Liquid Fusion of Heterogeneous Representations Towards General Salient Object Detection

Ke Chen[1], Ling Zhou[2], Guangqi Jiang[1], Gengshen Wu[3], Yi Liu[1*], and Shoukun Xu[1]

[1] School of Computer Science and Artificial Intelligence, Changzhou University, Changzhou, Jiangsu, 213159, China
[2] College of Computer Science and Artificial Intelligence, Fudan University, Shanghai, 200082, China
[3] Faculty of Data Science, City University of Macau, Avenida Padre Tomás Pereira Taipa, Macao, 999078, China
s24150812007@smail.cczu.edu.cn, lzhou24@m.fudan.edu.cn, gswu@cityu.edu.mo, {guangqijiang, liuyi0089, skxu}@cczu.edu.cn

Abstract. General Salient Object Detection (SOD) aims to identify and segment visually interesting objects from uni-modality or multi-modality scenes, recently advanced by cutting-edge State Space Models (SSMs). However, a critical limitation of current approaches is their neglect of the inherent spectral biases exhibited by different neural network paradigms. By digging to the dataset-level spectral analysis of Convolutional Neural Networks (CNNs) and SSMs, their semantic representations are inherently complementary based on their complementary frequency preferences. Inspired by this, we harmonize heterogeneous representations from SSMs and CNNs to bridge their spectral biases for general salient object detection. To this end, inspired by the dynamic information propagation of Liquid Neural Networks (LNNs), we introduce a liquid fusion to dynamically integrates features from two backbones, including VMamba and ConvNeXt, referred to Liquid Fusion Network (LFNet). Concretely, by treating the continuous VMamba features and ConvNeXt features as evolving states and exogenous stimulus, respectively, LFNet employs a dynamic gating mechanism for content-aware feature aggregation. Crucially, this state-stimulus paradigm enables to scale to multi-modal cues, resulting in flexibility in general SOD. Besides, a Saliency-Guided Upsampling (SGU) operator to propagate the features to the shallow layer, which leverages a spectral-spatial co-design to suppress upsampling artifacts while preserving semantics. Extensive experiments across five diverse tasks (RGB, RGB-D, RGB-T, VSOD, and VDT) demonstrate that LFNet achieves state-of-the-art performance, offering a superior trade-off between detection accuracy and model efficiency. Code has been released at https://github.com/cke520/LFNet.

Keywords: Salient Object Detection · Liquid fusion · Saliency-guided upsampling

* Corresponding author.

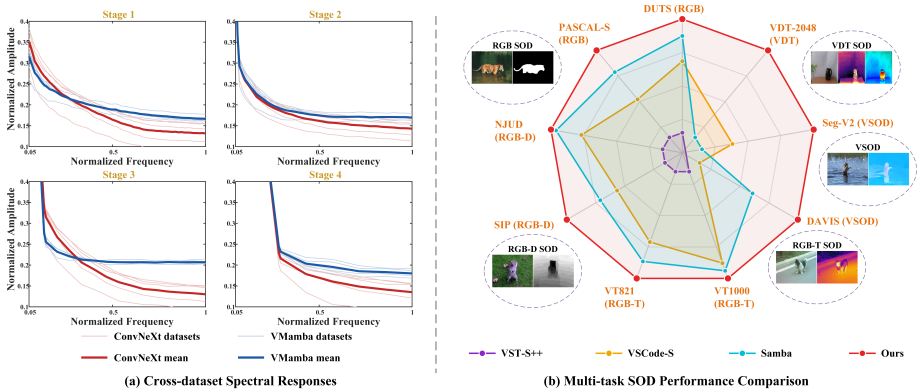

Fig. 1: Spectral response and multi-task performance. (a) Normalized frequency amplitude curves averaged over five representative datasets, i.e., PASCAL-S [31], NJUD [24], VT5000 [58], DAVIS [47], and VDT-2048 [54]. The intersecting patterns show that ConvNeXt (red) and VMamba (blue) have different energy preferences across frequency bands, necessitating their fusion for balanced full-spectrum perception. (b) Performance comparison (S_m) across five general SOD tasks, demonstrating our superior performance.

1 Introduction

Salient object detection, commonly referred to as SOD, is a fundamental computer vision task dedicated to identifying and precisely segmenting the most visually prominent objects within a scene. By effectively filtering out background clutter, this technique serves as an indispensable preliminary step for a multitude of downstream applications, including visual tracking [78], image editing [15], scene understanding [36, 52], autonomous driving [7, 37], and robotic navigation [51].

A robust SOD model must construct comprehensive representations that satisfy two competing requirements: capturing broad semantic contexts for object localization and preserving fine-grained structural details for precise boundary delineation. Recently, the field has progressed toward **general SOD** [18, 41], an ambitious framework encompassing single-modal RGB [68,81], dual-modal RGB-D/T [39,73] and Video SOD (VSOD) [49,55], as well as tri-modal Visible-Depth-Thermal (VDT) tasks [37, 40]. While Transformer-based methods [33, 35, 41] established strong baselines, their quadratic complexity limits efficiency. Consequently, State Space Models [14] (SSMs) like Mamba [13] have emerged as compelling alternatives. Notably, Samba [18] successfully adapted SSMs to general SOD, achieving state-of-the-art results by constructing efficient long-range dependencies with linear computational complexity.

From a signal processing perspective, the fundamental nature of SOD demands robust representations across the entire frequency spectrum. Current SSM-based approaches predominantly rely on engineering increasingly intricate

spatial scanning strategies (e.g., cross-shaped or continuous paths) to force a single architecture to capture all necessary multi-scale contexts. However, we argue that this direction overlooks the inherent architectural biases dictated by different neural processing mechanisms. To investigate this, we conduct a cross-dataset spectral analysis over representative benchmarks covering RGB, RGB-D, RGB-T, VSOD, and VDT SOD. Specifically, we extract hierarchical features across four scaling stages from a lightweight Convolutional Neural Network [19] (ConvNeXt-Pico [67]) and an SSM network [14] (VMamba-Small [38]), applying 2D Fast Fourier Transform (FFT [4]) to compute their radial frequency profiles.

As explicitly demonstrated in Fig. 1 (a), the spectral energy distributions of CNNs and SSMs exhibit distinct divergence and intersection patterns across different scaling stages (e.g., stage 1 and stage 3). This observed structural heterogeneity reveals that CNNs and SSMs encode spatial information in fundamentally different ways, exhibiting distinct spectral signatures. This factual spectral divergence mathematically proves that their representations are inherently complementary, and relying solely on a single paradigm inevitably creates representational blind spots.

Inspired by this natural complementarity, we harmonize heterogeneous representations from SSMs and CNNs to bridge their spectral biases for general salient object detection. To this end, inspired by the continuous-time dynamics of Liquid Neural Networks (LNNs) [16, 17], we recast them to a liquid fusion that dynamically integrate features from two backbones, including VMamba and ConvNeXt, which is referred to Liquid Fusion Network (LFNet). Concretely, by treating the continuous VMamba features and ConvNeXt features as evolving states and exogenous stimulus, respectively, LFNet employs a dynamic gating mechanism for content-aware feature aggregation. Crucially, this state-stimulus paradigm enables to scale to multi-modal cues. This ensures flexibility in general SOD beyond RGB SOD, including RGB-D SOD, RGB-T SOD, VSOD, and VDT SOD.

Besides, a Saliency-Guided Upsampling (SGU) operator to propagate the features to the shallow layer, which leverages a spectral-spatial co-design to suppress upsampling artifacts while preserving semantics. As evidenced by the quantitative comparisons in Fig. 1 (b), LFNet yields superior accuracy across diverse general SOD tasks.

In summary, our main contributions are four-fold:

- We propose LFNet, a novel general SOD framework that harmonizes heterogeneous representations from SSMs and CNNs to bridge their inherent spectral biases.
- We design a liquid fusion based on a continuous state-stimulus mechanism to dynamically aggregate heterogeneous features, serving as a highly scalable solution for multi-modal fusion.
- We develop a saliency-guided upsampling operator, which utilizes a dual-branch spectral-spatial design to effectively propagate semantics during resolution restoration.

– Extensive experiments demonstrate that LFNet achieves state-of-the-art performance across five diverse general SOD tasks.

2 Related Work

2.1 Salient Object Detection

To address complex real-world scenarios, Salient Object Detection (SOD) has evolved from single-image processing to a diverse set of tasks, encompassing single-modal RGB, dual-modal RGB-D/T, Video SOD (VSOD), and tri-modal Visible-Depth-Thermal (VDT) SOD.

RGB SOD. Early RGB SOD research was predominantly driven by CNN-based architectures [12, 50, 66, 68], which struggle with long-range global dependency modeling due to limited receptive fields. While recent Transformer-based approaches [33, 35, 41, 81] alleviate this by leveraging self-attention, their inherent quadratic computational complexity remains a critical bottleneck for high-resolution dense prediction tasks.

RGB-D and RGB-T SOD. To resolve visual ambiguities in challenging scenes, auxiliary modalities like depth (RGB-D [20, 56]) or thermal radiation (RGB-T [57,64]) are introduced. While effective, fusing these heterogeneous cues via intricate cross-modal attention [10] or dynamic guidance networks [73] often incurs significant computational overhead compared to rudimentary concatenation, highlighting the demand for more elegant integration mechanisms.

Video SOD (VSOD). VSOD extends detection into the temporal domain [21, 49], necessitating the joint modeling of spatial appearance and motion coherence. Existing approaches typically rely on precomputed optical flow or heavy spatiotemporal attention to capture inter-frame dependencies [34, 55, 75]. Unfortunately, both paradigms suffer from high latency and computational redundancy, highlighting the critical need for highly efficient spatiotemporal modeling.

VDT SOD. Addressing extreme scenarios, VDT SOD leverages the synergy of visible, depth, and thermal modalities for robust detection. However, simultaneously aligning and fusing three heterogeneous inputs drastically increases model complexity [37, 60], typically sacrificing inference speed for accuracy. Ultimately, as the field progresses toward a general SOD framework, developing a scalable architecture that seamlessly handles single, dual, and tri-modal inputs—without suffering from quadratic computational costs—has become a pressing necessity.

2.2 State Space Models

Motivated by the success of sequence modeling in natural language processing, State Space Models (SSMs) have emerged as highly efficient alternatives to Transformers, achieving global receptive fields with linear computational complexity through hardware-aware selective scanning. Pioneering this shift, Vim [80]

successfully adapted bidirectional SSMs for visual tasks, while subsequent architectures like VMamba [38] introduced refined visual state space blocks to enhance feature extraction across various domains, including semantic segmentation [62], object detection [79], and video understanding [2]. Within the specific domain of Salient Object Detection (SOD), Samba [18] pioneered the integration of SSMs to construct efficient contexts. To overcome the inherent 1D nature of SSMs when processing 2D visual data, a predominant trend involves engineering increasingly intricate scanning strategies, such as multi-directional diagonal paths [74], continuous 2D sequences [70], or Nested S-shaped Scanning (NSS) [27]. Despite these innovations, we argue that relying solely on increasingly complex scanning routes to force a single architecture to capture both broad semantics and fine-grained local textures often introduces significant architectural redundancy and optimization challenges. As revealed by our spectral analysis, pure SSMs suffer from inherent spectral biases that constrain their ability to represent high-frequency spatial features. This structural limitation highlights the fundamental difficulty of achieving comprehensive full-spectrum perception within a purely homogeneous SSM framework, thereby motivating our proposed heterogeneous hybrid paradigm that harmonizes SSMs with CNN-based local feature extraction.

2.3 Liquid Neural Networks

Liquid Neural Networks (LNNs) are continuous-time recurrent models governed by Liquid Time-Constant (LTC) dynamics [17]. Unlike traditional networks with fixed weights, LNNs offer highly expressive, input-dependent temporal adaptation [32]. Initially bottlenecked by numerical ODE solvers, the practical application of LNNs was revitalized by the Closed-form Continuous-time (CfC) network [16], which introduced a highly efficient analytical approximation. This breakthrough has propelled LNNs into diverse sequential and multimodal applications, including robotic control [25] and edge computing [6]. In this paper, we recognize that the continuous-time dynamics of LNNs perfectly align with the challenge of fusing heterogeneous spatial features. By adapting the LNN philosophy into our proposed Liquid Fusion Module (LFM), we move beyond static element-wise addition, treating global SSM representations and local CNN features as an evolving dynamic system. To the best of our knowledge, this work represents the first attempt to leverage liquid dynamics for constructing a cross-paradigm heterogeneous fusion framework in dense visual prediction tasks, such as general SOD.

3 Methodology

3.1 Preliminaries

Liquid Neural Networks and Closed-Form Dynamics. Unlike traditional recurrent models with fixed discrete dynamics, Liquid Neural Networks

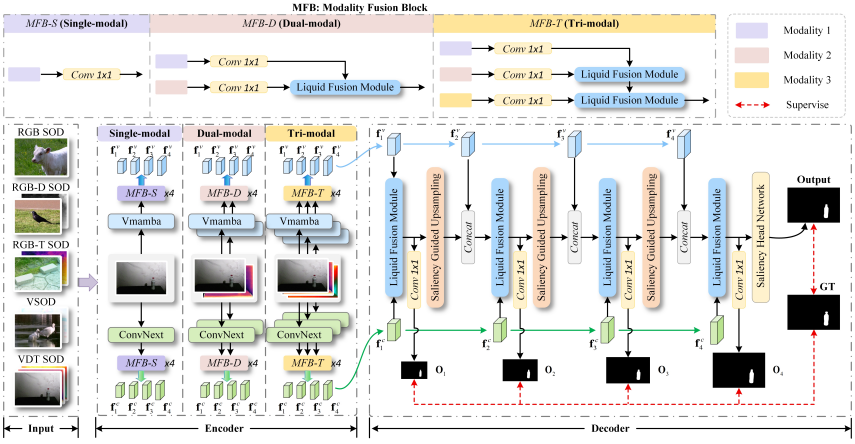

Fig. 2: Overall architecture of the proposed LFNet. Given varying input modalities, parallel VMamba and ConvNeXt backbones extract heterogeneous features, which are then harmonized by Modality Fusion Blocks (MFB) and progressively aggregated through a top-down decoder using Liquid Fusion Modules (LFMs) and Saliency-Guided Upsampling (SGU) operators with multi-scale supervision.

(LNNs) [17] exhibit continuous, input-dependent temporal behavior. The hidden state $\mathbf{x}(t)$ evolves under an exogenous input $\mathbf{I}(t)$ via a liquid time-constant (LTC) Ordinary Differential Equation (ODE):

$$\frac{d\mathbf{x}(t)}{dt} = -\left[w_\tau + f(\mathbf{x}, \mathbf{I}; \theta)\right] \odot \mathbf{x}(t) + A \odot f(\mathbf{x}, \mathbf{I}; \theta), \tag{1}$$

where w_τ represents the base time constant and $f(\cdot)$ denotes a nonlinearity. To bypass the computational bottleneck of stiff numerical solvers, Closed-form Continuous-depth (CfC) models [16] derive a bounded closed-form solution. Crucially, to mitigate vanishing gradients associated with exponential decay, CfC replaces the decay term with a learnable sigmoidal gate $\sigma(\cdot)$, effectively functioning as a dynamic interpolation mechanism.

Inspired by this continuous-time gating mechanism [16], we translate the temporal dynamics into the spatial domain. We formulate our Liquid Fusion dynamic as a weighted equilibrium between the evolving *memory state* (derived from the continuous SSM stream $\mathbf{h}$) and the *synaptic stimulus* (derived from the static CNN stream $\tilde{\mathbf{I}}$), defined as:

$$\mathbf{x}_{out} = (1 - \sigma) \odot \mathbf{h} + \sigma \odot \tilde{\mathbf{I}}, \tag{2}$$

where the dynamic permeability $\sigma \in (0, 1)$ regulates the injection of exogenous stimuli. In this context, high permeability ($\sigma \to 1$) facilitates rapid stimulus integration, while low permeability ($\sigma \to 0$) prioritizes the preservation of the inherent memory state.

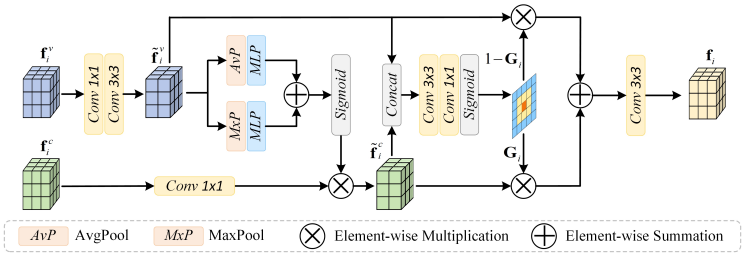

Fig. 3: Detailed architecture of the Liquid Fusion Module (LFM).

3.2 Overview

Inspired by the distinct yet complementary frequency preferences observed between VMamba and ConvNeXt, we propose LFNet to harmonize these heterogeneous representations for general SOD. As illustrated in Fig. 2, our Heterogeneous Hybrid Encoder extracts representations across four hierarchical stages: a VMamba stream captures continuous sequence-based semantics, while a parallel ConvNeXt stream preserves grid-based spatial inductive biases. To seamlessly integrate these heterogeneous cues, our decoder employs the novel Liquid Fusion Module (LFM), an LNN-inspired gated fusion module, for content-aware dynamic feature aggregation. Finally, the Saliency-Guided Upsampling (SGU) module progressively reconstructs high-resolution predictions, utilizing a spectral-spatial co-design to propagate semantics to shallow layers during resolution restoration.

3.3 Heterogeneous Hybrid Encoder and Modality Fusion

To leverage the structural heterogeneity revealed in our spectral analysis, we construct a Heterogeneous Hybrid Encoder integrating two distinct backbones. We employ a pre-trained VMamba-Small [38] to generate continuous state-space representations with linear complexity, yielding features $\mathbf{f}_i^v \in \mathbb{R}^{C_i \times H_i \times W_i}$ at stages $i \in \{1, 2, 3, 4\}$. Complementing this, a parallel ConvNeXt-Pico [67] stream introduces a distinct architectural inductive bias, ensuring the capture of comprehensive spectral cues without relying on complex scanning engineering. To handle diverse inputs, a Modality Fusion Block (*MFB*) within the ConvNeXt stream generates task-adapted features $\mathbf{f}_i^c$. The MFB applies a 1×1 projection for single-modal RGB inputs, uses a single LFM for dual-modal tasks (RGB-D/T, VSOD), and employs a cascaded LFM structure for tri-modal VDT scenarios, effectively absorbing auxiliary cues without the quadratic cost of cross-attention.

3.4 Liquid Fusion Module

Inspired by the dynamic equilibrium of Liquid Neural Networks [16], our LFM (detailed in Fig. 3) reformulates heterogeneous feature integration as a *state-stimulus* interaction process. We treat the VMamba feature $\mathbf{f}_i^v$ as an evolving

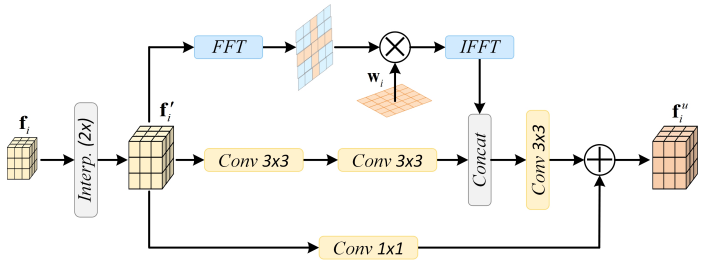

Fig. 4: Detailed architecture of the Saliency-Guided Upsampling (SGU).

memory state and the ConvNeXt feature $\mathbf{f}_i^c$ as an *exogenous stimulus*. Initially, the VMamba state is projected into an embedding $\tilde{\mathbf{f}}_i^v$ via sequential 1×1 and 3×3 convolutions.

Adaptive Channel Modulation. To adaptively filter task-irrelevant details, we modulate the ConvNeXt stimulus using a channel attention map derived from the VMamba state. This modulation is mathematically expressed as:

$$\tilde{\mathbf{f}}_i^c = Conv_{1\times1}(\mathbf{f}_i^c) \odot \sigma\left(\mathcal{M}(AvgP(\tilde{\mathbf{f}}_i^v)) + \mathcal{M}(MaxP(\tilde{\mathbf{f}}_i^v))\right), \tag{3}$$

where $\mathcal{M}$ denotes a shared Multi-Layer Perceptron (MLP), *AvgP* and *MaxP* represent spatial average and max pooling operations respectively, and σ designates the Sigmoid activation function.

Dynamic Spatial Gating. To achieve a dynamic balance between memory retention and stimulus injection, we first compute a spatial liquid gate $\mathbf{G}_i$ based on the concatenated features $[\tilde{\mathbf{f}}_i^v, \tilde{\mathbf{f}}_i^c]$, calculated as:

$$\mathbf{G}_i = \sigma\left(Conv_{1\times1}\left(Conv_{3\times3}([\tilde{\mathbf{f}}_i^v, \tilde{\mathbf{f}}_i^c])\right)\right). \tag{4}$$

Leveraging the theoretical foundation established in Eq. (2), we materialize the fusion process as a closed-form dynamic selection mechanism:

$$\mathbf{f}_i = Conv_{3\times3}\left((1 - \mathbf{G}_i) \odot \tilde{\mathbf{f}}_i^v + \mathbf{G}_i \odot \tilde{\mathbf{f}}_i^c\right). \tag{5}$$

Here, $\mathbf{G}_i$ serves as the dynamic permeability parameter (analogous to σ in Eq. (2)). Consequently, as $\mathbf{G}_i \to 1$, the network prioritizes the injection of grid-based spatial cues from ConvNeXt (stimulus); conversely, as $\mathbf{G}_i \to 0$, it prioritizes retaining the sequence-based context from VMamba (memory). This adaptive mechanism effectively reconciles the distinct architectural biases of the two streams.

3.5 Saliency-Guided Upsampling (SGU)

Standard bilinear upsampling often leads to spectral aliasing and blurred object boundaries. To preserve sharp details for SOD, our SGU module (illustrated in

Fig. 4) employs a dual-branch design to enforce consistency in both spatial and frequency domains.

Spectral-Spatial Co-Design. We first upscale the fused feature $\mathbf{f}_i$ via $2\times$ interpolation to yield the intermediate feature $\mathbf{f}_i'$. To preserve global shape integrity, the spectral branch transforms $\mathbf{f}_i'$ via Fast Fourier Transform (FFT) and modulates the spectrum with a learnable complex weight matrix $\mathbf{w}_i$, formulated as:

$$\mathbf{F}_i^{spec} = IFFT\left(FFT(\mathbf{f}_i') \odot \mathbf{w}_i\right), \tag{6}$$

where *IFFT* denotes the Inverse FFT. Concurrently, the spatial branch captures high-frequency edge cues via two stacked 3×3 convolutions to generate the spatial representation $\mathbf{F}_i^{spat}$.

Dual-Domain Fusion. The final reconstructed feature $\mathbf{f}_i^u$ integrates both domains alongside a residual connection, derived as:

$$\mathbf{f}_i^u = Conv_{3\times3}\left([\mathbf{F}_i^{spec}, \mathbf{F}_i^{spat}]\right) + Conv_{1\times1}(\mathbf{f}_i'). \tag{7}$$

By synergizing spatial precision with spectral consistency, SGU ensures that the decoding phase restores sharp boundaries without suffering from structural degradation.

3.6 Loss Function

To facilitate robust feature learning, we employ a deep supervision strategy. The total loss $\mathcal{L}_{total}$ is defined as a weighted combination of the Binary Cross Entropy (BCE) loss and the Intersection-over-Union (IoU) loss, applied to predictions at all scales:

$$\mathcal{L}_{total} = \sum_{k=1}^{4} \left(\mathcal{L}_{bce}(\mathbf{O}_k, \mathbf{GT}) + \mathcal{L}_{iou}(\mathbf{O}_k, \mathbf{GT})\right), \tag{8}$$

where $\mathbf{O}_k$ represents the prediction map at stage k and $\mathbf{GT}$ denotes the ground truth. This multi-scale supervision ensures that the model learns discriminative features progressively from coarse to fine levels.

4 Experiments and Results

4.1 Datasets and Metrics

To comprehensively evaluate the proposed LFNet, we conduct extensive experiments across five distinct tasks: RGB SOD, RGB-D SOD, RGB-T SOD, VSOD, and VDT SOD. For the RGB SOD task, we utilize five standard benchmark datasets: DUTS [65], DUT-O [71], HKU-IS [29], PASCAL-S [31], and ECSSD [69]. For RGB-D SOD, the evaluation is performed on NJUD [24], NLPR [46], SIP [8], STERE [44], and DUTLF-D [48]. Regarding RGB-T SOD, three benchmarks are employed: VT821 [63], VT1000 [59], and VT5000 [58]. For VSOD, we use DAVIS [47], DAVSOD-easy [9], FBMS [45], SegV2 [28],

Table 1: Quantitative comparison of our LFNet against other SOTA RGB SOD methods on five benchmark datasets. "-" indicates the result is not available. "↑" denotes larger is better. All metric values are in %. The best two results are highlighted in **red** and **blue**.

Method	Params	DUTS [65]			DUT-O [71]			HKU-IS [29]			PASCAL-S [31]			ECSSD [69]		
	(M)	$S_m\uparrow$	$F_m\uparrow$	$E_m\uparrow$	$S_m\uparrow$	$F_m\uparrow$	$E_m\uparrow$	$S_m\uparrow$	$F_m\uparrow$	$E_m\uparrow$	$S_m\uparrow$	$F_m\uparrow$	$E_m\uparrow$	$S_m\uparrow$	$F_m\uparrow$	$E_m\uparrow$
CNN-based Methods																
CSF-R2 [12]	36.53	89.0	86.9	92.9	83.8	77.5	86.9	-	-	-	86.3	83.9	88.5	93.1	94.2	96.0
EDN [68]	42.85	89.2	89.3	93.3	84.9	82.1	88.4	92.4	94.0	96.3	86.4	87.9	90.7	92.7	95.0	95.7
ICON-R [81]	33.09	89.0	87.6	93.1	84.5	79.9	88.4	92.0	93.1	96.0	86.2	84.4	88.8	92.8	94.3	96.0
MENet [66]	27.83	90.5	89.5	94.3	85.0	79.2	87.9	92.7	93.9	96.5	87.1	84.8	89.2	92.7	93.8	95.6
Transformer-based Methods																
ICON-S [81]	94.30	91.7	91.1	96.0	86.9	83.0	90.6	93.6	94.7	97.4	88.5	86.0	90.3	94.1	95.4	97.1
BBRF [42]	74.40	90.8	90.5	95.1	85.5	82.0	89.8	93.5	94.6	93.6	87.1	88.4	92.5	93.9	95.7	97.2
VST-S ++ [33]	74.90	90.9	89.7	94.7	85.9	81.3	89.0	93.2	94.1	96.9	88.0	85.9	90.1	93.9	95.1	96.9
VSCode-S [41]	74.72	92.6	92.2	96.0	87.7	84.0	91.2	94.0	95.1	97.4	88.7	86.4	90.4	94.9	95.9	97.4
Mamba-based Methods																
Samba [18]	49.59	**93.2**	**93.0**	**96.6**	**88.9**	**85.9**	**92.2**	**94.5**	**95.6**	**97.8**	**89.2**	**89.6**	**93.1**	**95.3**	**96.5**	**97.8**
Ours	43.23	**93.6**	**93.6**	**97.0**	**90.4**	**88.4**	**93.8**	**94.6**	**95.8**	**97.9**	**89.6**	**89.8**	**93.9**	**95.4**	**96.7**	**97.7**

and VOS [30]. Lastly, for VDT SOD, we employ the VDT-2048 [54] benchmark dataset.

To quantitatively assess the model performance, we adopt three widely used saliency metrics: structure-measure (S_m), maximum F-measure (F_m), and maximum enhanced-alignment measure (E_m). Furthermore, to evaluate the model size and computational complexity, we report the total number of parameters (Params).

4.2 Implementation Details

The proposed LFNet is implemented in PyTorch and trained on a single NVIDIA RTX 4090 D GPU. Following standard protocols, we utilize established training splits for all five tasks, rigorously removing duplicate samples in joint datasets to prevent data leakage. During training, all input images are resized to 512×512 and augmented using random flipping, cropping, and rotation. The model is optimized for 50 epochs with a batch size of 2 using the AdamW optimizer. The base learning rate is initialized at 1×10^{-4} with a weight decay of 0.05, while the pre-trained backbone networks use only five percent of this rate to preserve representational integrity. We apply a five-epoch linear warmup followed by a cosine annealing decay schedule. To enhance training efficiency and prevent gradient explosion, Automatic Mixed Precision (AMP) and gradient clipping (with a maximum norm threshold of 0.5) are integrated. The optimal model weights are selected based on the highest S-measure achieved on the validation set.

Table 2: Quantitative comparison of our LFNet against other SOTA RGB-D SOD methods on five benchmark datasets.

Method	Params	NJUD [24]			NLPR [46]			SIP [8]			STERE [44]			DUTLF-D [48]		
	(M)	$S_m\uparrow$	$F_m\uparrow$	$E_m\uparrow$	$S_m\uparrow$	$F_m\uparrow$	$E_m\uparrow$	$S_m\uparrow$	$F_m\uparrow$	$E_m\uparrow$	$S_m\uparrow$	$F_m\uparrow$	$E_m\uparrow$	$S_m\uparrow$	$F_m\uparrow$	$E_m\uparrow$
CNN-based Methods																
JL-DCF [11]	143.52	87.7	89.2	94.1	93.1	91.8	96.5	88.5	89.4	93.1	90.0	89.5	94.2	89.4	89.1	92.7
SP-Net [76]	67.88	92.5	92.8	95.7	92.7	91.9	96.2	89.4	90.4	93.3	90.7	90.6	94.9	89.5	89.9	93.3
DCF [22]	53.92	90.4	90.5	94.3	92.2	91.0	95.7	87.4	88.6	92.2	90.6	90.4	94.8	92.5	93.0	95.6
SPSN [26]	-	91.8	92.1	95.2	92.3	91.2	96.0	89.2	90.0	93.6	90.7	90.2	94.5	-	-	-
Transformer-based Methods																
CATNet [56]	262.73	93.2	93.7	96.0	93.8	93.4	97.1	91.0	92.8	95.1	92.0	92.2	95.8	95.2	95.8	97.5
VST-S ++ [33]	143.15	92.8	92.8	95.7	93.5	92.5	96.4	90.4	91.8	94.6	92.1	91.6	95.4	94.5	95.0	96.9
CPNet [20]	216.50	93.5	94.1	96.3	94.0	93.6	97.1	90.7	92.7	94.6	92.0	92.2	96.0	95.1	95.9	97.4
VSCode-S [41]	74.72	94.4	94.9	97.0	94.1	93.2	96.8	92.4	94.2	95.8	93.1	92.8	95.8	**96.0**	**96.7**	**98.0**
Mamba-based Methods																
Samba [18]	54.94	**94.9**	**95.6**	**97.5**	**94.7**	**94.1**	**97.6**	**93.1**	**94.9**	**96.6**	**93.5**	**93.3**	**96.3**	95.6	96.4	97.6
Ours	44.65	**95.0**	**95.8**	**97.8**	**94.9**	**94.3**	**97.7**	**94.5**	**96.1**	**97.7**	**93.6**	**93.5**	**96.4**	**96.3**	**97.3**	**98.4**

Table 3: Quantitative comparison of our LFNet against other SOTA RGB-T SOD methods on three benchmark datasets.

Method		***CNN-based***				***Transformer-based***				***Mamba-based***	
		MGAI [53]	TNet [5]	CGMDR [1]	SPNet [73]	VST-S ++ [33]	VSCode-S [41]	PCNet [64]	ConTri [57]	Samba [18]	**Ours**
Params (M)		87.09	87.04	-	104.03	143.15	74.72	-	96.31	54.94	44.65
VT821 [63]	$S_m\uparrow$	89.1	89.9	89.4	91.3	89.7	92.6	91.5	91.6	**93.4**	**94.1**
	$F_m\uparrow$	87.0	88.5	87.2	90.0	86.8	91.0	87.9	89.6	**92.7**	**93.2**
	$E_m\uparrow$	93.3	93.6	93.2	94.9	92.5	95.4	94.1	94.8	**96.5**	**96.8**
VT1000 [59]	$S_m\uparrow$	92.9	92.9	93.1	94.1	94.0	95.2	94.3	94.1	**95.3**	**95.4**
	$F_m\uparrow$	92.1	92.1	92.7	94.3	93.1	**94.7**	92.4	93.9	**95.6**	**95.6**
	$E_m\uparrow$	96.5	96.5	96.6	97.5	97.1	98.1	95.8	97.6	**98.3**	**98.2**
VT5000 [58]	$S_m\uparrow$	88.4	89.5	89.6	91.4	90.1	**92.5**	92.0	92.4	**92.8**	**92.8**
	$F_m\uparrow$	84.6	86.4	87.7	90.5	86.1	90.0	89.9	91.8	**91.9**	**92.0**
	$E_m\uparrow$	93.0	93.6	93.9	95.4	93.6	95.9	95.6	**96.3**	**96.3**	**96.3**

4.3 Comparison with State-of-the-Art Methods

Quantitative Evaluation. Since LFNet is designed as a heterogeneous hybrid framework for general SOD, we conduct extensive comparative experiments against existing state-of-the-art (SOTA) methods across five distinct tasks: RGB SOD, RGB-D SOD, RGB-T SOD, VSOD, and VDT SOD. The comprehensive quantitative results are presented in Tables 1, 2, 3, 4, and 5. These comparisons demonstrate that LFNet consistently outperforms leading CNN-based, Transformer-based, and Mamba-based competitors across all benchmark datasets, validating the superior efficacy of our proposed architecture.

Specifically, for the RGB SOD task (Table 1), our model achieves the best performance on nearly all metrics while maintaining a significantly lower parameter count compared to heavy Transformer-based methods. For instance, on the challenging DUTS dataset, LFNet surpasses the heavy transformer ICON-S by 1.9% in S_m and 2.5% in F_m, while using less than half of its parameters (43.23M vs.

Table 4: Quantitative comparison of our LFNet against other SOTA VSOD methods on five benchmark datasets.

Method	Params	DAVIS [47]			DAVSOD-easy [9]			FBMS [45]			SegV2 [28]			VOS [30]		
	(M)	$S_m\uparrow$	$F_m\uparrow$	$E_m\uparrow$	$S_m\uparrow$	$F_m\uparrow$	$E_m\uparrow$	$S_m\uparrow$	$F_m\uparrow$	$E_m\uparrow$	$S_m\uparrow$	$F_m\uparrow$	$E_m\uparrow$	$S_m\uparrow$	$F_m\uparrow$	$E_m\uparrow$
CNN-based Methods																
FSNet [21]	83.41	92.2	90.9	97.2	76.0	63.7	79.6	87.5	86.7	91.8	84.9	77.3	92.0	67.8	62.1	75.5
DCFNet [72]	69.56	91.4	89.9	97.0	72.9	61.2	78.1	88.3	85.3	91.0	90.3	87.0	95.3	83.8	77.3	86.1
UGPL [49]	-	91.1	89.5	96.8	73.2	60.2	77.1	89.7	88.4	93.9	86.7	82.8	93.8	75.1	68.5	81.1
MMNet [75]	50.81	89.6	87.7	96.3	74.8	63.6	79.7	88.3	86.5	92.1	90.3	88.3	95.5	-	-	-
Transformer-based Methods																
MGTNet [43]	150.91	92.5	91.9	97.6	76.5	65.3	80.0	90.0	88.1	92.9	90.3	86.1	94.6	81.4	72.7	81.9
UFO [55]	55.92	91.8	90.6	97.8	74.7	62.6	79.9	85.8	86.8	91.1	88.8	85.0	95.1	-	-	-
CoSTFormer [34]	-	92.3	90.6	97.8	77.9	66.7	81.9	86.9	86.1	91.3	87.4	81.3	94.3	79.1	70.8	81.1
VSCode-S [41]	74.72	93.6	92.2	97.3	80.0	71.0	83.5	90.5	90.2	93.9	**94.6**	93.7	98.4	-	-	-
Mamba-based Methods																
Samba [18]	54.94	**94.3**	**93.6**	**98.5**	**81.3**	**73.4**	**85.6**	**92.5**	**92.2**	**95.4**	94.3	**93.8**	**98.7**	**87.0**	**82.0**	**89.8**
Ours	44.65	**94.9**	**94.4**	**98.8**	**83.1**	**76.2**	**87.5**	**92.7**	**92.9**	**96.4**	**95.4**	**95.0**	**99.2**	**88.9**	**84.4**	**91.0**

Table 5: Quantitative comparison of our LFNet against other SOTA methods on the VDT-2048 benchmark dataset.

Method		***RGB-D***		***RGB-T***		***VDT***				***Mamba-based***	
		DPA [3]	SwinNet [39]	LSNet [77]	CAFC [23]	TMNet [61]	MFFNet [60]	DWFPR [40]	PWRF [37]	Samba [18]	**Ours**
Params (M)		92.39	124.72	4.57	-	-	103.24	-	-	60.28	46.07
VDT-2048 [54]	$S_m\uparrow$	81.8	91.9	88.8	91.3	93.3	93.6	**93.8**	93.2	**93.8**	**94.2**
	$F_m\uparrow$	45.9	72.9	74.3	85.4	89.7	90.1	90.1	91.0	**91.0**	**91.9**
	$E_m\uparrow$	68.6	89.6	92.0	97.7	98.9	**99.0**	**99.0**	98.8	**99.0**	**99.2**

94.30M). Even compared to the recent Mamba-based method Samba, our model achieves a consistent improvement of 0.4%–1.6% across five datasets, proving the advantage of our heterogeneous fusion over pure SSM modeling.

Regarding multi-modal tasks, including RGB-D SOD (Table 2) and RGB-T SOD (Table 3), LFNet sets new state-of-the-art standards. In RGB-D SOD, our method outperforms the strong competitor VST-S++ by a large margin (e.g., +2.2% S_m on NJUD) with only 31% of its parameters (44.65M vs. 143.15M). Similarly, in RGB-T SOD, our method surpasses Samba on all three benchmarks, achieving a notable 0.7% S_m gain on the VT821 dataset.

Furthermore, in video-based tasks (VSOD), our model demonstrates exceptional temporal modeling capabilities. As shown in Table 4, LFNet achieves significant gains over VSCode-S on the DAVSOD-easy dataset (+3.1% S_m, +5.2% F_m), highlighting its robustness in dynamic scenes.

In the tri-modal VDT SOD task (Table 5), LFNet reaches an S_m of 94.2%, establishing the new state-of-the-art. Notably, compared to the strongest competitor Samba, LFNet not only achieves superior segmentation accuracy across all five tasks but also benefits from a lower computational complexity. This in-

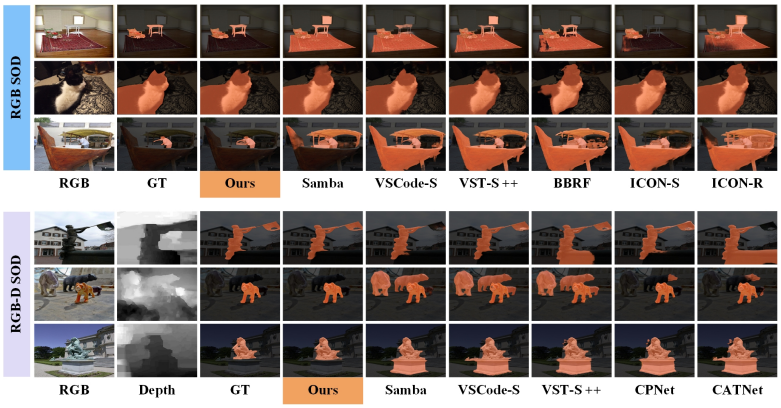

Fig. 5: Visual comparison against SOTA methods on RGB and RGB-D SOD tasks. The top three rows show RGB scenarios, while the bottom three rows display RGB-D scenarios with corresponding depth maps.

dicates that our heterogeneous framework can effectively capture multi-modal and temporal saliency cues.

Qualitative Evaluation. To visually demonstrate the superiority of LFNet, we display qualitative comparisons in Fig. 5. In RGB scenarios (top three rows), our model excels in handling complex topologies and low-contrast objects. Specifically, in the 1^{st} row, while competitors like Samba [18] and VST-S++ [33] fail to suppress the background gaps within the chair legs, LFNet cleanly delineates the object skeleton. In the 2^{nd} row, it accurately segments the black cat against a dark background, and in the 3^{rd} row, it recovers the complete boat structure despite severe internal occlusion, avoiding the fragmentation seen in ICON-R [81]. In RGB-D scenarios (bottom three rows), LFNet effectively leverages depth cues to resolve visual ambiguities. For instance, in the 4^{th} row, it distinguishes the statue from the complex building texture where others fail. Similarly, in the 5^{th} and 6^{th} rows, our method consistently produces sharp predictions for the toy animals and fountain statue, robustness against similar foreground-background colors and challenging lighting, outperforming depth-aware competitors like CPNet [20] and CATNet [56]. These results visually confirm that our heterogeneous hybrid paradigm effectively harmonizes global semantics with local details across diverse modalities. **More extensive visual comparisons across diverse scenarios can be found in the Supplementary Material.**

4.4 Ablation Study

To verify the contribution of each proposed component, we conduct thorough ablation studies across five representative datasets. As summarized in Table 6, our full architecture consistently outperforms all variants, validating the effectiveness of the heterogeneous hybrid design, the LFM, and the SGU operator.

Table 6: Ablation study of our LFNet on five datasets. All metric values are in %. **Bolded** results are the best.

Settings	*RGB SOD*						*RGB-D SOD*						*VDT*		
	DUTS [65]			DUT-O [71]			NJUD [24]			NLPR [46]			VDT-2048 [54]		
	$S_m \uparrow$	$F_m \uparrow$	$E_m \uparrow$	$S_m \uparrow$	$F_m \uparrow$	$E_m \uparrow$	$S_m \uparrow$	$F_m \uparrow$	$E_m \uparrow$	$S_m \uparrow$	$F_m \uparrow$	$E_m \uparrow$	$S_m \uparrow$	$F_m \uparrow$	$E_m \uparrow$
Ours (Full)	**93.6**	**93.6**	**97.0**	**90.4**	**88.4**	**93.8**	**95.0**	**95.8**	**97.8**	**94.9**	**94.3**	**97.7**	**94.2**	**91.9**	**99.2**
A: Ablation on Architecture															
A1: VMamba Only	92.0	91.7	96.4	89.5	87.1	93.6	94.3	94.7	97.3	94.2	93.5	97.4	91.4	87.4	97.9
A2: ConvNeXt Only	84.8	82.5	88.3	83.3	78.3	86.0	82.2	81.7	86.7	84.7	79.9	89.1	86.8	80.9	93.7
A3: Dual Stream Baseline	92.5	92.2	96.0	89.8	87.5	93.1	94.6	95.1	97.4	94.4	93.6	97.2	93.0	90.5	98.5
B: Ablation on Feature Fusion Strategy (Replacing LFM)															
B1: Additive Fusion	92.7	92.6	96.1	89.9	87.8	93.2	94.7	95.3	97.5	94.5	93.8	97.3	93.5	90.9	98.7
B2: Concatenation	93.0	92.8	96.3	90.1	88.0	93.4	94.8	95.5	97.6	94.7	94.0	97.5	93.7	91.2	98.9
B3: Cross-Attention	93.2	93.1	96.7	90.2	88.1	93.5	94.7	95.3	97.5	94.7	94.0	97.4	94.0	91.6	99.0
C: Ablation on Upsampling Strategy (Replacing SGU)															
C1: Bilinear	92.9	92.8	96.4	89.8	87.7	93.1	94.3	95.0	97.0	94.2	93.5	96.9	93.2	90.9	98.4
C2: Transposed Conv	93.2	93.1	96.7	90.1	88.0	93.4	94.6	95.4	97.4	94.6	93.9	97.3	93.5	91.3	98.7

Effectiveness of Architecture (Settings A). We evaluate the necessity of our heterogeneous hybrid design by comparing the full model with single-stream variants (A1, A2) and a naive dual-stream baseline (A3). The quantitative results demonstrate that relying solely on either the SSM-based stream (A1) or the CNN-based stream (A2) yields sub-optimal performance across all metrics. While the naive dual-stream baseline (A3) offers clear numerical improvements over both single streams by simply aggregating the two backbones, a noticeable performance gap persists compared to our full model. This confirms that the dual-stream architecture is inherently more effective than single streams, and our specific interaction design further maximizes their combined potential.

Effectiveness of Feature Fusion Strategy (Settings B). To validate our LFM, we replace it with standard operations: additive fusion (B1), simple concatenation (B2), and the widely-used cross-attention mechanism (B3). The metric comparisons consistently show that LFM achieves the highest scores across all five datasets. This numerical superiority indicates that our proposed dynamic fusion mechanism is a more effective strategy for aggregating heterogeneous features than conventional static fusion methods in dense prediction tasks.

Effectiveness of Upsampling Strategy (Settings C). The Saliency-Guided Upsampling (SGU) operator is designed to better propagate semantics across multi-level features during resolution restoration. Comparing SGU with standard bilinear interpolation (C1) and transposed convolution (C2), we observe consistent quantitative gains across all multi-modal tasks. These continuous metric improvements confirm that our specialized SGU effectively maintains semantic consistency during the upsampling process, leading to more accurate overall detection performance.

5 Conclusion

In this paper, we propose LFNet a heterogeneous hybrid framework integrating VMamba and ConvNeXt for general SOD, designed to bridge the inherent

spectral biases of single-paradigm networks. By revealing the complementary frequency preferences of VMamba and ConvNeXt, we developed a state-stimulus paradigm within the liquid fusion for dynamic feature aggregation. Complemented by the saliency-guided upsampling operator for effective feature propagation and boundary recovery, LFNet consistently establishes new state-of-the-art benchmarks across five general SOD tasks. Future work will focus on extending this liquid fusion paradigm to broader dense prediction scenarios and optimizing the architecture for lightweight edge applications.

Acknowledgements

The paper was supported in part by the National Natural Science Foundation of China under Grant 62571068, 62306048; in part by Science and Technology Development Fund, Macao SAR (Grant No: 0054/2025/RIB2); in part by the Qing Lan Project of Jiangsu Universities; in part by the Major Program of Jiangsu Higher Education Institutions Basic Science (Natural Science) Research under Grant 25KJA520001; in part by Changzhou Applied Basic Research Fund Project under Grant CJ20242060, CQ20230092, CJ20235036.

References

1. Chen, G., Shao, F., Chai, X., Chen, H., Jiang, Q., Meng, X., Ho, Y.S.: Cgmdrnet: Cross-guided modality difference reduction network for rgb-t salient object detection. IEEE Transactions on Circuits and Systems for Video Technology **32**(9), 6308–6323 (2022)
2. Chen, S., Luo, Y., Ma, Y., Qiao, Y., Wang, Y.: H-mba: Hierarchical mamba adaptation for multi-modal video understanding in autonomous driving. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 39, pp. 2212–2220 (2025)
3. Chen, Z., Cong, R., Xu, Q., Huang, Q.: Dpanet: Depth potentiality-aware gated attention network for rgb-d salient object detection. IEEE Transactions on Image Processing **30**, 7012–7024 (2020)
4. Chi, L., Jiang, B., Mu, Y.: Fast fourier convolution. Advances in Neural Information Processing Systems **33**, 4479–4488 (2020)
5. Cong, R., Zhang, K., Zhang, C., Zheng, F., Zhao, Y., Huang, Q., Kwong, S.: Does thermal really always matter for rgb-t salient object detection? IEEE Transactions on Multimedia **25**, 6971–6982 (2022)
6. Coskun, T., Vishwamith, H., Isik, M., Naoukin, J., Dikmen, I.C.: Exploring neuromorphic computing with loihi-2 for high-performance cfd simulations. In: 2025 IEEE High Performance Extreme Computing Conference (HPEC). pp. 1–7. IEEE (2025)
7. Ding, N., Zhang, C., Eskandarian, A.: Saliendet: A saliency-based feature enhancement algorithm for object detection for autonomous driving. IEEE Transactions on Intelligent Vehicles **9**(1), 2624–2635 (2023)
8. Fan, D.P., Lin, Z., Zhang, Z., Zhu, M., Cheng, M.M.: Rethinking rgb-d salient object detection: Models, data sets, and large-scale benchmarks. IEEE Transactions on neural networks and learning systems **32**(5), 2075–2089 (2020)

9. Fan, D.P., Wang, W., Cheng, M.M., Shen, J.: Shifting more attention to video salient object detection. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 8554–8564 (2019)
10. Fan, D.P., Zhai, Y., Borji, A., Yang, J., Shao, L.: Bbs-net: Rgb-d salient object detection with a bifurcated backbone strategy network. In: European conference on computer vision. pp. 275–292. Springer (2020)
11. Fu, K., Fan, D.P., Ji, G.P., Zhao, Q.: Jl-dcf: Joint learning and densely-cooperative fusion framework for rgb-d salient object detection. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 3052–3062 (2020)
12. Gao, S.H., Tan, Y.Q., Cheng, M.M., Lu, C., Chen, Y., Yan, S.: Highly efficient salient object detection with 100k parameters. In: European conference on computer vision. pp. 702–721. Springer (2020)
13. Gu, A., Dao, T.: Mamba: Linear-time sequence modeling with selective state spaces. arXiv preprint arXiv:2312.00752 (2023)
14. Gu, A., Goel, K., Ré, C.: Efficiently modeling long sequences with structured state spaces. arXiv preprint arXiv:2111.00396 (2021)
15. Hagiwara, A., Sugimoto, A., Kawamoto, K.: Saliency-based image editing for guiding visual attention. In: Proceedings of the 1st international workshop on pervasive eye tracking & mobile eye-based interaction. pp. 43–48 (2011)
16. Hasani, R., Lechner, M., Amini, A., Liebenwein, L., Ray, A., Tschaikowski, M., Teschl, G., Rus, D.: Closed-form continuous-time neural networks. Nature Machine Intelligence **4**(11), 992–1003 (2022)
17. Hasani, R., Lechner, M., Amini, A., Rus, D., Grosu, R.: Liquid time-constant networks. In: Proceedings of the AAAI conference on artificial intelligence. vol. 35, pp. 7657–7666 (2021)
18. He, J., Fu, K., Liu, X., Zhao, Q.: Samba: A unified mamba-based framework for general salient object detection. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 25314–25324 (2025)
19. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
20. Hu, X., Sun, F., Sun, J., Wang, F., Li, H.: Cross-modal fusion and progressive decoding network for rgb-d salient object detection. International Journal of Computer Vision **132**(8), 3067–3085 (2024)
21. Ji, G.P., Fu, K., Wu, Z., Fan, D.P., Shen, J., Shao, L.: Full-duplex strategy for video object segmentation. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 4922–4933 (2021)
22. Ji, W., Li, J., Yu, S., Zhang, M., Piao, Y., Yao, S., Bi, Q., Ma, K., Zheng, Y., Lu, H., et al.: Calibrated rgb-d salient object detection. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 9471–9481 (2021)
23. Jin, D., Shao, F., Xie, Z., Mu, B., Chen, H., Jiang, Q.: Cafcnet: Cross-modality asymmetric feature complement network for rgb-t salient object detection. Expert Systems with Applications **247**, 123222 (2024)
24. Ju, R., Ge, L., Geng, W., Ren, T., Wu, G.: Depth saliency based on anisotropic center-surround difference. In: 2014 IEEE international conference on image processing (ICIP). pp. 1115–1119. IEEE (2014)
25. Kao, P.: Robust flight navigation with liquid neural networks. Ph.D. thesis, Massachusetts Institute of Technology (2022)

26. Lee, M., Park, C., Cho, S., Lee, S.: Spsn: Superpixel prototype sampling network for rgb-d salient object detection. In: European conference on computer vision. pp. 630–647. Springer (2022)
27. Li, B., Zhao, H., Wang, W., Hu, P., Gou, Y., Peng, X.: Mair: A locality-and continuity-preserving mamba for image restoration. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 7491–7501 (2025)
28. Li, F., Kim, T., Humayun, A., Tsai, D., Rehg, J.M.: Video segmentation by tracking many figure-ground segments. In: Proceedings of the IEEE international conference on computer vision. pp. 2192–2199 (2013)
29. Li, G., Yu, Y.: Visual saliency based on multiscale deep features. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 5455–5463 (2015)
30. Li, J., Xia, C., Chen, X.: A benchmark dataset and saliency-guided stacked autoencoders for video-based salient object detection. IEEE Transactions on Image Processing **27**(1), 349–364 (2017)
31. Li, Y., Hou, X., Koch, C., Rehg, J.M., Yuille, A.L.: The secrets of salient object segmentation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 280–287 (2014)
32. Liu, M., Geißler, D., Bian, S., Zhou, B., Lukowicz, P.: Assessing the impact of sampling irregularity in time series data: Human activity recognition as a case study. arXiv preprint arXiv:2501.15330 (2025)
33. Liu, N., Luo, Z., Zhang, N., Han, J.: Vst++: Efficient and stronger visual saliency transformer. IEEE Transactions on Pattern Analysis and Machine Intelligence **46**(11), 7300–7316 (2024)
34. Liu, N., Nan, K., Zhao, W., Yao, X., Han, J.: Learning complementary spatial–temporal transformer for video salient object detection. IEEE Transactions on Neural Networks and Learning Systems **35**(8), 10663–10673 (2023)
35. Liu, N., Zhang, N., Wan, K., Shao, L., Han, J.: Visual saliency transformer. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 4722–4732 (2021)
36. Liu, Y., Xiong, Z., Yuan, Y., Wang, Q.: Transcending pixels: Boosting saliency detection via scene understanding from aerial imagery. IEEE Transactions on Geoscience and Remote Sensing **61**, 1–16 (2023)
37. Liu, Y., Li, C., Xu, S., Han, J.: Part-whole relational fusion towards multi-modal scene understanding. International Journal of Computer Vision **133**(7), 4483–4503 (2025)
38. Liu, Y., Tian, Y., Zhao, Y., Yu, H., Xie, L., Wang, Y., Ye, Q., Jiao, J., Liu, Y.: Vmamba: Visual state space model. Advances in neural information processing systems **37**, 103031–103063 (2024)
39. Liu, Z., Tan, Y., He, Q., Xiao, Y.: Swinnet: Swin transformer drives edge-aware rgb-d and rgb-t salient object detection. IEEE Transactions on Circuits and Systems for Video Technology **32**(7), 4486–4497 (2021)
40. Luo, Y., Shao, F., Mu, B., Chen, H., Li, Z., Jiang, Q.: Dynamic weighted fusion and progressive refinement network for visible-depth-thermal salient object detection. IEEE Transactions on Circuits and Systems for Video Technology **34**(11), 10662–10677 (2024)
41. Luo, Z., Liu, N., Zhao, W., Yang, X., Zhang, D., Fan, D.P., Khan, F., Han, J.: Vscode: General visual salient and camouflaged object detection with 2d prompt learning. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 17169–17180 (2024)

42. Ma, M., Xia, C., Xie, C., Chen, X., Li, J.: Boosting broader receptive fields for salient object detection. IEEE Transactions on Image Processing **32**, 1026–1038 (2023)
43. Min, D., Zhang, C., Lu, Y., Fu, K., Zhao, Q.: Mutual-guidance transformer-embedding network for video salient object detection. IEEE Signal Processing Letters **29**, 1674–1678 (2022)
44. Niu, Y., Geng, Y., Li, X., Liu, F.: Leveraging stereopsis for saliency analysis. In: 2012 IEEE conference on computer vision and pattern recognition. pp. 454–461. IEEE (2012)
45. Ochs, P., Malik, J., Brox, T.: Segmentation of moving objects by long term video analysis. IEEE transactions on pattern analysis and machine intelligence **36**(6), 1187–1200 (2013)
46. Peng, H., Li, B., Xiong, W., Hu, W., Ji, R.: Rgbd salient object detection: A benchmark and algorithms. In: European conference on computer vision. pp. 92–109. Springer (2014)
47. Perazzi, F., Pont-Tuset, J., McWilliams, B., Van Gool, L., Gross, M., Sorkine-Hornung, A.: A benchmark dataset and evaluation methodology for video object segmentation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 724–732 (2016)
48. Piao, Y., Ji, W., Li, J., Zhang, M., Lu, H.: Depth-induced multi-scale recurrent attention network for saliency detection. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 7254–7263 (2019)
49. Piao, Y., Lu, C., Zhang, M., Lu, H.: Semi-supervised video salient object detection based on uncertainty-guided pseudo labels. Advances in neural information processing systems **35**, 5614–5627 (2022)
50. Qin, X., Zhang, Z., Huang, C., Gao, C., Dehghan, M., Jagersand, M.: Basnet: Boundary-aware salient object detection. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 7479–7489 (2019)
51. Roberts, R., Ta, D.N., Straub, J., Ok, K., Dellaert, F.: Saliency detection and model-based tracking: a two part vision system for small robot navigation in forested environment. In: Unmanned Systems Technology XIV. vol. 8387, pp. 306–317. SPIE (2012)
52. Siris, A., Jiao, J., Tam, G.K., Xie, X., Lau, R.W.: Scene context-aware salient object detection. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 4156–4166 (2021)
53. Song, K., Huang, L., Gong, A., Yan, Y.: Multiple graph affinity interactive network and a variable illumination dataset for rgbt image salient object detection. IEEE Transactions on Circuits and Systems for Video Technology **33**(7), 3104–3118 (2022)
54. Song, K., Wang, J., Bao, Y., Huang, L., Yan, Y.: A novel visible-depth-thermal image dataset of salient object detection for robotic visual perception. IEEE/ASME Transactions on Mechatronics **28**(3), 1558–1569 (2022)
55. Su, Y., Deng, J., Sun, R., Lin, G., Su, H., Wu, Q.: A unified transformer framework for group-based segmentation: Co-segmentation, co-saliency detection and video salient object detection. IEEE Transactions on Multimedia **26**, 313–325 (2023)
56. Sun, F., Ren, P., Yin, B., Wang, F., Li, H.: Catnet: A cascaded and aggregated transformer network for rgb-d salient object detection. IEEE Transactions on Multimedia **26**, 2249–2262 (2023)
57. Tang, H., Li, Z., Zhang, D., He, S., Tang, J.: Divide-and-conquer: Confluent triple-flow network for rgb-t salient object detection. IEEE transactions on pattern analysis and machine intelligence **47**(3), 1958–1974 (2024)

58. Tu, Z., Ma, Y., Li, Z., Li, C., Xu, J., Liu, Y.: Rgbt salient object detection: A large-scale dataset and benchmark. IEEE Transactions on Multimedia **25**, 4163–4176 (2022)
59. Tu, Z., Xia, T., Li, C., Wang, X., Ma, Y., Tang, J.: Rgb-t image saliency detection via collaborative graph learning. IEEE Transactions on Multimedia **22**(1), 160–173 (2019)
60. Wan, B., Zhou, X., Sun, Y., Wang, T., Lv, C., Wang, S., Yin, H., Yan, C.: Mffnet: Multi-modal feature fusion network for vdt salient object detection. IEEE Transactions on Multimedia **26**, 2069–2081 (2023)
61. Wan, B., Zhou, X., Sun, Y., Zhu, Z., Wang, H., Yan, C., et al.: Tmnet: Triple-modal interaction encoder and multi-scale fusion decoder network for vdt salient object detection. Pattern Recognition **147**, 110074 (2024)
62. Wan, Z., Zhang, P., Wang, Y., Yong, S., Stepputtis, S., Sycara, K., Xie, Y.: Sigma: Siamese mamba network for multi-modal semantic segmentation. In: 2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV). pp. 1734–1744. IEEE (2025)
63. Wang, G., Li, C., Ma, Y., Zheng, A., Tang, J., Luo, B.: Rgb-t saliency detection benchmark: Dataset, baselines, analysis and a novel approach. In: Chinese Conference on Image and Graphics Technologies. pp. 359–369. Springer (2018)
64. Wang, K., Chen, K., Li, C., Tu, Z., Luo, B.: Alignment-free rgb-t salient object detection: A large-scale dataset and progressive correlation network. In: Proceedings of the AAAI conference on artificial intelligence. vol. 39, pp. 7780–7788 (2025)
65. Wang, L., Lu, H., Wang, Y., Feng, M., Wang, D., Yin, B., Ruan, X.: Learning to detect salient objects with image-level supervision. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 136–145 (2017)
66. Wang, Y., Wang, R., Fan, X., Wang, T., He, X.: Pixels, regions, and objects: Multiple enhancement for salient object detection. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10031–10040 (2023)
67. Woo, S., Debnath, S., Hu, R., Chen, X., Liu, Z., Kweon, I.S., Xie, S.: Convnext v2: Co-designing and scaling convnets with masked autoencoders. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 16133–16142 (2023)
68. Wu, Y.H., Liu, Y., Zhang, L., Cheng, M.M., Ren, B.: Edn: Salient object detection via extremely-downsampled network. IEEE Transactions on Image Processing **31**, 3125–3136 (2022)
69. Yan, Q., Xu, L., Shi, J., Jia, J.: Hierarchical saliency detection. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1155–1162 (2013)
70. Yang, C., Chen, Z., Espinosa, M., Ericsson, L., Wang, Z., Liu, J., Crowley, E.J.: Plainmamba: Improving non-hierarchical mamba in visual recognition. arXiv preprint arXiv:2403.17695 (2024)
71. Yang, C., Zhang, L., Lu, H., Ruan, X., Yang, M.H.: Saliency detection via graph-based manifold ranking. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 3166–3173 (2013)
72. Zhang, M., Liu, J., Wang, Y., Piao, Y., Yao, S., Ji, W., Li, J., Lu, H., Luo, Z.: Dynamic context-sensitive filtering network for video salient object detection. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 1553–1563 (2021)
73. Zhang, Z., Wang, J., Han, Y.: Saliency prototype for rgb-d and rgb-t salient object detection. In: Proceedings of the 31st ACM international conference on multimedia. pp. 3696–3705 (2023)

74. Zhao, S., Chen, H., Zhang, X., Xiao, P., Bai, L., Ouyang, W.: Rs-mamba for large remote sensing image dense prediction. IEEE Transactions on Geoscience and Remote Sensing **62**, 1–14 (2024)
75. Zhao, X., Liang, H., Li, P., Sun, G., Zhao, D., Liang, R., He, X.: Motion-aware memory network for fast video salient object detection. IEEE Transactions on Image Processing **33**, 709–721 (2024)
76. Zhou, T., Fu, H., Chen, G., Zhou, Y., Fan, D.P., Shao, L.: Specificity-preserving rgb-d saliency detection. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 4681–4691 (2021)
77. Zhou, W., Zhu, Y., Lei, J., Yang, R., Yu, L.: Lsnet: Lightweight spatial boosting network for detecting salient objects in rgb-thermal images. IEEE Transactions on Image Processing **32**, 1329–1340 (2023)
78. Zhou, Z., Pei, W., Li, X., Wang, H., Zheng, F., He, Z.: Saliency-associated object tracking. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 9866–9875 (2021)
79. Zhu, E., Chen, Z., Wang, D., Shi, H., Liu, X., Wang, L.: Unetmamba: An efficient unet-like mamba for semantic segmentation of high-resolution remote sensing images. IEEE Geoscience and Remote Sensing Letters **22**, 1–5 (2024)
80. Zhu, L., Liao, B., Zhang, Q., Wang, X., Liu, W., Wang, X.: Vision mamba: Efficient visual representation learning with bidirectional state space model. arXiv preprint arXiv:2401.09417 (2024)
81. Zhuge, M., Fan, D.P., Liu, N., Zhang, D., Xu, D., Shao, L.: Salient object detection via integrity learning. IEEE Transactions on Pattern Analysis and Machine Intelligence **45**(3), 3738–3752 (2022)