LEAF-Mamba: Local Emphatic and Adaptive Fusion State Space Model for RGB-D Salient Object Detection

Lanhu Wu
Dalian University of Technology
Dalian, China
lanhoong0406@gmail.com

Zilin Gao
Dalian University of Technology
Dalian, China
gzl@mail.dlut.edu.cn

Hao Fei*
National University of Singapore
Singapore
haofei37@nus.edu.sg

Mong-Li Lee
National University of Singapore
Singapore
dcsleeml@nus.edu.sg

Wynne Hsu
National University of Singapore
Singapore
dcshsuw@nus.edu.sg

ABSTRACT

RGB-D salient object detection (SOD) aims to identify the most conspicuous objects in a scene with the incorporation of depth cues. Existing methods mainly rely on CNNs, limited by the local receptive fields, or Vision Transformers that suffer from the cost of quadratic complexity, posing a challenge in balancing performance and computational efficiency. Recently, state space models (SSM), Mamba, have shown great potential for modeling long-range dependency with linear complexity. However, directly applying SSM to RGB-D SOD may lead to deficient local semantics as well as the inadequate cross-modality fusion. To address these issues, we propose a **L**ocal **E**mphatic and **A**daptive **F**usion state space model (**LEAF-Mamba**) that contains two novel components: 1) a local emphatic state space module (LE-SSM) to capture multi-scale local dependencies for both modalities. 2) an SSM-based adaptive fusion module (AFM) for complementary cross-modality interaction and reliable cross-modality integration. Extensive experiments demonstrate that the LEAF-Mamba consistently outperforms 16 state-of-the-art RGB-D SOD methods in both efficacy and efficiency. Moreover, our method can achieve excellent performance on the RGB-T SOD task, proving a powerful generalization ability. Our code is publicly available at https://github.com/LanhooNg/LEAF-Mamba.

CCS CONCEPTS

• **Computing methodologies** → **Interest point and salient region detections**.

KEYWORDS

RGB-D Salient Object Detection, State Space Model, Local Emphatic, Adaptive Fusion

*Corresponding Author.

ACM MM '25, October 27–31, 2025, Dublin, Ireland

ACM ISBN 978-1-4503-XXXX-X/2018/06
https://doi.org/XXXXXXX.XXXXXXX

ACM Reference Format:
Lanhu Wu, Zilin Gao, Hao Fei, Mong-Li Lee, and Wynne Hsu. 2025. LEAF-Mamba: Local Emphatic and Adaptive Fusion State Space Model for RGB-D Salient Object Detection. In *Proceedings of Proceedings of the 33rd ACM International Conference on Multimedia (ACM MM '25).* ACM, New York, NY, USA, 10 pages. https://doi.org/XXXXXXX.XXXXXXX

1 INTRODUCTION

Salient object detection (SOD) is one of the fundamental vision computing tasks, aiming to pinpoint the most prominent objects in an image. Yet RGB SOD [32, 45, 58, 62] may struggle in challenging scenarios such as complex backgrounds and similar appearances between objects and their surroundings. Thus, depth data, with affluent spatial structure information, is naturally utilized as a supplementary input in addition to the RGB image for accurate saliency prediction, resulting in the task of RGB-D SOD [37, 72]. Numerous prior methods have been proposed for RGB-D SOD. Early methods predominantly rely on the convolutional neural networks (CNNs) for single-modality representation and cross-modality fusion, focusing on discriminative modeling [7, 68], feature fusion [16, 21, 28, 67], information optimization [2, 23], model lightweighting [44, 61, 71]. However, CNN-based methods inherently suffer from the limited receptive field of convolutional operation, posing challenges for capturing long-range dependencies. To overcome this, a series of Transformer-based methods [6, 9, 47] are developed, leveraging the self-attention mechanism [52] for global context modeling, thus achieving the state-of-the-art (SoTA). Nevertheless, Transformers can be constrained by high computational complexity due to the quadratic growth of resources with the increase in tokens, sacrificing efficiency. While some attempts [35, 41] improve the efficiency by reducing the dimension of processing features, they compromise the extent of the receptive fields. Therefore, achieving ***high performance*** meanwhile maintaining ***model efficiency*** becomes the key bottleneck of RGB-D SOD. Figure 1 compares the existing RGB-D SOD research in these two dimensions.

Recently, the newly merged state space models (SSMs), especially Mamba [19], have shown great potential for modeling long-range dependency with linear complexity, achieving excellent performance with prominent efficiency advantage in various visual tasks, such as image classification [34, 75], video analysis [29] and pathological diagnosis [39, 66]. One may directly integrate SSM backbone

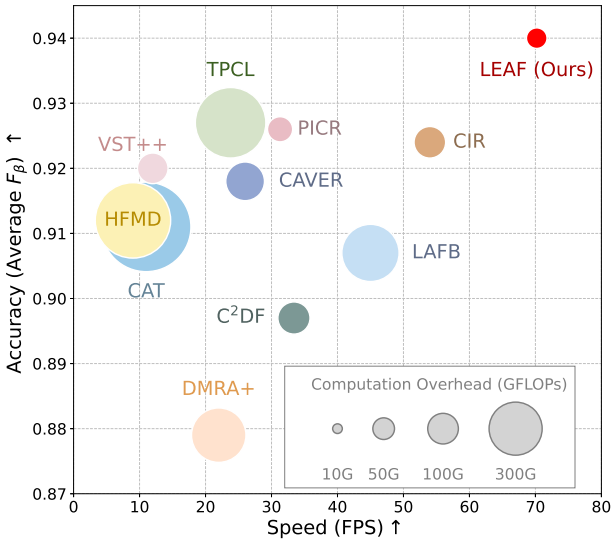

Figure 1: The comparisons with baselines on RGB-D SOD, with respect to efficacy and efficiency. The accuracy indicates the average F_β on NJUD [25], NLPR [42] and STERE [40].

for RGB-D SOD, by modeling the 2D selective scan (SS2D) mechanism to process vision data, i.e., VMamba [34]. Despite bridging 1D array scanning and 2D plane traversal, it struggles with maintaining the proximity of adjacent tokens, which is critical for local representation modeling in RGB-D SOD. While the windowed selective scan strategy might be helpful [22, 54], it still fails to capture the multi-scale information due to the fixed window size. Besides, current SSM-based methods tend to treat the RGB and depth features equally during the process of intermediate fusion [18, 53], while, unfortunately, they largely ignore the cross-modality complementarity and single-modality reliability evidently.

This work is dedicated to addressing both the efficacy and efficiency bottlenecks in existing RGB-D SOD. We introduce a novel *Local Emphatic and Adaptive Fusion* SSM system, namely **LEAF-Mamba**, as shown in Figure 2. **First**, a local emphatic state space module (LE-SSM) is proposed to enrich multi-scale local information in intermediate features of each modality via the multi-scale windowed 2D selective scan (MSW-SS2D). Different from existing SS2D [34], our MSW-SS2D adopts a four-scale windowed scanning mechanism in four ways for spatial domain traversal. As such, adjacent tokens are fully aggregated in multi-scale windows for local modeling without extra computational cost. **Second**, an SSM-based adaptive fusion module (AFM) is devised for cross-modality interaction and integration at multiple stages. Specifically, the AFM incorporates a cross-modality second-order pooling (CSoP) layer to compute the modality-specific similarity between RGB and depth features. Based on this, the AFM selects the discriminative regions for cross-modality interaction as well as the similar regions for cross-modality fusion under the paradigm of SSM. In this way, our system achieves complementary interplay and reliable integration of two modalities in an attentive manner. Attributing to the comprehensive feature representation, robust cross-modality fusion and efficient SSM, our LEAF-Mamba delivers great performance with a relatively low computational cost, as depicted in Figure 1.

Experimentally, we validate our method on seven RGB-D SOD benchmarks including NJUD [25], NLPR [42], SIP [15], STERE [40], SSD [74], LFSD [30] and DUT-D [43], where the results demonstrate its superiority over 16 SoTA methods in terms of both efficacy and efficiency. Particularly, LEAF-Mamba reduces the MAE by 13.2% and 8.16% on SSD and LFSD with only 18.1 GFLOPs and an astonishing real-time speed of 70.2 FPS. Besides, we conduct detailed ablations to verify the effectiveness of the proposed LE-SSM and AFM in multi-scale local enhancement and selective cross-modality fusion, respectively. Moreover, we extend our method to the RGB-T SOD task and further demonstrate its prominent generalizability.

To sum up, in this paper we propose a novel RGB-D SOD system (**LEAF-Mamba**) based on the SSM technique, where our main contributions are threefold:

- We devise a novel local emphatic state space module (LE-SSM) which performs a four-scale windowed selective scan to enrich multi-scale local information with low computational cost.
- We introduce an SSM-based adaptive fusion module (AFM) with a modality-specific selective mechanism, which dynamically interacts the complementary cues and fuses the reliable content in RGB and depth features.
- Our system not only sets new records on 7 RGB-D SOD benchmarks, but also achieves a low computation overhead of 18.1 GFLOPs and a real-time inference speed of 70.2 FPS. Also, it shows prominent generalizability on RGB-T SOD.

2 RELATED WORK

2.1 RGB-D Salient Object Detection

RGB-D SOD combines RGB images with depth cues to identify the most conspicuous objects in a scene. With the development of deep learning technology, CNN-based methods are firstly proposed for RGB-D SOD. Zhang *et al.* [69] introduce a complimentary interaction module to discriminatively select useful representation from the RGB and depth data. Ji *et al.* [23] propose a depth calibration and fusion framework to calibrate the depth map and realize an efficient fusion of two modalities. Wu *et al.* [61] present an efficient RGB-D SOD method based on mobile network and an implicit depth restoration technique to strengthen the mobile backbones.

However, due to the confined receptive field of CNN, these methods are deficient in extracting global context. To this end, some Transformer-based methods are proposed, which realize long-range modeling by self-attention mechanism. Liu *et al.* [35] introduce a triplet Transformer embedding module for modeling high-level features. Pang *et al.* [41] propound a view-mixed Transformer to excavate the global cues in intra-modal features and simplify the cross-modal interaction and alignment. Cong *et al.* [9] set forth a CNN-assisted Transformer architecture with point-aware interaction and CNN-induced refinement. Despite their promising results, these methods typically suffer from the quadratic scaling inherent in the self-attention mechanism, particularly for the dual-modality scenario. Different from them, our proposed method benefits from the linear overhead of Mamba, making a trade-off between effectiveness and efficiency, as shown in Figure 1.

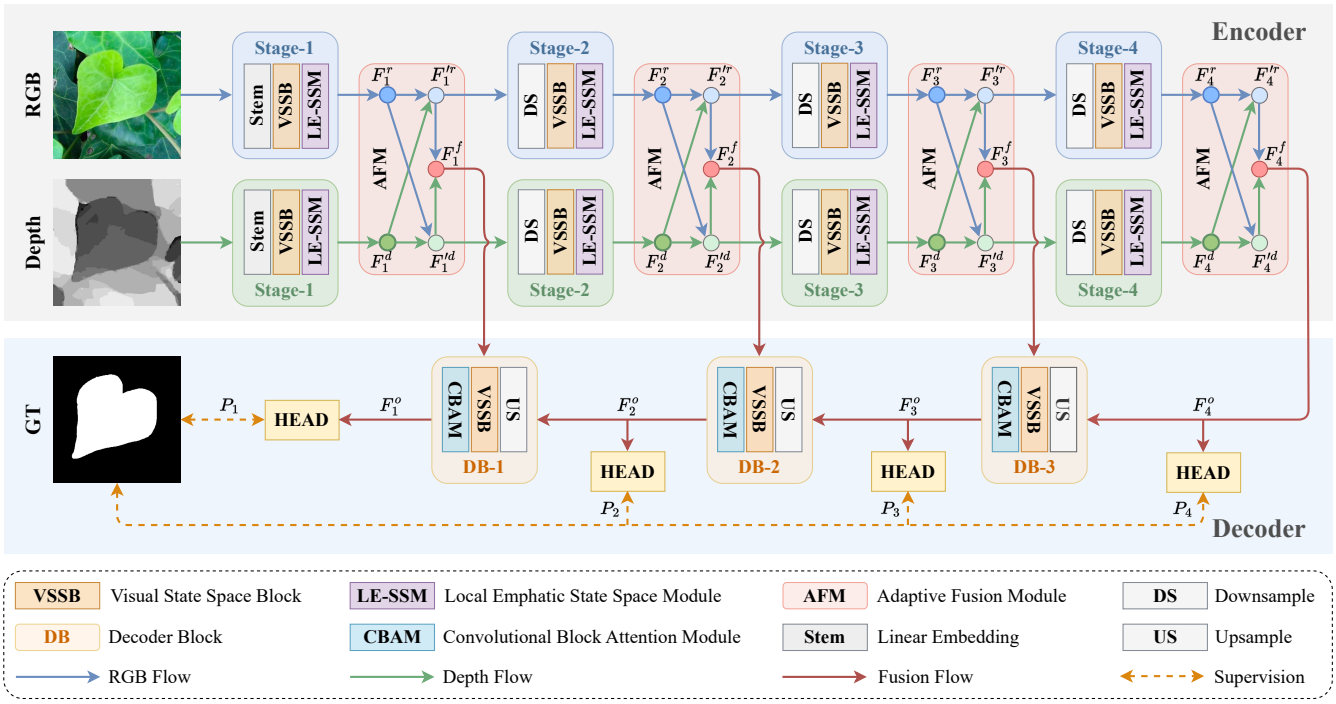

Figure 2: The whole pipeline of our proposed LEAF-Mamba, which consists of two main components: local emphatic state space module (LE-SSM) and adaptive fusion module (AFM). Please refer to Section 3 for details.

2.2 State Space Models

State space models (SSMs) [20, 46], with linear complexity, have emerged as compelling alternatives to Transformers for modeling long-range dependency. Recently, Gu *et al.* [19] propose the selective state space model, Mamba, which demonstrates superior performance over Transformers in NLP. Inspired by its remarkable performance, researchers extend it to the domain of computer vision. Zhu *et al.* [75] integrate SSM with bidirectional scanning, making each patch related to another. Liu *et al.* [34] extend the scanning in both horizontal and vertical directions to further interpret spatial relationships. However, they elongate the distance between adjacent tokens, overlooking the preservation of local 2D dependency. To this end, Huang *et al.* [22] introduce a local scanning strategy that divides images into distinct windows to capture local dependencies while maintaining a global perspective. Despite its effectiveness, it relies on the fixed-size window for local modeling, limiting the diversity of local dependencies. In contrast, our proposed LE-SSM adopts a four-way four-scale windowed scanning strategy, which enriches the multi-scale local information.

SSMs have been preliminarily employed and explored in a wide range of multi-modal tasks. Wan *et al.* [53] introduce a Siamese Mamba network for multi-modal semantic segmentation with a fusion module. Gao *et al.* [18] propose a multi-scale feature fusion Mamba for multi-source remote sensing image classification. However, they treat dual modalities equally during the cross-modality fusion, ignoring their complementarity and reliability. Conversely, our SSM-based AFM selectively interacts with the complementary cues and fuses the reliable messages in RGB and depth features.

3 METHODOLOGY

In this section, we first introduce some essential concepts of the state space model. Then, we provide a detailed description of our LEAF-Mamba, including its overall framework and module design. Figure 2 illustrates the overall architecture of LEAF-Mamba.

3.1 Preliminaries

State Space Models. SSM is a linear time-invariant system that maps an input sequence $x(t) \in \mathbb{R}^N$ to an output sequence $y(t) \in \mathbb{R}^N$. They are mathematically represented by the following linear ordinary differential equations:

$$\begin{aligned} h'(t) &= \boldsymbol{A}h(t) + \boldsymbol{B}x(t), \\ y(t) &= \boldsymbol{C}h(t) + \boldsymbol{D}x(t), \end{aligned} \tag{1}$$

where $h(t) \in \mathbb{R}^N$ indicates a hidden state, $h'(t) \in \mathbb{R}^N$ refers to the time derivative of $h(t)$, and N is the number of states. Additionally, $\boldsymbol{A} \in \mathbb{R}^{N\times N}$ is the state transition matrix, $\boldsymbol{B} \in \mathbb{R}^{N\times 1}$, $\boldsymbol{C} \in \mathbb{R}^{1\times N}$ are projection matrices, and $\boldsymbol{D} \in \mathbb{R}^{N\times 1}$ is a residual connection.

SSMs are continuous-time models, and are challenging to incorporate into deep learning networks. To address this, discrete versions of SSMs are proposed. The ordinary differential equations are discretized by the zero-order hold rule. A timescale parameter Δ is introduced to convert the continuous parameters $\boldsymbol{A}$ and $\boldsymbol{B}$ into discrete parameters $\overline{\boldsymbol{A}}$ and $\overline{\boldsymbol{B}}$, respectively, as:

$$\begin{aligned} \overline{\boldsymbol{A}} &= \exp(\Delta \boldsymbol{A}), \\ \overline{\boldsymbol{B}} &= (\Delta \boldsymbol{A})^{-1}(\exp(\Delta \boldsymbol{A}) - \boldsymbol{I}) \cdot \Delta \boldsymbol{B} \approx \Delta \boldsymbol{B}. \end{aligned} \tag{2}$$

The matrix $\overline{B}$ can be approximated by applying a first-order Taylor expansion to the term involving the matrix exponential. After discretization, the SSM system can be reformulated as:

$$\begin{aligned} h_t &= \overline{A}h_{t-1} + \overline{B}x_t, \\ y_t &= Ch_t + Dx_t. \end{aligned} \tag{3}$$

Further, the models compute output through a global convolution.

$$\begin{aligned} \overline{K} &= (C\overline{B}, C\overline{AB}, \ldots, C\overline{A}^{L-1}\overline{B}), \\ y &= x * \overline{K}, \end{aligned} \tag{4}$$

where L is the length of the input sequence x, and $\overline{K} \in \mathbb{R}^L$ is a structured convolutional kernel.

Selective Scan Mechanism. Traditional SSMs adopt a linear time-invariant framework, wherein the projection matrices remain fixed and unaffected by variations in the input sequence. However, this static configuration results in a lack of attention on individual elements within the sequence. To overcome this limitation, Mamba [19] introduces a selective scan mechanism where the parameter matrices become input-dependent. In this way, SSMs can better model the complex interactions present in long sequences through the transformation into linear time-varying systems.

3.2 Overview of LEAF-Mamba

The overall framework of our proposed LEAF-Mamba is shown in Figure 2, which follows a standard encoder-decoder architecture. The encoder is a dual-stream structure built upon VMamba [34], which yields multi-stage features from RGB and depth images. To enrich multi-scale local semantics, the last VMamba block of each stage is substituted with our local emphatic state space module (LE-SSM), resulting in local-enhanced dual-modality features $\{F_i^r\}_{i=1}^4$ and $\{F_i^d\}_{i=1}^4$. Then, they are fed into the adaptive fusion module (AFM) for attentive cross-modality interaction and integration. Specifically, AFM communicates the complementary cues in $\{F_i^r\}_{i=1}^4$ and $\{F_i^d\}_{i=1}^4$ for inter-enhanced features $\{F_i'^r\}_{i=1}^4$ and $\{F_i'^d\}_{i=1}^4$ which act as the inputs for next stage. Meanwhile, AFM fuses the reliable content of $\{F_i'^r\}_{i=1}^4$ and $\{F_i'^d\}_{i=1}^4$ for RGB-D features $\{F_i^f\}_{i=1}^4$. The multi-stage RGB-D features are put into an SSM-based FPN [31] decoder for multi-level predictions $\{P_i\}_{i=1}^4$ with deep supervision, where the CBAM [59] is adopted to facilitate the spatial- and channel-wise variation of the processing features. We take P_1 as the final prediction map.

3.3 Local Emphatic State Space Module

Early vision mamba methods [34, 65, 75] always flatten 2D plane into 1D array along rows and columns, disrupting the proximity of adjacent tokens. Although recent works [10, 22, 54] adopt the windowed scan strategy for local modeling, they fail to capture the multi-scale semantic cues due to the fixed window size. To solve this problem, we propose a local emphatic state space module (LE-SSM), which captures multi-view local dependencies via a multi-scale windowed 2D selective scan (MSW-SS2D) mechanism.

The illustration of LE-SSM is provided in Figure 3 (a). Unlike the Mamba [19] used in NLP, the LE-SSM consists of a single network branch with two residual modules, mimicking the architecture of Transformer block [12]. Meanwhile, the S6 block of Mamba

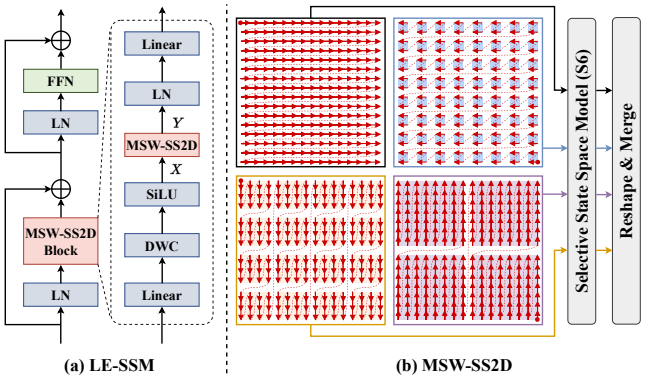

Figure 3: Illustration of the local emphatic state space module (LE-SSM) and multi-scale windowed 2D selective scan (MSW-SS2D) mechanism.

is substituted with the newly proposed multi-scale windowed 2D selective scan (MSW-SS2D), shown in Figure 3 (b). Specifically, given an input as X, MSW-SS2D first unfolds it into sequences along four distinct traversal paths, i.e., horizontal (H) and vertical (V) directions along with their flipped counterparts (HF and VF). Each of them adopts the windowed selective scan strategy with a unique window size. In this paper, the standard/flipped horizontal scan adopts the window size of 1/2 while the standard/flipped vertical scan adopts the window size of 4/8, denoted as H_1, HF_2, V_4 and VF_8. Notably, the matching between directions and window sizes can be set randomly because the rotation augmentation during the training phase can make them fully connected. Then each sequence is processed in parallel using a separate S6 block, and the resultant sequences are reshaped and merged to form the output Y. The whole procedure of MSW-SS2D can be formulated as:

$$Y = \sum\nolimits_{\mathrm{Scan} \in \mathcal{S}} \mathrm{Reshape}(\mathrm{S6}(\mathrm{Scan}(X))), \tag{5}$$

where $\mathcal{S} = \{\mathrm{H}_1, \mathrm{HF}_2, \mathrm{V}_4, \mathrm{VF}_8\}$, denoting the set of traversal paths.

In this case, MSW-SS2D achieves a four-way, four-scale windowed selective scan, which enables the LE-SSM to effectively extract the multi-scale local information without extra computational cost compared to SS2D [34].

3.4 Adaptive Fusion Module

Current SSM-based multi-modal methods [18, 53] typically treat the cross-modality features equally during the fusion process, which overlooks their complementarity and reliability. To this end, we design an adaptive fusion module (AFM) to dynamically interact with the complementary cues and fuse the reliable content in RGB and depth features under the paradigm of SSM.

The whole structure of AFM is illustrated in Figure 4 (a), consisting of two sequential processes: cross-modality interaction and cross-modality fusion. Both of them follow the architecture of VMamba block. For cross-modality interaction, the same-stage RGB and depth features F^r and F^d are first fed into a cross-modality second-order pooling (CSoP) layer to compute the modality-specific similarity maps S^r and S^d. Then they are reversed to produce the

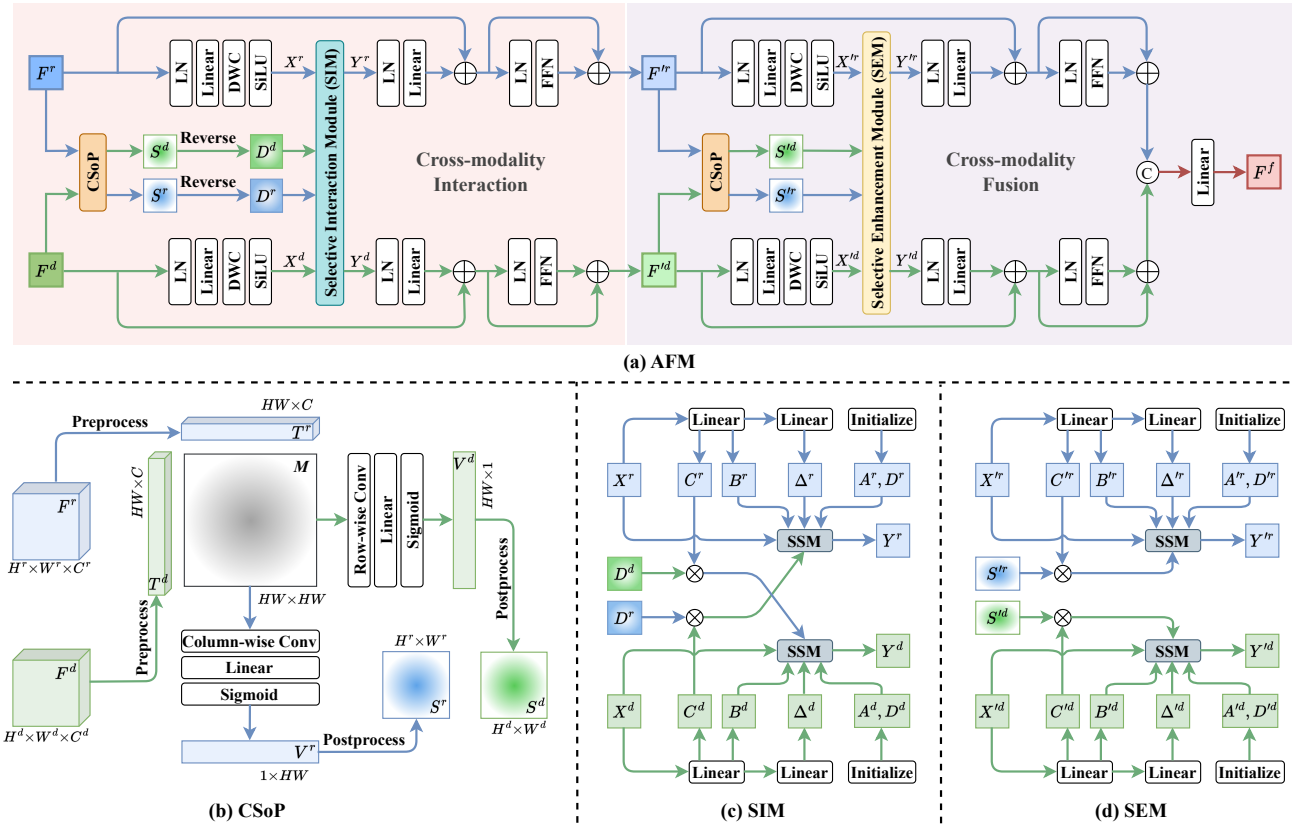

Figure 4: Illustration of the adaptive fusion module (AFM) and its components.

distance maps D^r and D^d, which are utilized to select the discriminative regions of the processing RGB and depth features X^r and X^d for complementary cross-modality interaction in the selective interaction module (SIM). For cross-modality fusion, the interacted dual-modality features F'^r and F'^d are also fed into a CSoP layer to measure their similarity S'^r and S'^d. Then, a selective enhancement module (SEM) is incorporated to weight the similar regions for subsequent reliable cross-modality fusion F^f. We now depict the technical details of CSoP, SIM and SEM, respectively.

Cross-modality Second-order Pooling. The structure of CSoP is illustrated in Figure 4 (b). Take the CSoP in cross-modality interaction as an example. We first preprocess the cross-modality features F^r and F^d with 1×1 convolution ($\mathrm{C}_{1\times1}$) and downsampling (DS) for reducing the number of their channels and spatial size to a fixed $H \times W \times C$, so as to decrease the computational cost of the following operations. Then we flatten them into tokens T^r and T^d along the spatial dimension and compute their covariance matrix M. The above process can be formulated as:

$$\begin{aligned} T^r &= \mathrm{Flatten}(\mathrm{DS}(\mathrm{C}_{1\times1}(F^r))), \\ T^d &= \mathrm{Flatten}(\mathrm{DS}(\mathrm{C}_{1\times1}(F^d))), \\ M_{i,j} &= \mathrm{Cov}(T_i^d, T_j^r). \end{aligned} \tag{6}$$

After that, we separately perform row-wise convolution ($\mathrm{C}^{\mathrm{row}}$) and column-wise convolution ($\mathrm{C}^{\mathrm{col}}$) on the covariance matrix M for modality-specific pooling. Each of them is followed by a linear layer and a sigmoid function (σ) for channel scaling and nonlinear activation respectively, outputting the weighted vectors V^r and V^d of length HW, formulated as:

$$\begin{aligned} V^r &= \sigma(\mathrm{Linear}(\mathrm{C}^{\mathrm{col}}(M))), \\ V^d &= \sigma(\mathrm{Linear}(\mathrm{C}^{\mathrm{row}}(M))). \end{aligned} \tag{7}$$

Finally, we postprocess the weighted vectors V^r and V^d by reshaping them into $H \times W$ matrices and upsampling (US) them to the original spatial size for modality-specific similarity S^r and S^d, written as:

$$\begin{aligned} S^r &= \mathrm{US}(\mathrm{Reshape}(V^r)), \\ S^d &= \mathrm{US}(\mathrm{Reshape}(V^d)). \end{aligned} \tag{8}$$

As such, our CSoP extends the traditional point-to-point similarity to a points-to-points manner, which further explores the relevance of cross-modality features from a global perspective.

Selective Interaction Module. The structure of SIM is illustrated in Figure 4 (c). As demonstrated in [11], Transformers are SSMs, with $\boldsymbol{Q}$, $\boldsymbol{K}$ and $\boldsymbol{V}$ corresponding to $\boldsymbol{C}$, $\boldsymbol{B}$ and the input x. Inspired by the cross-attention mechanism [3] that interchanges $\boldsymbol{Q}$ for information communication, we swap the $\boldsymbol{C}$ matrices of RGB and depth features in the process of selective scan for cross-modality interaction. Moreover, the distance maps D^r and D^d are derived by reversing S^r and S^d, namely $1 - S^r$ and $1 - S^d$. By such means, D^r

Table 1: Quantitative comparisons with state-of-the-arts on seven benchmark datasets. ↑ (↓) denotes higher the better (lower the better). The best and second-best results are shown in Red and Blue, respectively. – means not available.

Backbone		CNN							Transformer									SSM
Method		DMRA+ [24]	C2DF [70]	DCMF [55]	CIR [8]	DIF [67]	MFUR [17]	LAFB [54]	MITF [4]	PICR [9]	CAVER [41]	CAT [47]	TPCL [60]	HFMD [36]	EM-T [5]	VST++ [33]	DCT [38]	LEAF – (Ours)
Publish		TIP	TMM	TIP	TIP	TIP	KBS	TCSVT	TCSVT	MM	TIP	TMM	TMM	TIM	TNNLS	TPAMI	TIP	
Year		2022	2022	2022	2022	2023	2024	2024	2022	2023	2023	2023	2023	2024	2024	2024	2024	–
NJUD	F_β ↑	.882	.899	.915	.928	.906	**.937**	.919	.926	.931	.923	.929	.930	.923	.935	.927	.934	**.945**
	S_α ↑	.905	–	.913	.925	–	.920	–	.923	.927	.920	**.937**	.926	.927	.931	.926	.932	**.940**
	E_ξ ↑	.914	.919	.948	–	.923	–	.924	.957	–	.951	.933	.959	.956	**.961**	.957	.959	**.967**
	M ↓	.044	.038	.043	.035	.037	.035	.028	.030	.029	.031	**.025**	.028	.028	**.027**	.031	.031	**.025**
NLPR	F_β ↑	.880	.899	.906	.924	.906	.930	.905	.928	.928	.921	.916	.930	.913	**.934**	.922	.923	**.939**
	S_α ↑	.926	–	.922	.933	–	.931	–	.933	.935	.929	.939	.936	.931	**.940**	.934	.934	**.945**
	E_ξ ↑	.952	.958	.954	–	.960	–	.958	.968	–	.961	.968	**.970**	.964	**.970**	.966	.965	**.976**
	M ↓	.026	.021	.029	.023	.020	.022	.021	.018	.019	.022	.018	**.017**	.021	**.017**	.020	.023	**.016**
SIP	F_β ↑	.863	–	–	.896	.873	.910	.902	.913	–	.906	.918	**.922**	.896	.920	.917	.910	**.935**
	S_α ↑	.852	–	–	.888	–	.890	–	.899	–	.893	**.913**	.902	.886	.903	.903	.899	**.920**
	E_ξ ↑	.906	–	–	–	.915	–	.937	.940	–	.933	.944	**.946**	.930	.944	**.946**	.942	**.950**
	M ↓	.060	–	–	.052	.051	.049	.041	.040	–	.042	**.034**	.035	.044	.039	.038	.038	**.032**
STERE	F_β ↑	.875	.892	.906	.914	.894	.919	.896	.910	.920	.911	.902	.922	.901	**.926**	.911	.919	**.935**
	S_α ↑	.903	–	.910	917	–	.920	–	.909	.921	.914	**.925**	.920	.900	**.925**	.913	.922	**.933**
	E_ξ ↑	.920	.927	.946	–	.930	–	.930	.953	–	.949	.935	**.960**	.943	**.958**	.952	.955	**.958**
	M ↓	.043	.038	.043	.038	.036	.040	.037	.034	.031	.033	.030	.029	.040	**.028**	.035	.035	**.026**
SSD	F_β ↑	.824	.848	.867	–	–	–	.860	.862	–	.854	–	–	.871	.875	**.883**	–	**.904**
	S_α ↑	.868	–	.882	–	–	–	–	.877	–	.874	–	–	.887	.885	**.896**	–	**.918**
	E_ξ ↑	.911	.911	.921	–	–	–	.922	.914	–	.924	–	–	.934	.935	**.944**	–	**.953**
	M ↓	.049	.047	.053	–	–	–	.041	.047	–	.043	–	–	**.038**	.039	**.038**	–	**.033**
LFSD	F_β ↑	.861	.863	.875	.883	.875	**.891**	–	.876	.894	.886	.884	.888	.883	–	.887	–	**.908**
	S_α ↑	.871	–	.878	.875	–	.863	–	.874	.888	.882	**.894**	.892	.880	–	.888	–	**.907**
	E_ξ ↑	.902	.883	.909	–	.907	–	–	.911	–	.921	.908	**.926**	.915	–	.915	–	**.934**
	M ↓	.069	.065	.068	.068	.055	.075	–	.063	.053	.056	.051	**.049**	.059	–	.060	–	**.045**
DUT-D	F_β ↑	.911	.934	.932	.938	.940	.948	.930	.934	.951	.942	.951	**.956**	.951	–	.948	.952	**.958**
	S_α ↑	.919	–	.928	.932	–	.943	–	.937	.943	.931	**.953**	.946	.950	–	.943	.948	**.952**
	E_ξ ↑	.948	.958	.958	–	.958	–	.957	.960	–	.964	.971	**.974**	.971	–	.966	.969	**.973**
	M ↓	.035	.025	.035	.029	.025	.027	.027	.025	**.020**	.028	**.020**	**.020**	**.019**	–	.022	.023	**.019**
Params (M)↓		63.0	47.5	58.9	103.2	31.6	–	453.0	127.5	111.9	93.8	262.6	129.5	431.6	–	85.4	80.0	84.5
FLOPs (G)↓		126.3	44.1	–	42.6	24.9	–	139.7	24.1	27.0	63.9	341.8	212.0	242.2	–	40.0	49.0	18.1
FPS↑		22.0	33.4	–	54.0	–	–	45.0	–	21.3	26.0	11.0	23.8	9.0	–	12.0	–	70.2

and D^d can focus on the complementary information from the counterparts, and thus are utilized to weight $\boldsymbol{C^d}$ and $\boldsymbol{C^r}$ respectively. The whole process of SIM can be formulated as follows.

$$\begin{aligned} \overline{\boldsymbol{A^r}} &= \exp(\Delta^r \boldsymbol{A^r}),\ \overline{\boldsymbol{A^d}} = \exp(\Delta^d \boldsymbol{A^d}), \\ \overline{\boldsymbol{B^r}} &= \Delta^r \boldsymbol{B^r},\ \overline{\boldsymbol{B^d}} = \Delta^d \boldsymbol{B^d}, \\ h_t^r &= \overline{\boldsymbol{A^r}} h_{t-1}^r + \overline{\boldsymbol{B^r}} X_t^r,\ h_t^d = \overline{\boldsymbol{A^d}} h_{t-1}^d + \overline{\boldsymbol{B^d}} X_t^d, \\ Y_t^r &= (D^r \boldsymbol{C^d}) h_t^r + \boldsymbol{D^r} X_t^r,\ Y_t^d = (D^d \boldsymbol{C^r}) h_t^d + \boldsymbol{D^d} X_t^d, \end{aligned} \tag{9}$$

where $\boldsymbol{B}$, $\boldsymbol{C}$ and Δ are projected from the input X. $\boldsymbol{A}$ and $\boldsymbol{D}$ are randomly initialized. t denotes the time step.

Selective Enhancement Module. As shown in Figure 4 (d), the design philosophy of the SEM is very similar to that of the SIM. To be specific, the similarity maps S'^r and S'^d are separately utilized to weight $\boldsymbol{C'^r}$ and $\boldsymbol{C'^d}$, which enhances the reliability of the single-modality features for subsequent fusion. The process of SEM can be represented as:

$$\begin{aligned} \overline{\boldsymbol{A'^r}} &= \exp(\Delta'^r \boldsymbol{A'^r}),\ \overline{\boldsymbol{A'^d}} = \exp(\Delta'^d \boldsymbol{A'^d}), \\ \overline{\boldsymbol{B'^r}} &= \Delta'^r \boldsymbol{B'^r},\ \overline{\boldsymbol{B'^d}} = \Delta'^d \boldsymbol{B'^d}, \\ h_t'^r &= \overline{\boldsymbol{A'^r}} h_{t-1}'^r + \overline{\boldsymbol{B'^r}} X_t'^r,\ h_t'^d = \overline{\boldsymbol{A'^d}} h_{t-1}'^d + \overline{\boldsymbol{B'^d}} X_t'^d, \\ Y_t'^r &= (S'^r \boldsymbol{C'^r}) h_t'^r + \boldsymbol{D'^r} X_t'^r,\ Y_t'^d = (S'^d \boldsymbol{C'^d}) h_t'^d + \boldsymbol{D'^d} X_t'^d. \end{aligned} \tag{10}$$

Thanks to AFM, the overall model can dynamically build up the global relationship between cross-modality features, in turn achieving more comprehensive cross-modality interaction and fusion.

4 EXPERIMENTS

4.1 Experimental Setup

In the following, we state the datasets, evaluation metrics and training implementation. More details can be found in the Appendix.

Datasets. To evaluate the performance of our method, we conduct experiments on seven widely used RGB-D SOD datasets, including NJUD [25], NLPR [42], STERE [40], SIP [15], SSD [74], LFSD [30] and DUT-D [43]. For a fair comparison, we follow the same training settings as [35, 44], where 1485 samples from the NJUD, 700 samples from NLPR and 800 samples from DUT-D are selected as the training set. The rest of NJUD, NLPR and DUT-D, and all of STERE, SIP, SSD, LFSD are used for testing.

Evaluation Metrics. We adopt four commonly used metrics including F-measure (F_β) [1], S-measure (S_α) [13], E-measure (E_ξ) [14] and mean absolute error (M) to quantitatively evaluate the performance. For M, lower value is better. For others, higher is better.

Implementation Details. We adopt the VMamba-T [34] as encoder network, which is pre-trained on ImageNet-1K [27]. The network input resolution is 256×256, following [41, 67]. The size of

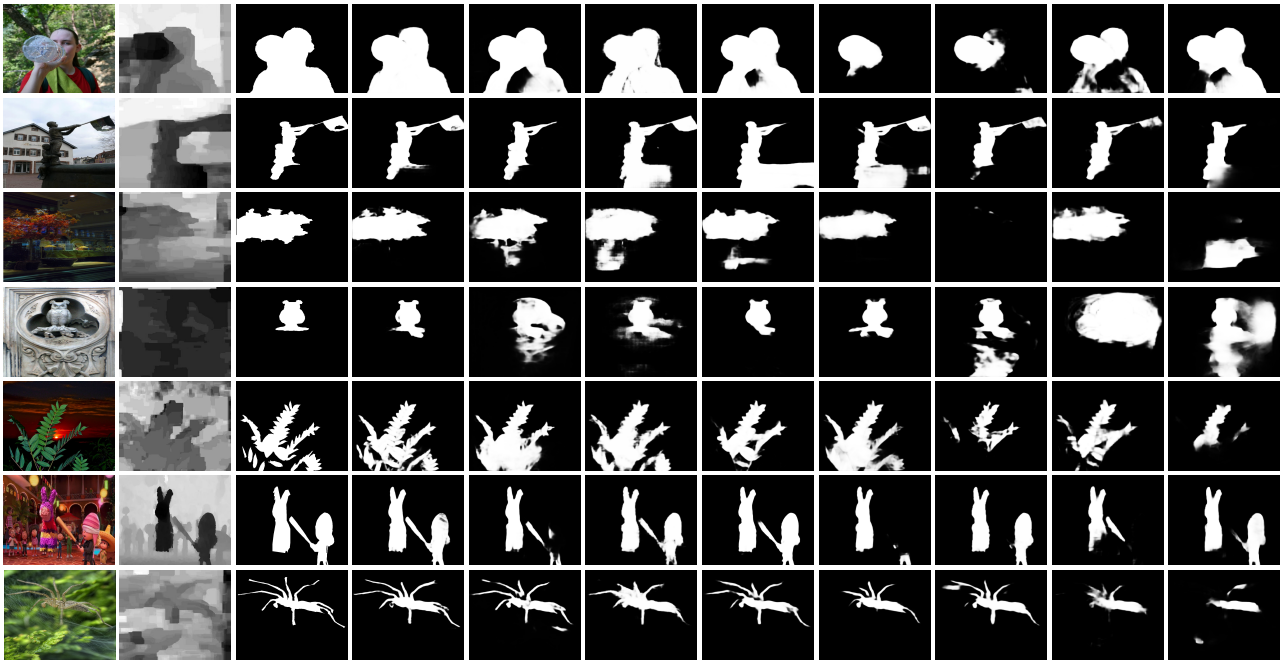

Figure 5: Qualitative results of our LEAF-Mamba and other representative methods.

$\{H, W, C\}$ in CSoP is fixed at $\{8, 8, 96\}$. We use pixel position-aware loss [58] for multi-level supervision to pay different attention to the hard and easy pixels. Adam [26] algorithm serves as our optimizer.

4.2 Comparison with State-of-the-arts

We compare our LEAF-Mamba with recent 16 state-of-the-art methods, in terms of quantitative and qualitative aspects.

Quantitative evaluation. Table 1 shows the quantitative results of our methods against other 16 state-of-the-art methods on seven benchmark datasets. It can be seen that our network outperforms other advanced models across all the datasets in terms of most evaluation metrics, which demonstrates the superiority of our method in efficacy. For instance, compared with the second best method VST++ [33], our method improves the F_β and MAE by 2.38% and 13.2% respectively on SSD dataset. Meanwhile, we compare our method with other state-of-the-art models in terms of parameters, FLOPs and FPS to evaluate the model size, computation overhead and inference speed, respectively. As shown in the bottom of Table 1, the LEAF-Mamba achieves the lowest FLOPs of 18.1G and the highest FPS of 70.2, with comparable parameters of 84.5M, validating its advances in computational efficiency.

Qualitative evaluation. In Figure 5, we visualize some challenging scenes and results generated by our method and other top-ranking models. As we can see, our model can accurately segment objects in varying scales including large object (Row 1), middle object (Row 2-3), small object (Row 4) and multiple objects (Row 6), which demonstrates the effectiveness of our LE-SSM in extracting multi-scale information. Meanwhile, for objects with complex backgrounds (Row 3 and 7) or low-quality depth maps (Row 4), our method is able to generate fine predictions that are more consistent with the ground truth, benefiting from the promising cross-modality interaction achieved by our proposed AFM.

Table 2: Ablation analyses of each component on the NJUD and SSD datasets. bold: top-1 results.

No.	Configuration	*NJUD*		*SSD*	
		$F_\beta \uparrow$	$M \downarrow$	$F_\beta \uparrow$	$M \downarrow$
#1	Baseline	.917	.030	.851	.044
#2	+ LE-SSM	.931	.028	.872	.039
#3	+ AFM	.940	.026	.891	.035
#4	+ LE-SSM + AFM (LEAF)	**.945**	**.025**	**.904**	**.033**

4.3 Ablation Studies

We conduct ablation studies to verify the effectiveness of two main components (LE-SSM and AFM) in the LEAF-Mamba on the NJUD and SSD datasets, and choose F_β and MAE for evaluation. The quantitative results are summarized in Table 2.

Components ablation. In this part, we firstly evaluate the effectiveness of the key components (namely, LE-SSM and AFM) in our LEAF-Mamba. Results are reported in Table 2. Our baseline (No. #1) employs a two-stream VMamba-based network where the dual-modality features with the same resolution are added directly. Detailed baseline architecture is presented in the Appendix. Compared with baseline, our LE-SSM/AFM obtains numerical improvements, e.g., 2.47%/4.7% and 11.4%/20.0% increase on SSD in terms of F_β and MAE, respectively. Furthermore, when concurrently adopting

Table 3: Quantitative results of various scanning strategies for local enhancement (LE). bold: top-1 results.

No.	Scanning Strategy	*NJUD*		*SSD*	
		$F_\beta \uparrow$	$M \downarrow$	$F_\beta \uparrow$	$M \downarrow$
#1	SS2D	.917	.030	.851	.044
#5	Continuous scan [65]	.919	.030	.856	.043
#6	Fixed windowed scan [22]	.925	.029	.863	.041
#2	MSW-SS2D (Ours)	**.931**	**.028**	**.872**	**.039**

Table 4: Ablation analyses of AFM. bold: top-1 results.

No.	CSoP	SIM	SEM	*NJUD*		*SSD*	
				$F_\beta \uparrow$	$M \downarrow$	$F_\beta \uparrow$	$M \downarrow$
#1				.917	.030	.851	.044
#7		✓		.924	.029	.863	.042
#8			✓	.921	.030	.858	.043
#9		✓	✓	.927	.028	.867	.041
#10	✓	✓		.933	.027	.879	.038
#11	✓		✓	.926	.028	.872	.040
#3	✓	✓	✓	**.940**	**.026**	**.891**	**.035**

LE-SSM and AFM, our LEAF-Mamba raises the gains to 6.23% and 25%, showing the synergy between the two modules.

Scanning strategy in LE-SSM. We assess the effectiveness of various scanning strategies in LE-SSM. Specifically, we compare our MSW-SS2D with SS2D [34], continuous scan [65] and fixed windowed scan [22]. The detailed illustration of them can be found in the Appendix. The results are presented in Table 3. It can be observed that, our MSW-SS2D strategy surpasses all counterparts with a clear margin, demonstrating the positive gains of comprehensively modeling local dependencies in a multi-scale manner.

Design of AFM. In the following, we evaluate the effect of core components in our AFM, i.e., CSoP, SIM and SEM. The results are reported in Table 4. In the absence of CSoP, independently applying the SIM (or SEM) achieves performance boost with 1.41% (0.82%) and 4.54% (2.27%) on the SSD dataset in terms of F_β and M metrics. Moreover, adopting them together yields impressive performance gains of 1.88% and 6.81%. These observations admit the effectiveness of SSM mechanism in dual-modality interaction and single-modality enhancement. Based on these findings, we conduct in-depth exploration on the role of CSoP-based selective mechanism for SIM and SEM, respectively (No. #10 & #11). All of these results typically make further improvements, benefiting from the adaptive weights from CSoP. Ultimately, our AFM (with CSoP, SIM and SEM) achieves the best performance. More detailed empirical studies on CSoP, SIM and SEM are provided in the Appendix.

4.4 Application to RGB-T SOD

To validate the generalization ability of LEAF-Mamba, we extend it to the RGB-Thermal (RGB-T) SOD task and conduct experiments on three public RGB-T SOD datasets, i.e., VT821 [56], VT1000 [51], and VT5000 [50]. Following [41, 49], the training set contains 2500 images from VT5000, with the remaining images used for testing.

Table 5: Quantitative comparisons with recent RGB-T SOD methods on three benchmarks. The best two results are shown in Red and Blue. – means not avaliable.

	Method	CAVER [24]	LSNet [73]	CMDBIF [64]	XMSNet [63]	LAFB [57]	ConTriNet [48]	LEAF (Ours)
	Publish	TIP	TIP	TCSVT	MM	TCSVT	TPAMI	–
	Year	2023	2023	2023	2023	2024	2024	–
VT821	$F_\beta \uparrow$	.877	.827	.855	.859	.843	**.878**	**.885**
	$S_\alpha \uparrow$	.898	.877	.882	.906	–	**.915**	**.926**
	$E_\xi \uparrow$	.928	.911	.927	.929	.915	**.940**	**.943**
	$M \downarrow$	.027	.033	.032	.028	.043	**.022**	**.020**
VT1000	$F_\beta \uparrow$	**.939**	.887	.914	.903	.905	.918	**.926**
	$S_\alpha \uparrow$	.938	.924	.927	.936	–	**.941**	**.945**
	$E_\xi \uparrow$	.949	.936	**.967**	.945	.945	.954	**.962**
	$M \downarrow$	**.017**	.022	.019	.018	.018	**.015**	**.015**
VT5000	$F_\beta \uparrow$	.882	.827	.869	.871	.857	**.898**	**.893**
	$S_\alpha \uparrow$	.899	.876	.886	.907	–	**.923**	**.919**
	$E_\xi \uparrow$	.941	.916	.937	.939	.931	**.956**	**.958**
	$M \downarrow$	.028	.036	.032	.028	.030	**.020**	**.021**

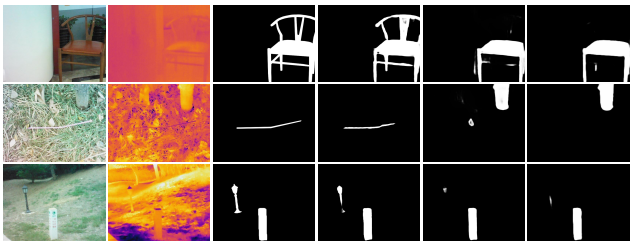

Figure 6: Visual comparisons on RGB-T SOD.

We compare our network with 6 recent RGB-T SOD methods and show the quantitative results in Table 5. It can be seen that our model achieves overall competitive performance on the three benchmarks, e.g., with an improvement of 1.20% and 9.09% on VT821 in term of S_α and MAE. Additionally, the visual comparisons shown in Figure 6 indicate that LEAF-Mamba excels in challenging scenarios such as low-quality thermal images, complex backgrounds and multi-scale objects, further demonstrating its effectiveness and prominent generalizability on multi-modality tasks.

5 CONCLUSION

In this paper, we propose a novel SSM-based system (LEAF-Mamba) to address both efficacy and efficiency bottlenecks in existing RGB-D salient object detection. We develop a local emphatic state space module equipped with the multi-scale windowed scanning strategy to capture multi-scale local dependencies in the process of feature extraction. We also design an SSM-based adaptive fusion module to dynamically select the discriminative regions of two modalities for complementary cross-modality interaction, as well as the similar ones for reliable cross-modality fusion. Experimental results on seven benchmarks show superior performance of our method over 16 state-of-the-art models in both accuracy and efficiency. Furthermore, extensive ablation studies demonstrate the effectiveness of each proposed component.

REFERENCES

[1] Radhakrishna Achanta, Sheila Hemami, Francisco Estrada, and Sabine Susstrunk. 2009. Frequency-tuned salient region detection. In *CVPR*. 1597–1604.

[2] Chenglizhao Chen, Jipeng Wei, Chong Peng, Weizhong Zhang, and Hong Qin. 2020. Improved saliency detection in RGB-D images using two-phase depth estimation and selective deep fusion. *IEEE Transactions on Image Processing* 29 (2020), 4296–4307.

[3] Chun-Fu Richard Chen, Quanfu Fan, and Rameswar Panda. 2021. CrossViT: Cross-attention multi-scale vision transformer for image classification. In *ICCV*. 357–366.

[4] Gang Chen, Feng Shao, Xiongli Chai, Hangwei Chen, Qiuping Jiang, Xiangchao Meng, and Yo-Sung Ho. 2023. Modality-Induced Transfer-Fusion Network for RGB-D and RGB-T Salient Object Detection. *IEEE Transactions on Circuits and Systems for Video Technology* 33, 4 (2023), 1787–1801. https://doi.org/10.1109/TCSVT.2022.3215979

[5] Geng Chen, Qingyue Wang, Bo Dong, Ruitao Ma, Nian Liu, Huazhu Fu, and Yong Xia. 2024. EM-Trans: Edge-aware multimodal transformer for RGB-D salient object detection. *IEEE Transactions on Neural Networks and Learning Systems* 36, 2 (2024), 3175–3188.

[6] Hao Chen, Feihong Shen, Ding Ding, Yongjian Deng, and Chao Li. 2024. Disentangled cross-modal transformer for RGB-D salient object detection and beyond. *IEEE Transactions on Image Processing* (2024).

[7] Shuhan Chen and Yun Fu. 2020. Progressively guided alternate refinement network for RGB-D salient object detection. In *ECCV*. Springer, 520–538.

[8] Runmin Cong, Qinwei Lin, Chen Zhang, Chongyi Li, Xiaochun Cao, Qingming Huang, and Yao Zhao. 2022. CIR-Net: Cross-modality interaction and refinement for RGB-D salient object detection. *IEEE Transactions on Image Processing* 31 (2022), 6800–6815.

[9] Runmin Cong, Hongyu Liu, Chen Zhang, Wei Zhang, Feng Zheng, Ran Song, and Sam Kwong. 2023. Point-aware interaction and cnn-induced refinement network for RGB-D salient object detection. In *ACM MM*. 406–416.

[10] Trung Dinh Quoc Dang, Huy Hoang Nguyen, and Aleksei Tiulpin. 2024. LoG-VMamba: Local-Global Vision Mamba for Medical Image Segmentation. In *ACCV*. 548–565.

[11] Tri Dao and Albert Gu. 2024. Transformers are SSMs: generalized models and efficient algorithms through structured state space duality. In *ICML*. 10041–10071.

[12] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. 2021. An image is worth 16x16 words: Transformers for image recognition at scale. *ICLR* (2021).

[13] Deng-Ping Fan, Ming-Ming Cheng, Yun Liu, Tao Li, and Ali Borji. 2017. Structure-measure: A new way to evaluate foreground maps. In *ICCV*. 4548–4557.

[14] Deng-Ping Fan, Cheng Gong, Yang Cao, Bo Ren, Ming-Ming Cheng, and Ali Borji. 2018. Enhanced-alignment measure for binary foreground map evaluation. *IJCAI* (2018), 698–704.

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

[16] Deng-Ping Fan, Yingjie Zhai, Ali Borji, Jufeng Yang, and Ling Shao. 2020. BBS-Net: RGB-D salient object detection with a bifurcated backbone strategy network. In *ECCV*. Springer, 275–292.

[17] Zhengqian Feng, Wei Wang, Wang Li, Gang Li, Min Li, and Mingle Zhou. 2024. MFUR-Net: Multimodal feature fusion and unimodal feature refinement for RGB-D salient object detection. *Knowledge-Based Systems* 299 (2024), 112022.

[18] Feng Gao, Xuepeng Jin, Xiaowei Zhou, Junyu Dong, and Qian Du. 2025. MSF-Mamba: Multi-Scale Feature Fusion State Space Model for Multi-Source Remote Sensing Image Classification. *IEEE Transactions on Geoscience and Remote Sensing* (2025).

[19] Albert Gu and Tri Dao. 2023. Mamba: Linear-time sequence modeling with selective state spaces. *COLM* (2023).

[20] Albert Gu, Karan Goel, and Christopher Ré. 2021. Efficiently modeling long sequences with structured state spaces. *ICLR* (2021).

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

[22] Tao Huang, Xiaohuan Pei, Shan You, Fei Wang, Chen Qian, and Chang Xu. 2024. LocalMamba: Visual state space model with windowed selective scan. *arXiv preprint arXiv:2403.09338* (2024).

[23] Wei Ji, Jingjing Li, Shuang Yu, Miao Zhang, Yongri Piao, Shunyu Yao, Qi Bi, Kai Ma, Yefeng Zheng, Huchuan Lu, et al. 2021. Calibrated RGB-D salient object detection. In *CVPR*. 9471–9481.

[24] Wei Ji, Ge Yan, Jingjing Li, Yongri Piao, Shunyu Yao, Miao Zhang, Li Cheng, and Huchuan Lu. 2022. DMRA: Depth-induced multi-scale recurrent attention network for RGB-D saliency detection. *IEEE Transactions on Image Processing* 31 (2022), 2321–2336.

[25] Ran Ju, Ling Ge, Wenjing Geng, Tongwei Ren, and Gangshan Wu. 2014. Depth saliency based on anisotropic center-surround difference. In *ICIP*. IEEE, 1115–1119.

[26] Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014).

[27] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. 2012. Imagenet classification with deep convolutional neural networks. *NeurIPS* 25 (2012).

[28] Gongyang Li, Zhi Liu, Linwei Ye, Yang Wang, and Haibin Ling. 2020. Cross-modal weighting network for RGB-D salient object detection. In *ECCV*. Springer, 665–681.

[29] Kunchang Li, Xinhao Li, Yi Wang, Yinan He, Yali Wang, Limin Wang, and Yu Qiao. 2024. VideoMamba: State space model for efficient video understanding. In *ECCV*. Springer, 237–255.

[30] Nianyi Li, Jinwei Ye, Yu Ji, Haibin Ling, and Jingyi Yu. 2014. Saliency detection on light field. In *CVPR*. 2806–2813.

[31] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. 2017. Feature pyramid networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2117–2125.

[32] Jiang-Jiang Liu, Qibin Hou, Zhi-Ang Liu, and Ming-Ming Cheng. 2022. PoolNet+: Exploring the potential of pooling for salient object detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 45, 1 (2022), 887–904.

[33] Nian Liu, Ziyang Luo, Ni Zhang, and Junwei Han. 2024. VST++: Efficient and stronger visual saliency transformer. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2024).

[34] Yue Liu, Yunjie Tian, Yuzhong Zhao, Hongtian Yu, Lingxi Xie, Yaowei Wang, Qixiang Ye, Jianbin Jiao, and Yunfan Liu. 2024. VMamba: Visual state space model. *NeurIPS* 37 (2024), 103031–103063.

[35] Zhengyi Liu, Yuan Wang, Zhengzheng Tu, Yun Xiao, and Bin Tang. 2021. TriTransNet: RGB-D salient object detection with a triplet transformer embedding network. In *ACM MM*. 4481–4490.

[36] Yi Luo, Feng Shao, Zhengxuan Xie, Huizhi Wang, Hangwei Chen, Baoyang Mu, and Qiuping Jiang. 2024. HFMDNet: Hierarchical fusion and multi-level decoder network for RGB-D salient object detection. *IEEE Transactions on Instrumentation and Measurement* (2024).

[37] Vijay Mahadevan and Nuno Vasconcelos. 2009. Saliency-based discriminant tracking. In *CVPR*. IEEE, 1007–1013.

[38] Ao Mou, Yukang Lu, Jiahao He, Dingyao Min, Keren Fu, and Qijun Zhao. 2024. Salient object detection in RGB-D videos. *IEEE Transactions on Image Processing* (2024).

[39] Ali Nasiri-Sarvi, Vincent Quoc-Huy Trinh, Hassan Rivaz, and Mahdi S Hosseini. 2024. Vim4Path: Self-supervised vision mamba for histopathology images. In *CVPR*. 6894–6903.

[40] Yuzhen Niu, Yujie Geng, Xueqing Li, and Feng Liu. 2012. Leveraging stereopsis for saliency analysis. In *CVPR*. IEEE, 454–461.

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

[42] Houwen Peng, Bing Li, Weihua Xiong, Weiming Hu, and Rongrong Ji. 2014. RGBD salient object detection: A benchmark and algorithms. In *ECCV*. Springer, 92–109.

[43] Yongri Piao, Wei Ji, Jingjing Li, Miao Zhang, and Huchuan Lu. 2019. Depth-induced multi-scale recurrent attention network for saliency detection. In *CVPR*. 7254–7263.

[44] Yongri Piao, Zhengkun Rong, Miao Zhang, Weisong Ren, and Huchuan Lu. 2020. A2Dele: Adaptive and attentive depth distiller for efficient RGB-D salient object detection. In *CVPR*. 9060–9069.

[45] Xuebin Qin, Zichen Zhang, Chenyang Huang, Chao Gao, Masood Dehghan, and Martin Jagersand. 2019. BasNet: Boundary-aware salient object detection. In *CVPR*. 7479–7489.

[46] Jimmy TH Smith, Andrew Warrington, and Scott W Linderman. 2023. Simplified state space layers for sequence modeling. *ICLR* (2023).

[47] Fuming Sun, Peng Ren, Bowen Yin, Fasheng Wang, and Haojie Li. 2023. CATNet: A cascaded and aggregated transformer network for RGB-D salient object detection. *IEEE Transactions on Multimedia* 26 (2023), 2249–2262.

[48] Hao Tang, Zechao Li, Dong Zhang, Shengfeng He, and Jinhui Tang. 2024. Divide-and-Conquer: Confluent Triple-Flow Network for RGB-T Salient Object Detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2024).

[49] Zhengzheng Tu, Zhun Li, Chenglong Li, Yang Lang, and Jin Tang. 2021. Multi-interactive dual-decoder for RGB-thermal salient object detection. *IEEE Transactions on Image Processing* 30 (2021), 5678–5691.

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

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

[52] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all

you need. *NeurIPS* 30 (2017).

[53] Zifu Wan, Pingping Zhang, Yuhao Wang, Silong Yong, Simon Stepputtis, Katia Sycara, and Yaqi Xie. 2024. Sigma: Siamese Mamba network for multi-modal semantic segmentation. In *WACV*.

[54] Chuanzhi Wang, Jun Huang, Mingyun Lv, Huafei Du, Yongmei Wu, and Ruiru Qin. 2024. A local enhanced mamba network for hyperspectral image classification. *International Journal of Applied Earth Observation and Geoinformation* 133 (2024), 104092.

[55] Fengyun Wang, Jinshan Pan, Shoukun Xu, and Jinhui Tang. 2022. Learning Discriminative Cross-Modality Features for RGB-D Saliency Detection. *IEEE Transactions on Image Processing* 31 (2022), 1285–1297.

[56] Guizhao Wang, Chenglong Li, Yunpeng Ma, Aihua Zheng, Jin Tang, and Bin Luo. 2018. RGB-T saliency detection benchmark: Dataset, baselines, analysis and a novel approach. In *Image and graphics technologies and applications: 13th conference on image and graphics technologies and applications, IGTA 2018, Beijing, China, April 8–10, 2018, revised selected papers 13*. Springer, 359–369.

[57] Kunpeng Wang, Zhengzheng Tu, Chenglong Li, Cheng Zhang, and Bin Luo. 2024. Learning adaptive fusion bank for multi-modal salient object detection. *IEEE Transactions on Circuits and Systems for Video Technology* (2024).

[58] Jun Wei, Shuhui Wang, and Qingming Huang. 2020. F^3Net: fusion, feedback and focus for salient object detection. In *AAAI*, Vol. 34. 12321–12328.

[59] Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon. 2018. CBAM: Convolutional block attention module. In *ECCV*. 3–19.

[60] Jiesheng Wu, Fangwei Hao, Weiyun Liang, and Jing Xu. 2023. Transformer fusion and pixel-level contrastive learning for RGB-D salient object detection. *IEEE Transactions on Multimedia* 26 (2023), 1011–1026.

[61] Yu-Huan Wu, Yun Liu, Jun Xu, Jia-Wang Bian, Yu-Chao Gu, and Ming-Ming Cheng. 2022. MobileSal: Extremely efficient RGB-D salient object detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 44, 12 (2022), 10261–10269.

[62] Zhe Wu, Li Su, and Qingming Huang. 2019. Cascaded partial decoder for fast and accurate salient object detection. In *CVPR*. 3907–3916.

[63] Zongwei Wu, Jingjing Wang, Zhuyun Zhou, Zhaochong An, Qiuping Jiang, Cédric Demonceaux, Guolei Sun, and Radu Timofte. 2023. Object segmentation by mining cross-modal semantics. In *ACM MM*. 3455–3464.

[64] Zhengxuan Xie, Feng Shao, Gang Chen, Hangwei Chen, Qiuping Jiang, Xiangchao Meng, and Yo-Sung Ho. 2023. Cross-modality double bidirectional interaction and fusion network for RGB-T salient object detection. *IEEE Transactions on Circuits and Systems for Video Technology* 33, 8 (2023), 4149–4163.

[65] Chenhongyi Yang, Zehui Chen, Miguel Espinosa, Linus Ericsson, Zhenyu Wang, Jiaming Liu, and Elliot J Crowley. 2024. PlainMamba: Improving non-hierarchical mamba in visual recognition. *arXiv preprint arXiv:2403.17695* (2024).

[66] Shu Yang, Yihui Wang, and Hao Chen. 2024. Mambamil: Enhancing long sequence modeling with sequence reordering in computational pathology. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 296–306.

[67] Shunyu Yao, Miao Zhang, Yongri Piao, Chaoyi Qiu, and Huchuan Lu. 2023. Depth injection framework for RGBD salient object detection. *IEEE Transactions on Image Processing* 32 (2023), 5340–5352.

[68] Miao Zhang, Sun Xiao Fei, Jie Liu, Shuang Xu, Yongri Piao, and Huchuan Lu. 2020. Asymmetric two-stream architecture for accurate RGB-D saliency detection. In *ECCV*. Springer, 374–390.

[69] Miao Zhang, Weisong Ren, Yongri Piao, Zhengkun Rong, and Huchuan Lu. 2020. Select, supplement and focus for RGB-D saliency detection. In *CVPR*. 3472–3481.

[70] Miao Zhang, Shunyu Yao, Beiqi Hu, Yongri Piao, and Wei Ji. 2022. C^2DFNet: Criss-cross dynamic filter network for RGB-D salient object detection. *IEEE Transactions on Multimedia* 25 (2022), 5142–5154.

[71] Wenbo Zhang, Ge-Peng Ji, Zhuo Wang, Keren Fu, and Qijun Zhao. 2021. Depth quality-inspired feature manipulation for efficient RGB-D salient object detection. In *ACM MM*. 731–740.

[72] Jia-Xing Zhao, Yang Cao, Deng-Ping Fan, Ming-Ming Cheng, Xuan-Yi Li, and Le Zhang. 2019. Contrast prior and fluid pyramid integration for RGBD salient object detection. In *CVPR*. 3927–3936.

[73] Wujie Zhou, Yun Zhu, Jingsheng Lei, Rongwang Yang, and Lu Yu. 2023. LSNet: Lightweight spatial boosting network for detecting salient objects in RGB-thermal images. *IEEE Transactions on Image Processing* 32 (2023), 1329–1340.

[74] Chunbiao Zhu and Ge Li. 2017. A three-pathway psychobiological framework of salient object detection using stereoscopic technology. In *ICCV workshop*. 3008–3014.

[75] Lianghui Zhu, Bencheng Liao, Qian Zhang, Xinlong Wang, Wenyu Liu, and Xinggang Wang. 2024. Vision Mamba: Efficient Visual Representation Learning with Bidirectional State Space Model. In *ICML*. PMLR, 62429–62442.

Appendix

This Appendix provides additional details and results that complement the main manuscript, which are omitted due to page limitations. The contents are organized as follows:

- Dataset Specification in § A;
- Metrics Description in § B;
- Experimental Implementation in § C;
- Architecture of Baseline in § D;
- Scanning Strategy in § E;
- Additional Experiments in § F;
- Feature Visualization in § G;
- Failure Cases in § H.

A DATASETS

We conduct extensively experiments on ten multi-modality salient object detection (SOD) datasets, including seven RGB-Depth (RGB-D) benchmarks and three RGB-Thermal (RGB-T) ones. The prior datastes includes NJUD [25], NLPR [42], STERE [40], SIP [15], SSD [74], LFSD [30], and DUT-D [43]; and the latter ones involve VT821 [56], VT1000 [51] and VT5000 [50]. The characteristics of each dataset are summarized below.

RGB-D Datasets. NJUD [25] contains 1,985 pairs of RGB and depth images collected from the Internet, 3D movies, and stereo photographs. Dataset NLPR [42] includes 1,000 RGB-D pairs covering a variety of indoor and outdoor scenes. STERE [40] consists of 1,000 stereoscopic images sourced from Flickr, NVIDIA 3D Vision Live, and the Stereoscopic Image Gallery. SIP [15] is a high-resolution dataset comprising 929 image pairs captured in outdoor environments with complex lighting and diverse human poses. SSD [74] includes 80 samples from both indoor and outdoor scenes. LFSD [30] provides 100 RGB-D image pairs designed for saliency detection in light field images. DUT-D [43] consists of 1,200 RGB-D pairs, with 800 captured indoors and 400 captured outdoors. We follow the training protocol as in previous works [43] [44] and [35]. Specifically, 1,485 samples from NJUD, 700 samples from NLPR, and 800 samples from DUT-D are used for training. The remaining samples from NJUD, NLPR, and DUT-D, along with all samples from STERE, SIP, SSD, and LFSD, are used for testing.

RGB-T Datasets. VT821 [56] contains 821 manually aligned RGB-T image pairs; VT1000 [51], comprising 1,000 image pairs captured in relatively simple scenes with well-aligned sensors; and VT5000 [50], which includes 5,000 high-resolution and diverse image pairs with minimal misalignment. Following the training protocol of [41, 49], we use 2,500 image pairs from VT5000 for training, while the remaining samples from VT5000, along with all images from VT821 and VT1000, are used for evaluation.

B PERFORMANCE METRIC

We employ four popular metrics to assess the performance, following [5, 47]. F-measure (F_β) [1] is a region-based similarity metric based on precision and recall. S-measure (S_α) [13] focuses on region-aware and object-aware structural similarities between the saliency map and the ground truth. E-measure (E_ξ) [14] is characterized as both image-level statistics and local pixel matching. Mean absolute error (MAE, M) measures the average difference between the prediction and the ground truth in the pixel level. The lower value is better for M and the higher is better for others.

C TRAINING IMPLEMENTATION

Our proposed LEAF-Mamba is implemented using the PyTorch toolkit and trained on a PC equipped with a single NVIDIA RTX 4090 GPU. VMamba-T [34] is employed as the encoder network, initialized with weights pre-trained on ImageNet-1K [27]. The size of $\{H, W, C\}$ in the CSoP module is fixed at $\{8, 8, 96\}$. Following [41, 67], all images are uniformly resized to 256×256 during both training and inference. During training phase, random flipping and random rotation are employed for data augmentation to alleviate overfitting. The Adam algorithm [26] serves as the optimizer with a mini-batch size of 8. The initial learning rate is set to 1e-4 and decayed by a factor of 10 every 60 epochs. The training process runs for a total of 200 epochs.

D ARCHITECTURE OF BASELINE

The detailed architecture of baseline is illustrated in Figure 7. We adopt the two-stream VMamba-T as encoder which is initialized by the parameters pre-trained on ImageNet-1K. The RGB and depth features with the same resolution are added directly for cross-modality fusion to generate RGB-D features. The multi-stage RGB-D features are delivered into an SSM-based FPN-like decoder for prediction. Each decoder block contains one upsample layer, one VMamba block and one CBAM.

E SCANNING STRATEGY

We illustrate various scanning strategy counterparts in Figure 8. As shown in figure, SS2D [34] performs the row-wise and column-wise scanning along the same direction. Different from SS2D, continuous scan [65] reverses the scanning direction in adjacent lines, leading to a Zigzag scanning path, which partially reserves the local proximity. In addition, fixed windowed scan [22] splits the feature map into several local regions and performs SS2D-like scanning in each window. In contrast to these previous works, our MSW-SS2D concurrently considers multi-scale semantics in once four-way scanning. Therefore, our MSW-SS2D is potential to capture multi-scale local dependencies. Equipped with the MSW-SS2D, our LEAF-Mamba achieves superior performance on both the quantitative and qualitative experiments in Table 3 and Figure 5 of the manuscript, demonstrating its advance in detecting multi-scale salient objects.

F ADDITIONAL EXPERIMENTS

In this section, we elaborately validate the effects of some components in our method, including CSoP, SIM and SEM. The experiments are conducted on NJUD and SSD datasets and F_β and M are chosen for evaluation. The identical configuration are denoted in the same color. The best result is highlighted in **bold**.

F.1 Ablation Study on CSoP

Similarity. Initially, we evaluate the devise of CSoP, including the similarity function and pooling approach. Results are presented

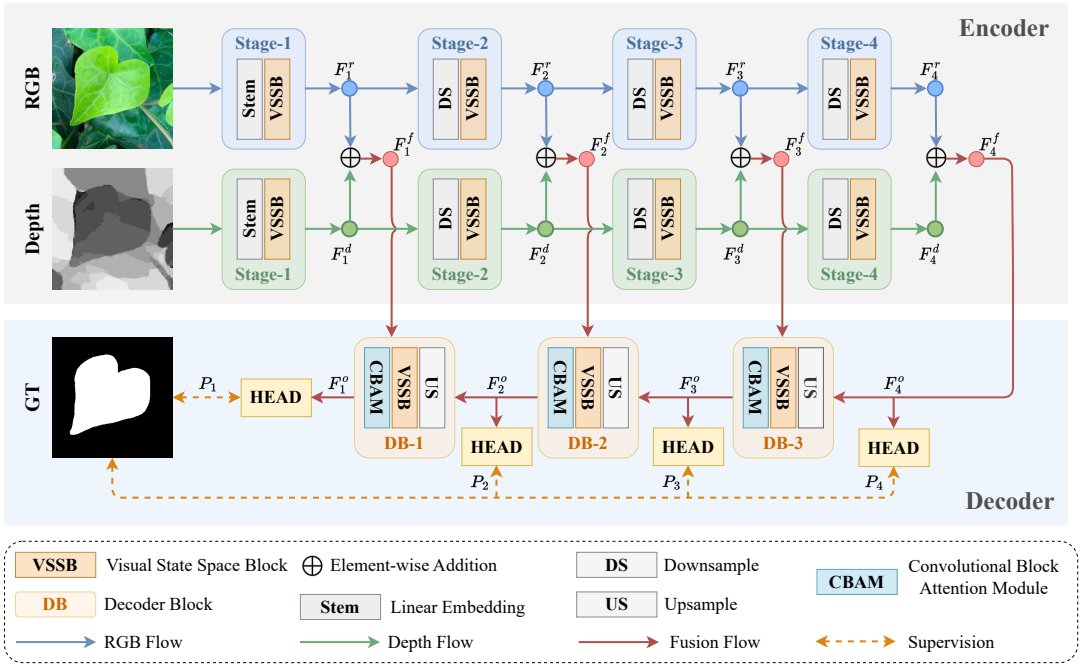

Figure 7: Architecture of the baseline.

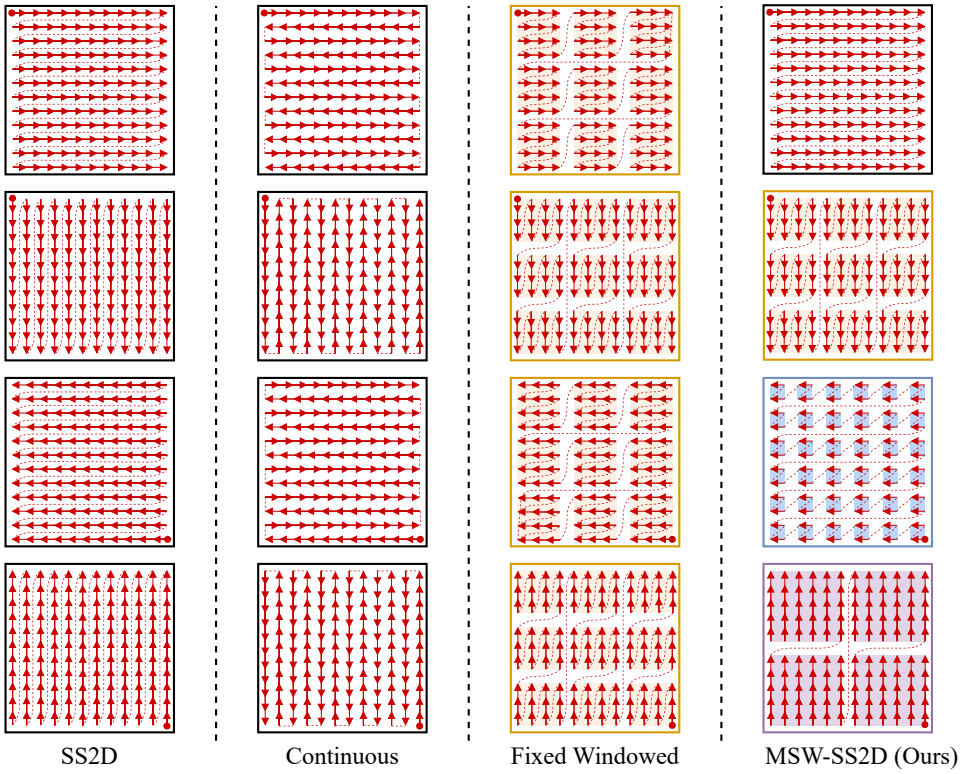

Figure 8: Various scanning strategy counterparts.

Table 6: Ablation analyses of CSoP module on the NJUD and SSD datasets. bold: top-1 results. Numbers shown in gray indicates to those referenced in the manuscript.

No.	Similarity	*NJUD*		*SSD*	
		$F_\beta \uparrow$	$M \downarrow$	$F_\beta \uparrow$	$M \downarrow$
#A1 (#1)	–	.917	.030	.851	.044
#A2	Cosine	.932	.027	.877	.038
#A3 (#3)	Covariance (AFM)	**.940**	**.026**	**.891**	**.035**
No.	Pooling	*NJUD*		*SSD*	
		$F_\beta \uparrow$	$M \downarrow$	$F_\beta \uparrow$	$M \downarrow$
#A1 (#1)	–	.917	.030	.851	.044
#A4	AvgPool	.933	.027	.880	.038
#A5	MaxPool	.935	.027	.883	.037
#A3 (#3)	ConvPool (AFM)	**.940**	**.026**	**.891**	**.035**

in Table 6. In the first pannel, we compare two similarity implementation: a point-to-point cosine similarity (No. #A2) and a global cross-covariance matrix (No. #A3, indicating Eqn. (7) in manuscript). To be specific, for the RGB token and depth token $T^r \in \mathbb{R}^{HW\times C}$ and $T^d \in \mathbb{R}^{HW\times C}$, the cosine similarity between T_i^r and T_i^d can be written as:

$$M_i^{Cos} = \mathcal{F}\left(T_i^d, T_i^r\right) = \frac{\left\langle T_i^d, T_i^r \right\rangle}{\|T_i^d\| \cdot \|T_i^r\|}, . \tag{11}$$

where $\mathcal{F}(\cdot)$ denotes the cosine similarity function, notation $\langle\cdot,\cdot\rangle$ and $\|\cdot\|$ indicate inner production and ℓ_2 normalization operation respectively. And the cross-covariance similarity for T_i^r and T_j^d can be formulated as:

$$M_{i,j}^{Cov} = \mathcal{H}\left(T_i^d, T_j^r\right) = \frac{1}{C-1}\left\langle T_i^d - \bar{T}_i^d, T_j^r - \bar{T}_j^r \right\rangle \tag{12}$$

$$\bar{T}_i^d = \frac{1}{C}\sum_{c=1}^{C} T_{i,c}^d,\ \bar{T}_j^r = \frac{1}{C}\sum_{c=1}^{C} T_{j,c}^r, . \tag{13}$$

where notation $\mathcal{H}(\cdot)$ denotes the cross-covariance similarity function. From the results, we can observe that, considering similarity in CSoP obtains performance gains over baseline (No. #A1). In addition, covariance similarity is superior to the cosine one with an advance of 1.59% and 7.89% on F_β and M metric of SSD dataset. Thus, covariance similarity is used as the default setting in the CSoP.

Pooling. In the second pannel of Table 6, we assess various pooling functions for the covariance similarity matrix M^{Cov}. Among all pooling implementations, the convolutional pooling achieves the best performance. It leads average/max pooling over 1.25%/0.91% and 7.89%/0.54% on F_β and M. Due to its superiority, convolution pooling serves as the default configuration for CSoP.

F.2 Ablation Study on SIM

In this part, we evaluate the devise of SIM. To be specific, we compare the various swap objects in cross-modality interaction operation, namely X, $\boldsymbol{B}$ and $\boldsymbol{C}$. As shown the results in Table 7, considering the cross-modality interaction (No. #A6&A7&A8) obtains novel performance gains over baseline (No. #A1). Moreover, among all

Table 7: Ablation analyses of swap feature in SIM on the NJUD and SSD datasets. bold: top-1 results. Numbers shown in gray indicates to those referenced in the manuscript.

No.	Cross-modality Interaction	Swap	*NJUD*		*SSD*	
			$F_\beta \uparrow$	$M \downarrow$	$F_\beta \uparrow$	$M \downarrow$
#A1 (#1)	×	–	.917	.030	.851	.044
#A6	✓	X	.919	.030	.855	.044
#A7	✓	$\boldsymbol{B}$	.921	**.029**	.857	.043
#A8 (#7)	✓	$\boldsymbol{C}$	**.924**	**.029**	**.863**	**.042**

Table 8: Ablation analyses of weighting object in SEM on the NJUD and SSD datasets. bold: top-1 results. Numbers shown in gray indicates to those referenced in the manuscript.

No.	Weighting	*NJUD*		*SSD*	
		$F_\beta \uparrow$	$M \downarrow$	$F_\beta \uparrow$	$M \downarrow$
#A1 (#1)	–	.917	.030	.851	.044
#A9	X	.925	**.028**	.864	.042
#A10	$\boldsymbol{B}$	.923	.029	.867	.041
#A11 (#11)	$\boldsymbol{C}$	**.926**	**.028**	**.872**	**.040**

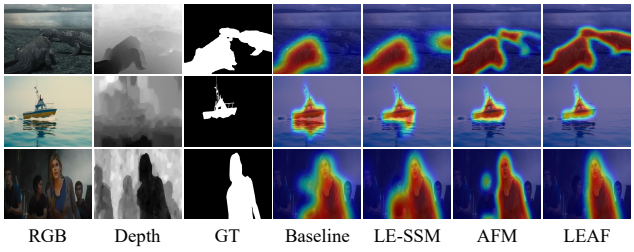

Figure 9: Visualization analyses of the LEAF-Mamba.

cross-modality interaction implementation, swapping $\boldsymbol{C}$ obtains the best performance. Based on these findings, our SIM module select $\boldsymbol{C}$ to perform modality knowledge promotion.

F.3 Ablation Study on SEM

In addition, we consider various weighting objects in SEM. We report the results of weighting various candidates (X, $\boldsymbol{B}$ and $\boldsymbol{C}$) in Table 8. As we can see, considering the different roles and reliability of each features generally performs better than the equally regarding approach (baseline, No. #A1). Among various weighting options, $\boldsymbol{C}$ lags its counterparts X/$\boldsymbol{B}$ with a clear margin of 0.93%/0.58% and 4.76%/2.44% in F_β and M on the SSD dataset. Accordingly, weighting on $\boldsymbol{C}$ is set to the default configuration in our SEM module.

G FEATURE VISUALIZATION

To provide a more comprehensive evaluation of LEAF-Mamba, we present several failure cases in Fig.10. In the first row, when both RGB and depth inputs are of low quality, the model struggles to extract insightful knowledge from either modality, resulting in inaccurate predictions. Moreover, certain challenging scenes also

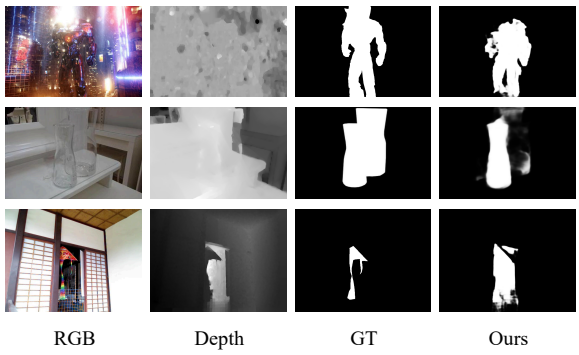

Figure 10: Failure cases illustration.

present inherent difficulties, such as transparent glass in row #2 and a partially occluded kite in row #3.

H FAILURE CASES

To provide a more comprehensive evaluation of LEAF-Mamba, we present several failure cases in Fig.10. In the first row, when both RGB and depth inputs are of low quality, the model struggles to extract insightful knowledge from either modality, resulting in inaccurate predictions. Moreover, certain challenging scenes also present inherent difficulties, such as transparent glass in row #2 and a partially occluded kite in row #3.