# HyPSAM: Hybrid Prompt-driven Segment Anything Model for RGB-Thermal Salient Object Detection

Ruichao Hou, *Member, IEEE*, Xingyuan Li, Tongwei Ren, *Member, IEEE*, Dongming Zhou, Gangshan Wu, *Member, IEEE* and Jinde Cao, *Fellow, IEEE*



***Abstract*—RGB-thermal salient object detection (RGB-T SOD) aims to identify prominent objects by integrating complementary information from RGB and thermal modalities. However, learning the precise boundaries and complete objects remains challenging due to the intrinsic insufficient feature fusion and the extrinsic limitations of data scarcity. In this paper, we propose a novel hybrid prompt-driven segment anything model (HyPSAM), which leverages the zero-shot generalization capabilities of the segment anything model (SAM) for RGB-T SOD. Specifically, we first propose a dynamic fusion network (DFNet) that generates high-quality initial saliency maps as visual prompts. DFNet employs dynamic convolution and multi-branch decoding to facilitate adaptive cross-modality interaction, overcoming the limitations of fixed-parameter kernels and enhancing multi-modal feature representation. Moreover, we propose a plug-and-play refinement network (P2RNet) which serves as a general optimization strategy to guide SAM in refining saliency maps by using hybrid prompts. The text prompt ensures reliable modality input, while the mask and box prompts enable precise salient object localization. Extensive experiments on three public datasets demonstrate that our method achieves state-of-the-art performance. Notably, HyPSAM has remarkable versatility, seamlessly integrating with different RGB-T SOD methods to achieve significant performance gains, thereby highlighting the potential of prompt engineering in this field. The code and results of our method are available at: https://github.com/milotic233/HyPSAM.**

***Index Terms*—RGB-thermal, salient object detection, dynamic convolution, hybrid prompts, segment anything model.**

## I. Introduction

SALIENT object detection (SOD) aims to identify and segment the most visually attractive and prominent objects within a given scene [1], which is foundational for numerous downstream applications, such as image retrieval [2], visual object tracking [3], and camouflaged object detection [4]. While unimodal SOD methods [5]–[8] may suffer performance degradation in challenging scenarios, *e.g.*, low illumination and cluttered backgrounds, recent studies have incorporated the complementary thermal infrared spectrum to enhance object perception, giving rise to the RGB-thermal (RGB-T) SOD task.

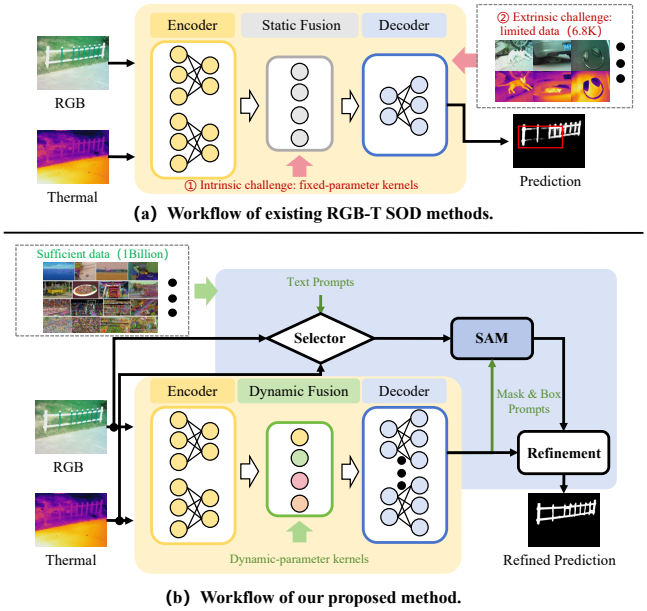



Fig. 1: Comparisons of the workflows of existing RGB-T SOD methods and our method. (a) Existing methods typically adopt the encoder-decoder paradigm, incorporating well-designed fusion strategies. (b) Our method designs a dynamic fusion network combined with hybrid prompts to guide SAM for accurate saliency predictions.

Existing works in RGB-T SOD, whether employing single stream [9], dual stream [10]–[36] and triple stream architectures [37], share a common goal of exploring complementarity from dual modalities to improve the multi-modal representation. As illustrated in Fig. 1(a), most of them follow a similar workflow, but two key challenges remain unresolved. First, the inherent differences between RGB and thermal images complicate the extraction of complementary features. Existing fusion mechanisms often rely on static convolutional kernels and complex attention designs, which may lack adaptability to diverse scenes. Second, high-quality RGB-T annotations are costly to obtain, limiting available training data and making models prone to overfitting or poor generalization. As shown in Fig. 2, these issues frequently lead to incomplete object detection and blurred boundaries, especially in complex scenarios.

In this paper, we propose a novel hybrid prompt-driven

This work was supported by the National Natural Science Foundation of China (62072232), the Key R&D Project of Jiangsu Province (BE2022138), the Fundamental Research Funds for the Central Universities (021714380026), the program B for Outstanding Ph.D. candidate of Nanjing University, and the Collaborative Innovation Center of Novel Software Technology and Industrialization. *(Corresponding authors: Tongwei Ren)*

Ruichao Hou, Xingyuan Li, Tongwei Ren, Gangshan Wu are with the State Key Laboratory for Novel Software Technology, Nanjing University, Nanjing 210008, China (e-mail: rchou@nju.edu.cn; lixy@smail.nju.edu.cn; rentw@nju.edu.cn; gswu@nju.edu.cn).

Dongming Zhou is with the School of Information Science and Engineering, Yunnan University, Kunming 650091, China (e-mail: zhoudm@ynu.edu.cn).

Jinde Cao is with the School of Mathematics, Southeast University, Nanjing 210096, China (e-mail: jdcao@seu.edu.cn).

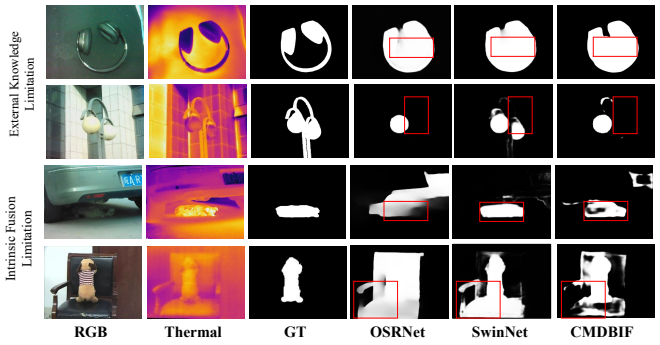



Fig. 2: Qualitative comparison illustrating two key limitations: external knowledge limitation (top two rows) leads to incomplete semantic understanding, while intrinsic fusion limitation (bottom two rows) causes inaccurate modality fusion. Red boxes highlight the incorrect saliency regions.

segment anything model (HyPSAM) that leverages the zero-shot generalization capabilities of the segment anything model (SAM) [38] for RGB-T SOD. As shown in Fig.1(b), HyPSAM integrates a coarse-to-fine framework by combining a robust saliency prediction network with a SAM-based refinement network. However, directly applying SAM to RGB-T SOD is nontrivial due to two key limitations. First, as a general-purpose model, SAM relies on task-specific and well-designed prompts to accurately identify salient objects. Second, being trained solely on RGB images, SAM lacks inherent support for multi-modal inputs.

To this end, we introduce two core components. First, a dynamic fusion network (DFNet) is designed to improve intrinsic saliency detection by adaptively merging RGB and thermal features via context-aware dynamic convolutions and a multi-branch decoder. Second, a plug-and-play refinement network (P2RNet) is proposed to adapt SAM through hybrid prompts and modality-aware input selection. Specifically, a CLIP-based modality selector is used to choose the most informative modality based on scene semantics. Meanwhile, multi-level geometric prompts, derived from the initial saliency map, are introduced to guide SAM in refining predictions. Extensive experiments on three public RGB-T benchmarks demonstrate that HyPSAM outperforms state-of-the-art methods and its seamless integration with existing RGB-T SOD methods without task-specific training.

The main contributions are summarized as follows:

- We propose a novel hybrid prompt-driven framework that breaks the performance bottleneck by improving intrinsic feature fusion and embedding external rich semantic knowledge.
- We propose a dynamic fusion network that improves detection accuracy by employing context-aware dynamic convolutions and multi-branch decoding. It enables adaptive cross-modality interaction, overcoming the limitations of fixed-parameter kernels and improving feature fusion.
- We propose a plug-and-play refinement network to alleviate the limitations of data scarcity. By utilizing hybrid prompts, comprising text, masks, and boxes, it adapts SAM for RGB-T SOD tasks without additional training, achieving precise results and improved generalization.

## II. Related Work

### A. RGB-T Salient Object Detection

RGB-T SOD aims to enhance scene understanding by incorporating the advantages of RGB and thermal modality, thus accurately segmenting the common saliency regions. Recent advancements in this field have been largely driven by deep learning, with methods broadly categorized into three types based on architecture: single-stream, dual-stream, and triple-stream [9], [12]–[37].

Among these, dual-stream architectures are the most prevalent due to their effectiveness in capturing cross-modal interactions. For example, Ma *et al.* [13] proposed a modality complementary fusion network that mitigates the negative impact of low-quality modalities from both global and local perspectives. Liu *et al.* [12] designed a dual-stream encoder based on the Transformer, which improves detection performance by optimizing cross-modal features through spatial and channel recalibration modules. More recently, Jin *et al.* [36] proposed a lightweight local and global perception network that integrates convolutional inductive bias with Transformer-based global modeling to achieve efficient multimodal fusion.

Single-stream architectures are particularly advantageous for real-time applications. For example, OSRNet [9] aggregated multi-modal information at the early fusion stage using operations such as cascading and element-wise computation, achieving faster inference.

Three-stream architectures, such as MDBIFNet [37], utilize the third stream as an additional layer of interaction, enhancing the integration of complementary data from both modalities. The interaction branch strengthens cross-modal correlations, improving the overall performance of saliency detection.

Nevertheless, most existing RGB-T SOD methods rely heavily on static fusion modules with fixed parameters, which often struggle to adaptively capture the complex and dynamic modality relationships encountered in real-world scenarios. In contrast, our proposed framework introduces a dynamic fusion mechanism and incorporates hybrid prompt strategies to guide the foundation model, effectively alleviating modality bias and significantly enhancing generalization performance across diverse conditions.

### B. Dynamic Multi-Modal Fusion

Dynamic fusion techniques have become increasingly important for multi-modal saliency detection due to their ability to adaptively model uncertain modality relationships. Unlike static fusion schemes with fixed parameters, dynamic fusion adjusts feature interactions based on input content, allowing more flexible and robust feature representation.

A major research direction is dynamic convolution, which generates input-dependent convolutional kernels to adaptively modulate feature extraction [39]–[41]. For example, Pang *et al.* [42] proposed a hierarchical dynamic filtering network for cross-modal feature fusion, which significantly improves RGB-D salient object detection. Zhang *et al.* [43] developed an interlaced dynamic filtering network by decoupling dynamic convolution, dynamically combining discriminative RGB-D

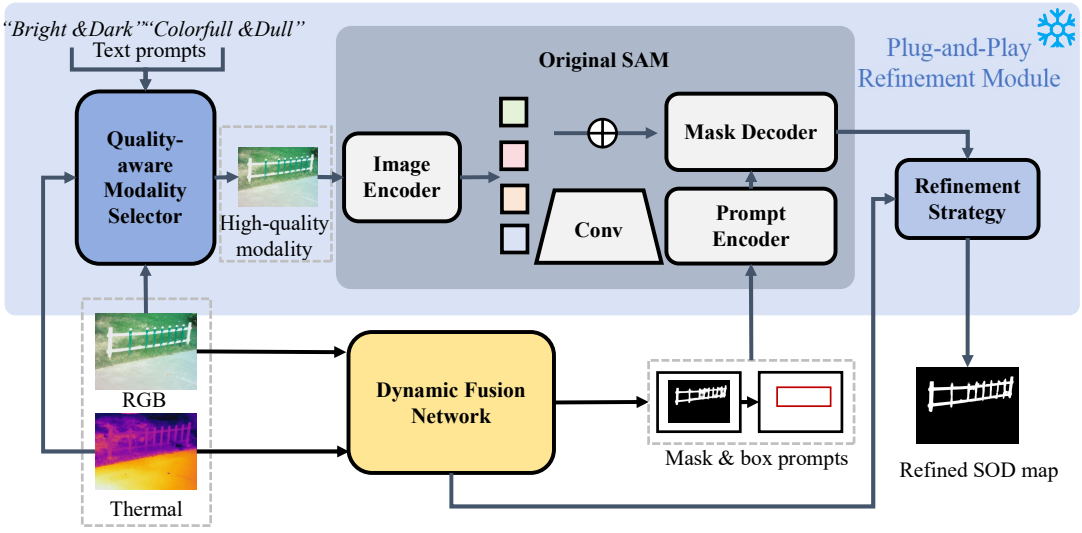



Fig. 3: The framework of our proposed HyPSAM. We first feed the RGB-T image pairs into the DFNet to generate the saliency map, which is then used in conjunction with the augmentation strategy to create hybrid prompts. Additionally, the image pairs are processed by the quality-aware modality selector to select the most effective modality as input for SAM. Finally, the mask output from SAM and the saliency map are fused by the refinement strategy to obtain the final refined results.

features to boost detection performance. Beyond convolution, dynamic learning strategies have also been explored. For example, Jin *et al.* [44] proposed a dual-stage self-paced learning framework combined with depth emphasis for underwater salient object detection.

Nonetheless, many dynamic filters remain RGB-centric and construct dynamic convolutional kernels along restricted dimensions, which limits their robustness. In contrast, our method adopts a simple yet effective dynamic convolution mechanism applied symmetrically to both modalities, which alleviates the issues caused by degraded quality.

### C. SAM and Its Application

SAM [38], trained on the SA-1B dataset with over one billion masks, has demonstrated exceptional visual representation. As a foundational model for image segmentation, it is renowned for its zero-shot transfer capabilities, allowing users to perform segmentation through simple prompts.

Numerous studies have explored SAM's applications across various imaging domains. For example, Ma *et al.* [45] proposed MedSAM, which is designed for universal medical image segmentation, supported by a large-scale medical image dataset. Wang *et al.* [46] applied SAM to remote sensing, creating SAMRS, an architecture for generating large-scale remote sensing segmentation datasets. Furthermore, SAM has also been extended to multi-modal vision tasks. For example, Yu *et al.* [47] proposed a depth-aware camouflage object detection and segmentation model that leverages the zero-shot capabilities of SAM to achieve precise segmentation in the RGB-D domain. Fang *et al.* [48] developed a new RGB-T crowd-counting method using semantic maps from SAM to distinguish between foreground and background for guiding cross-modal feature fusion. Wu *et al.* [49] proposed SAGE, a multi-modality image fusion framework that distills semantic priors from SAM to balance visual quality and downstream task adaptability without relying on SAM during inference. Zhai *et al.* [50] propose a SAM-guided label optimization and progressive cross-modal cross-scale fusion framework for weakly supervised RGB-T salient object detection using scribble annotations.

While most existing works adopt adapter-based modifications or label generation, our HyPSAM directly utilizes SAM by integrating prompts with RGB-T saliency priors through a plug-and-play refinement framework without any additional task-specific training.

## III. Methodology

The framework of HyPSAM is illustrated in Fig. 3, where DFNet and P2RNet are integrated to enhance generalization and improve saliency detection in complex scenarios. RGB-T image pairs are first processed by DFNet to generate hybrid prompts. Meanwhile, a quality-aware modality selector automatically determines the more reliable modality as the sole input to SAM, ensuring compatibility with its input format. Notably, SAM is used without any fine-tuning, with all components (image encoder, prompt encoder, and mask decoder) kept completely frozen. Hybrid prompts are injected exclusively through SAM's standard prompt encoder, fully leveraging its inherent prompt-driven segmentation capability without modifying its architecture. Finally, a refinement strategy is applied to produce the final saliency map.

### A. Dynamic Fusion Network

Figure 4 illustrates the architecture of DFNet. We first utilize the Swin Transformer [51] as the backbone to construct a symmetric dual-stream encoder, which combines the strengths of both Transformer and CNN, demonstrating robust capabilities in global and local feature representation. To reduce computational complexity, $1 \times 1$ convolutions compress the feature channels to 64, and only the last four

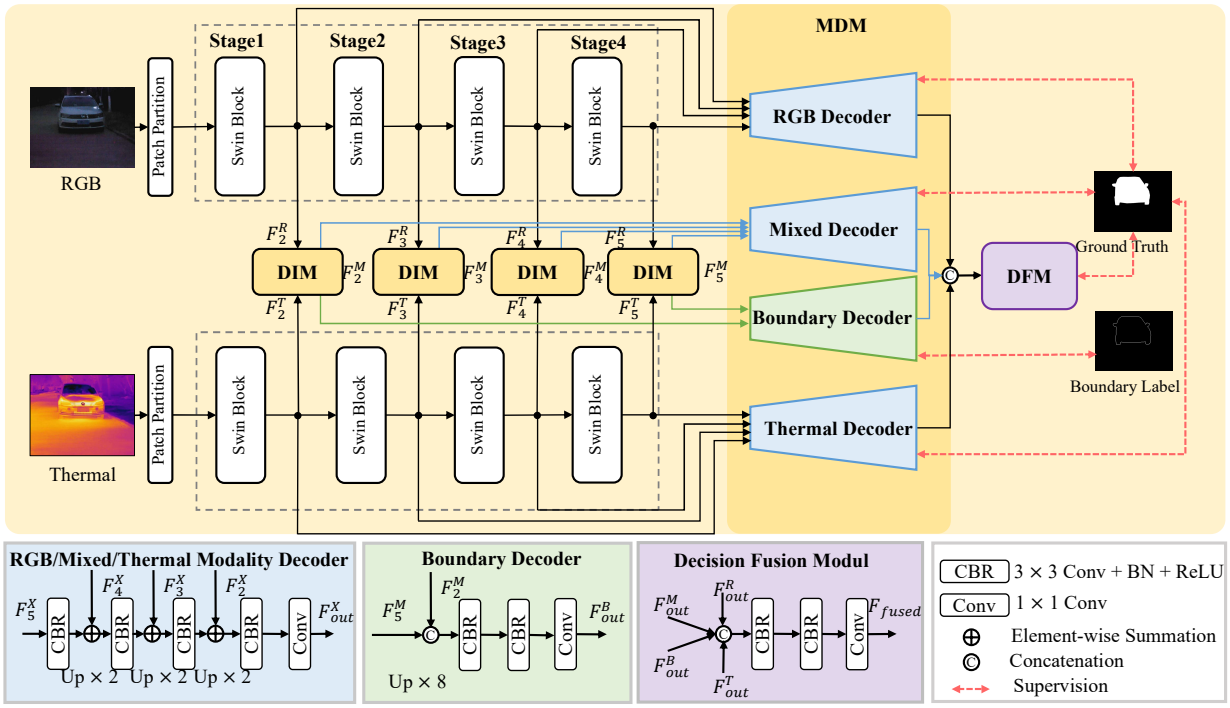



Fig. 4: The architecture of DFNet. We feed the RGB-T image pairs into a Swin Transformer-based backbone to extract features. These features are then enhanced through the dynamic interaction module. Next, the multi-branch decoding module establishes dedicated branches for each modality to decode their respective features effectively. Finally, the decision fusion module combines the predictions from multiple branches, generating the saliency map.

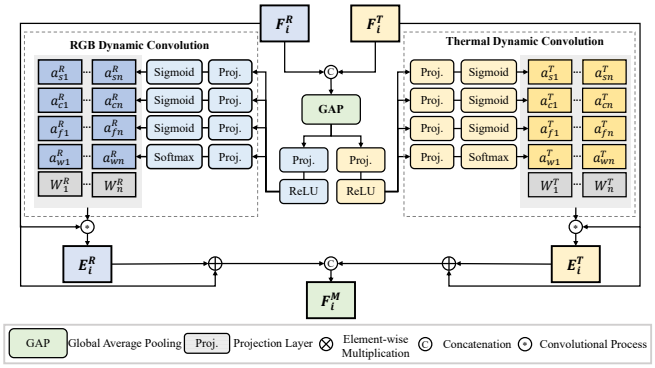



Fig. 5: Details of the dynamic interaction module in DFNet.

layers of features are considered for interaction and decoding. Mathematically, the hierarchical RGB-T features are denoted as $\{F_i^R | i = 2, 3, 4, 5\}$ and $\{F_i^T | i = 2, 3, 4, 5\}$, respectively. Next, we design the dynamic interaction module to enhance multi-modal features $\{F_i^M | i = 2, 3, 4, 5\}$ by dynamically integrating RGB and thermal cues, improving robustness in complex scenarios. The multi-branch decoding module predicts single-modal and mixed-modal saliency results and boundary details. Finally, the decision fusion module merges the predictions from multiple branches and generates the initial saliency map.

*1) Dynamic Interaction Module:*

To address the limitations of fixed convolutional kernels, we design a dynamic interaction module that adaptively enhances features for each modality. Unlike prior dynamic methods [42], [43] that typically produce filters conditioned on fused features to enhance the RGB stream alone, our method treats both RGB and thermal inputs equally, avoiding modality dominance bias. Inspired by the omni-dimensional dynamic convolution (ODConv) [52], we employ multi-dimensional attention to generate dynamic convolutional weights tailored to each modality, as shown in Fig. 5.

Specifically, we first concatenate the modality-specific features to obtain the fused representation:

$$F_i^f = \text{Concat}(F_i^R, F_i^T). \tag{1}$$

where $F_i^R$ and $F_i^T$ denote the features from the RGB and thermal modalities, respectively, and Concat$(\cdot)$ indicates channel-wise concatenation.

Next, global contextual information is extracted by applying global average pooling (GAP) followed by a linear projection and non-linear activation:

$$V_i^x = \text{ReLU}(\text{Proj}^x(\text{GAP}(F_i^f))), \quad x \in \{R, T\}, \tag{2}$$

where GAP$(\cdot)$ denotes global average pooling, Proj$(\cdot)$ is linear projection layer, ReLU$(\cdot)$ is non-linear activation function.

The context vector $V_i$ is then used to compute dimension-specific attention scalars for each modality through multiple parallel projection branches. For modality $x \in \{R, T\}$, the attention weights across four dimensions are calculated as:

$$\begin{cases} a_{sn}^x = \text{Sigmoid}(\text{Proj}_s^x(V_i^x)) \\ a_{cn}^x = \text{Sigmoid}(\text{Proj}_c^x(V_i^x)) \\ a_{fn}^x = \text{Sigmoid}(\text{Proj}_f^x(V_i^x)) \\ a_{wn}^x = \text{Softmax}(\text{Proj}_w^x(V_i^x)) \end{cases}, \quad n = 1, 2, \ldots, N \tag{3}$$

where $N$ is the number of convolutional kernels, and Sigmoid or Softmax normalize the weights across corresponding dimensions, $a^x_{wn}$, $a^x_{fn}$, $a^x_{cn}$, and $a^x_{sn}$ correspond to kernel, filter, channel, and spatial attention weights, respectively.

The dimension-specific attention scalars are employed to adaptively reweight the convolutional parameters, enhancing cross-modal feature interactions in a fine-grained and effective manner. The context-aware dynamic convolution for each modality is then formulated as:

$$E^x_i = \sum_{n=1}^{N} \left(a^x_{wn} \odot a^x_{fn} \odot a^x_{cn} \odot a^x_{sn} \odot W^x_n\right) * F^x_i, \quad (4)$$

where $F_i$ denotes input features, $E_i$ represents the enhanced features, $*$ indicates the convolution operation, while $\odot$ represents the multiplication along different dimensions. Here, $n$ denotes the index of the $n$-th convolutional kernel, with each kernel $W_n$ being modulated by the four attention dimensions before aggregation.

Finally, the enhanced mixed features are obtained by fusing the original input and enhanced features of both modalities:

$$F^M_i = \text{Concat}(E^R_i + F^R_i,\ E^T_i + F^T_i), \quad (5)$$

where $E^R_i$ and $E^T_i$ indicate the enhanced features of RGB and thermal modalities, respectively.

*2) Multi-branch Decoding Module:* To address the challenges posed by modality quality discrepancies and false positive detection, the multi-branch decoding module enhances saliency prediction by decoupling the RGB and thermal modalities while preserving discriminative features. Unlike single-decoder methods, our module independently processes RGB, thermal, and mixed-modality features, enabling tailored predictions for each modality. By leveraging hierarchical fusion, our module progressively aggregates deep and shallow layer features, utilizing their complementary benefits to improve prediction accuracy. The decoding process can be formulated as follows:

$$\tilde{F}^x_3 = \text{Up}(\text{CBR}(\text{Up}(\text{CBR}(F^x_5)) \oplus F^x_4)) \oplus F^x_3, \quad (6)$$

$$F^x_{out} = \text{CBR}(\text{Up}(\text{CBR}(\tilde{F}^x_3)) \oplus F^x_2), x \in \{R, T, M\}, \quad (7)$$

where $F^x_{out}$ and $\tilde{F}^x_3$ indicate the output and $3rd$ layer decoding features, CBR means the sub-network consists of Convolution, Batch Normalization, and ReLU operations, Up denotes bilinear interpolation, $\oplus$ is the element-wise addition, $x \in \{R, T, M\}$ represents the RGB, thermal and mixed-modality, respectively.

Moreover, to preserve the detailed edges in the saliency results, we develop a lightweight boundary decoder. Inspired by previous work [53], [54], we use the Canny operator for label decoupling, separating binary labels into boundary and content parts, and enabling supervised training on boundary details. Given that shallow features contain rich edge details but are also susceptible to noise, we fuse deep features from the $5th$ layer with shallow features from the $2nd$ layer to construct cross-level fused features for precise edge prediction. The boundary decoder is defined as follows:

$$F^B_{out} = \text{CBR}_{\times 2}(\text{Concat}(\text{Up}(F^M_5), F^M_2))), \quad (8)$$

where $F^B_{out}$ is the boundary decoding features, $\text{CBR}_{\times 2}$ indicates the CBR is stacked two times.

*3) Decision Fusion Module:* In addition, we introduce a decision fusion module to produce a final saliency map that aggregates the decoding features from multiple branches. This module is jointly trained with the decoders described above. The details of the architecture are illustrated in Fig. 3. We first concatenate the output features from the four branches. Then, we apply convolutional layers to fuse different features. The fused saliency result can be described as follows:

$$F_f = \text{CBR}_{\times 2}(\text{Concat}(F^M_{out}, F^R_{out}, F^T_{out}, F^B_{out})). \quad (9)$$

*4) Loss Function:* To strengthen the ability of the multi-branch decoding module to capture discriminative features, different loss functions are assigned to each branch for supervision. For the boundary prediction, class imbalance is a critical challenge due to the sparse nature of edge pixels in the overall image. To address this, the dice loss [55] is used to supervise the boundary decoder. Additionally, the other decoders are supervised by the hybrid loss [53]:

$$\ell_{hyb} = \ell_{bce} + \ell_{ssim} + \ell_{iou}, \quad (10)$$

where $\ell_{bce}$, $\ell_{ssim}$ and $\ell_{iou}$ denote BCE loss [56], SSIM loss [57] and IoU loss [58], respectively. The total loss $\mathcal{L}$ is expressed as follows:

$$\mathcal{L} = \ell^R_{hyb} + \ell^T_{hyb} + \ell^M_{hyb} + \ell^B_{dice} + \ell^F_{hyb}, \quad (11)$$

where $\ell^R_{hyb}$, $\ell^T_{hyb}$, $\ell^M_{hyb}$, $\ell^B_{dice}$ and $\ell^F_{hyb}$ indicate RGB, thermal, mixed-modality, boundary and fusion loss, respectively.

### B. Plug-and-Play Refinement Network

Although the proposed DFNet has achieved excellent performance on publicly available RGB-T datasets, its generalizability is restricted by the limited scale and diversity of the training data. Consequently, it still encounters misclassification in challenging or unseen scenarios. Such inaccuracies in saliency maps hinder the practical application of SOD methods, as precise and reliable outputs are essential for real-world interaction.

Existing approaches [59], [60] often incorporate the SAM encoder as a backbone or design adapters to enhance feature extraction. Nevertheless, these manners usually incur additional training costs. To address this problem, we propose a P2RNet, which serves as a general optimization network to further refine coarse saliency maps, as shown in Fig. 3 (blue background). P2RNet aims to simply leverage SAM from multiple perspectives without further modifying its weights or architecture. The refinement process begins with RGB-T image pairs being processed by a quality-aware modality selector, which evaluates the quality of both modalities and selects the optimal one for further processing. Based on this evaluation, the augmentation strategy generates high-quality visual prompts, including masks and bounding boxes, derived from the initial saliency map. These hybrid prompts guide the SAM to precisely distinguish the saliency object. The final saliency maps are refined by incorporating both the

initial predictions and segmentation masks, resulting in sharper boundaries and more complete object structures.

*1) Quality-aware Modality Selector:* Since RGB-T images are well-aligned and represent the same scene, both modalities share inherent semantic similarities, allowing SAM to process thermal images to some extent. Therefore, we attempt to feed the reliable quality modality into SAM and achieve good segmentation results without any task-specific fine-tuning. Nevertheless, image degradation caused by various factors makes accurate quality assessment challenging, thus, a simple binary classifier explicitly trained on limited labeled datasets may fail to reliably evaluate image quality (IQA). Inspired by advances in image quality assessment, we directly employ CLIP-IQA [61] as a quality-aware modality selector, as shown in Fig. 6. CLIP is a foundational model trained in contrastive learning, primarily used for the cross-modal alignment of language and images. CLIP-IQA applies the large model to the IQA task, considering both quality and abstract perception. It calculates the cosine similarity between the image features and antonym prompts and then uses Softmax to compute the final quality score.

$$s_i = \frac{f \odot t_i}{\|f\| \cdot \|t_i\|}, i \in \{1, 2\}, \tag{12}$$

where $f$ is the image features and $t_i$ are the features from antonym prompts, $\odot$ is the dot product and $\|\cdot\|$ denotes $\ell_2$ norm.

$$s = \frac{e^{s_1}}{e^{s_1} + e^{s_2}}, \tag{13}$$

where $s_1$ and $s_2$ are the scores corresponding to the positive and negative quality descriptions, respectively. The exponentiation operation amplifies the difference between the two similarity scores, making the output score more sensitive to the dominant attribute.

To enhance reliability, we adopt two pairs of antonym prompts, *i.e.*, "bright and dark" and "colorful and dull", to reduce the ambiguity of the prompt. The "bright and dark" prompt assesses the illumination conditions of the scene. If the confidence score $s_\alpha$ exceeds the threshold $\tau$, the RGB image is determined to be more reliable. The "colorful and dull" prompt evaluates the richness of the scene information. If the quality score $s_\beta$ exceeds the threshold $\theta$, the RGB image is of higher quality. Otherwise, the thermal modality is selected. The effective combination of two text prompts allows for a robust selection of the most suitable modality.

$$m_{re} = \begin{cases} m_r & \text{if } s_\alpha > \tau \text{ or } s_\beta > \theta, \\ m_t & \text{others.} \end{cases} \tag{14}$$

*2) Prompt Augmentation Strategy:* SAM is an interactive segmentation model that supports diverse types of prompts, including points, boxes, text, and masks. Previous work, such as SAMAug [62], has demonstrated that augmenting point-based prompts can boost segmentation performance, underscoring the potential of visual prompt engineering. However, relying exclusively on sparse point prompts may result in incomplete objects, particularly for complex or irregular object shapes.

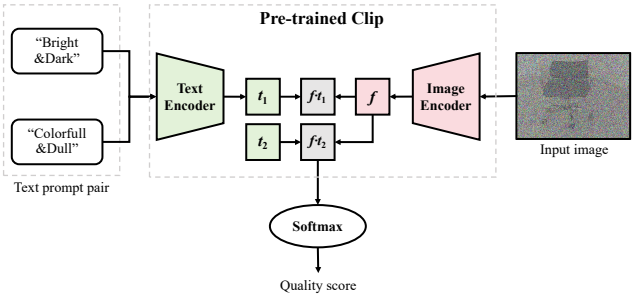



Fig. 6: Details of the quality-aware modality selector.

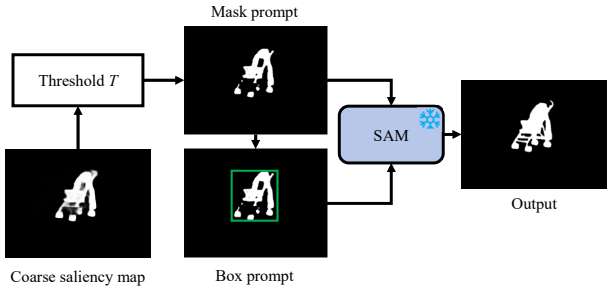



Fig. 7: Illustration of saliency prompts generation.

To address this limitation, we propose a hybrid prompt augmentation strategy that combines both mask and box prompts, providing comprehensive global and local guidance for salient object representation. Specifically, we begin with a coarse saliency map, which serves as an initial approximation of object regions. Given that this map may contain noise and exhibit ambiguous boundaries, especially near object edges, we apply a binarization operation with a threshold $T$ to produce a refined binary mask $S_t$, which helps suppress uncertain regions and reduces the impact of blurred contours on SAM's segmentation results. We then use $S_t$ as a mask prompt by integrating it into the image embedding via element-wise addition, allowing SAM to attend to spatially informative regions. Furthermore, we extract bounding boxes around the $S_t$ to generate box prompts. These box-level cues impose additional spatial constraints, which enhance boundary precision and improve segmentation robustness. The entire hybrid prompt generation process is illustrated in Fig. 7.

*3) Refinement Strategy:* While hybrid prompts improve SAM's segmentation accuracy, complex saliency regions can still lead to ambiguities, such as hollow objects. To address this issue, we propose a refinement strategy designed to preserve both the salient regions while sharpening edges. Therefore, we merge the coarse saliency map $S_t$ with the segmentation mask $S_g$ to produce the refined map $S_r$, maintaining the integrity of the objects. The refinement process is described as follows:

$$S_r(x, y) = \max\{S_t(x, y), S_g(x, y)\}, \tag{15}$$

where the maximum operation is applied element-wise across the two maps, combining their strengths for optimal saliency representation.

## IV. Experiments

### A. Datasets and Evaluation Metrics

We train and test our method on three well-known RGB-T benchmark datasets, namely VT821 [28], VT1000 [29], and VT5000 [30]. Specifically, the VT821 dataset includes 821 pairs of well-aligned RGB-T images and their corresponding ground-truth labels, covering approximately 60 scenes, and some samples have added noise interference to enhance the challenge of the dataset. The VT1000 dataset expands the VT821 dataset, containing 1,000 pairs of RGB-T images and ground-truth labels collected in real-life scenarios. The VT5000 dataset consists of 5,000 pairs of RGB-T images, with significantly enhanced diversity in objects and scenes, which includes 13 challenging attributes, *i.e.*, big salient object (BSO), center bias (CB), cross image boundary (CIB), image clutter (IC), low illumination (LI), multiple salient objects (MSO), out of focus (OF), small salient object (SSO), similar appearance (SA), thermal cross (TC), bad weather (BW), bad RGB (bRGB), and bad thermal (bT), respectively.

Following the common setting in these methods [12], [23], we select 2,500 image pairs from VT5000 as the training set, while the remaining 2,500 image pairs in VT5000, along with the VT821 and VT1000 datasets, constitute the testing set. We evaluate the performance of the model using widely adopted metrics [63]–[67], including mean F-measure ($F_{avg}$), maximum F-measure ($F_{max}$), weighted F-measure ($F_w$), MAE ($\mathcal{M}$), E-measure ($E_m$), S-measure ($S_m$), and Precision-Recall curve.

The F-measure is a comprehensive metric that evaluates the balance between precision and recall, defined as:

$$F_m = (\beta^2 + 1) \cdot \frac{precision \cdot recall}{\beta^2 \cdot precision + recall}, \tag{16}$$

where $\beta^2$ is empirically set to 0.3 to emphasize the precision over recall.

The $\mathcal{M}$ quantifies the average absolute error between the predicted map and ground truth, defined as:

$$\mathcal{M} = \frac{1}{W \times H} \sum_{x=1}^{W} \sum_{y=1}^{H} |S(x,y) - G(x,y)|, \tag{17}$$

where $S$ and $G$ denote the predicted saliency map and GT, respectively. $W$ and $H$ represent the width and height, respectively. $(x, y)$ denotes the pixel coordinates.

$E_m$ evaluates both global and local similarity between the predicted saliency map and the ground truth, defined as:

$$E_m = \frac{1}{W \times H} \sum_{x=1}^{W} \sum_{y=1}^{H} \varphi\big(S(x,y), G(x,y)\big), \tag{18}$$

where $\varphi$ denotes the matrix entry.

$S_m$ evaluates the structural similarity between the ground truth and the predicted saliency map, defined as:

$$S_m = \alpha S_o + (1 - \alpha) S_r, \tag{19}$$

where $S_o$ and $S_r$ represent the object-aware structural similarity and the region-aware structural similarity, respectively. The weight $\alpha$ is set to 0.5 to balance these two components.

### B. Implementation Details

HyPSAM is implemented on the PyTorch platform with an Intel i9 13900K CPU and dual NVIDIA RTX 4090 GPUs. The input samples are resized to a resolution of $384 \times 384$. During the training phase, the training samples are augmented in various ways, including random flipping, rotating, and clipping. The backbone of DFNet is initialized with a pre-trained SwinV2-B network [51]. Additionally, we train the network for 50 epochs with a batch size of 8 using the SGD optimizer. The learning rate for the backbone is set to $5e^{-3}$, while the learning rate of other parameters is set to $5e^{-2}$ and scheduled by the cosine strategy. The thresholds $\tau$ and $\theta$ in the quality-aware selector are set to 0.01 and 0.85, respectively. The remaining parameter settings follow CLIP-IQA. We select ViT-H as the pre-trained weights for SAM.

### C. Comparison with the State-of-the-art Methods

To evaluate the performance of the proposed HyPSAM, we compare it with 23 state-of-the-art RGB-T methods on public benchmarks, namely, SGDL [29], CSRNet [32], CGFNet [19], SwinNet [12], ADF [30], TNet [15], OSRNet [9], ACMANet [17], MCFNet [13], CMDBIF [37], CAVER [14], ADNet [23], WGOFNet [33], UMINet [26], TCINet [68], LAFB [34], SACNet [35], DSCDNet [69], FFANet [70], PATNet [71], ISMNet [72], ConTriNet [73] and KAN-SAM [74]. For a fair comparison, the saliency maps used for the comparison are provided by published works, and the metrics are calculated using the same evaluation toolbox [27].

*1) Quantitative Comparison:* Table I presents the comprehensive quantitative comparison of our method against the state-of-the-art approaches on RGB-T benchmarks. The results demonstrate that HyPSAM achieves significant performance improvements across all benchmarks, surpassing existing methods by a substantial margin.

To provide a deeper analysis, we compare our HyPSAM with representative encoder-decoder-based methods. Specifically, compared to OSRNet [9], a representative single-stream method, the $F_w$ metric, a critical measure for evaluating segmentation accuracy, shows significant improvements with HyPSAM, increasing by 10.4%, 5.3%, and 10.2% on the VT5000, VT1000, and VT821 datasets, respectively. These experimental results demonstrate that our method effectively integrates complementary information rather than simplistic single-stream fusion. In comparison to SwinNet [12], a dual-stream architecture optimized for edge details, HyPSAM achieves notable improvements in $F_w$, $E_m$, and $S_m$ on the VT5000 dataset by 6.5%, 2.1%, and 1.9%, respectively. These results demonstrate the superiority of our approach in enhancing dynamic feature interaction and preserving object boundary details. Compared to the triple-stream CMDBIF [37], HyPSAM improved the $S_m$ by 4.5%, 2.7%, and 5.0% on three benchmarks, respectively. This indicates that the saliency objects predicted by HyPSAM exhibit superior structural integrity. Moreover, compared with recent advanced methods *e.g.*, ISMNet [72], and ConTriNet [73], our method adaptively fuses complementary cues and absorbs semantic features provided by SAM, yielding satisfactory

TABLE I: Quantitative comparison with different RGB-T methods. The best and second-best results are highlighted in **bold** and underline. ↑ (↓) denotes larger is better (smaller is better). '−' indicates the code or result is not available.

| Methods | Pub.Info | VT5000 | | | | | | VT1000 | | | | | | VT821 | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | $F_{avg}$ ↑ | $F_{max}$ ↑ | $F_w$ ↑ | $\mathcal{M}$ ↓ | $E_m$ ↑ | $S_m$ ↑ | $F_{avg}$ ↑ | $F_{max}$ ↑ | $F_w$ ↑ | $\mathcal{M}$ ↓ | $E_m$ ↑ | $S_m$ ↑ | $F_{avg}$ ↑ | $F_{max}$ ↑ | $F_w$ ↑ | $\mathcal{M}$ ↓ | $E_m$ ↑ | $S_m$ ↑ |
| SGDLL [29] | TMM20 | 0.672 | 0.737 | 0.558 | 0.089 | 0.824 | 0.750 | 0.764 | 0.807 | 0.652 | 0.090 | 0.856 | 0.787 | 0.731 | 0.780 | 0.583 | 0.085 | 0.846 | 0.764 |
| CSRNet [32] | TCSVT21 | 0.811 | 0.857 | 0.796 | 0.042 | 0.905 | 0.868 | 0.877 | 0.918 | 0.878 | 0.024 | 0.925 | 0.918 | 0.831 | 0.880 | 0.821 | 0.038 | 0.909 | 0.885 |
| CGFNet [19] | TCSVT22 | 0.851 | 0.887 | 0.831 | 0.035 | 0.922 | 0.883 | 0.906 | 0.936 | 0.900 | 0.023 | 0.944 | 0.923 | 0.845 | 0.885 | 0.829 | 0.038 | 0.912 | 0.881 |
| SwinNet [12] | TCSVT22 | 0.865 | 0.915 | 0.846 | 0.026 | 0.942 | 0.912 | 0.896 | 0.948 | 0.894 | 0.018 | 0.947 | 0.938 | 0.847 | 0.903 | 0.818 | 0.030 | 0.926 | 0.904 |
| ADFF [30] | TMM22 | 0.778 | 0.863 | 0.722 | 0.048 | 0.891 | 0.864 | 0.847 | 0.923 | 0.804 | 0.034 | 0.921 | 0.91 | 0.717 | 0.804 | 0.627 | 0.077 | 0.843 | 0.810 |
| TNet [15] | TMM22 | 0.846 | 0.895 | 0.840 | 0.033 | 0.927 | 0.895 | 0.889 | 0.937 | 0.895 | 0.021 | 0.937 | 0.929 | 0.842 | 0.904 | 0.841 | 0.030 | 0.919 | 0.899 |
| OSRNet [9] | TIM22 | 0.823 | 0.866 | 0.807 | 0.040 | 0.908 | 0.875 | 0.892 | 0.929 | 0.891 | 0.022 | 0.935 | 0.926 | 0.814 | 0.862 | 0.801 | 0.043 | 0.896 | 0.875 |
| ACMANet [17] | KBS22 | 0.858 | 0.890 | 0.823 | 0.033 | 0.932 | 0.887 | 0.904 | 0.933 | 0.889 | 0.021 | 0.945 | 0.927 | 0.837 | 0.873 | 0.807 | 0.035 | 0.914 | 0.883 |
| MCFNet [13] | AI22 | 0.848 | 0.886 | 0.836 | 0.033 | 0.924 | 0.887 | 0.902 | 0.939 | 0.906 | 0.019 | 0.944 | 0.932 | 0.844 | 0.889 | 0.835 | 0.029 | 0.918 | 0.891 |
| CMDBIF [37] | TCSVT23 | 0.868 | 0.892 | 0.846 | 0.032 | 0.933 | 0.886 | 0.914 | 0.931 | 0.909 | 0.019 | 0.952 | 0.927 | 0.856 | 0.887 | 0.837 | 0.032 | 0.923 | 0.882 |
| CAVER [14] | TIP23 | 0.856 | 0.897 | 0.849 | 0.028 | 0.935 | 0.899 | 0.906 | 0.945 | 0.912 | 0.016 | 0.949 | 0.938 | 0.854 | 0.897 | 0.846 | 0.026 | 0.928 | 0.897 |
| ADNet [23] | MMA23 | 0.893 | 0.924 | 0.884 | 0.022 | 0.953 | 0.924 | 0.916 | 0.952 | 0.920 | 0.015 | 0.952 | 0.944 | 0.869 | 0.915 | 0.860 | 0.024 | 0.930 | 0.915 |
| WGOFNet [33] | TOMM24 | 0.883 | 0.912 | 0.873 | 0.025 | 0.945 | 0.911 | 0.919 | 0.946 | 0.922 | 0.016 | 0.951 | 0.940 | 0.875 | 0.911 | 0.868 | 0.025 | 0.934 | 0.908 |
| UMINet [26] | TVC24 | 0.831 | 0.877 | 0.820 | 0.035 | 0.919 | 0.882 | 0.892 | 0.935 | 0.896 | 0.021 | 0.941 | 0.926 | 0.791 | 0.849 | 0.782 | 0.054 | 0.879 | 0.905 |
| TCINet [68] | TCE24 | 0.905 | 0.927 | 0.900 | **0.019** | 0.959 | 0.925 | 0.925 | 0.943 | 0.928 | 0.014 | 0.956 | 0.944 | 0.882 | 0.910 | 0.879 | 0.021 | 0.942 | 0.915 |
| LAFB [34] | TCSVT24 | 0.857 | 0.893 | 0.841 | 0.030 | 0.931 | 0.893 | 0.905 | 0.937 | 0.905 | 0.018 | 0.945 | 0.932 | 0.843 | 0.884 | 0.817 | 0.034 | 0.915 | 0.884 |
| SACNet [35] | TMM24 | 0.901 | 0.922 | 0.888 | 0.021 | 0.957 | 0.917 | 0.923 | 0.949 | 0.927 | 0.014 | 0.958 | 0.942 | 0.868 | 0.904 | 0.859 | 0.025 | 0.932 | 0.906 |
| DSCDNet [69] | TCE24 | 0.888 | - | 0.881 | 0.023 | 0.949 | 0.918 | 0.921 | - | 0.927 | 0.014 | 0.955 | 0.946 | 0.876 | - | 0.873 | 0.022 | 0.940 | 0.915 |
| FFANet [70] | PR24 | 0.886 | - | - | 0.021 | 0.953 | 0.918 | 0.918 | - | - | 0.014 | 0.955 | 0.943 | 0.855 | - | - | 0.027 | 0.926 | 0.905 |
| PATNet [71] | KBS24 | 0.883 | 0.916 | 0.879 | 0.023 | 0.946 | 0.917 | 0.910 | 0.948 | 0.920 | 0.015 | 0.951 | 0.940 | 0.870 | 0.914 | 0.872 | 0.024 | 0.933 | 0.910 |
| ISMNet [72] | TCSVT25 | 0.885 | - | 0.876 | 0.025 | 0.945 | 0.913 | 0.922 | - | 0.924 | 0.014 | 0.954 | 0.942 | 0.886 | - | 0.881 | 0.021 | 0.945 | 0.917 |
| ConTriNet [73] | TPAMI25 | 0.898 | 0.927 | 0.895 | 0.020 | 0.956 | 0.923 | 0.917 | 0.943 | 0.923 | 0.015 | 0.953 | 0.941 | 0.878 | 0.914 | 0.875 | 0.022 | 0.940 | 0.916 |
| KAN-SAM [74] | ICME25 | 0.909 | 0.931 | 0.905 | 0.020 | 0.957 | 0.927 | 0.930 | 0.947 | 0.934 | 0.013 | 0.958 | 0.946 | 0.883 | 0.911 | 0.880 | 0.025 | 0.932 | 0.915 |
| DFNet | - | 0.899 | 0.933 | 0.898 | 0.020 | 0.958 | 0.930 | 0.920 | 0.956 | 0.930 | 0.013 | 0.955 | 0.950 | 0.879 | 0.925 | 0.881 | 0.022 | 0.940 | 0.926 |
| HyPSAM | - | **0.928** | **0.939** | **0.911** | **0.019** | **0.963** | **0.931** | **0.946** | **0.957** | **0.944** | **0.011** | **0.965** | **0.954** | **0.914** | **0.930** | **0.903** | **0.020** | **0.948** | **0.932** |

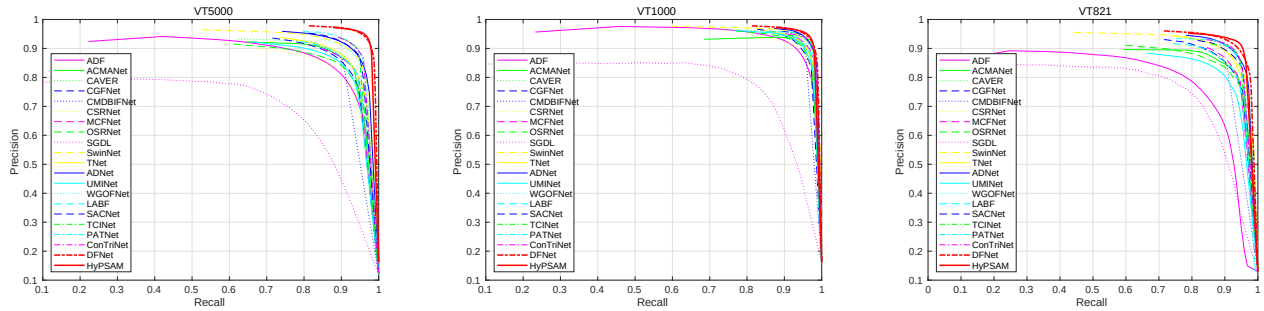



Fig. 8: Precision (vertical axis) - Recall (horizontal axis) curves comparison with the state-of-the-art RGB-T SOD methods on VT5000, VT1000, and VT821 datasets. The red solid curves show that our method outperforms the existing models.

detection results. In Fig. 8, we display the Precision-Recall curves on three RGB-T SOD datasets, which show that our method can obtain the highest precision on the three datasets.

*2) Attribute-based Quantitative Comparison:* To further verify the ability of different methods to handle complex scenarios, we conduct experiments on 13 challenging scenarios in the VT5000 dataset. We evaluate the $F_w$ of HyPSAM and 16 state-of-the-art methods. The comparison results for different challenging attributes are presented in Table II. The results demonstrate that HyPSAM consistently outperforms competing methods across all challenging outdoor scenarios.

Specifically, compared to the second-best method, ConTriNet [73], our model achieves significant performance gains of 3.2%, 2.1%, and 1.8% in MSO, bRGB, and bT scenarios, respectively. These gains highlight the ability of HyPSAM to resist interference, effectively mitigating the impact of challenging conditions such as low-light environments or low-quality input data. The superior performance can be attributed to context-aware dynamic convolutions, which adaptively handle multi-modal feature interactions to provide robust saliency prompts. Additionally, by dynamically selecting the highest-quality modality and integrating hybrid prompts, HyPSAM ensures robust detection even in the presence of degraded or ambiguous inputs.

*3) Qualitative Comparison:* We present representative results of the proposed method and nine RGB-T SOD methods in various challenging scenarios, as shown in Fig.9. In scenarios shown in Fig.9 (a) and (b), the object is not prominent in the thermal images, and other methods fail to effectively suppress interference from the thermal infrared modality, leading to incomplete objects, such as the back of the chair being inaccurately segmented. The inability to filter irrelevant thermal cues compromises object integrity. Figure 9 (c)–(f) depict scenes with significant challenges, where existing methods struggle to fully exploit complementary information from both modalities, leading to ambiguous object boundaries and hollow regions within salient objects. In the cases shown in Fig.9 (g) and (h), low illumination and noisy environments make it difficult to segment the pineapple and dog completely. The interference from poor-quality RGB images hinders the detection process. CAVER [14] may overly rely on the RGB modality, failing to

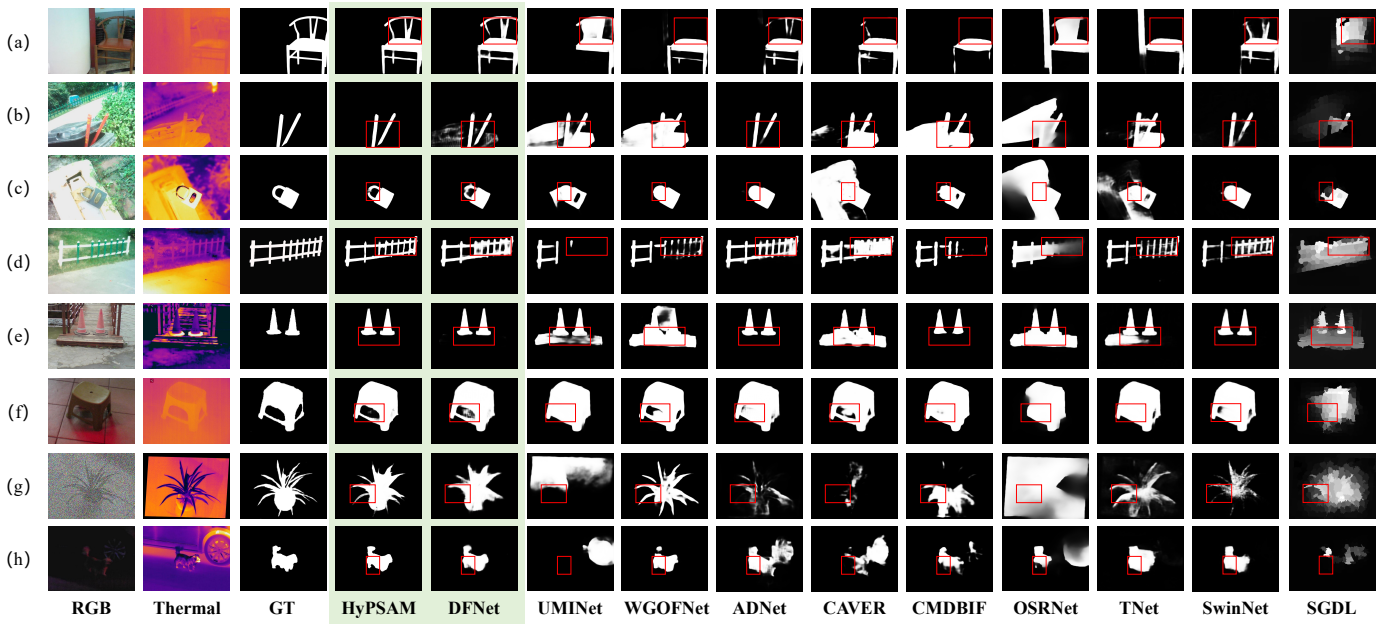



Fig. 9: Qualitative comparison with nine state-of-the-art methods. We select eight RGB-T image pairs with diverse challenges from three datasets for saliency map comparison. From left to right, the columns are RGB images, thermal images, ground truths, and the results of the ten methods. The key qualitative differences are highlighted with red bounding boxes.

TABLE II: Performance comparison (F-measure, $F_w$ ↑) on 13 challenging attributes of the VT5000. The best and second-best results are highlighted in **bold** and underline. ↑ (↓) denotes larger is better (smaller is better).

| Attributes | Pub.Info | BSO | CB | CIB | IC | LI | MSO | OF | SSO | SA | TC | BW | bRGB | bT |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| SGDL [29] | TMM20 | 0.559 | 0.522 | 0.490 | 0.482 | 0.530 | 0.513 | 0.566 | 0.565 | 0.428 | 0.473 | 0.443 | 0.568 | 0.575 |
| CSRNet [32] | TCSVT21 | 0.824 | 0.767 | 0.776 | 0.746 | 0.813 | 0.760 | 0.800 | 0.718 | 0.749 | 0.753 | 0.697 | 0.798 | 0.805 |
| CGFNet [19] | TCSVT22 | 0.837 | 0.825 | 0.808 | 0.792 | 0.834 | 0.786 | 0.816 | 0.768 | 0.808 | 0.806 | 0.750 | 0.836 | 0.837 |
| SwinNet [12] | TCSVT22 | 0.879 | 0.835 | 0.865 | 0.820 | 0.870 | 0.819 | 0.843 | 0.738 | 0.825 | 0.827 | 0.786 | 0.846 | 0.850 |
| ADF [30] | TMM22 | 0.785 | 0.713 | 0.749 | 0.701 | 0.739 | 0.705 | 0.711 | 0.495 | 0.677 | 0.673 | 0.683 | 0.725 | 0.732 |
| TNet [15] | TMM22 | 0.845 | 0.835 | 0.816 | 0.793 | 0.844 | 0.822 | 0.822 | 0.799 | 0.825 | 0.819 | 0.767 | 0.844 | 0.845 |
| OSRNet [9] | TIM22 | 0.838 | 0.801 | 0.798 | 0.764 | 0.819 | 0.782 | 0.793 | 0.707 | 0.764 | 0.769 | 0.766 | 0.810 | 0.815 |
| ACMANet [17] | KBS22 | 0.852 | 0.807 | 0.816 | 0.784 | 0.850 | 0.775 | 0.795 | 0.696 | 0.770 | 0.786 | 0.777 | 0.827 | 0.831 |
| MCFNet [13] | AI22 | 0.858 | 0.827 | 0.839 | 0.788 | 0.858 | 0.805 | 0.820 | 0.762 | 0.813 | 0.806 | 0.804 | 0.840 | 0.842 |
| CMDBIF [37] | TCSVT23 | 0.851 | 0.833 | 0.821 | 0.803 | 0.850 | 0.800 | 0.829 | 0.816 | 0.827 | 0.827 | 0.777 | 0.851 | 0.851 |
| CAVER [14] | TIP23 | 0.870 | 0.844 | 0.858 | 0.810 | 0.863 | 0.817 | 0.825 | 0.780 | 0.815 | 0.827 | 0.813 | 0.854 | 0.853 |
| ADNet [23] | MMA23 | 0.893 | 0.883 | 0.884 | 0.861 | 0.863 | 0.868 | 0.844 | 0.851 | 0.848 | 0.882 | 0.851 | 0.892 | 0.886 |
| WGOFNet [33] | TOMM24 | 0.889 | 0.865 | 0.874 | 0.837 | 0.880 | 0.845 | 0.863 | 0.844 | 0.851 | 0.849 | 0.790 | 0.877 | 0.878 |
| UMINet [26] | TVC24 | 0.840 | 0.814 | 0.827 | 0.769 | 0.834 | 0.789 | 0.808 | 0.742 | 0.810 | 0.781 | 0.787 | 0.824 | 0.829 |
| SACNet [35] | TMM24 | 0.899 | 0.822 | 0.890 | 0.868 | 0.888 | 0.862 | 0.862 | 0.846 | 0.861 | 0.883 | 0.832 | 0.891 | 0.890 |
| ConTriNet [73] | TPAMI25 | 0.906 | 0.890 | 0.899 | 0.874 | 0.894 | 0.864 | 0.878 | 0.863 | 0.878 | 0.885 | 0.835 | 0.898 | 0.898 |
| **DFNet** | - | 0.905 | 0.894 | 0.896 | 0.869 | 0.900 | 0.874 | 0.876 | 0.871 | 0.878 | 0.894 | 0.855 | 0.900 | 0.898 |
| **HyPSAM** | - | **0.920** | **0.911** | **0.911** | **0.887** | **0.917** | **0.896** | **0.891** | **0.893** | **0.895** | **0.911** | **0.871** | **0.919** | **0.916** |

exploit thermal information under these conditions, resulting in suboptimal detection. Qualitative comparisons further verify that the proposed method effectively refines object details, preventing false positive detections in various challenging scenarios.

### *D. Ablation Studies*

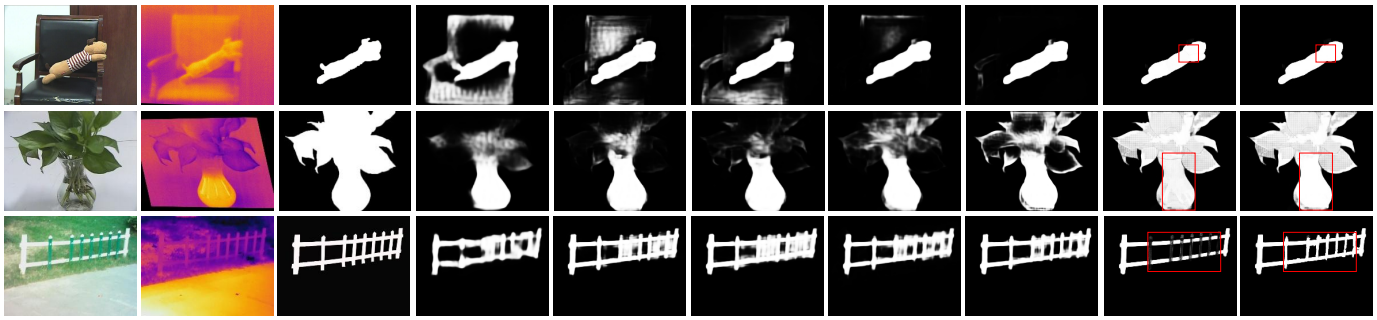
We investigate the importance and contributions of different modules within the proposed method on three benchmarks. The ablation studies are mainly divided into two parts: component ablation and prompt ablation.

*1) Effectiveness of Components:* Table III shows the quantitative results of different components. The first row represents the baseline, where shallow and deep features from the backbone are simply merged via element-wise addition. The second row evaluates the impact of the dynamic interaction module (DIM), which facilitates cross-modal fusion by simultaneously exploring different dimensions' relations across modalities. The third row incorporates the multi-branch decoding module (MDM), which disentangles modality-specific and shared features to reduce the impact of modality bias, yielding further performance gains. The fourth row shows the results of applying the decision fusion module (DFM), which can adaptively fuse the cross-modal complementary information at the decision level. The fifth row demonstrates the effectiveness of DFNet, where all components work synergistically to exploit the various relations among multi-modal features at different dimensions, achieving sufficient feature fusion, visualization examples of ablation studies, as shown in Fig.10.

We also perform ablation studies on P2RNet, focusing on the quality-aware modality selector (QMS), prompt

TABLE III: Ablation analysis of components on three datasets. The best results are highlighted in **bold**.

| No. | Settings | VT5000 | | | | VT1000 | | | | VT821 | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | $F_w \uparrow$ | $\mathcal{M} \downarrow$ | $E_m \uparrow$ | $S_m \uparrow$ | $F_w \uparrow$ | $\mathcal{M} \downarrow$ | $E_m \uparrow$ | $S_m \uparrow$ | $F_w \uparrow$ | $\mathcal{M} \downarrow$ | $E_m \uparrow$ | $S_m \uparrow$ |
| 1 | Baseline | 0.867 | 0.024 | 0.942 | 0.913 | 0.911 | 0.016 | 0.946 | 0.938 | 0.850 | 0.027 | 0.929 | 0.905 |
| 2 | Baseline + DIM | 0.894 | 0.020 | 0.955 | 0.928 | 0.929 | 0.013 | 0.957 | 0.949 | 0.877 | 0.024 | 0.936 | 0.922 |
| 3 | Baseline + MDM | 0.888 | 0.021 | 0.950 | 0.926 | 0.926 | 0.014 | 0.950 | 0.948 | 0.871 | 0.024 | 0.931 | 0.921 |
| 4 | Baseline + MDM + DFM | 0.888 | 0.021 | 0.951 | 0.927 | 0.926 | 0.014 | 0.951 | 0.949 | 0.879 | 0.022 | 0.937 | 0.924 |
| 5 | Baseline + DIM + MDM + DFM (DFNet) | 0.898 | 0.020 | 0.958 | 0.930 | 0.930 | 0.013 | 0.955 | 0.950 | 0.881 | 0.022 | 0.940 | 0.926 |
| 6 | SAM + random prompts | 0.368 | 0.221 | 0.591 | 0.617 | 0.564 | 0.139 | 0.701 | 0.735 | 0.462 | 0.169 | 0.668 | 0.682 |
| 7 | DFNet + PAS + SAM-RGB | 0.905 | 0.022 | 0.961 | 0.928 | 0.940 | 0.014 | 0.963 | 0.952 | 0.892 | 0.023 | 0.946 | 0.927 |
| 8 | DFNet + PAS + SAM-T | 0.884 | 0.025 | 0.958 | 0.916 | 0.922 | 0.017 | 0.960 | 0.939 | 0.852 | 0.030 | 0.934 | 0.900 |
| 9 | DFNet + PAS + QMS | 0.906 | 0.021 | 0.962 | 0.929 | 0.940 | 0.014 | 0.963 | 0.952 | 0.895 | 0.023 | 0.948 | 0.928 |
| 10 | DFNet + PAS + QMS + RS (HyPSAM) | **0.911** | **0.019** | **0.963** | **0.931** | **0.944** | **0.011** | **0.965** | **0.954** | **0.903** | **0.020** | **0.948** | **0.932** |



(a) (b) (c) (d) (e) (f) (g) (h) (i) (j)

Fig. 10: Visualization examples of ablation studies. (a) RGB image. (b) Thermal image. (c) Ground truth. (d) Baseline. (e) Baseline+DIM. (f) Baseline+MDM. (g) Baseline+MDM+DFM. (h) DFNet. (i) HyPSAM (w/o RS). (j) HyPSAM (full model).

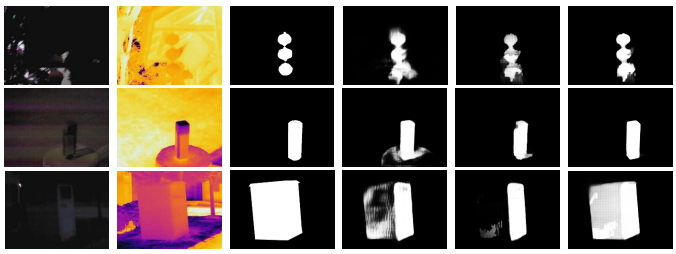



Fig. 11: Comparison of SAM predictions guided by different single-modality inputs.

augmentation strategy (PAS), and refinement strategy (RS), with results also summarized in Table III. As a baseline, we report the performance of SAM used independently with random prompts on RGB inputs (Row 6), aiming to isolate the effect of prompt engineering and emphasize the contribution of DFNet. To further validate the necessity of the QMS module, we analyze the segmentation performance of SAM when guided by different single-modality inputs, as illustrated in Fig. 11. SAM performs reasonably well in well-lit scenes with RGB inputs, but its effectiveness significantly deteriorates under low-light or visually ambiguous conditions. Notably, although SAM is pretrained solely on RGB images, it can still yield meaningful segmentation results on thermal inputs. This is because SAM relies not only on color information but also on structural priors, edge cues, and learned visual semantics, which can still be partially preserved in aligned thermal images. The ablation results clearly demonstrate that QMS enhances robustness by adaptively selecting the most informative modality, thereby avoiding noisy or misleading

TABLE IV: Comparison with different prompts on the VT5000 dataset. The best results are highlighted in **bold**.

| No. | Prompt Type | $F_w \uparrow$ | $\mathcal{M} \downarrow$ | $E_m \uparrow$ | $S_m \uparrow$ |
|---|---|---|---|---|---|
| 1 | Point | 0.832 | 0.063 | 0.912 | 0.875 |
| 2 | Box | 0.903 | 0.022 | 0.957 | 0.927 |
| 3 | Mask | 0.549 | 0.091 | 0.942 | 0.816 |
| 4 | Point + Box | 0.881 | 0.028 | 0.943 | 0.915 |
| 5 | Point + Mask | 0.823 | 0.064 | 0.919 | 0.876 |
| 6 | Box + Mask | **0.911** | **0.019** | **0.963** | **0.931** |
| 7 | Point + Box + Mask | 0.871 | 0.031 | 0.944 | 0.911 |

TABLE V: Comparison with different refinement strategies on the VT5000 dataset. The best results are highlighted in **bold**.

| No. | Refinement Strategies | $F_w \uparrow$ | $\mathcal{M} \downarrow$ | $E_m \uparrow$ | $S_m \uparrow$ |
|---|---|---|---|---|---|
| 1 | Add | 0.900 | 0.021 | 0.959 | **0.934** |
| 2 | Max | 0.911 | **0.019** | **0.963** | 0.931 |
| 3 | CRF | 0.879 | **0.019** | 0.929 | 0.891 |
| 4 | F-BRS | 0.907 | **0.019** | 0.962 | 0.930 |
| 5 | Weighted Add | 0.898 | 0.021 | 0.958 | 0.933 |
| 6 | Morphological Refinement | **0.912** | **0.019** | 0.962 | 0.926 |

guidance. The RS can fully integrate the initial saliency map and segmentation map, ensuring that both global object structures and fine-grained details are accurately captured, leading to a notable improvement in overall performance.

*2) Effectiveness of Prompts:* We conduct prompt ablation studies to further investigate the impact of different prompts, as summarized in Table IV. We first evaluate three types of individual prompts: points, boxes, and masks, where the point represents the centroids of the mask. Their results are presented in the first three rows of Table IV. Visualization examples of different prompts as shown in Fig.12. It can be

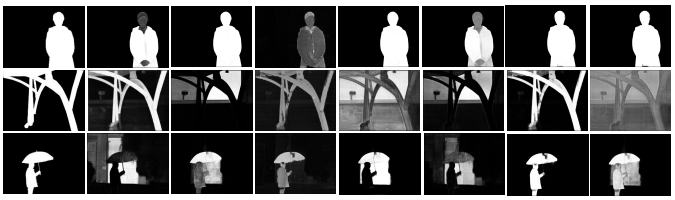



Fig. 12: Visualization examples of different prompts. (a) Ground truth. (b) Point. (c) Box. (d) Mask. (e) Point+Box. (f) Point+Mask. (g) Box+Mask. (h) Point+Box+Mask.

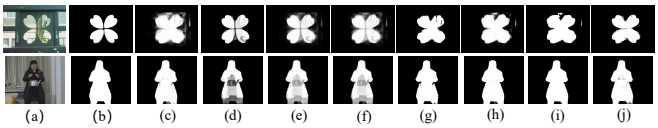



Fig. 13: Visualization examples of different refinement strategies. (a) RGB. (b) Ground truth. (c) DFnet. (d) HyPSAM (w/o RS). (e) Add. (f) Weighted Add. (g) CRF. (h) F-BRS. (i)Morphological Refinement. (j) Max.

observed that the box prompt yields superior performance due to its ability to provide global structured spatial information for guiding segmentation effectively. In contrast, point prompts are too sparse, offering limited contextual information, which leads to suboptimal results. Surprisingly, the mask prompt performs the worst. When used alone, it often causes SAM to generate soft probability maps with background noise and even over-segment the object. The mask serves as a dense and strict prior, making SAM closely follow it. As a result, any noise or inaccurate edges in the input mask are retained or even amplified, leading to poorer segmentation quality.

We then explore the combination of these prompts to investigate their complementary effects. The combination of all prompts does not yield competitive results. This may be due to conflicting information introduced by point prompts, which can create inconsistencies when combined with richer spatial information from boxes and masks. Instead, the combination of box and mask prompts achieves the best performance, as the box defines clear regions of interest, while the mask refines object boundaries, thereby striking an effective balance between global and local guidance.

*3) Effectiveness of Refinement Strategies:* Table V compares six refinement strategies on the VT5000 dataset, including add, max, CRF, F-BRS [82], entropy-based weighted fusion, and morphological refinement, with visual examples shown in Fig. 13. The advantage of max fusion lies in its ability to retain the most confident activations from both the initial saliency map and the SAM-predicted mask. Since SAM often produces fine-grained results that split semantically unified objects (*e.g.*, a person holding a kettle) into separate instances, additive or weighted fusion may amplify redundant or ambiguous boundaries. While F-BRS and morphological refinement achieve performance comparable to max fusion, the latter is significantly more efficient and simpler to implement. Therefore, we adopt max fusion as the default refinement strategy in our framework.

TABLE VI: Comparison with different methods on unaligned RGB-T datasets. The best results are highlighted in **bold**.

| | Metric | SwinNet [12] | TNet [15] | OSRNet [9] | MCFNet [13] | SACNet [35] | DFNet Ours | HyPSAM Ours |
|---|---|---|---|---|---|---|---|---|
| un-VT5000 | $F_w \uparrow$ | 0.823 | 0.806 | 0.571 | 0.757 | 0.876 | 0.880 | **0.895** |
| | $\mathcal{M} \downarrow$ | 0.031 | 0.038 | 0.106 | 0.044 | 0.023 | 0.023 | **0.022** |
| | $E_m \uparrow$ | 0.923 | 0.910 | 0.770 | 0.905 | 0.949 | 0.947 | **0.955** |
| | $S_m \uparrow$ | 0.899 | 0.879 | 0.724 | 0.864 | 0.911 | 0.922 | **0.922** |
| un-VT1000 | $F_w \uparrow$ | 0.890 | 0.877 | 0.701 | 0.833 | 0.923 | 0.925 | **0.942** |
| | $\mathcal{M} \downarrow$ | 0.018 | 0.025 | 0.077 | 0.028 | 0.014 | 0.014 | **0.013** |
| | $E_m \uparrow$ | 0.938 | 0.927 | 0.825 | 0.929 | 0.954 | 0.950 | **0.961** |
| | $S_m \uparrow$ | 0.936 | 0.920 | 0.800 | 0.914 | 0.941 | 0.948 | **0.951** |
| un-VT821 | $F_w \uparrow$ | 0.799 | 0.788 | 0.575 | 0.741 | 0.857 | 0.864 | **0.880** |
| | $\mathcal{M} \downarrow$ | 0.036 | 0.047 | 0.086 | 0.044 | 0.026 | 0.025 | **0.023** |
| | $E_m \uparrow$ | 0.905 | 0.889 | 0.790 | 0.899 | 0.929 | 0.928 | **0.936** |
| | $S_m \uparrow$ | 0.888 | 0.873 | 0.733 | 0.867 | 0.905 | 0.917 | **0.920** |

TABLE VII: Comparison with state-of-the-art methods on RGB-Nir datasets. The best results are highlighted in **bold**.

| Methods | RGBN $F_{avg} \uparrow$ | $F_{max} \uparrow$ | $F_w \uparrow$ | $\mathcal{M} \downarrow$ | $E_m \uparrow$ | $S_m \uparrow$ |
|---|---|---|---|---|---|---|
| RC [75] | 0.664 | 0.736 | 0.442 | 0.148 | 0.810 | 0.724 |
| DCL [76] | 0.779 | 0.838 | 0.660 | 0.076 | 0.881 | 0.796 |
| SOD8s+ [77] | 0.803 | 0.850 | 0.745 | 0.061 | 0.894 | 0.828 |
| DFNet | 0.936 | 0.956 | 0.931 | 0.017 | 0.976 | 0.946 |
| HyPSAM | **0.957** | **0.963** | **0.946** | **0.015** | **0.978** | **0.951** |

### *E. Application to Other SOD Tasks*

To evaluate the generalization capability of HyPSAM, we extend its application to various salient object detection tasks across unaligned RGB-T, RGB-D, and RGB-NIR modalities. HyPSAM achieves consistently strong performance. On unaligned RGB-T datasets [35] (Table VI), our method outperforms recent methods across all evaluation metrics, demonstrating its robustness to cross-modal misalignment. For RGB-D datasets [83]–[85] (Table VII), HyPSAM achieves competitive or superior performance compared to several state-of-the-art models, including Transformer-based approaches, highlighting its semantic transferability and effective depth adaptation. On the RGB-NIR dataset [77] (Table VIII), HyPSAM attains the best results with an $F_w$ of 0.946 and a $\mathcal{M}$ of 0.015, further validating its ability to generalize across unseen modalities.

TABLE VIII: Comparison with state-of-the-art methods on RGB-D datasets. The best results are highlighted in **bold**.

| | Metric | MMNet [78] | SwinNet [12] | CATNet [79] | CPNet [80] | BTNet [81] | DFNet Ours | HyPSAM Ours |
|---|---|---|---|---|---|---|---|---|
| NLPR | $F_w \uparrow$ | 0.889 | 0.908 | 0.912 | 0.918 | 0.917 | 0.916 | **0.925** |
| | $\mathcal{M} \downarrow$ | 0.024 | 0.018 | 0.018 | **0.016** | **0.016** | 0.017 | **0.016** |
| | $E_m \uparrow$ | 0.950 | 0.967 | 0.967 | 0.970 | 0.969 | 0.970 | **0.971** |
| | $S_m \uparrow$ | 0.250 | 0.941 | 0.940 | 0.940 | 0.941 | **0.944** | **0.944** |
| NJU2K | $F_w \uparrow$ | 0.900 | 0.922 | 0.919 | 0.923 | 0.917 | **0.927** | **0.931** |
| | $\mathcal{M} \downarrow$ | 0.038 | 0.027 | 0.026 | 0.025 | 0.025 | **0.022** | 0.023 |
| | $E_m \uparrow$ | 0.919 | 0.934 | 0.933 | 0.935 | 0.932 | 0.939 | **0.942** |
| | $S_m \uparrow$ | 0.911 | 0.935 | 0.932 | 0.935 | 0.932 | **0.940** | **0.940** |
| STERE | $F_w \uparrow$ | 0.880 | 0.893 | 0.894 | 0.895 | 0.902 | 0.899 | **0.915** |
| | $\mathcal{M} \downarrow$ | 0.045 | 0.033 | 0.030 | 0.029 | 0.028 | 0.029 | **0.026** |
| | $E_m \uparrow$ | 0.924 | 0.929 | 0.936 | 0.933 | 0.940 | 0.936 | **0.945** |
| | $S_m \uparrow$ | 0.891 | 0.919 | 0.921 | 0.920 | 0.927 | 0.929 | **0.934** |

TABLE IX: The generalization of HyPSAM is tested on three benchmarks, encompassing three distinct architectural paradigms. The best results are highlighted in **bold**. † indicates the results are refined by HyPSAM.

| Types | Methods | VT5000 | | | | | | VT1000 | | | | | | VT821 | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | $F_{avg}\uparrow$ | $F_{max}\uparrow$ | $F_w\uparrow$ | $\mathcal{M}\downarrow$ | $E_m\uparrow$ | $S_m\uparrow$ | $F_{avg}\uparrow$ | $F_{max}\uparrow$ | $F_w\uparrow$ | $\mathcal{M}\downarrow$ | $E_m\uparrow$ | $S_m\uparrow$ | $F_{avg}\uparrow$ | $F_{max}\uparrow$ | $F_w\uparrow$ | $\mathcal{M}\downarrow$ | $E_m\uparrow$ | $S_m\uparrow$ |
| Single-stream | OSRNet [9] | 0.823 | 0.866 | 0.807 | 0.040 | 0.908 | 0.875 | 0.892 | 0.929 | 0.891 | 0.022 | 0.935 | 0.926 | 0.814 | 0.862 | 0.801 | 0.043 | 0.896 | 0.875 |
| | OSRNet† | **0.857** | **0.874** | **0.829** | **0.037** | **0.918** | **0.882** | **0.919** | **0.932** | **0.910** | **0.020** | **0.947** | **0.932** | **0.849** | **0.867** | **0.826** | **0.036** | **0.912** | **0.882** |
| Dual-stream | CSRNet [32] | 0.811 | 0.857 | 0.796 | 0.042 | 0.905 | 0.868 | 0.877 | **0.918** | 0.878 | 0.024 | 0.925 | 0.918 | 0.831 | 0.880 | 0.821 | 0.038 | 0.909 | 0.885 |
| | CSRNet† | **0.843** | **0.863** | **0.814** | **0.039** | **0.914** | **0.872** | **0.903** | **0.918** | **0.893** | **0.022** | **0.937** | **0.924** | **0.862** | **0.882** | **0.843** | **0.032** | **0.920** | **0.892** |
| | CGFNet [19] | 0.851 | 0.887 | 0.831 | 0.035 | 0.922 | 0.883 | 0.906 | 0.936 | 0.900 | 0.023 | 0.944 | 0.923 | 0.845 | 0.885 | 0.829 | 0.038 | 0.912 | 0.881 |
| | CGFNet† | **0.879** | **0.895** | **0.847** | **0.033** | **0.927** | **0.889** | **0.926** | **0.938** | **0.913** | **0.021** | **0.950** | **0.930** | **0.876** | **0.893** | **0.847** | **0.035** | **0.919** | **0.889** |
| | SwinNet [12] | 0.865 | 0.915 | 0.846 | 0.026 | 0.942 | 0.912 | 0.896 | 0.948 | 0.894 | 0.018 | 0.947 | 0.938 | 0.847 | 0.903 | 0.818 | 0.030 | 0.926 | 0.904 |
| | SwinNet† | **0.910** | **0.924** | **0.888** | **0.023** | **0.950** | **0.918** | **0.940** | **0.952** | **0.935** | **0.013** | **0.961** | **0.948** | **0.895** | **0.915** | **0.876** | **0.024** | **0.934** | **0.914** |
| | ADF [30] | 0.778 | 0.863 | 0.722 | 0.048 | 0.891 | 0.864 | 0.847 | 0.923 | 0.804 | 0.034 | 0.921 | 0.910 | 0.717 | **0.804** | 0.627 | 0.077 | **0.843** | 0.810 |
| | ADF† | **0.850** | **0.871** | **0.812** | **0.038** | **0.909** | **0.873** | **0.914** | **0.930** | **0.898** | **0.024** | **0.940** | **0.925** | **0.766** | 0.798 | **0.726** | **0.058** | 0.835 | **0.822** |
| | TNet [15] | 0.846 | 0.895 | 0.840 | 0.033 | 0.927 | 0.895 | 0.889 | 0.937 | 0.895 | 0.021 | 0.937 | 0.929 | 0.842 | 0.904 | 0.841 | 0.030 | 0.919 | 0.899 |
| | TNet† | **0.885** | **0.901** | **0.860** | **0.031** | **0.937** | **0.901** | **0.922** | **0.939** | **0.914** | **0.019** | **0.952** | **0.935** | **0.889** | **0.908** | **0.872** | **0.027** | **0.932** | **0.911** |
| | ACMANet [17] | 0.858 | 0.890 | 0.823 | 0.033 | **0.932** | 0.887 | 0.904 | 0.933 | 0.889 | 0.021 | 0.945 | 0.927 | 0.837 | 0.873 | 0.807 | 0.035 | 0.914 | 0.883 |
| | ACMANet† | **0.878** | **0.895** | **0.854** | **0.030** | **0.932** | **0.895** | **0.921** | **0.934** | **0.917** | **0.018** | **0.948** | **0.935** | **0.856** | **0.876** | **0.839** | **0.029** | **0.918** | **0.890** |
| | MCFNet [13] | 0.848 | **0.886** | 0.836 | 0.033 | 0.924 | 0.887 | 0.902 | **0.939** | 0.906 | 0.019 | 0.944 | 0.932 | 0.844 | **0.889** | 0.835 | 0.029 | **0.918** | 0.891 |
| | MCFNet† | **0.868** | 0.884 | **0.846** | **0.032** | **0.927** | **0.892** | **0.926** | 0.938 | **0.921** | **0.016** | **0.952** | **0.938** | **0.863** | 0.884 | **0.846** | **0.028** | **0.918** | **0.896** |
| | CAVER [14] | 0.856 | 0.897 | 0.849 | 0.028 | 0.935 | 0.899 | 0.906 | 0.945 | 0.912 | **0.016** | 0.949 | 0.938 | 0.854 | 0.897 | 0.846 | **0.026** | 0.928 | 0.897 |
| | CAVER† | **0.884** | **0.903** | **0.866** | **0.029** | **0.939** | **0.906** | **0.929** | **0.946** | **0.926** | **0.016** | **0.955** | **0.943** | **0.881** | **0.899** | **0.866** | **0.026** | **0.929** | **0.908** |
| | ADNet [23] | 0.893 | 0.924 | 0.884 | 0.022 | 0.953 | 0.922 | 0.916 | **0.952** | 0.920 | 0.015 | 0.952 | 0.944 | 0.869 | 0.915 | 0.860 | 0.024 | 0.930 | 0.915 |
| | ADNet† | **0.918** | **0.930** | **0.899** | **0.021** | **0.958** | **0.923** | **0.938** | 0.950 | **0.934** | **0.014** | **0.960** | **0.947** | **0.901** | **0.919** | **0.888** | **0.021** | **0.941** | **0.922** |
| | WGOFNet [33] | 0.883 | 0.912 | 0.873 | 0.025 | 0.945 | 0.911 | 0.919 | **0.946** | 0.922 | 0.016 | 0.951 | 0.940 | 0.875 | 0.911 | 0.868 | 0.025 | 0.934 | 0.908 |
| | WGOFNet† | **0.900** | **0.913** | **0.883** | **0.024** | **0.948** | **0.914** | **0.933** | 0.944 | **0.930** | **0.014** | **0.956** | **0.944** | **0.896** | **0.912** | **0.883** | **0.023** | **0.939** | **0.916** |
| | UMINet [26] | 0.831 | 0.877 | 0.820 | 0.035 | 0.919 | 0.882 | 0.892 | 0.935 | 0.896 | 0.021 | 0.941 | 0.926 | 0.791 | 0.849 | 0.782 | **0.054** | 0.879 | 0.858 |
| | UMINet† | **0.864** | **0.882** | **0.839** | **0.034** | **0.926** | **0.892** | **0.922** | **0.937** | **0.916** | **0.018** | **0.951** | **0.936** | **0.827** | **0.853** | **0.804** | **0.054** | **0.887** | **0.868** |
| Triple-stream | CMDBIF [37] | 0.868 | 0.892 | 0.846 | 0.032 | 0.933 | 0.886 | 0.914 | **0.931** | 0.909 | 0.019 | 0.952 | 0.927 | 0.856 | 0.887 | 0.837 | 0.032 | 0.923 | 0.882 |
| | CMDBIF† | **0.882** | **0.896** | **0.857** | **0.031** | **0.936** | **0.895** | **0.924** | **0.931** | **0.918** | **0.017** | **0.953** | **0.936** | **0.871** | **0.889** | **0.850** | **0.031** | **0.924** | **0.893** |
| | ConTriNet [73] | 0.898 | 0.927 | 0.895 | 0.020 | 0.956 | 0.923 | 0.917 | 0.943 | 0.923 | 0.015 | 0.953 | 0.941 | 0.878 | 0.914 | 0.875 | 0.022 | 0.940 | 0.916 |
| | ConTriNet† | **0.918** | **0.928** | **0.904** | **0.020** | **0.959** | **0.926** | **0.934** | **0.945** | **0.933** | **0.014** | **0.957** | **0.945** | **0.899** | **0.916** | **0.890** | **0.020** | **0.944** | **0.923** |

TABLE X: Performance comparison of SAMv2 and ETAM in our framework on the RGB-T dataset.

| Models | VT5000 | | | | | | VT1000 | | | | | | VT821 | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | $F_{avg}\uparrow$ | $F_{max}\uparrow$ | $F_w\uparrow$ | $\mathcal{M}\downarrow$ | $E_m\uparrow$ | $S_m\uparrow$ | $F_{avg}\uparrow$ | $F_{max}\uparrow$ | $F_w\uparrow$ | $\mathcal{M}\downarrow$ | $E_m\uparrow$ | $S_m\uparrow$ | $F_{avg}\uparrow$ | $F_{max}\uparrow$ | $F_w\uparrow$ | $\mathcal{M}\downarrow$ | $E_m\uparrow$ | $S_m\uparrow$ |
| HyPSAM (SAMv2) | 0.928 | 0.939 | 0.911 | 0.019 | 0.963 | 0.931 | 0.946 | 0.957 | 0.944 | 0.011 | 0.965 | 0.954 | 0.914 | 0.930 | 0.903 | 0.020 | 0.948 | 0.932 |
| HyPSAM (ETAM) | 0.927 | 0.934 | 0.916 | 0.018 | 0.964 | 0.929 | 0.946 | 0.949 | 0.944 | 0.011 | 0.966 | 0.950 | 0.911 | 0.921 | 0.902 | 0.019 | 0.950 | 0.926 |

### *F. Generalization Performance of HyPSAM*

To evaluate the generalization capability of our framework, we apply it to three types of RGB-T SOD models by generating hybrid prompts from their saliency maps. As shown in Table IX, where "†" denotes results after optimization, most metrics across all methods improve substantially. In addition, we observe a clear positive correlation between the accuracy of the initial saliency maps and the performance of the optimized models. In other words, the more accurate the initial maps are, the greater the improvements after optimization. These results demonstrate the effectiveness of the hybrid prompt-driven approach in refining segmentation outcomes.

### *G. Generalization with Alternative Segmentation Models*

To further validate the generality of our proposed HyPSAM framework, we replace the SAMv2 in P2RNet with the efficient track anything model (ETAM) [86], a lightweight segmentation model. Table X reports the performance comparison. While ETAM achieves reasonable accuracy, SAMv2 yields better overall results. These results confirm that our framework is compatible with different segmentation backbones and can benefit from stronger base models.

### *H. Complexity Analysis*

Model complexity is typically measured by parameters, FLOPs, and FPS. As shown in Table XI, DFNet strikes a favorable balance between performance and complexity. Each ODConv module is lightweight and computationally efficient, requiring only 0.64 ms per image with 2.385M FLOPs and 25.8K parameters. The full version of HyPSAM (SAMv2), which incorporates the original SAM with a ViT-H backbone, comprises 817.6M parameters and 3033.4 GFLOPs. Despite the high computational cost, the significant performance gains in segmentation justify the overhead. To enhance efficiency, we also introduce a lightweight variant, HyPSAM (ETAM), which reduces the parameter count to 208.6M and FLOPs to 394.3G, while maintaining competitive accuracy and achieving a real-time inference speed of 25 FPS.

TABLE XI: Comparison of the complexity of some recent publicly available state-of-the-art RGB-T SOD methods.

| Method | Backbone | FLOPs (G) | Params. (M) | FPS |
|---|---|---|---|---|
| ADF [30] | VGG-16 | 128.2 | 83.1 | 7 |
| MIDD [21] | VGG-16 | 216.6 | 52.4 | 22 |
| CGFNet [19] | VGG-16 | 345.1 | 66.4 | 13 |
| OSRNet [9] | VGG-16 | 34.3 | 15.6 | 53 |
| CAVER [14] | ResNet-50d | 44.4 | 55.8 | 27 |
| SwinNet [12] | Swin-B | 124.3 | 198.7 | 10 |
| WGOFNet [33] | PVT | 48.5 | 61.8 | 11 |
| ADNet [23] | Swin-B+MobileViT | 56.7 | 93.2 | 43 |
| DFNet | Swin-B | 74.1 | 133.3 | 41 |
| HyPSAM (SAMv2) | Swin-B+ViT-H | 3033.4 | 817.6 | 3 |
| HyPSAM (ETAM) | Swin-B+EfficientViT | 394.3 | 208.6 | 25 |

### *I. Failure Cases and Analysis*

To better analyze the limitations of HyPSAM, we present failure cases in Fig. 14, compared with existing methods including OSRNet [9], SwinNet [12], and CMDBIF [37]. The first two rows show that although the QMS correctly identifies the more reliable modality, SAM still fails in thermal inputs with low target-background contrast, producing artifacts or even incorrect object recognition. In the third row, HyPSAM tends to over-segment semantically unified objects due to its fine-grained instance sensitivity, resulting in unnecessary internal separation. In the last row, the foreground object exhibits low contrast with the background, making it difficult for HyPSAM to distinguish foreground boundaries accurately. These cases demonstrate that while HyPSAM significantly outperforms prior methods in structure preservation and modality adaptation, certain limitations persist in scenarios with subtle foreground-background variations and complex object details. Moreover, since SAM is pre-trained exclusively on large-scale RGB data, a domain gap remains when generalizing to challenging thermal scenarios, particularly under weak contrast. Future work may focus on enhancing prompt semantics and refining instance integration to further improve robustness in complex scenarios.

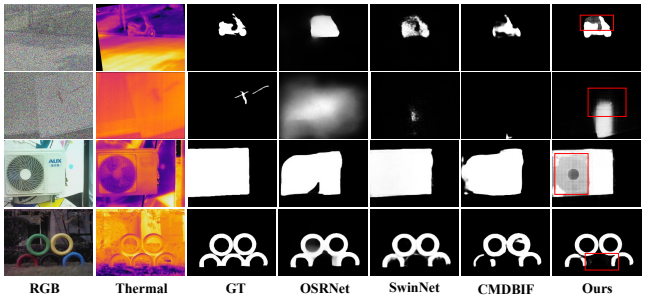



Fig. 14: Visual results of our HyPSAM and other advanced methods in some typical failure cases.

## V. Conclusion

In this paper, we proposed a HyPSAM, a novel hybrid prompt-driven segment anything model for RGB-T SOD, which consists of key components: DFNet and P2RNet. To address the intrinsic insufficient feature fusion, DFNet incorporated a dynamic fusion module coupled with a multi-branch decoding module to facilitate adaptive cross-modality interaction and generate robust saliency maps. To mitigate the extrinsic limitations of data scarcity, P2RNet leveraged a comprehensive hybrid prompting strategy that synergistically combines texts, boxes, and masks to drive SAM to accurately segment the saliency object. The initial saliency map and the segmentation map were then fused to produce the refined result with complete structures and clear boundaries. Extensive experiments demonstrated the superiority of our proposed method on multi-modal benchmarks. Furthermore, HyPSAM exhibits strong versatility and generalization by seamlessly integrating with different methods, achieving substantial performance gains without additional training.

## References

[1] L. Itti, C. Koch, and E. Niebur, "A model of saliency-based visual attention for rapid scene analysis," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 20, no. 11, pp. 1254–1259, 1998.
[2] H. Wang, Z. Li, Y. Li, B. B. Gupta, and C. Choi, "Visual saliency guided complex image retrieval," *Pattern Recogn. Lett.*, vol. 130, pp. 64–72, 2020.
[3] X. Wang, X. Shu, S. Zhang, B. Jiang, Y. Wang, Y. Tian, and F. Wu, "Mfgnet: Dynamic modality-aware filter generation for rgb-t tracking," *IEEE Trans. Multimedia*, vol. 25, pp. 4335–4348, 2022.
[4] X. Zhou, Z. Wu, and R. Cong, "Decoupling and integration network for camouflaged object detection," *IEEE Trans. Multimedia*, vol. 26, pp. 7114–7129, 2024.
[5] H. Zhou, B. Qiao, L. Yang, J. Lai, and X. Xie, "Texture-guided saliency distilling for unsupervised salient object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 7257–7267.
[6] M. Ma, C. Xia, C. Xie, X. Chen, and J. Li, "Boosting broader receptive fields for salient object detection," *IEEE Trans. Image Process.*, vol. 32, pp. 1026–1038, 2023.
[7] L. Zhang and Q. Zhang, "Salient object detection with edge-guided learning and specific aggregation," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 34, no. 1, pp. 534–548, 2023.
[8] Z. Wang, Y. Zhang, Y. Liu, D. Zhu, S. A. Coleman, and D. Kerr, "Elwnet: An extremely lightweight approach for real-time salient object detection," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 33, no. 11, pp. 6404–6417, 2023.
[9] F. Huo, X. Zhu, Q. Zhang, Z. Liu, and W. Yu, "Real-time one-stream semantic-guided refinement network for rgb-thermal saliency object detection," *IEEE Trans. Instrum. Meas.*, vol. 71, pp. 1–12, 2022.
[10] R. Ju, L. Ge, W. Geng, T. Ren, and G. Wu, "Depth saliency based on anisotropic center-surround difference," in *Proc. Int. Conf. Image Process.*, 2014, pp. 1115–1119.
[11] J. Wang, Z. Zhang, N. Yu, and Y. Han, "Progressive expansion for semi-supervised bi-modal salient object detection," *Pattern Recogn.*, vol. 157, pp. 110868, 2025.
[12] Z. Liu, Y. Tan, Q. He, and Y. Xiao, "Swinnet: Swin transformer drives edge-aware rgb-d and rgb-t salient object detection," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 7, pp. 4486–4497, 2021.
[13] S. Ma, K. Song, H. Dong, H. Tian, and Y. Yan, "Modal complementary fusion network for rgb-t salient object detection," *Appl. Intell.*, vol. 53, no. 8, pp. 9038–9055, 2023.
[14] Y. Pang, X. Zhao, L. Zhang, and H. Lu, "Caver: Cross-modal view-mixed transformer for bi-modal salient object detection," *IEEE Trans. Image Process.*, vol. 32, pp. 892–904, 2023.
[15] R. Cong, K. Zhang, C. Zhang, F. Zheng, Y. Zhao, Q. Huang, and S. Kwong, "Does thermal really always matter for rgb-t salient object detection?," *IEEE Trans. Multimedia*, vol. 25, pp. 6971–6982, 2022.
[16] J. Wang, K. Song, Y. Bao, Y. Yan, and Y. Han, "Unidirectional rgb-t saliency object detection with intertwined driving of encoding and fusion," *Eng. Appl. Artif. Intell.*, vol. 114, pp. 105162, 2022.
[17] C. Xu, Q. Li, Q. Zhou, X. Jiang, D. Yu, and Y. Zhou, "Asymmetric cross-modal activation network for rgb-t salient object detection," *Knowl.-Based Syst.*, vol. 258, pp. 110047, 2022.
[18] H. Zhou, C. Tian, Z. Zhang, C. Li, Y. Ding, Y. Xie, and Z. Li, "Position-aware relation learning for rgb-thermal salient object detection," *IEEE Trans. Image Process.*, vol. 32, pp. 2593–2607, 2023.

[19] J. Wang, K. Song, Y. Bao, L. Huang, and Y. Yan, "Cgfnet: Cross-guided fusion network for rgb-t salient object detection," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 5, pp. 2949–2961, 2021.

[20] Q. Zhang, T. Xiao, N. Huang, D. Zhang, and J. Han, "Revisiting feature fusion for rgb-t salient object detection," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 31, no. 5, pp. 1804–1818, 2020.

[21] Z. Tu, Z. Li, C. Li, Y. Lang, and J. Tang, "Multi-interactive dual-decoder for rgb-thermal salient object detection," *IEEE Trans. Image Process.*, vol. 30, pp. 5678–5691, 2021.

[22] J. Wu, W. Zhou, X. Qian, J. Lei, L. Yu, and T. Luo, "Menet: Lightweight multimodality enhancement network for detecting saliency objects in rgb-thermal images," *Neurocomputing*, vol. 527, pp. 119–129, 2023.

[23] Y. Fang, R. Hou, J. Bei, T. Ren, and G. Wu, "Adnet: An asymmetric dual-stream network for rgb-t salient object detection," in *Proc. ACM Int. Conf. Multimedia in Asia*, 2023, pp. 1–7.

[24] H. Wang, K. Song, L. Huang, H. Wen, and Y. Yan, "Thermal images-aware guided early fusion network for cross-illumination rgb-t salient object detection," *Eng. Appl. Artif. Intell.*, vol. 118, pp. 105640, 2023.

[25] W. Zhou, F. Sun, Q. Jiang, R. Cong, and J.-N. Hwang, "Wavenet: Wavelet network with knowledge distillation for rgb-t salient object detection," *IEEE Trans. Image Process.*, vol. 32, pp. 3027–3039, 2023.

[26] L. Gao, P. Fu, M. Xu, T. Wang, and B. Liu, "Uminet: A unified multi-modality interaction network for rgb-d and rgb-t salient object detection," *The Vis. Comput.*, vol. 40, no. 3, pp. 1565–1582, 2024.

[27] X. Jiang, Y. Hou, H. Tian, and L. Zhu, "Mirror complementary transformer network for rgb-thermal salient object detection," *IET Comput. Vis.*, vol. 18, no. 1, pp. 15–32, 2024.

[28] G. Wang, C. Li, Y. Ma, A. Zheng, J. Tang, and B. Luo, "Rgb-t saliency detection benchmark: Dataset, baselines, analysis and a novel approach," in *Image and Graphics Technologies and Applications: IGTA 2018, Beijing, China*. 2018, pp. 359–369, Springer.

[29] Z. Tu, T. Xia, C. Li, X. Wang, Y. Ma, and J. Tang, "Rgb-t image saliency detection via collaborative graph learning," *IEEE Trans. Multimedia*, vol. 22, no. 1, pp. 160–173, 2019.

[30] Z. Tu, Y. Ma, Z. Li, C. Li, J. Xu, and Y. Liu, "Rgbt salient object detection: A large-scale dataset and benchmark," *IEEE Trans. Multimedia*, vol. 25, pp. 4163–4176, 2022.

[31] F. Sun, W. Zhou, L. Ye, and L. Yu, "Hierarchical decoding network based on swin transformer for detecting saliency objects in rgb-t images," *IEEE Signal Process. Lett.*, vol. 29, pp. 1714–1718, 2022.

[32] F. Huo, X. Zhu, L. Zhang, Q. Liu, and Y. Shu, "Efficient context-guided stacked refinement network for rgb-t salient object detection," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 5, pp. 3111–3124, 2021.

[33] J. Wang, G. Li, J. Shi, and J. Xi, "Weighted guided optional fusion network for rgb-t salient object detection," *ACM Trans. Multimed. Comput. Commun. Appl.*, vol. 20, no. 5, pp. 1–20, 2024.

[34] K. Wang, Z. Tu, C. Li, C. Zhang, and B. Luo, "Learning adaptive fusion bank for multi-modal salient object detection," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 34, pp. 7344–7358, 2024.

[35] K. Wang, D. Lin, C. Li, Z. Tu, and B. Luo, "Alignment-free rgbt salient object detection: Semantics-guided asymmetric correlation network and a unified benchmark," *IEEE Trans. Multimedia*, vol. 26, pp. 10692–10707, 2024.

[36] D. Jin, F. Shao, Z. Xie, B. Mu, and H. Chen, "Rethinking lightweight rgb–thermal salient object detection with local and global perception network," *IEEE Internet Things J.*, vol. 12, no. 11, pp. 18056–18069, 2025.

[37] Z. Xie, F. Shao, G. Chen, H. Chen, Q. Jiang, X. Meng, and Y.-S. Ho, "Cross-modality double bidirectional interaction and fusion network for rgb-t salient object detection," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 33, no. 8, pp. 4149–4163, 2023.

[38] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo, et al., "Segment anything," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 4015–4026.

[39] X. Jia, B. De Brabandere, T. Tuytelaars, and L. V. Gool, "Dynamic filter networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2016, vol. 29.

[40] J. Wu, D. Li, Y. Yang, C. Bajaj, and X. Ji, "Dynamic filtering with large sampling field for convnets," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 185–200.

[41] J. Zhou, V. Jampani, Z. Pi, Q. Liu, and M.-H. Yang, "Decoupled dynamic filter networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 6647–6656.

[42] Y. Pang, L. Zhang, X. Zhao, and H. Lu, "Hierarchical dynamic filtering network for rgb-d salient object detection," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 235–252.

[43] M. Zhang, S. Yao, B. Hu, Y. Piao, and W. Ji, "C2dfnet: Criss-cross dynamic filter network for rgb-d salient object detection," *IEEE Trans. Multimedia*, vol. 25, pp. 5142–5154, 2022.

[44] J. Jin, Q. Jiang, Q. Wu, B. Xu, and R. Cong, "Underwater salient object detection via dual-stage self-paced learning and depth emphasis," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 35, no. 3, pp. 2147–2160, 2025.

[45] J. Ma, Y. He, F. Li, L. Han, C. You, and B. Wang, "Segment anything in medical images," *Nat. Commun.*, vol. 15, no. 1, pp. 654, 2024.

[46] D. Wang, J. Zhang, B. Du, M. Xu, L. Liu, D. Tao, and L. Zhang, "Samrs: Scaling-up remote sensing segmentation dataset with segment anything model," in *Proc. Adv. Neural Inf. Process. Syst.*, 2024, vol. 36.

[47] Z. Yu, X. Zhang, L. Zhao, Y. Bin, and G. Xiao, "Exploring deeper segment anything model with depth perception for camouflaged object detection," *arXiv preprint arXiv:2407.12339*, 2024.

[48] Y. Fang, Y. Shi, J. Bei, and T. Ren, "Semantic-guided rgb-thermal crowd counting with segment anything model," in *Proc. Int. Conf. Multimedia Retrieval*, 2024, pp. 570–578.

[49] Guanyao Wu, Haoyu Liu, Hongming Fu, Yichuan Peng, Jinyuan Liu, Xin Fan, and Risheng Liu, "Every SAM drop counts: Embracing semantic priors for multi-modality image fusion and beyond," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2025, pp. 17882–17891.

[50] S. Zhai, C. Liu, Z. Tu, C. Li, and L. Gao, "Weakly supervised rgb–t salient object detection via sam-guided label optimization and progressive cross-modal cross-scale fusion," *Inf. Fusion*, vol. 120, pp. 103048, 2025.

[51] Z. Liu, H. Hu, Y. Lin, Z. Yao, Z. Xie, Y. Wei, J. Ning, Y. Cao, Z. Zhang, L. Dong, et al., "Swin transformer v2: Scaling up capacity and resolution," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 12009–12019.

[52] C. Li, A. Zhou, and A. Yao, "Omni-dimensional dynamic convolution," in *Proc. Int. Conf. Learn. Representations*, 2022.

[53] X. Qin, Z. Zhang, C. Huang, C. Gao, M. Dehghan, and M. Jagersand, "Basnet: Boundary-aware salient object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 7479–7489.

[54] J. Wei, S. Wang, Z. Wu, C. Su, Q. Huang, and Q. Tian, "Label decoupling framework for salient object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 13025–13034.

[55] R. Zhao, B. Qian, X. Zhang, Y. Li, R. Wei, Y. Liu, and Y. Pan, "Rethinking dice loss for medical image segmentation," in *Proc. IEEE Int. Conf. Data Mining*, 2020, pp. 851–860.

[56] P.-T. De Boer, D. P. Kroese, S. Mannor, and R. Y. Rubinstein, "A tutorial on the cross-entropy method," *Annals of Operations Research*, vol. 134, pp. 19–67, 2005.

[57] Z. Wang, E. P. Simoncelli, and A. C. Bovik, "Multiscale structural similarity for image quality assessment," in *The Thrity-Seventh Asilomar Conference on Signals, Systems & Computers*, 2003, pp. 1398–1402.

[58] G. Mattyus, W. Luo, and R. Urtasun, "Deeproadmapper: Extracting road topology from aerial images," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 3438–3446.

[59] S. Gao, P. Zhang, T. Yan, and H. Lu, "Multi-scale and detail-enhanced segment anything model for salient object detection," in *Proc. ACM Int. Conf. Multimedia*, 2024, pp. 9894–9903.

[60] B. Xu, Q. Jiang, X. Zhao, C. Lu, H. Liang, and R. Liang, "Multidimensional exploration of segment anything model for weakly supervised video salient object detection," *IEEE Trans. Circuits Syst. Video Technol.*, pp. 1–1, 2024.

[61] J. Wang, K. C. Chan, and C. C. Loy, "Exploring clip for assessing the look and feel of images," in *Proc. AAAI Conf. Artif. Intell.*, 2023, vol. 37, pp. 2555–2563.

[62] H. Dai, C. Ma, Z. Yan, Z. Liu, E. Shi, Y. Li, P. Shu, X. Wei, L. Zhao, Z. Wu, et al., "Samaug: Point prompt augmentation for segment anything model," *arXiv preprint arXiv:2307.01187*, 2023.

[63] R. Margolin, L. Zelnik-Manor, and A. Tal, "How to evaluate foreground maps?," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2014, pp. 248–255.

[64] F. Perazzi, P. Krahenbuhl, Y. Pritch, and A. Hornung, "Saliency filters: Contrast based filtering for saliency region detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2012, pp. 733–740.

[65] D.-P. Fan, C. Gong, Y. Cao, B. Ren, M.-M. Cheng, and A. Borji, "Enhanced-alignment measure for binary foreground map evaluation," in *Proc. Int. Joint Conf. Artif. Intell.*, 2018, pp. 698–704.

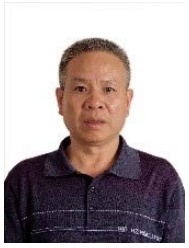
[66] D.-P. Fan, M.-M. Cheng, Y. Liu, T. Li, and A. Borji, "Structure-measure: A new way to evaluate foreground maps," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 4548–4557.

[67] R. Achanta, S. Hemami, F. Estrada, and S. Susstrunk, "Frequency-tuned saliency region detection," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2009, pp. 1597–1604.
[68] C. Lv, X. Zhou, B. Wan, S. Wang, Y. Sun, J. Zhang, and C. Yan, "Transformer-based cross-modal integration network for rgb-t salient object detection," *IEEE Trans. Consum. Electron.*, vol. 70, no. 2, pp. 4741–4755, 2024.
[69] X. Yu, X. Cheng, Y. Liu, and Z. Zheng, "A dual-stream cross-domain integration network for rgb-t salient object detection," *IEEE Trans. Consum. Electron.*, pp. 1–13, 2024.
[70] H. Zhou, C. Tian, Z. Zhang, C. Li, Y. Xie, and Z. Li, "Frequency-aware feature aggregation network with dual-task consistency for rgb-t salient object detection," *Pattern Recogn.*, vol. 146, pp. 110043, 2024.
[71] M. Jiang, J. Ma, J. Chen, Y. Wang, and X. Fang, "Patnet: Patch-to-pixel attention-aware transformer network for rgb-d and rgb-t salient object detection," *Knowl.-Based Syst.*, vol. 291, pp. 111597, 2024.
[72] J. Wang, G. Li, H. Yu, J. Xi, J. Shi, and X. Wu, "Intra-modality self-enhancement mirror network for rgb-t salient object detection," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 35, no. 3, pp. 2513–2525, 2025.
[73] H. Tang, Z. Li, D. Zhang, S. He, and J. Tang, "Divide-and-conquer: Confluent triple-flow network for rgb-t salient object detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 47, no. 3, pp. 1958–1974, 2025.
[74] X. Li, R. Hou, T. Ren, and G. Wu, "Kan-sam: Kolmogorov-arnold network guided segment anything model for rgb-t salient object detection," *arXiv preprint arXiv:2504.05878*, 2025.
[75] M.-M. Cheng, N. J. Mitra, X. Huang, P. H. S. Torr, and S.-M. Hu, "Global contrast based salient region detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 3, pp. 569–582, 2014.
[76] G. Li and Y. Yu, "Deep contrast learning for salient object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 478–487.
[77] S. Song, Z. Miao, H. Yu, J. Fang, K. Zheng, C. Ma, and S. Wang, "Deep domain adaptation based multi-spectral salient object detection," *IEEE Trans. Multimedia*, vol. 24, pp. 128–140, 2022.
[78] W. Gao, G. Liao, S. Ma, G. Li, Y. Liang, and W. Lin, "Unified information fusion network for multi-modal rgb-d and rgb-t salient object detection," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 4, pp. 2091–2106, 2021.
[79] F. Sun, P. Ren, B. Yin, F. Wang, and H. Li, "Catnet: A cascaded and aggregated transformer network for rgb-d salient object detection," *IEEE Trans. Multimedia*, vol. 26, pp. 2249–2262, 2024.
[80] X. Hu, F. Sun, J. Sun, F. Wang, and H. Li, "Cross-modal fusion and progressive decoding network for rgb-d salient object detection," *Int. J. Comput. Vis.*, vol. 132, no. 8, pp. 3067–3085, 2024.
[81] P. Ren, T. Bai, and F. Sun, "Bio-inspired two-stage network for efficient rgb-d salient object detection," *Neural Networks*, p. 107244, 2025.
[82] Konstantin Sofiiuk, Ilia Petrov, Olga Barinova, and Anton Konushin, "f-brs: Rethinking backpropagating refinement for interactive segmentation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 8623–8632.
[83] Y. Niu, Y. Geng, X. Li, and F. Liu, "Leveraging stereopsis for saliency analysis," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2012, pp. 454–461.
[84] R. Ju, L. Ge, W. Geng, T. Ren, and G. Wu, "Depth saliency based on anisotropic center-surround difference," in *Proc. IEEE Int. Conf. Image Process.*, 2014, pp. 1115–1119.
[85] H. Peng, B. Li, W. Xiong, W. Hu, and R. Ji, "Rgbd salient object detection: A benchmark and algorithms," in *Proc. Eur. Conf. Comput. Vis.*, 2014, pp. 92–109.
[86] Y. Xiong, C. Zhou, X. Xiang, L. Wu, C. Zhu, Z. Liu, S. Suri, B. Varadarajan, R. Akula, F. Iandola, et al., "Efficient track anything," 2024, arXiv preprint.

**Ruichao Hou** (Member, IEEE) received his Ph.D degree from the Department of Computer Science and Technology, Nanjing University in 2023. He is currently an Assistant Researcher at the Software Institute of Nanjing University. His research mainly focuses on multi-modal object detection and tracking.

**Xingyuan Li** received the B.S. degree from the School of Artificial Intelligence at Nanjing University, Nanjing, China, where he is currently pursuing the Ph.D. degree at the School of Artificial Intelligence at Nanjing University. His current research interests include RGB-T salient object detection.

**Tongwei Ren** (Member, IEEE) received the B.S., M.E., and Ph.D. degrees from Nanjing University, Nanjing, China, in 2004, 2006, and 2010, respectively. He joined Nanjing University in 2010, and at present he is a professor. His research interests mainly include multimedia computing and its real-world applications. He has published more than 40 papers in top-tier journals and conferences. He was a recipient of the best paper candidate awards of ICIMCS 2014, PCM 2015, and MMAsia 2020, and he was in the champion teams of ECCV 2018 PIC challenge, MM 2019 VRU challenge, MM 2020 DVU challenge, MM 2022 DVU challenge and MM 2023 DVU challenge.



**Dongming Zhou** received the B.S. and M.S. degrees in industry automatization from the Department of Automatic Control Engineering, Huazhong University of Science and Technology, Wuhan, China, in 1985 and 1988, respectively, and the Ph. D. degree in circuitry and system from the School of Information Science and Technology, Fudan University, Shanghai, China, in 2004. In 2008, he was a Visiting Scholar with York University, Toronto, ON, Canada. He is currently a Professor at the School of Information Science and Engineering, Yunnan University, Kunming, China. His current research interests include biomedical engineering, computational intelligence, complex systems, neural networks, and their applications.

**Gangshan Wu** (Member, IEEE) received the B.Sc., M.S., and Ph.D. degrees from the Department of Computer Science and Technology, Nanjing University, Nanjing, China, in 1988, 1991, and 2000, respectively. He is currently a Professor with the School of Computer Science, Nanjing University. His current research interests include computer vision, multimedia content analysis, multimedia information retrieval, digital museum, and large-scale volumetric data processing.

**Jinde Cao** (Fellow, IEEE) received the B.S. degree in mathematics/applied mathematics from Anhui Normal University, Wuhu, China, in 1986, the M.S. degree in mathematics/applied mathematics from Yunnan University, Kunming, China, in 1989, and the Ph.D. degree in mathematics/applied mathematics from Sichuan University, Chengdu, China, in 1998. He is currently an Endowed Chair Professor, the Dean of the School of Mathematics, the Director of the Jiangsu Provincial Key Laboratory of Networked Collective Intelligence of China and the Research Center for Complex Systems and Network Sciences, Southeast University, Nanjing, China.

Prof. Cao is elected as a member of the Academy of Europe and the European Academy of Sciences and Arts, a Foreign Member of the Russian Academy of Natural Sciences and the Lithuanian Academy of Sciences, a fellow of the Pakistan Academy of Sciences and the African Academy of Sciences, and an IASCYS Academician. He was a recipient of the National Innovation Award of China, the Obada Prize, and the Highly Cited Researcher Award in Engineering, Computer Science, and Mathematics by Thomson Reuters/Clarivate Analytics.