Salient Object Detection," *IEEE Transactions on Circuits and Systems for Video Technology* 32, no. 5 (2021): 3111–3124, https://doi.org/10.1109/tcsvt.2021.3102268.

24. F. Huo, X. Zhu, Q. Zhang, Z. Liu, and W. Yu, "Real-Time One-Stream Semantic-Guided Refinement Network for RGB-Thermal Salient Object Detection," *IEEE Transactions on Instrumentation and Measurement* 71 (2022): 1–12, https://doi.org/10.1109/tim.2022.3185323.

25. Y. Pang, X. Zhao, L. Zhang, and H. Lu, "CAVER: Cross-Modal View-Mixed Transformer for Bi-Modal Salient Object Detection," *IEEE Transactions on Image Processing* 32 (2023): 892–904, https://doi.org/10.1109/tip.2023.3234702.

26. C. Xu, Q. Li, Q. Zhou, X. Jiang, D. Yu, and Y. Zhou, "Asymmetric Cross-Modal Activation Network for RGB-T Salient Object Detection," *Knowledge-Based Systems* 258 (2022): 110047, https://doi.org/10.1016/j.knosys.2022.110047.

27. Y. Wang, C. Dongye, and W. Zhao, "Cross-Collaboration Weighted Fusion Network for RGB-T Salient Detection," in *International Conference on Intelligent Computing* (2024), 301–312.

28. A. Gu, K. Goel, and C. Re, "Efficiently Modeling Long Sequences With Structured State Spaces," in *International Conference on Learning Representations* (2022), 1–27.

29. J. T. Smith, A. Warrington, and S. Linderman, "Simplified State Space Layers for Sequence Modeling," in *The Eleventh International Conference on Learning Representations* (2023).

30. L. Zhu, B. Liao, Q. Zhang, X. Wang, W. Liu, and X. Wang, "Vision Mamba: Efficient Visual Representation Learning With Bidirectional State Space Model," in *Forty-First International Conference on Machine Learning* (2024).

31. Y. Liu, Y. Tian, Y. Zhao, et al., "VMamba: Visual State Space Model," preprint, arXiv:2401.10166 (2024).

32. S. Li, H. Singh, and A. Grover, "Mamba-ND: Selective State Space Modeling for Multi-Dimensional Data," *Lecture Notes in Computer Science* (2024): 75–92, preprint, arXiv:2402.05892, https://doi.org/10.1007/978-3-031-73414-4_5.

33. Z. Zhao and P. He, "YOLO-Mamba: Object Detection Method for Infrared Aerial Images," *Signal, Image and Video Processing* 18, no. 12 (2024): 1–11, https://doi.org/10.1007/s11760-024-03507-4.

34. Z. Wang, C. Li, H. Xu, and X. Zhu, "Mamba YOLO: SSMs-Based YOLO for Object Detection," preprint, arXiv:2406.05835 (2024).

35. J. Ma, F. Li, and B. Wang, "U-Mamba: Enhancing Long-Range Dependency for Biomedical Image Segmentation," preprint, arXiv:2401.04722 (2024).

36. Z. Xing, T. Ye, Y. Yang, G. Liu, and L. Zhu, "SegMamba: Long-Range Sequential Modeling Mamba for 3D Medical Image Segmentation," *Lecture Notes in Computer Science* (2024): 578–588, https://doi.org/10.1007/978-3-031-72111-3_54.

37. T. Huang, X. Pei, S. You, F. Wang, C. Qian, and C. Xu, "LocalMamba: Visual State Space Model With Windowed Selective Scan," preprint, arXiv:2403.09338 (2024).

38. H. Ding, C. Liu, S. Wang, and X. Jiang, "VLT: Vision-Language Transformer and Query Generation for Referring Segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence* 45, no. 6 (2023): 7900–7916, https://doi.org/10.1109/tpami.2022.3217852.

39. Z. Wu, L. Su, and Q. Huang, "Cascaded Partial Decoder for Fast and Accurate Salient Object Detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2019), 3907–3916.

40. X. Xie, Y. Cui, T. Tan, X. Zheng, and Z. Yu, "FusionMamba: Dynamic Feature Enhancement for Multimodal Image Fusion With Mamba," *Visual Intelligence* 2, no. 1 (2024): 37, https://doi.org/10.1007/s44267-024-00072-9.

41. L. Yuan, Y. Chen, T. Wang, et al., "Tokens-to-Token ViT: Training Vision Transformers From Scratch on ImageNet," in *Proceedings of the IEEE International Conference on Computer Vision* (2021), 558–567.

42. J. Wei, S. Wang, Z. Wu, C. Su, Q. Huang, and Q. Tian, "Label Decoupling Framework for Salient Object Detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2020), 13025–13034.

43. R. Achanta, S. Hemami, F. Estrada, and S. Susstrunk, "Frequency-Tuned Salient Region Detection," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition* (2009), 1597–1604.

44. R. Margolin, L. Zelnik-Manor, and A. Tal, "How to Evaluate Foreground Maps?," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2014), 248–255.

45. D.-P. Fan, C. Gong, Y. Cao, B. Ren, M.-M. Cheng, and A. Borji, "Enhanced-Alignment Measure for Binary Foreground Map Evaluation," in *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence* (2018), 698–704, preprint, arXiv:1805.10421, https://doi.org/10.24963/ijcai.2018/97.

46. D.-P. Fan, M.-M. Cheng, Y. Liu, T. Li, and A. Borji, "Structure-Measure: A New Way to Evaluate Foreground Maps," in *Proceedings of the IEEE International Conference on Computer Vision* (2017), 4548–4557.

47. F. Perazzi, P. Krähenbühl, Y. Pritch, and A. Hornung, "Saliency Filters: Contrast Based Filtering for Salient Region Detection," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition* (2012), 733–740.

48. L. Gao, P. Fu, M. Xu, T. Wang, and B. Liu, "UMINet: A Unified Multi-Modality Interaction Network for RGB-D and RGB-T Salient Object Detection," *Visual Computer* 40, no. 3 (2024): 1565–1582, https://doi.org/10.1007/s00371-023-02870-6.

49. R. Cong, K. Zhang, C. Zhang, et al., "Does Thermal Really Always Matter for RGB-T Salient Object Detection?," *IEEE Transactions on Multimedia* 25 (2022): 6971–6982, https://doi.org/10.1109/tmm.2022.3216476.

50. S. Ma, K. Song, H. Dong, H. Tian, and Y. Yan, "Modal Complementary Fusion Network for RGB-T Salient Object Detection," *Applied Intelligence* 53, no. 8 (2023): 9038–9055, https://doi.org/10.1007/s10489-022-03950-1.

51. J. Wang, K. Song, Y. Bao, L. Huang, and Y. Yan, "CGFNet: Cross-Guided Fusion Network for RGB-T Salient Object Detection," *IEEE Transactions on Circuits and Systems for Video Technology* 32, no. 5 (2021): 2949–2961, https://doi.org/10.1109/tcsvt.2021.3099120.

52. Z. Xie, F. Shao, G. Chen, et al., "Cross-Modality Double Bidirectional Interaction and Fusion Network for RGB-T Salient Object Detection," *IEEE Transactions on Circuits and Systems for Video Technology* 33, no. 8 (2023): 4149–4163, https://doi.org/10.1109/tcsvt.2023.3241196.

53. Y. Piao, W. Ji, J. Li, M. Zhang, and H. Lu, "Depth-Induced Multi-Scale Recurrent Attention Network for Saliency Detection," in *Proceedings of the IEEE/CVF International Conference on Computer Vision* (2019), 7254–7263.

54. H. Peng, B. Li, W. Xiong, W. Hu, and R. Ji, "RGBD Salient Object Detection: A Benchmark and Algorithms," in *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6–12, 2014, Proceedings, Part III 13* (Springer, 2014), 92–109.

55. R. Ju, L. Ge, W. Geng, T. Ren, and G. Wu, "Depth Saliency Based on Anisotropic Center-Surround Difference," in *2014 IEEE International Conference on Image Processing* (IEEE, 2014), 1115–1119.

56. Y. Niu, Y. Geng, X. Li, and F. Liu, "Leveraging Stereopsis for Saliency Analysis," in *2012 IEEE Conference on Computer Vision and Pattern Recognition* (IEEE, 2012), 454–461.

57. Y. Cheng, H. Fu, X. Wei, J. Xiao, and X. Cao, "Depth Enhanced Saliency Detection Method," in *Proceedings of International Conference on Internet Multimedia Computing and Service* (2014), 23–27.

58. D.-P. Fan, Z. Lin, Z. Zhang, M. Zhu, and M.-M. Cheng, "Rethinking RGB-D Salient Object Detection: Models, Data Sets, and Large-Scale Benchmarks," *IEEE Transactions on Neural Networks and Learning Systems* 32, no. 5 (2020): 2075–2089, https://doi.org/10.1109/tnnls.2020.2996406.

59. G. Li, Z. Liu, and H. Ling, "ICNet: Information Conversion Network for RGB-D Based Salient Object Detection," *IEEE Transactions on Image Processing* 29 (2020): 4873–4884, https://doi.org/10.1109/tip.2020.2976689.

60. H. Chen, Y. Deng, Y. Li, T.-Y. Hung, and G. Lin, "RGBD Salient Object Detection via Disentangled Cross-Modal Fusion," *IEEE Transactions on Image Processing* 29 (2020): 8407–8416, https://doi.org/10.1109/tip.2020.3014734.

61. X. Wang, S. Li, C. Chen, Y. Fang, A. Hao, and H. Qin, "Data-Level Recombination and Lightweight Fusion Scheme for RGB-D Salient Object Detection," *IEEE Transactions on Image Processing* 30 (2020): 458–471, https://doi.org/10.1109/tip.2020.3037470.

62. M. Zhang, W. Ren, Y. Piao, Z. Rong, and H. Lu, "Select, Supplement and Focus for RGB-D Saliency Detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2020), 3472–3481.

63. J. Zhang, D.-P. Fan, Y. Dai, et al., "UC-Net: Uncertainty Inspired RGB-D Saliency Detection via Conditional Variational Autoencoders," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2020), 8582–8591.

64. K. Fu, D.-P. Fan, G.-P. Ji, and Q. 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* (2020), 3052–3062.

65. W. Ji, J. Li, M. Zhang, Y. Piao, and H. Lu, "Accurate RGB-D Salient Object Detection via Collaborative Learning," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XVIII 16* (Springer, 2020), 52–69.

66. X. Zhao, L. Zhang, Y. Pang, H. Lu, and L. Zhang, "A Single Stream Network for Robust and Real-Time RGB-D Salient Object Detection," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXII 16* (Springer, 2020), 646–662.

67. N. Huang, Y. Yang, D. Zhang, Q. Zhang, and J. Han, "Employing Bilinear Fusion and Saliency Prior Information for RGB-D Salient Object Detection," *IEEE Transactions on Multimedia* 24 (2021): 1651–1664, https://doi.org/10.1109/tmm.2021.3069297.

68. W.-D. Jin, J. Xu, Q. Han, Y. Zhang, and M.-M. Cheng, "CDNet: Complementary Depth Network for RGB-D Salient Object Detection," *IEEE Transactions on Image Processing* 30 (2021): 3376–3390, https://doi.org/10.1109/tip.2021.3060167.

69. G. Li, Z. Liu, M. Chen, Z. Bai, W. Lin, and H. Ling, "Hierarchical Alternate Interaction Network for RGB-D Salient Object Detection," *IEEE Transactions on Image Processing* 30 (2021): 3528–3542, https://doi.org/10.1109/tip.2021.3062689.

70. Q. Chen, Z. Liu, Y. Zhang, K. Fu, Q. Zhao, and H. Du, "RGB-D Salient Object Detection via 3D Convolutional Neural Networks," in *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 35 (2021), 1063–1071, https://doi.org/10.1609/aaai.v35i2.16191.

71. P. Sun, W. Zhang, H. Wang, S. Li, and X. Li, "Deep RGB-D Saliency Detection With Depth-Sensitive Attention and Automatic Multi-Modal Fusion," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2021), 1407–1417.

72. W. Gao, G. Liao, S. Ma, G. Li, Y. Liang, and W. Lin, "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, no. 4 (2021): 2091–2106, https://doi.org/10.1109/tcsvt.2021.3082939.