SAMSOD: Rethinking SAM Optimization for RGB-T Salient Object Detection

Zhengyi Liu, Xinrui Wang, Xianyong Fang, Zhengzheng Tu, Linbo Wang*

***Abstract*—RGB-T salient object detection (SOD) aims to segment attractive objects by combining RGB and thermal infrared images. To enhance performance, the Segment Anything Model has been fine-tuned for this task. However, the imbalance convergence of two modalities and significant gradient difference between high- and low- activations are ignored, thereby leaving room for further performance enhancement. In this paper, we propose a model called *SAMSOD*, which utilizes unimodal supervision to enhance the learning of non-dominant modality and employs gradient deconfliction to reduce the impact of conflicting gradients on model convergence. The method also leverages two decoupled adapters to separately mask high- and low-activation neurons, emphasizing foreground objects by enhancing background learning. Fundamental experiments on RGB-T SOD benchmark datasets and generalizability experiments on scribble supervised RGB-T SOD, fully supervised RGB-D SOD datasets and full-supervised RGB-D rail surface defect detection all demonstrate the effectiveness of our proposed method.** [1]

***Index Terms*—multi-modal, salient object detection, gradient, activation, SAM.**

I. Introduction

RGB-T salient object detection (SOD) [1], [2] aims to identify and segment the most attractive objects with the help of RGB and thermal infrared images, especially in certain extreme conditions, such as nighttime, foggy and rainy environment. Thermal infrared images can provide strong contrast information on temperature changes, thus compensating for the shortcomings of RGB images due to high noise, low light, and high exposure, making RGB-T SOD a topic of practical significance.

Segment Anything Model (SAM) [3] [4] has emerged as a powerful segmentation model [5]–[7] by leveraging the advanced transformer architecture and being trained on a large volume of data. The common practice of fine-tuning SAM for RGB-T segmentation task [8], [9] is shown in Fig 1 (a). RGB and thermal images are respectively fed into two encoders, which are a frozen and shared SAM encoder along with two groups of finetuned adapters [10], [11], [12]. The extracted features are fused and fed into the full-tuned SAM decoder to obtain the prediction result.

This work is supported by National Natural Science Foundation of China under Grant 62376005 (Corresponding author: Linbo Wang).

Zhengyi Liu, Xinrui Wang, Xianong Fang, Zhengzheng Tu, and Linbo Wang are with Key Laboratory of Intelligent Computing and Signal Processing of Ministry of Education, School of Computer Science and Technology, Anhui University, Hefei, China(e-mail: liuzywen@ahu.edu.cn, 2325687760@qq.com, fangxianyong@ahu.edu.cn, zhengzhengahu@163.com, wanglb@ahu.edu.cn)

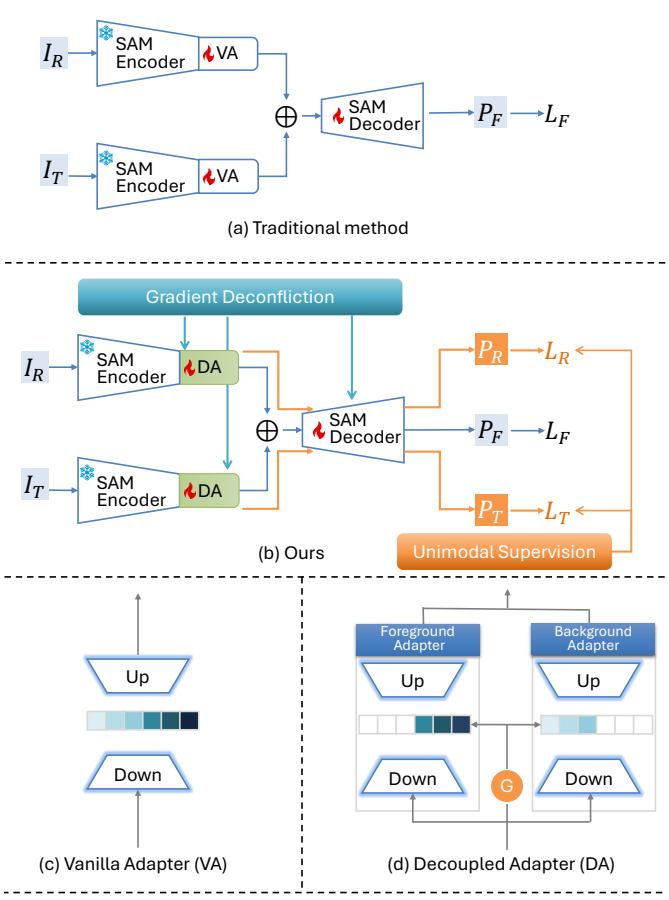


Fig. 1. The framework comparison between traditional methods and ours and their adapters.

However, imbalance convergence of two modalities and significant gradient difference between high- and low- activations usually affect SAM's optimization.

Imbalance convergence of two modalities occurs in the scenario with dominant modality [13] [14]. Concretely, the gradient of decoder will affect those of two encoders according to the backward propagation rule. Once an encoder for dominant modality and the decoder converge, another encoder for non-dominant modality will be under-optimized, leading to poor multi-modal learning performance.

To address the imbalance convergence, we introduce unimodal supervision to enhance the learning of each modality, as shown in the orange flow of Fig 1 (b). RGB feature, thermal feature, and fusion feature are all fed into the shared SAM decoder. Three predictions are supervised by the ground truth.

Unfortunately, gradient conflict may appear in the encoders and the decoder. The gradients of unimodal supervision and fusion supervision mutually impact the RGB encoder and thermal encoder, and the ones of all supervision collaboratively influence the decoder. Once gradient conflict occurs, the training of the model may fail to converge to an ideal solution. Therefore, a gradient deconfliction strategy is applied to mitigate the impact of inconsistent gradient directions in the RGB encoder, the thermal encoder, and the decoder, respectively, as shown in the cyan flow of Fig 1 (b).

The significant gradient difference between high- and low-activations occurs in the vanilla adapter. Concretely, in the optimization process of SAM, the adapter (Fig 1 (c)), a bottleneck structure which consists of a down-sampling, an activation, and a up-sampling, is responsible for learning modal-specific features. The neurons after activation can be divided into two groups. High-activation neurons provide important information related with the learning object while low-activation neurons encode background cues [15]. The derivative of the activation function for high-activation neurons is close to 1, allowing the gradient to be fully backpropagated, whereas for low-activation neurons, the derivative is close to 0, making effective parameter updates impossible. In other words, the model tends to focus on learning the foreground while ignoring the background, causing insufficient background learning. However, in many cases, learning the background can, in turn, guide the learning of the foreground. For example, when a silver object appears in a kitchen background, it is more likely to be recognized as a pot rather than a wheel.

To address significant gradient difference between high- and low- activations, decoupled adapters are designed to synchronously strengthen the learning of foreground and background, as shown in Fig 1 (d). Foreground adapter is responsible for the learning of high-activation neurons, while background adapter is in charge of the training of low-activation neurons. By separately masking high- and low- activations, gradient interference from high-activation regions on the low-activation pathway is reduced, ensuring that low-activation features are adequately updated during training. Furthermore, the high- and low- activation levels are determined by a learnable gate represented by the orange 'G', eliminating the need to manually set the thresholds for activation.

Our contributions can be summarized as follows:

- We rethink SAM optimization for the RGB-T salient object detection task and propose that the issues of imbalance convergence between the two modalities and significant gradient difference between high- and low-activations significantly affect optimization performance.
- To address the imbalanced convergence of the two modalities, unimodal supervision and subsequent gradient deconfliction are employed to enhance training robustness, particularly by reinforcing the learning of weaker modality and preventing the model from training instability.
- To address significant gradient difference between high- and low- activations, decoupled adapters are utilized to simultaneously emphasize on the learning of foreground and background.
- The method achieves state-of-the-art performance on three RGB-T SOD datasets, and obtains the consistent and excellent results on weakly supervised RGB-T and full supervised RGB-D SOD, and full-supervised RGB-D rail surface defect detection, demonstrating the generalization ability.

II. Related Works

A. RGB-T Salient Object Detection

The salient object detection relying solely on RGB images remains susceptible to imaging conditions such as illumination variations, rainy or foggy weather, etc. Incorporating the thermal infrared sensors can provide temperature information to solve these problems to a certain extent.

Existing methods focus on model structures to achieve multi-modal fusion. MITF-Net [16] employs structural similarity multi-modal fusion, cross-level attention aggregation, and edge guidance. CGMDRNet [17] achieves intrinsic consistency feature fusion via reducing the modality differences. CMDBIF-Net [18] introduces an additional interactive branch and double bidirectional interaction among three encoder branches. CAVER [19] constructs a transformer framework and aligns two modalities from spatial and channel views. HRTransNet [20] introduces HRFormer backbones and utilizes the attention mechanism to improve the performance. PATNet [21] designs sophisticated modules to emphasize on patch-level and pixel-level complementarity, pursuing the completeness and detailed refinement. TNet [22] analyzes the role of thermal modality and controls the fusion between two modalities via illuminance score. XMSNet [23] addresses sensor noise and misalignment issue by mining the cross-modal semantics. LAFB [24] constructs an adaptive fusion bank. Some methods, such as FFANet [25], DSCDNet [26], and WaveNet [27] address multi-modal fusion from frequency perspective. ConTriNet [28] designs a unified encoder and three specific decoders to obtain modal-specific and modal-complementary information. SACNet [29] proposes a new detection task for original unaligned RGB-T image pairs and gives an alignment-free solution. SPDE [30] tackles salient object detection task of underwater scenes via dual-stage self-paced learning and adaptive depth emphasis.

In the paper, we propose a SAM based RGB-T model. For optimizing SAM to RGB-T model, we analyze the mechanism of multi-modal learning in encoder-decoder framework, and introduce unimodal supervision to reduce the modality dominant, and further mitigate the conflict within encoder and decoder via gradient deconfliction.

B. SAM's Optimization for Multi-modal Task

SAM has played a crucial role in the segmentation tasks of RGB images [31], remote-sensing images [32], medical images [33], [34]. Due to the scarcity of paired data, RGB-T foundation model is currently challenging. Therefore, fine-tuning SAM for RGB-T multi-modal task is a more feasible option. There are two mainstream solutions. One is auxiliary modality injection method. SE-Adapter [35] feeds event modality into SAM via attention-like adapter. Sammese [8]

injects the multi-modal information via a multi-modal adapter and further generates the useful prompts to help SAM segment saliency-related regions. OpenRSS [36] finetunes SAM via thermal information prompt and dynamic low-rank adaptation in open-vocabulary RGB-T semantic segmentation task. GoPT [37] further emphasizes on the grouping prompt tuning. The other is two-stream parallel paradigm. SSFam [38] uses two finetuned SAMs to extract multi-modal features, achieving a scribble supervised SOD method. CPAL [9] uses the bi-directional cross-prompting tuning to optimize the foundation model.

The aforementioned methods elaborately design model structure and finetuning strategy. However, imbalance convergence optimization of two modalities is not addressed. In the paper, we rethink SAM's optimization for two-stream segmentation model.

C. Imbalance Optimization in Multi-modal Learning

Modality dominance in multi-modal learning has received widespread attention. One modality can end up dominating the learning process, restricting the effective use of information from other modalities and resulting in suboptimal model performance. OGM-GE [13] points out optimization imbalance issue in multi-modal classification task and proposes to modulate the gradient of each modality via contribution discrepancy to the learning objective. PMR [39] introduces prototype to address the same issue. AGM [40] improves OGM-GE to accommodate complex fusion strategies and more modalities. ReGrad [41] reduces the influence of dominant modality via modulating both conflicted and unconflicted gradients. CGGM [42] focuses on both the magnitude and direction of gradient. MLA [43] uses an alternating unimodal learning process to minimize interference between modalities. GDNet [44] decouples gradient into probabilistic distribution to alleviate the gradient coupling between modalities, promoting the optimization of each modality. MMPareto [14] proposes the optimization in multitask-like multimodal framework considering both gradient direction and gradient magnitude via Pareto idea which finds a trade-off gradient beneficial for all objectives. DRL [45] leverages varying degrees of re-initialization on the encoders of different modalities to unlock the full potential of each modality. Fuller [46] calibrates gradient across tasks and modalities, respectively. Wei et al [47] achieve a balanced integration of modalities by addressing the sample-level modality discrepancy.

Inspired by aforementioned methods, we analyse modality dominant issue in the two-stream encoder-decoder framework for segmentation task and give the corresponding solution.

III. Method

To adapt SAM for the RGB-T salient object detection task, SAM is optimized from two perspectives, resulting in a model called *SAMSOD*, as shown in Fig 2. It introduces the unimodal supervision (orange flow) and gradient deconfliction (cyan flow) to address the imbalance convergence issue of RGB and thermal modalities. Moreover, it designs the decoupled adapters (green block) to solve the significant gradient difference between high- and low- activations issue.

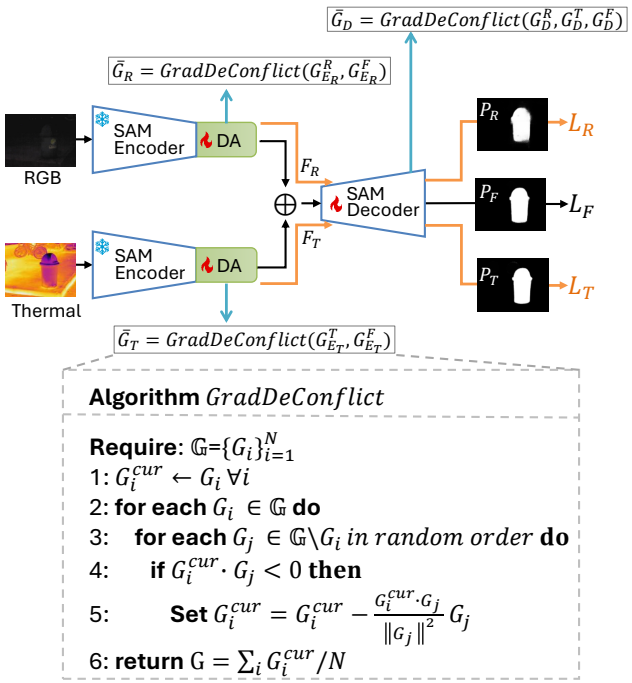


Fig. 2. The pipeline of the proposed *SAMSOD*.

A. Imbalance Convergence and Solution

Existing methods address RGB-T salient object detection task via the encoder-decoder framework. Specifically, as shown in the black flow of Fig 2, the two-stream encoder takes as input the paired RGB and thermal images $\{I_R, I_T\}$ and outputs multi-scale RGB features F_R and thermal features F_T, respectively.

$$\begin{aligned} F_R &= \mathcal{E}_R(I_R; \theta_R) \\ F_T &= \mathcal{E}_T(I_T; \theta_T) \end{aligned} \tag{1}$$

where $\mathcal{E}_R$ and $\mathcal{E}_T$ are the RGB encoder and thermal encoder. Here SAM2 [4] is adopted as the encoder due to its advanced performance in image segmentation task. Following the common practice, SAM2 encoder is frozen and shared, along with two groups of adapters with the corresponding parameters θ_R and θ_T to be finetuned.

Then the two features F_R and F_T are fused and fed into the decoder to generate prediction P_F.

$$P_F = \mathcal{D}(F_R + F_T; \theta_{\mathcal{D}}) \tag{2}$$

where $\mathcal{D}$ is the decoder with parameter $\theta_{\mathcal{D}}$.

Further, the loss between the prediction P_F and ground truth GT are calculated.

$$\mathcal{L}_F = \mathcal{L}_{wiou}(P_F, GT) + \mathcal{L}_{wbce}(P_F, GT) \tag{3}$$

where $\mathcal{L}_{wiou}$ and $\mathcal{L}_{wbce}$ are the weighted intersection-over-union (IoU) loss and the weighted binary cross-entropy (BCE) loss [48], $\mathcal{L}_F$ is called the fusion loss.

According to chain rule, the gradients of two encoders are represented as:

$$\begin{aligned} G_R &= \frac{\partial \mathcal{L}_F}{\partial \theta_R} = \frac{\partial \mathcal{L}_F}{\partial P_F} \cdot \frac{\partial P_F}{\partial \theta_{\mathcal{D}}} \cdot \frac{\partial \theta_{\mathcal{D}}}{\partial F_R} \cdot \frac{\partial F_R}{\partial \theta_R} \\ G_T &= \frac{\partial \mathcal{L}_F}{\partial \theta_T} = \frac{\partial \mathcal{L}_F}{\partial P_F} \cdot \frac{\partial P_F}{\partial \theta_{\mathcal{D}}} \cdot \frac{\partial \theta_{\mathcal{D}}}{\partial F_T} \cdot \frac{\partial F_T}{\partial \theta_T} \end{aligned} \tag{4}$$

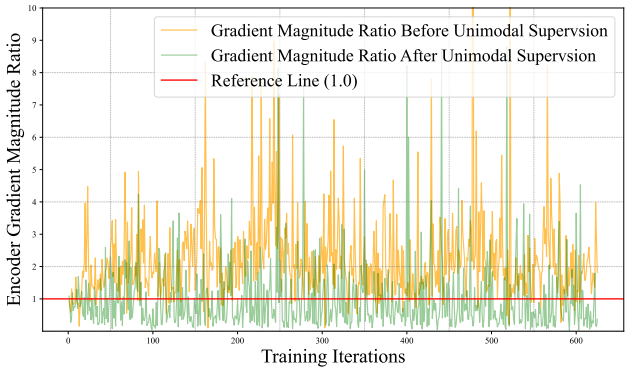


Fig. 3. Gradient magnitude ratio between RGB encoder and thermal encoder.

Due to the differences between the RGB and thermal modalities, the last two terms of Formula 4 are different, resulting in $G_R \neq G_T$. The modality with the larger gradient tends to converge earlier, while the one with the smaller gradient may not reach convergence, leading to an imbalanced convergence problem. Fig 3 draws the gradient relation between G_R and G_T. Obviously, the yellow curve is above the red horizontal line representing a ratio of 1, indicating G_R is greater than G_T in most cases. The parameters of the RGB encoder are updated more quickly than those of the thermal encoder, as shown by the following parameter update formula:

$$\begin{aligned} \theta_R^{t+1} &= \theta_R^t - \eta * G_R \\ \theta_T^{t+1} &= \theta_T^t - \eta * G_T \end{aligned} \tag{5}$$

where η is learning rate, and t is training iteration step. The difference in parameter update speed indicates the RGB modal dominates the training of the whole model and the thermal modal is under-optimized so that two modalities are not well utilized, which achieves sub-optimal performance.

To address the challenge, we introduce unimodal supervision to enhance the optimization of thermal modal, as shown in the orange flow of Fig 2. Concretely, the shared decoder is used to decode the RGB feature and thermal feature to obtain unimodal prediction P_R and P_T.

$$P_R = \mathcal{D}(F_R; \theta_{\mathcal{D}}), P_T = \mathcal{D}(F_T; \theta_{\mathcal{D}}) \tag{6}$$

Accordingly, two unimodal losses are calculated via:

$$\begin{aligned} \mathcal{L}_R &= \mathcal{L}_{wiou}(P_R, GT) + \mathcal{L}_{wbce}(P_R, GT) \\ \mathcal{L}_T &= \mathcal{L}_{wiou}(P_T, GT) + \mathcal{L}_{wbce}(P_T, GT) \end{aligned} \tag{7}$$

The total loss is defined as:

$$\mathcal{L} = \mathcal{L}_F + \mathcal{L}_R + \mathcal{L}_T \tag{8}$$

From Fig 2, we can observe that the encoder part is traversed by two flows: one in black and the other in orange, while the decoder part is traversed by three flows: one in black and two in orange. Formally, in the encoder $\mathcal{E}_R$ part, the gradient $G^R_{\mathcal{E}_R}$ from the unimodal loss $\mathcal{L}_R$ and the gradient $G^F_{\mathcal{E}_R}$ from the fusion loss $\mathcal{L}_F$ jointly influence parameter update, in the encoder $\mathcal{E}_T$ part, the gradient $G^T_{\mathcal{E}_T}$ from the unimodal loss $\mathcal{L}_T$ and the gradient $G^F_{\mathcal{E}_T}$ from the fusion loss $\mathcal{L}_F$ jointly influence parameter update, and in the decoder $\mathcal{D}$ part, the gradient $G^R_{\mathcal{D}}$ from RGB stream loss, the gradient $G^T_{\mathcal{D}}$ from thermal stream loss, and the gradient $G^F_{\mathcal{D}}$ from fusion stream jointly influence parameter update.

According to chain rule, the gradients of two encoders are updated as:

$$\begin{aligned} G_R &= G^F_{\mathcal{E}^R} + G^R_{\mathcal{E}^R} = \frac{\partial \mathcal{L}^F}{\partial \theta^R} + \frac{\partial \mathcal{L}^R}{\partial \theta^R} \\ G_T &= G^F_{\mathcal{E}^T} + G^T_{\mathcal{E}^T} = \frac{\partial \mathcal{L}^F}{\partial \theta^T} + \frac{\partial \mathcal{L}^T}{\partial \theta^T} \end{aligned} \tag{9}$$

By the help of two unimodal losses, the gradients of RGB encoder and thermal encoder are influenced by both the fusion loss and their own respective losses indicated by the second term of Formula 9. Each encoder is guided to extract its own most discriminative features, without overly relying on the fusion loss, thereby alleviating the convergence imbalance. The green curve in Fig 3 gives the gradient magnitude ratio between RGB and thermal encoders after unimodal supervision. The green curve is below the yellow curve and oscillate around 1, indicating the gradient magnitude of RGB encoder and the gradient magnitude of thermal encoder are mostly equal. This also reveals that their convergence imbalance has been alleviated.

Similarly, the gradient of the decoder is updated as:

$$G_{\mathcal{D}} = G^F_{\mathcal{D}} + G^R_{\mathcal{D}} + G^T_{\mathcal{D}} = \frac{\partial \mathcal{L}^F}{\partial \theta^{\mathcal{D}}} + \frac{\partial \mathcal{L}^R}{\partial \theta^{\mathcal{D}}} + \frac{\partial \mathcal{L}^T}{\partial \theta^{\mathcal{D}}} \tag{10}$$

The Formulas 9 and 10 show that both G_R and G_T consist of two gradient terms each, while $G_{\mathcal{D}}$ consists of three gradient terms. Take $G_{\mathcal{D}}$ as an example, if the directions of $G^R_{\mathcal{D}}$, $G^T_{\mathcal{D}}$, and $G^F_{\mathcal{D}}$ are not consistent, the overall gradient $G_{\mathcal{D}}$ is no longer the optimal optimization direction of the three gradients, but rather a compromised direction resulting from their cancellation. It is called gradient conflict. Fig 4 draws the cosine similarity of different gradients among G^R_D, G^T_D, and G^F_D. Obviously, the cosine similarities of gradients are not always larger than 0, indicating there are some opposite gradient direction, which hinders effective optimization and slows down convergence. To this end, a gradient deconfliction is applied to alleviate the conflict of gradients to yield three updated gradients $\bar{G}_R$, $\bar{G}_T$, and $\bar{G}_{\mathcal{D}}$.

$$\bar{G}_R = \textit{GradDeConflict}(G^R_{\mathcal{E}^R}, G^F_{\mathcal{E}^R}) \tag{11}$$

$$\bar{G}_T = \textit{GradDeConflict}(G^T_{\mathcal{E}^T}, G^F_{\mathcal{E}^T}) \tag{12}$$

$$\bar{G}_{\mathcal{D}} = \textit{GradDeConflict}(G^R_{\mathcal{D}}, G^T_{\mathcal{D}}, G^F_{\mathcal{D}}) \tag{13}$$

where *GradDeConflict*($\mathbb{G}$) denotes a gradient deconfliction strategy on a set of gradients $\mathbb{G}$ following PCGrad [49], as shown in the bottom of Fig 2. Concretely, for a given gradient $G_i \in \mathbb{G}$, the current gradient G_i^{cur} is initialized as

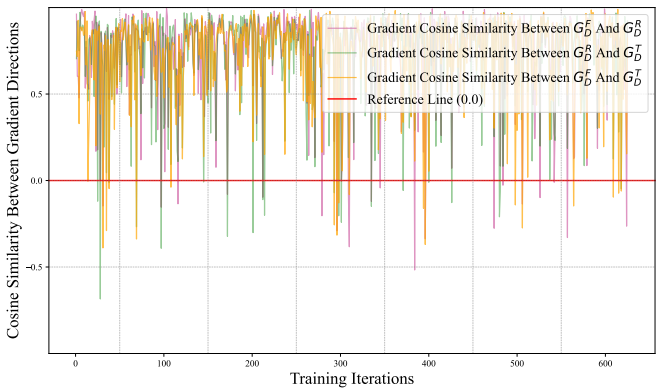


Fig. 4. The cosine similarity between gradients.

G_i. Then, the cosine similarity between current gradient G_i^{cur} and another gradient $G_j \in \mathbb{G}(j \neq i)$ is estimated.

$$S_{ij} = \frac{G_i^{cur} \cdot G_j}{\|G_i^{cur}\| \|G_j\|}, \tag{14}$$

If the cosine similarity is less than 0, it indicates a conflict in the optimization directions of the two gradients. G_i^{cur} subtracts its projection of G^j, otherwise, remains unchanged.

$$\begin{cases} G_i^{cur} = G_i^{cur} - \dfrac{G_i^{cur} \cdot G_j}{\|G_j\|^2} G_j, & if\ S_{ij} < 0 \\ G_i^{cur} = G_i^{cur}, & else \end{cases} \tag{15}$$

The potential conflicts between G_i^{cur} and all other gradients are eliminated in the same way.

The process is repeated among all the gradients. Finally, all the updated gradients are averaged to generate the final gradient G.

$$G = \sum_i G_i^{cur} / N \tag{16}$$

After eliminating the gradient conflicts in all the encoders and the decoder, the parameters of two encoders and the decoder are updated as:

$$\begin{aligned} \theta_R^{t+1} &= \theta_R^t - \eta * \bar{G}_R \\ \theta_T^{t+1} &= \theta_T^t - \eta * \bar{G}_T \\ \theta_{\mathcal{D}}^{t+1} &= \theta_{\mathcal{D}}^t - \eta * \bar{G}_{\mathcal{D}} \end{aligned} \tag{17}$$

B. Significant Gradient Difference Between High- and Low-Activations and Solution

To adapt SAM to RGB-T salient object detection task, the adapters are commonly used in practice. Generally, for a specific input feature $x \in \mathbb{R}^{b \times n \times d}$, a vanilla adapter is a bottleneck structure, which includes a down-projection layer with parameters $W_{down} \in \mathbb{R}^{d \times \hat{d}}$, a GeLU activation $\sigma(\cdot)$, and an up-projection layer with parameters $W_{up} \in \mathbb{R}^{\hat{d} \times d}$, where $\hat{d}$ is the bottleneck middle dimension and satisfies $\hat{d} \ll d$.

$$\tilde{x} = W_{up} \cdot \sigma(W_{down} \cdot x) \tag{18}$$

Let z=$W_{down} \cdot x$ and h=$\sigma(z)$, the derivative of the total loss $\mathcal{L}$ with respect to W_{down} is denoted as:

$$\frac{\partial \mathcal{L}}{\partial W_{down}} = \frac{\partial L}{\partial \tilde{x}} \cdot \frac{\partial \tilde{x}}{\partial h} \cdot \frac{\partial h}{\partial z} \cdot \frac{\partial z}{\partial W_{down}} = \frac{\partial L}{\partial \tilde{x}} \cdot W_{up}^T \cdot \sigma'(z) \cdot x^T \tag{19}$$

where the first term is the derivative of the total loss with respect to the output of adapter $\tilde{x}$, the second term is the transpose of the up-projection weights, the third term is the derivative of the GeLU activation function on z, and the fourth term is the transpose of the input x. From the formula, we can observe that the derivative of the GeLU activation function on z determines the gradient of W_{down}. Concretely, $\sigma'(z)$ is defined as:

$$\sigma'(z) \approx \begin{cases} 1, & z \gg 0 \\ 0.5, & z \approx 0 \\ 0, & z \ll 0 \end{cases} \tag{20}$$

where $\sigma'(z) \approx 1$ represents that the signal propagates freely without any suppression when $z \gg 0$, allowing the network to update the gradients normally, $\sigma'(z) \approx 0.5$ represents that the signal propagation is reduced by half when $z \approx 0$, and the network's gradient flow is partially suppressed, and $\sigma'(z) \approx 0$ represents that the signal barely passes through when $z \ll 0$, meaning the gradients of these neurons are nearly zero, making effective parameter updates impossible. In other words, the high-activation features ($z \gg 0$) dominate gradient propagation, overshadowing low-activation information ($z \ll 0$). Furthermore, high-activation neurons generally highlight the outline of salient objects which is important for salient object detection, while low-activation part falls more on background for the task, which is can be seen from Fig 5. Moreover, the learning of background can provide robust context-aware cues to distinguish salient foreground objects from the background. Therefore, to reduce gradient interference from high-activation regions on the low-activation pathway and ensure that low-activation features are adequately updated during training, we decouple the adapter into a foreground adapter and a background adapter to synchronously enhance the learning of foreground and background.

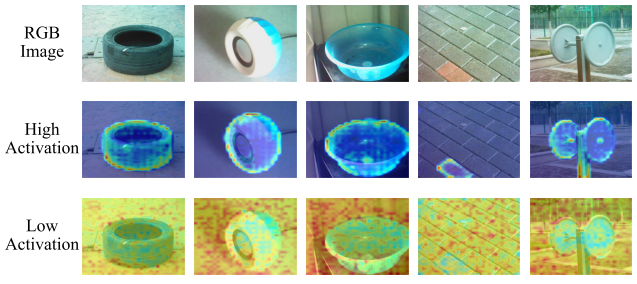


Fig. 5. High- and low- activation maps of RGB images.

Concretely, the foreground adapter learns high-activation and ignores the others by dropout operation, and the background adapter learns low-activation and dropout the others.

$$\begin{aligned} \tilde{x} = W_{up}^{for} \cdot \textit{TopK}(\sigma(W_{down}^{for} \cdot x), P_{for}, \textit{True}) \\ + W_{up}^{back} \cdot \textit{TopK}(\sigma(W_{down}^{back} \cdot x), P_{back}, \textit{False}) \end{aligned} \tag{21}$$

where $TopK(input, k, largest)$ retains the largest k channels and dropouts the others for non-linear-transformed activated feature *input* when *largest* is *True*, and retains the smallest k channels and dropouts the others when *largest* is *False*, and P_{for} and P_{back} are used to determine the extent of foreground and background activation levels, W_{down}^{for} and W_{up}^{for} are parameters of foreground adapter, while W_{down}^{back} and W_{up}^{back} are parameters of background adapter.

In the foreground adapter, low-activation neurons are set as 0, indicating the difference of $\sigma'(z)$ between high-activation and low-activation is in the range of $[0.5, 1]$ according to the Formula 20. Similarly, in the background adapter, high-activation neurons are set as 0, the difference of $\sigma'(z)$ is in the range of $[0, 0.5]$. Compared with a vanilla adapter whose difference is in the range of $[0, 1]$, the two decoupled adapters reduce gradient differences, preventing high-activation neurons from suppressing low-activation ones. This allows both high- and low-activation neurons to be sufficiently learned.

To adaptively determine the extent of foreground and background activation levels, P_{for} and P_{back} are defined as:

$$\begin{aligned} P_{for} &= \left\lfloor \hat{d} \times (\alpha_{for} + \beta \cdot G_{for}) \right\rfloor \\ P_{back} &= \left\lfloor \hat{d} \times (\alpha_{back} + \beta \cdot G_{back}) \right\rfloor \end{aligned} \tag{22}$$

where $\lfloor \cdot \rfloor$ is floor function, $\hat{d}$ is bottleneck dimension with 32 channels, α_{for} and α_{back} are predefined activation ratio for the foreground adapter and background adapter, β is a temperature scale, and G_{for} and G_{back} are learnable activation ratios of foreground and background from a gate. Specifically, a linear layer with weight matrix $A \in \mathbb{R}^{n \times d}$ and bias $b \in \mathbb{R}^n$ and a softmax function are used to calculate the weights of the gate.

$$[G_{for}, G_{back}] = Softmax(Ax + b) \tag{23}$$

By separately masking high- and low- activations, this approach reduces gradient disparities and avoid the phenomenon where strong activations dominate while weak activations are ignored.

IV. Experiments

A. Dataset and Evaluation Metrics

Three RGB-T salient object detection datasets are used to assess the performance of the proposed method. The VT821 dataset [50] comprises 821 manually aligned image pairs. The VT1000 dataset [51] encompasses 1,000 RGB-T image pairs captured using highly synchronized RGB and thermal cameras. The VT5000 dataset [2] includes 5,000 pairs of high-resolution, diverse, and low-deviation RGB-T images. To ensure a fair comparison, we employ the same training dataset according to the usual settings, which consists of 2,500 image pairs selected from the VT5000 dataset. The remaining image pairs are reserved for testing purposes.

Four evaluation metrics are utilized, including S-measure (S) [52], max F-measure (F_β) [53], max E-measure (E_ξ) [54], and mean absolute error (M) [55].

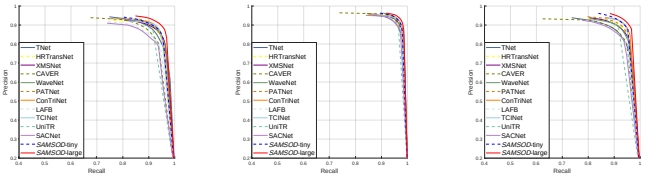
B. Implementation Details

Our method is implemented based on PyTorch on a PC with an NVIDIA RTX 3090 GPU. The input image size is 512×512. During the training process, photometric data augmentation is utilized for the RGB-T input data. We leverage the Adam optimizer to train our model. The max training epoch is set to 45 and the learning rate is 1e-3. Based on the large and tiny version of SAM2, two versions of the model are implemented: *SAMSOD*-large and *SAMSOD*-tiny. For *SAMSOD*-large, the batch size is set to 4, and the training process takes approximately 15 hours. In contrast, *SAMSOD*-tiny uses a batch size of 8 and requires around 4 hours for training. Empirically, the initial threshold α_{for} of foreground adapter is 0.45, the initial threshold α_{back} of background adapter is 0.35. The temperature scale β is 0.1. For testing, the sum of three parallel saliency predictions is the final prediction.

C. Comparison Experiments

1) ***Comparison Methods****:* We compare our method with some RGB-T SOD algorithms, including TNet [22], HRTransNet [20], CAVER [19], WaveNet [27], PRLNet [56], PATNet [21], MAGNet [57], FFANet [25], ConTriNet [28], LAFB [24], ISMNet [58], DSCDNet [26], TCINet [59], UniTR [60], and SACNet [29]. For a fair comparison, we either use the performance results reported in the original papers or evaluate the saliency maps provided by the authors using the same evaluation tools.

2) ***Quantitative Analysis****:* From Table I, it is clear that *SAMSOD*-large achieves the best results regardless of individual dataset or overall performance. In fact, by introducing unimodal supervision and further employing gradient deconfliction strategy, the two modalities are well utilized to jointly achieve the better segmentation performance. Meanwhile, the proposed decoupled adapters finetuned on the frozen SAM2 encoder enhances the understanding of foreground objects by studying the background, leading to a more accurate segmentation of the foreground region. In addition, the comparison of PR curves in Fig 6 further validates the superiority of *SAMSOD*-large, because our PR curve is closer to the top-right corner indicating that both precision and recall are highest.



(a) VT821 dataset (b) VT1000 dataset (c) VT5000 dataset

Fig. 6. P-R curves comparison of different models on three RGB-T datasets.

3) ***Qualitative Analysis****:* Fig 7 presents the visualization comparison among saliency maps of *SAMSOD*-large and competitors in various complex scenarios, including complex background (1^{st} and 2^{nd} rows), multiple objects (3^{rd} and 4^{th} rows), small objects (5^{th} and 6^{th} rows), fine-grained objects (7^{th} and 8^{th} rows), objects in low light environment (9^{th}

TABLE I
Quantitative comparisons on three RGB-T SOD datasets and overall performance. '-' indicates that the corresponding cost is not reported in the original paper. The best results are bold.

Methods	Source	Overall				VT821				VT1000				VT5000				Cost		
		S↑	F_β ↑	E_ξ ↑	M↓	S↑	F_β ↑	E_ξ ↑	M↓	S↑	F_β ↑	E_ξ ↑	M↓	S↑	F_β↑	E_ξ ↑	M↓	Params (M)↓	FLOPs (G)↓	FPS (Hz)↑
TNet [22]	TMM22	.904	.894	.944	.030	.899	.888	.938	.030	.929	.930	.966	.021	.895	.881	.937	.033	87.0	39.7	49.6
HRTransNet [20]	TCSVT22	.917	.909	.958	.023	.906	.888	.941	.026	.938	.941	.975	.017	.912	.903	.956	.025	58.9	**17.3**	14.9
CAVER [19]	TIP23	.908	.894	.949	.025	.898	.877	.934	.027	.938	.939	.973	.017	.899	.882	.944	.028	93.8	31.6	28.5
WaveNet [27]	TIP23	.919	.907	.955	.023	.912	.895	.943	.024	.945	.945	.977	.015	.911	.896	.950	.026	**30.2**	26.7	7.7
PRLNet [56]	TIP23	.926	.878	.946	.022	.917	.860	.932	.025	.944	.902	.951	.016	.921	.875	.948	.023	-	-	-
PATNet [21]	KBS24	.921	.924	.952	.021	.914	.914	.938	.024	.941	.948	.964	.015	.916	.917	.951	.023	95.9	51.1	-
MAGNet [57]	KBS24	.916	.882	.942	.023	.909	.865	.930	.026	.938	.913	.949	.016	.909	.875	.943	.025	-	-	-
FFANet [25]	PR24	.921	.888	.948	.021	.905	.855	.926	.027	.943	.918	.955	.014	.918	.886	.953	.021	364.3	206.5	-
ConTriNet [28]	TPAMI24	.926	.899	.952	.019	.915	.878	.940	.022	.941	.918	.954	.015	.923	.898	.956	.020	96.3	126.9	-
LAFB [24]	TCSVT24	.910	.899	.950	.025	.900	.884	.940	.028	.932	.930	.969	.018	.904	.891	.945	.027	453.0	139.7	45.0
ISMNet [58]	TCSVT24	.920	.894	.947	.022	.917	.886	.945	**.021**	.942	.922	.954	.014	.913	.885	.945	.025	114.3	100.4	6.2
DSCDNet [26]	TCE24	.924	.893	.949	.021	.915	.876	.940	.022	.946	.921	.955	.014	.918	.888	.949	.023	92.3	134.1	-
TCINet [59]	TCE24	.926	.910	.965	.018	.914	.886	.951	**.021**	.942	.928	.976	.014	.924	.910	.965	.019	88.2	91.9	25.9
UniTR [60]	TMM24	.913	.914	.958	.021	.901	.898	.941	.025	.938	.939	.975	.014	.907	.910	.956	.023	72.0	-	**133.3**
SACNet [29]	TMM24	.921	.892	.952	.020	.906	.859	.932	.025	.942	.927	.958	.014	.917	.888	.957	.021	327.7	-	27.0
***SAMSOD*-tiny**	-	.925	.921	.964	.020	.913	.895	.950	.022	.942	.942	.976	.015	.923	.921	.964	.021	32.7	60.4	23.1
***SAMSOD*-large**	-	**.935**	**.933**	**.969**	**.017**	**.924**	**.916**	**.956**	**.021**	**.948**	**.949**	**.981**	**.012**	**.934**	**.933**	**.969**	**.017**	224.1	418.0	6.4

and 10^{th} rows), and objects in strong noise condition (11^{th} and 12^{th} rows). From the results we can observe that our method produces saliency maps that most closely resemble the ground truth across aforementioned scenes. For example, in the first two rows, foreground objects are well recognized by learning the background. In the last four rows, although the RGB modality provides limited information in low-light and high-noise environments, thermal images remain effectively utilized, demonstrating the effectiveness of unimodal supervision.

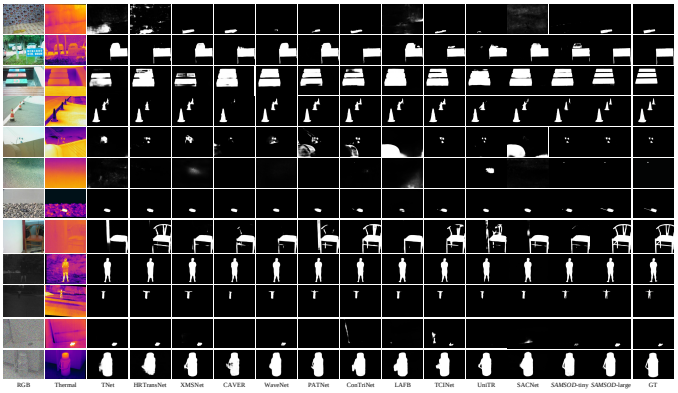


Fig. 7. Qualitative comparisons of visualization in complex background (1^{st} and 2^{nd} rows), multiple objects (3^{rd} and 4^{th} rows), small objects (5^{th} and 6^{th} rows), fine-grained objects (7^{th} and 8^{th} rows), objects in low light environment (9^{th} and 10^{th} rows), and objects in strong noise condition (11^{th} and 12^{th} rows)

*4) **Failure Cases Analysis**:* Fig 8 presents the failure cases. The first two rows show the poor performance when handling hollow objects. Our method primarily focuses on optimizing two modalities while avoiding over-reliance on either one. The segmentation accuracy relies on the fine-tuned SAM, which is based on the Vision Transformer (ViT) architecture. While ViT excels at capturing global context, it is less sensitive to the hollow objects. Additionally, the RGB-T salient object detection datasets are not specifically designed for hollow object segmentation, further limiting the model's ability in this regard. A potential solution for improving the segmentation ability of hollow objects is to incorporate edge-aware or structure-aware modules to address this limitation. The last two rows display the subpar results in local overexposure scenes. Local overexposure alters the overall semantics of the RGB image, leading to incorrect recognition of salient objects, such as the white background and red mat in the third row, and the hollow circle in the fourth row. A potential solution is to incorporate diverse exposure augmentation strategies to improve model robustness.

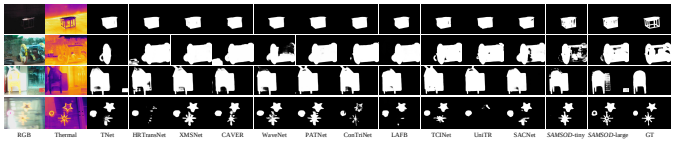


Fig. 8. Failure cases involving hollow objects (1^{st} and 2^{nd} rows) and local overexposure scenes (3^{rd} and 4^{th} rows).

*5) **Cost Analysis**:* The last column in Table I shows the computation cost comparison with the state-of-the-art methods. Specifically, *SAMSOD*-large contains $224.1M$ parameters, requires $418.0G$ FLOPs, and runs at 6.4 FPS. It must be acknowledged that while our method achieves good performance, the computational cost is also considerable. To further meet the requirements of real-time applications, we provide a tiny version which adopts SAM2-tiny in the encoder. It achieves top-three performance while significantly improving real-time efficiency. Specifically, *SAMSOD*-tiny contains $32.7M$ parameters, requires $60.4G$ FLOPs, and runs at 23.1 FPS. *SAMSOD*-large is better suited for scenarios requiring high accuracy, while *SAMSOD*-tiny offers a more balanced trade-off between performance and computational cost.

D. Ablation Study

1) Optimization Effectiveness Analysis: Table II shows the effectiveness of two optimization solutions. The first row shows the baseline model SAM with vanilla adapters. By introducing the first optimization solution to address the imbalance convergence issue, the performance in the second row shows a significant improvement. It benefits from unimodal supervision and gradient deconfliction, which enhance the learning of both modalities while also preventing the model from converging to a suboptimal solution, thereby improving robustness. By introducing the second optimization solution to address the significant gradient difference between high- and low- activations issue, the performance in the third row show a notable

improvement from the baseline due to decoupled adapters which mitigate the gradient difference between foreground and background, preventing the background from being ignored during learning. The comparison between the two groups of optimizations shows that the first group of optimization leads to more significant improvements. At last, the two groups of optimizations together achieve the best results shown in the fourth row.

TABLE II
Ablation study on optimizations for addressing the imbalance convergence between two modalities and significant gradient difference between high- and low- activations.

SAM	Unimodal Supervsion	w/o Vanilla Adapters	VT821				VT1000				VT5000			
w/ Vanilla Adapters	Gradient Deconfliction	w/ Decoupled Adapters	$S\uparrow$	$F_\beta\uparrow$	$E_\xi\uparrow$	$M\downarrow$	$S\uparrow$	$F_\beta\uparrow$	$E_\xi\uparrow$	$M\downarrow$	$S\uparrow$	$F_\beta\uparrow$	$E_\xi\uparrow$	$M\downarrow$
✓			.915	.899	.946	.025	.940	.939	.976	.015	.923	.918	.961	.021
✓	✓		.920	.907	.951	.021	.947	.948	.980	.013	.932	.929	.967	.018
✓		✓	.918	.904	.952	.023	.941	.942	.981	.014	.925	.921	.964	.020
✓	✓	✓	**.924**	**.916**	**.956**	**.021**	**.948**	**.949**	**.981**	**.012**	**.934**	**.933**	**.969**	**.017**

TABLE III
Ablation study on the effectiveness of optimization in addressing the imbalance convergence issue.

Methods	VT821				VT1000				VT5000			
	S↑	$F_\beta\uparrow$	$E_\xi\uparrow$	M↓	S↑	$F_\beta\uparrow$	$E_\xi\uparrow$	M↓	S↑	$F_\beta\uparrow$	$E_\xi\uparrow$	M↓
Only RGB (SAM+**VA**)	.916	.900	.941	.024	.945	.947	.978	.016	.924	.915	.959	.023
RGB-T (SAM+**VA**)	.915	.899	.946	.025	.940	.939	.976	.015	.923	.918	.961	.021
Add Unimodal Supervision	.919	.902	.947	.023	.945	.943	.979	.013	.930	.925	.966	.019
Add Gradient Deconfliction	.920	.907	.951	**.021**	.947	.948	.980	.013	.932	.929	.967	.018
RGB-T (SAM+**DA**)	.918	.904	.952	.023	.941	.942	.981	.014	.925	.921	.964	.020
Add Unimodal Supervision	.922	.906	.952	.022	.946	.948	.979	.014	.930	.928	.965	.019
Add Gradient Deconfliction	**.924**	**.916**	**.956**	**.021**	**.948**	**.949**	**.981**	**.012**	**.934**	**.933**	**.969**	**.017**

TABLE IV
Generalizability experiment on scribble supervised RGB-T salient object detection dataset.

Methods	Source	VT821				VT1000				VT5000			
		S↑	$F_\beta\uparrow$	$E_\xi\uparrow$	M↓	S↑	$F_\beta\uparrow$	$E_\xi\uparrow$	M↓	S↑	$F_\beta\uparrow$	$E_\xi\uparrow$	M↓
WSVSOD [61]	CVPR21	.822	.762	.875	.052	.886	.870	.935	.035	.812	.763	.880	.055
DENet [62]	TIP22	.813	.732	.839	.054	.906	.891	.947	.028	.839	.794	.895	.048
RGBTScribble [63]	ICME23	.895	.878	.942	.027	.925	.922	.964	.020	.877	.859	.933	.033
LGR [64]	TCSVT24	.893	.870	.941	.034	.932	.918	.964	.019	.887	.866	.939	.030
Ours (*SAMSOD*-large)	-	**.915**	**.896**	**.950**	**.022**	**.944**	**.945**	**.979**	**.014**	**.916**	**.908**	**.958**	**.023**

2) Optimization Effectiveness about Imbalance Convergence: Table III shows the optimization effectiveness about imbalance convergence issue. The first row 'Only RGB' refers to a single-stream model that uses the RGB modality without incorporating the thermal modality. The second row 'RGB-T (SAM+VA)' refers to a double-stream model that uses two modalities. Both are constructed on SAM model with vanilla adapters. By the comparison, we find double-steam model is not absolutely better than single-stream model, suggesting that the contribution of the thermal modality has not been fully utilized. By introducing the unimodal supervision, the thermal modal is well utilized so that the gradient ratio between RGB and thermal encoders is reduced, which can be seen the green curve from Fig 3. Meanwhile, the third row also shows a corresponding improvement, with only 1 out of the 12 metrics decreasing. Furthermore, gradient deconfliction further improve the performance of salient object detection by reducing potential gradient conflict. The second group of Table III also demonstrates that the unimodal supervision and gradient deconfliction progressively enhance the performance based on decoupled adapters instead of vanilla adapters. Fig 9 gives the visualization. In scenarios with blurring RGB images, low light, strong light, shadows, and occlusions, unimodal supervision enhances the learning of the thermal infrared modality, while gradient deconfliction prevents the model from getting stuck in a suboptimal solution, together achieving performance close to the ground truth.

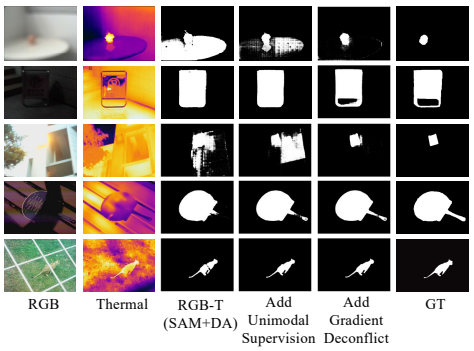


Fig. 9. The visualization of the ablation study on unimodal supervision and gradient deconfliction in scenarios with blurring RGB images (1^{st}), low light (2^{nd}), strong light (3^{rd}), shadows (4^{th}), and occlusions (5^{th}).

3) Optimization Effectiveness about the Mitigation of Significant Gradient Difference: Fig 10 shows the decoupled adapters give a clearer object boundary compared with the vanilla adapter. The quantitative comparison about decoupled adapters and vanilla adapters can be seen in the comparison between the first and third or second and fourth rows of Table II. By zeroing out low- and high- activation neurons in the foreground adapter and background adapter respectively, the gradient interference from high-activation regions on the low-activation pathway is reduced, allowing low-activation features to be adequately updated during training.

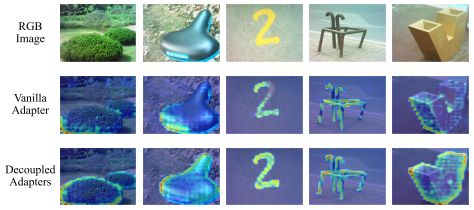


Fig. 10. The visual comparison between vanilla adapters (VA) and decoupled adapters (DA).

E. Generalizability Experiments

Generalizability experiments are conducted on scribble-supervised RGB-T SOD dataset, full-supervised RGB-D SOD dataset, and full-supervised RGB-D rail surface defect detection. The first is used to test the generalizability from full supervision to weakly supervision, while the last two are used to test the generalizability from RGB-T task to RGB-D task.

In the first generalizability experiment, RGBT-S dataset [63] is used to train and test. The comparison methods include WSVSOD [61], DENet [62], RGBTScribble [63], and LGR [64]. From Table IV, our method achieves the best performance in scribble supervised task. The excellent results are attributed to the resolution of imbalance convergence of two modalities and significant gradient difference between high- and low- activations existing in two-stream SOD frameworks, which are not covered by the compared methods.

In the second generalizability experiment, NLPR [65], NJU2K [66], STERE [67], and SIP [68] are used to train and test. The concrete partition of training and testing samples follows the common practice. The comparison methods include JL-DCF [69], SwinNet [70], PCIR-Net [71], HRTransNet [20], HiDANet [72], CATNet [73], TSVT [74], HFMDNet [75], MAGNet [57], EM-Trans [76], DFormer [77], TPCL [78], and DPPNet [79]. From Table V, our method achieves the comparative performance in RGB-D SOD task. The evaluation metric shows a noticeable improvement, especially on STERE and SIP dataset.

TABLE V
Generalizability experiment on fully supervised RGB-D salient object detection dataset.

Methods	Source	NLPR				NJU2K				STERE				SIP			
		S↑	$F_\beta\uparrow$	$E_\xi\uparrow$	M↓	S↑	$F_\beta\uparrow$	$E_\xi\uparrow$	M↓	S↑	$F_\beta\uparrow$	$E_\xi\uparrow$	M↓	S↑	$F_\beta\uparrow$	$E_\xi\uparrow$	M↓
JL-DCF [69]	TPAMI21	.926	.917	.964	.023	.911	.913	.948	.040	.911	.907	.949	.039	.892	.900	.949	.046
SwinNet [70]	TCSVT22	.941	.936	.974	.018	.935	.938	.963	.027	.919	.918	.956	.033	.911	.927	.950	.035
PCIR-Net [71]	ACM MM23	.935	.931	.970	.019	.927	.931	.958	.029	.920	.920	.957	.031	.899	.915	.939	.040
HRTransNet [20]	TCSVT23	.942	.936	.974	.016	.933	.939	.963	.026	.921	.919	.956	.030	.909	.929	.949	.035
HiDANet [72]	TIP23	.930	.929	.961	.021	.926	.939	.954	.029	.911	.921	.946	.035	.892	.919	.927	.043
CATNet [73]	TMM23	.940	.934	.972	.018	.932	.937	.960	.026	.921	.922	.958	.030	.911	.928	.952	.034
TSVT [74]	PR24	.935	.924	.966	.021	.917	.917	.939	.037	.912	.915	.946	.035	.905	.916	.944	.037
HFMDNet [75]	TIM24	.938	.933	.971	.017	.937	.944	.966	**.023**	.918	.920	.957	.031	.886	.905	.930	.044
MAGNet [57]	KBS24	.938	.931	.969	.017	.928	.935	.962	.027	.922	.920	.956	.029	.907	.924	.947	.036
EM-Trans [76]	TNNLS24	.940	.934	.970	.017	.931	.935	.961	.027	.925	.926	.958	.028	.903	.920	.944	.039
DFormer [77]	ICLR24	.942	.936	.973	.016	.937	.943	**.967**	**.023**	.923	.920	.956	.030	.915	.930	.953	.032
TPCL [78]	TMM24	.935	.930	.970	.017	.925	.930	.959	.028	.916	.917	.956	.031	.900	.917	.943	.038
DPPNet [79]	TMM25	**.944**	.939	**.978**	.016	.934	.939	.965	.026	.922	.921	.956	.032	.907	.924	.946	.038
Ours (*SAMSOD*-large)	-	.941	**.942**	.974	**.015**	**.938**	**.945**	.966	**.023**	**.930**	**.931**	**.962**	**.025**	**.920**	**.937**	**.955**	**.029**

In the third generalizability experiment, RGB-D rail surface defect detection on NEU RSDDS-AUG dataset [80] with 1,500 training pairs and 362 testing ones is adopted. From Table VI, our method achieves excellent performance, matching that of the best method, DSSNet [81]. SAM and its optimization for two proposed issues are the main reasons for the performance improvement.

TABLE VI
Generalizability experiment on NEU RSDDS-AUG dataset [80] for RGB-D rail surface defect detection.

Methods	Source	$S_m\uparrow$	$maxE_m\uparrow$	$maxF_m\uparrow$	$meanE_m\uparrow$	$meanF_m\uparrow$	$MAE\downarrow$
CLANet [80]	TMech22	.834	.921	.877	.912	.863	.069
DRERNet [82]	SPL22	.844	.933	.891	.929	.878	.059
FHENet [83]	TIM23	.836	.926	.881	.921	.869	.064
CSANet [84]	SPL23	.861	.941	.904	.928	.885	.058
MENet [85]	TCSVT23	.856	.939	.899	.931	.884	.057
SA2F [86]	TII23	.845	.927	.884	.923	.870	.061
PENet [87]	TIM23	.859	.940	.905	-	-	.054
SAINet [88]	OLEN24	.849	.936	.888	.934	.886	.055
DSSNet [81]	TITS24	.866	.946	.911	**.942**	**.897**	.049
Ours (*SAMSOD*-large)	-	**.875**	**.947**	**.915**	**.942**	.896	**.048**

V. Conclusion

In this paper, we propose a model called *SAMSOD* to achieve RGB-T salient object detection task. *SAMSOD* is a SAM-based method addressing two optimization issues. To play an equal role of RGB and thermal modalities and avoid imbalance convergence and modality conflict, unimodal supervision and gradient deconfliction are proposed. Furthermore, the vanilla adapters in the SAM are replaced with our proposed decoupled adapters. The gradient interference from high-activation regions on the low-activation pathway is reduced, allowing the foreground and background to be updated simultaneously and preventing the learning of the background from being weakened. The final results on RGB-T datasets prove our method's effectiveness. A lightweight version is also provided to meet real-time demand.

References

[1] W. Wang, Q. Lai, H. Fu, J. Shen, H. Ling, and R. Yang, "Salient object detection in the deep learning era: An in-depth survey," *TPAMI*, vol. 44, no. 6, pp. 3239–3259, 2021.

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

[3] 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 *CVPR*, 2023, pp. 4015–4026.

[4] N. Ravi, V. Gabeur, Y.-T. Hu, R. Hu, C. Ryali, T. Ma, H. Khedr, R. Rädle, C. Rolland, L. Gustafson, E. Mintun, J. Pan, K. V. Alwala, N. Carion, C.-Y. Wu, R. Girshick, P. Dollar, and C. Feichtenhofer, "SAM 2: Segment anything in images and videos," in *ICLR*, 2025, pp. 1–44.

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

[6] P. Zhang, T. Yan, Y. Liu, and H. Lu, "Fantastic animals and where to find them: Segment any marine animal with dual SAM," in *CVPR*, 2024, pp. 2578–2587.

[7] S. Lian, Z. Zhang, H. Li, W. Li, L. T. Yang, S. Kwong, and R. Cong, "Diving into underwater: Segment anything model guided underwater salient instance segmentation and a large-scale dataset," in *ICML*, 2024, pp. 29 545–29 559.

[8] K. Wang, D. Lin, C. Li, Z. Tu, and B. Luo, "Adapting segment anything model to multi-modal salient object detection with semantic feature fusion guidance," *arXiv preprint arXiv:2408.15063*, 2024.

[9] Y. Liu, P. Wu, M. Wang, and J. Liu, "CPAL: Cross-prompting adapter with LoRAs for RGB+X semantic segmentation," *TCSVT*, pp. 5858–5871, 2025.

[10] S. Chen, C. Ge, Z. Tong, J. Wang, Y. Song, J. Wang, and P. Luo, "AdaptFormer: Adapting vision transformers for scalable visual recognition," *NeurIPS*, vol. 35, pp. 16 664–16 678, 2022.

[11] T. Chen, L. Zhu, C. Deng, R. Cao, Y. Wang, S. Zhang, Z. Li, L. Sun, Y. Zang, and P. Mao, "SAM-Adapter: Adapting segment anything in underperformed scenes," in *ICCV*, 2023, pp. 3367–3375.

[12] T. Chen, A. Lu, L. Zhu, C. Ding, C. Yu, D. Ji, Z. Li, L. Sun, P. Mao, and Y. Zang, "SAM2-Adapter: Evaluating & adapting segment anything 2 in downstream tasks: Camouflage, shadow, medical image segmentation, and more," *arXiv preprint arXiv:2408.04579*, 2024.

[13] X. Peng, Y. Wei, A. Deng, D. Wang, and D. Hu, "Balanced multimodal learning via on-the-fly gradient modulation," in *CVPR*, 2022, pp. 8238–8247.

[14] Y. Wei and D. Hu, "MMPareto: Boosting multimodal learning with innocent unimodal assistance," in *ICML*, 2024, pp. 52 559–52 572.

[15] T. Li, Z. Wen, Y. Li, and T. S. Lee, "Emergence of shape bias in convolutional neural networks through activation sparsity," *NeurIPS*, vol. 36, pp. 71 755–71 766, 2023.

[16] G. Chen, F. Shao, X. Chai, H. Chen, Q. Jiang, X. Meng, and Y.-S. Ho, "Modality-induced transfer-fusion network for RGB-D and RGB-T salient object detection," *TCSVT*, vol. 33, no. 4, pp. 1787–1801, 2022.

[17] G. Chen, F. Shao, X. Chai, H. Chen, and Q. Jiang, "CGMDRNet: Cross-guided modality difference reduction network for RGB-T salient object detection," *TCSVT*, vol. 32, no. 9, pp. 6308–6323, 2022.

[18] 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," *TCSVT*, vol. 33, no. 8, pp. 4149–4163, 2023.

[19] Y. Pang, X. Zhao, L. Zhang, and H. Lu, "CAVER: Cross-modal view-mixed transformer for bi-modal salient object detection," *TIP*, vol. 32, pp. 892–904, 2023.

[20] B. Tang, Z. Liu, Y. Tan, and Q. He, "HRTransNet: HRFormer-driven two-modality salient object detection," *TCSVT*, vol. 33, no. 2, pp. 728–742, 2022.

[21] 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," *KBS*, vol. 291, pp. 1–13, 2024.

[22] 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?" *TMM*, pp. 6971–6982, 2022.

[23] Z. Wu, J. Wang, Z. Zhou, Z. An, Q. Jiang, C. Demonceaux, G. Sun, and R. Timofte, "Object segmentation by mining cross-modal semantics," in *ACM MM*, 2023, pp. 3455–3464.

[24] K. Wang, Z. Tu, C. Li, C. Zhang, and B. Luo, "Learning adaptive fusion bank for multi-modal salient object detection," *TCSVT*, pp. 7344–7358, 2024.

[25] 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," *PR*, vol. 146, pp. 1–14, 2024.

[26] X. Yu, X. Cheng, Y. Liu, and Z. Zheng, "A dual-stream cross-domain integration network for RGB-T salient object detection," *TCE*, pp. 1–13, 2024.

[27] W. Zhou, F. Sun, Q. Jiang, R. Cong, and J.-N. Hwang, "WaveNet: Wavelet network with knowledge distillation for RGB-T salient object detection," *TIP*, vol. 32, pp. 3027–3039, 2023.

[28] H. Tang, Z. Li, D. Zhang, S. He, and J. Tang, "Divide-and-Conquer: Confluent triple-flow network for RGB-T salient object detection," *TPAMI*, pp. 1958–1974, 2025.

[29] 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," *TMM*, pp. 10 692–10 707, 2024.

[30] J. Jin, Q. Jiang, Q. Wu, B. Xu, and R. Cong, "Underwater salient object detection via dual-stage self-paced learning and depth emphasis," *TCSVT*, pp. 2147–2160, 2025.

[31] X. Huang, J. Wang, Y. Tang, Z. Zhang, H. Hu, J. Lu, L. Wang, and Z. Liu, "Segment and caption anything," in *CVPR*, 2024, pp. 13 405–13 417.

[32] Z. Yan, J. Li, X. Li, R. Zhou, W. Zhang, Y. Feng, W. Diao, K. Fu, and X. Sun, "RingMo-SAM: A foundation model for segment anything in multimodal remote-sensing images," *TGRS*, vol. 61, pp. 1–16, 2023.

[33] Y. Gao, W. Xia, D. Hu, W. Wang, and X. Gao, "DeSAM: Decoupled segment anything model for generalizable medical image segmentation," in *MICCAI*. Springer, 2024, pp. 509–519.

[34] Y. Shen, J. Li, X. Shao, B. Inigo Romillo, A. Jindal, D. Dreizin, and M. Unberath, "FastSAM3D: An efficient segment anything model for 3d volumetric medical images," in *MICCAI*. Springer, 2024, pp. 542–552.

[35] B. Yao, Y. Deng, Y. Liu, H. Chen, Y. Li, and Z. Yang, "SAM-Event-Adapter: Adapting segment anything model for Event-RGB semantic segmentation," in *ICRA*. IEEE, 2024, pp. 9093–9100.

[36] G. Zhao, J. Huang, X. Yan, Z. Wang, J. Tang, Y. Ou, X. Hu, and T. Peng, "Open-vocabulary RGB-Thermal semantic segmentation," in *ECCV*. Springer, 2024, pp. 304–320.

[37] Q. He, "Prompting multi-modal image segmentation with semantic grouping," in *AAAI*, vol. 38, no. 3, 2024, pp. 2094–2102.

[38] Z. Liu, S. Deng, X. Wang, L. Wang, X. Fang, and B. Tang, "SSFam: Scribble supervised salient object detection family," *TMM*, vol. 27, pp. 1988–2000, 2025.

[39] Y. Fan, W. Xu, H. Wang, J. Wang, and S. Guo, "PMR: Prototypical modal rebalance for multimodal learning," in *CVPR*, 2023, pp. 20 029–20 038.

[40] H. Li, X. Li, P. Hu, Y. Lei, C. Li, and Y. Zhou, "Boosting multi-modal model performance with adaptive gradient modulation," in *ICCV*, 2023, pp. 22 214–22 224.

[41] X. Lin, S. Wang, R. Cai, Y. Liu, Y. Fu, W. Tang, Z. Yu, and A. Kot, "Suppress and rebalance: Towards generalized multi-modal face anti-spoofing," in *CVPR*, 2024, pp. 211–221.

[42] Z. Guo, T. Jin, J. Chen, and Z. Zhao, "Classifier-guided gradient modulation for enhanced multimodal learning," *NeurIPS*, vol. 37, pp. 133 328–133 344, 2025.

[43] X. Zhang, J. Yoon, M. Bansal, and H. Yao, "Multimodal representation learning by alternating unimodal adaptation," in *CVPR*, 2024, pp. 27 456–27 466.

[44] S. Wei, C. Luo, X. Ma, and Y. Luo, "Gradient decoupled learning with unimodal regularization for multimodal remote sensing classification," *TGRS*, pp. 1–12, 2024.

[45] Y. Wei, S. Li, R. Feng, and D. Hu, "Diagnosing and re-learning for balanced multimodal learning," in *ECCV*. Springer, 2024, pp. 71–86.

[46] Z. Huang, S. Lin, G. Liu, M. Luo, C. Ye, H. Xu, X. Chang, and X. Liang, "Fuller: Unified multi-modality multi-task 3D perception via multi-level gradient calibration," in *ICCV*, 2023, pp. 3502–3511.

[47] Y. Wei, R. Feng, Z. Wang, and D. Hu, "Enhancing multimodal cooperation via sample-level modality valuation," in *CVPR*, 2024, pp. 27 338–27 347.

[48] J. Wei, S. Wang, and Q. Huang, "F^3Net: Fusion, Feedback and Focus for Salient Object Detection," in *AAAI*, vol. 34, 2020, pp. 12 321–12 328.

[49] T. Yu, S. Kumar, A. Gupta, S. Levine, K. Hausman, and C. Finn, "Gradient surgery for multi-task learning," *NeurIPS*, vol. 33, pp. 5824–5836, 2020.

[50] 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 *IGTA*. Springer, 2018, pp. 359–369.

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

[52] D.-P. Fan, M.-M. Cheng, Y. Liu, T. Li, and A. Borji, "Structure-measure: A new way to evaluate foreground maps," in *ICCV*, 2017, pp. 4548–4557.

[53] R. Achanta, S. Hemami, F. Estrada, and S. Susstrunk, "Frequency-tuned salient region detection," in *CVPR*. IEEE, 2009, pp. 1597–1604.

[54] D.-P. Fan, C. Gong, Y. Cao, B. Ren, M.-M. Cheng, and A. Borji, "Enhanced-alignment measure for binary foreground map evaluation," in *IJCAI*, 2018, pp. 698–704.

[55] F. Perazzi, P. Krähenbühl, Y. Pritch, and A. Hornung, "Saliency filters: Contrast based filtering for salient region detection," in *CVPR*. IEEE, 2012, pp. 733–740.

[56] 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," *TIP*, pp. 2593–2607, 2023.

[57] M. Zhong, J. Sun, P. Ren, F. Wang, and F. Sun, "MAGNet: Multi-scale awareness and global fusion network for RGB-D salient object detection," *KBS*, vol. 299, pp. 1–13, 2024.

[58] 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," *TCSVT*, pp. 2513–2525, 2024.

[59] 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," *TCE*, pp. 4741–4755, 2024.

[60] R. Guo, X. Ying, Y. Qi, and L. Qu, "UniTR: A unified TRansformer-based framework for co-object and multi-modal saliency detection," *TMM*, pp. 7622–7635, 2024.

[61] W. Zhao, J. Zhang, L. Li, N. Barnes, N. Liu, and J. Han, "Weakly supervised video salient object detection," in *CVPR*, 2021, pp. 16 826–16 835.

[62] Y. Xu, X. Yu, J. Zhang, L. Zhu, and D. Wang, "Weakly supervised RGB-D salient object detection with prediction consistency training and active scribble boosting," *TIP*, vol. 31, pp. 2148–2161, 2022.

[63] Z. Liu, X. Huang, G. Zhang, X. Fang, L. Wang, and B. Tang, "Scribble-supervised RGB-T salient object detection," in *ICME*, 2023, pp. 2369–2374.

[64] Y. Wang, L. Zhang, P. Zhang, Y. Zhuge, J. Wu, H. Yu, and H. Lu, "Learning local-global representation for scribble-based RGB-D salient object detection via Transformer," *TCSVT*, pp. 11 592–11 604, 2024.

[65] H. Peng, B. Li, W. Xiong, W. Hu, and R. Ji, "RGBD salient object detection: A benchmark and algorithms," in *ECCV*. Springer, 2014, pp. 92–109.

[66] R. Ju, L. Ge, W. Geng, T. Ren, and G. Wu, "Depth saliency based on anisotropic center-surround difference," in *ICIP*. IEEE, 2014, pp. 1115–1119.

[67] Y. Niu, Y. Geng, X. Li, and F. Liu, "Leveraging stereopsis for saliency analysis," in *CVPR*. IEEE, 2012, pp. 454–461.

[68] D.-P. Fan, Z. Lin, Z. Zhang, M. Zhu, and M.-M. Cheng, "Rethinking RGB-D salient object detection: Models, data sets, and large-scale benchmarks," *TNNLS*, vol. 32, no. 5, pp. 2075–2089, 2020.

[69] K. Fu, D.-P. Fan, G.-P. Ji, Q. Zhao, J. Shen, and C. Zhu, "Siamese network for RGB-D salient object detection and beyond," *TPAMI*, vol. 44, no. 9, pp. 5541–5559, 2021.

[70] Z. Liu, Y. Tan, Q. He, and Y. Xiao, "SwinNet: Swin transformer drives edge-aware RGB-D and RGB-T salient object detection," *TCSVT*, vol. 32, no. 7, pp. 4486–4497, 2022.

[71] R. Cong, H. Liu, C. Zhang, W. Zhang, F. Zheng, R. Song, and S. Kwong, "Point-aware interaction and CNN-induced refinement network for RGB-D salient object detection," in *ACM MM*, 2023, pp. 406–416.

[72] Z. Wu, G. Allibert, F. Meriaudeau, C. Ma, and C. Demonceaux, "HiDANet: RGB-D salient object detection via hierarchical depth awareness," *TIP*, vol. 32, pp. 2160–2173, 2023.

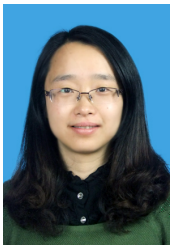
[73] F. Sun, P. Ren, B. Yin, F. Wang, and H. Li, "CATNet: A cascaded and aggregated Transformer network for RGB-D salient object detection," *TMM*, pp. 2249–2262, 2023.

[74] L. Gao, B. Liu, P. Fu, and M. Xu, "TSVT: Token Sparsification Vision Transformer for robust RGB-D salient object detection," *PR*, vol. 148, pp. 1–14, 2024.

[75] Y. Luo, F. Shao, Z. Xie, H. Wang, H. Chen, B. Mu, and Q. Jiang, "HFMDNet: Hierarchical fusion and multi-level decoder network for RGB-D salient object setection," *TIM*, pp. 1–15, 2024.

[76] G. Chen, Q. Wang, B. Dong, R. Ma, N. Liu, H. Fu, and Y. Xia, "EM-Trans: Edge-aware multimodal transformer for rgb-d salient object detection," *TNNLS*, vol. 36, no. 2, pp. 3175–3188, 2024.

[77] B. Yin, X. Zhang, Z.-Y. Li, L. Liu, M.-M. Cheng, and Q. Hou, "DFormer: Rethinking RGBD representation learning for semantic segmentation," in *ICLR*, 2024, pp. 1–23.
[78] J. Wu, F. Hao, W. Liang, and J. Xu, "Transformer fusion and pixel-level contrastive learning for RGB-D salient object detection," *TMM*, vol. 26, pp. 1011–1026, 2024.
[79] J. Yuan, Y. Wang, Z. Wang, Q. Xu, B. Veeravalli, and X. Yang, "DPPNet: A depth pixel-wise potential-aware network for RGB-D salient object detection," *TMM*, pp. 4256–4268, 2025.
[80] J. Wang, K. Song, D. Zhang, M. Niu, and Y. Yan, "Collaborative learning attention network based on RGB image and depth image for surface defect inspection of no-service rail," *TMECH*, vol. 27, no. 6, pp. 4874–4884, 2022.
[81] J. Wang, G. Li, G. Qiu, G. Ma, J. Xi, and N. Yu, "Depth-assisted semi-supervised RGB-D rail surface defect inspection," *TITS*, vol. 25, no. 7, pp. 8042–8052, 2024.
[82] J. Wu, W. Zhou, W. Qiu, and L. Yu, "Depth repeated-enhancement RGB network for rail surface defect inspection," *SPL*, vol. 29, pp. 2053–2057, 2022.
[83] W. Zhou and J. Hong, "FHENet: Lightweight feature hierarchical exploration network for real-time rail surface defect inspection in RGB-D images," *TIM*, vol. 72, pp. 1–8, 2023.
[84] J. Yang, W. Zhou, R. Wu, and M. Fang, "CSANet: Contour and semantic feature alignment fusion network for rail surface defect detection," *SPL*, vol. 30, pp. 972–976, 2023.
[85] W. Zhou, J. Hong, W. Yan, and Q. Jiang, "Modal evaluation network via knowledge distillation for no-service rail surface defect detection," *TCSVT*, vol. 34, no. 5, pp. 3930–3942, 2023.
[86] L. Huang and A. Gong, "Surface defect detection for no-service rails with skeleton-aware accurate and fast network," *TII*, vol. 20, no. 3, pp. 4571–4581, 2023.
[87] B. Wang, W. Zhou, W. Yan, Q. Jiang, and R. Cong, "PENet-KD: Progressive enhancement network via knowledge distillation for rail surface defect detection," *TIM*, vol. 72, pp. 1–11, 2023.
[88] Y. Yan, X. Jia, K. Song, W. Cui, Y. Zhao, C. Liu, and J. Guo, "Specificity autocorrelation integration network for surface defect detection of no-service rail," *OLEN*, vol. 172, pp. 1–10, 2024.



Zhengyi Liu is a professor in School of Computer Science and Technology, Anhui University, China. She received her Ph.D. degree from Anhui University, China in 2007. Her research interests include deep learning and computer vision.

Xinrui Wang is a M.S. Candidate of Anhui University. He received his B.S. from Hefei University, China in 2023. His research interests include deep learning and computer vision.

Xianyong Fang is a professor in School of Computer Science and Technology, Anhui University, China. He received his Ph.D. degree from Zhejiang University, China in 2005. His research interests include computer graphics, virtual reality, computer vision, pattern recognition, multimedia, and human-computer interaction.

Zhengzheng Tu is a professor in School of Computer Science and Technology, Anhui University, China. She received her Ph.D. degree from Anhui University, China in 2015. Her research interests include salient object detection and the other computer vision researches.

Linbo Wang is an associate professor in School of Computer Science and Technology, Anhui University, China. He received his Ph.D. degree from Nanjing University, China in 2014. His research interests include deep learning, computer vision, and image processing.