Global-Local Multi-scale Fusion Network for RGB-D Salient Object Detection

Abstract

In recent years, RGB-D salient object detection has attracted widespread attention due to its potential advantages in complex scenes. However, two critical challenges remain unresolved: (1) how to efficiently fuse RGB and depth features, and (2) how to fully exploit multi-level cross-modal information during the decoding stage. To address these issues, we propose a Global-Local Multi-scale Fusion Network (GMFNet). Specifically, a dual-stream Pyramid Vision Transformer (PVT) is adopted as the encoder to extract multi-level features from RGB and depth images. Then, the proposed Global-Local Interaction Module (GLM) employs cross-modal gating and global-local collaborative attention to achieve efficient cross-modal feature fusion, thereby enhancing the robustness and discriminability of feature representations. In the decoding stage, the Multi-scale Refinement Module (MRM) integrates high-level semantics with low-level features via multi-scale spatial attention and channel interaction, improving boundary and detail detection. Extensive experiments demonstrate that GMFNet significantly outperforms 16 state-of-the-art methods on five public RGB-D datasets and exhibits strong generalization on RGB-T SOD tasks.

***Keywords:** Salient Object Detection, RGB-D, Cross-modal, Multi-scale.*

1. Introduction

Salient Object Detection aims to detect the most visually attractive regions in an image and accurately segment them [3, 2]. Owing to the powerful learning and representation capabilities of Convolutional Neural Networks (CNNs), deep learning-based SOD methods have achieved superior performance compared to handcrafted feature-based approaches. SOD has been successfully applied to various visual tasks, including visual object tracking [44], semantic segmentation [47], and video analysis [17]. However, in certain challenging scenarios—such as low contrast between salient objects and the background, complex backgrounds, or the presence of multiple prominent objects—it is difficult to achieve optimal performance using only a single modality of data. In recent years, thanks to the development and popularization of depth sensors, combining RGB images with depth maps for SOD (RGB-D SOD) to further enhance saliency detection performance has gradually attracted significant research attention and rapidly become a hot topic. For the RGB-D salient object detection task, the primary challenge is to effectively integrate RGB and depth features to achieve complementary information fusion. Specifically, RGB images contain rich color and semantic information, while depth maps provide abundant structural and layout information. Some studies [11, 34] directly integrate depth and RGB images into a four-channel image and input it into the network. Some existing works [18, 32, 53] employ addition, multiplication, or concatenation operations to aggregate RGB and depth features. However, these methods fail to account for the independence of the two modalities and cannot fully leverage the complementary relationship between RGB appearance and depth geometric structure, thus failing to generate discriminative cross-modal features.

Another key challenge is how to fully exploit the interrelationships between features at the same level and across different levels after cross-modal fusion during the decoding phase, thereby improving the model's detection performance. It is well known that features at different levels possess distinct attributes: high-level features contain more semantic knowledge, while low-level features retain more detailed information. Therefore, bridging the gap between multi-modal features across different levels to obtain a final, powerful salient feature representation is crucial. Cong et al. [12] proposed a gated fusion architecture that dynamically assigns different weights to depth and RGB images. However, in the process of merging features from different levels, this approach relies solely on simple addition and cannot effectively utilize the latent relationships among features at each level. Zhao et al. [58] proposed a Fold-ASPP module based on Atrous Spatial Pyramid Pooling (ASPP) to aggregate contextual information. Nevertheless, due to imprecise selection of dilation rates, the feature maps may not adequately represent their content. Furthermore, naively stacking features of different scales can introduce redundancy and noise.

To address the aforementioned issues, we propose the Global-Local Multi-scale Fusion Network (GMFNet), which consists of an encoder, a Global-Local Interaction Module (GLM), and a Multi-scale feature Refinement Module (MRM). Specifically, we first adopt a dual-stream, sym-

metric Pyramid Vision Transformer (PVT) [48] as the encoder to extract features from RGB and depth images separately. PVT combines the local modeling advantages of CNNs with the long-range dependency modeling capabilities of Transformers, effectively expanding the receptive field and enhancing the diversity of feature representation. Subsequently, the GLM module achieves bidirectional guidance between RGB and depth features through a cross-modal gating mechanism. It further incorporates global-local collaborative attention to enhance the perception of the overall salient regions while preserving edge and structural details, thereby significantly improving the discriminability and robustness of cross-modal features. Finally, in the decoding stage, the MRM module is employed to fuse high-level semantic features with low-level detail features. This module adaptively models salient regions under different receptive fields through multi-scale spatial attention and channel interaction mechanisms, and combines residual convolution to enhance the stability of information integration, enabling the network to simultaneously capture both the global contours and local details of salient objects. Extensive experimental results demonstrate that GMFNet exhibits superior performance and competitiveness compared to 16 mainstream RGB-D methods on five public datasets. Our main contributions are as follows:

- We propose a Global-Local Multi-scale Fusion Network (GMFNet) for the RGB-D SOD task. The network comprises an encoder, a Global-Local Interaction Module (GLM), and a Multi-scale feature Refinement Module (MRM). Extensive experimental results on benchmark datasets validate the effectiveness of the proposed method.

- We design a Global-Local Interaction Module (GLM) that achieves deep fusion of RGB and depth features through cross-modal gating and global-local collaborative attention, thereby enhancing the robustness and discriminability of feature representations.

- To address the feature redundancy caused by scale variations, we propose the Multi-scale Refinement Module (MRM), which leverages multi-scale spatial attention and channel interaction mechanisms to preserve fine-grained details while enhancing the complementary nature of multi-scale semantics.

2. Related Work

2.1. RGB-D Salient Object Detection

Unlike RGB images, depth images provide positional information of objects in three-dimensional space, which is crucial for enhancing scene understanding and helps achieve better performance in SOD. To utilize multi-modal features more effectively, some researchers have investigated multi-modal feature fusion. Qu et al. [37] studied salient feature fusion by optimizing feature integration to improve model performance. [56] introduced a feature processing method that allows explicit control and enhancement of depth features during cross-modal fusion. Moreover, reducing unreliable features in depth maps has been proven to be an effective strategy for improving multi-modal fusion efficiency [10]. To avoid the impact of unreliable information in depth images, Chen et al. [9] developed a triplet encoding network to address the unreliability caused by noise in depth images. Although efficient multi-modal feature fusion methods can yield good results, a geometric mismatch problem still exists between depth and RGB features. To address this, Yin et al. [52] constructed a backbone network capable of simultaneously processing RGB and depth images. Unlike general network frameworks, Cong et al. [12] introduced a three-stream network to capture the interactions among RGB, depth, and RGB-D modalities. [23] developed a collaborative learning framework capable of simultaneously achieving edge detection, depth estimation, and SOD.

2.2. Cross-modal Feature Fusion

In the RGB-D SOD task, RGB images contain abundant color and texture information, while depth images primarily focus on 3D layout and spatial positional information. Effectively combining the complementary information inherent in RGB and depth features to achieve cross-modal feature fusion has always been a key focus of RGB-D SOD research. Extensive work has been conducted on this issue. Zhao et al. [57] designed a consistency-difference aggregation structure to achieve cross-modal and cross-level fusion through multi-path integration. Chen et al. [8] performed pre-fusion using a 3D Convolutional Neural Network during the encoding stage and executed deep fusion in the decoder stage. Chen and Li [6] designed a progressive dual-stream network that explores cross-modal and cross-level complementarity using cross-modal residual functions and complementary perceptual supervision. Fu et al. [20] mined useful complementary features through a Siamese network. Pang et al. [32] generated dynamic filters with varying receptive fields via a densely connected structure to achieve depth-guided fusion. Chen et al. [10] introduced depth potential perception to simulate the latent potential of depth images and fused features in the later stages of the network to integrate cross-modal complementarity. Liang et al. [30] designed a multi-modal interactive attention unit to filter and enhance cross-modal features along the channel dimension. Feng et al. [19] designed a depth-interleaved backbone network to propagate information from different modalities between encoders, thereby achieving fusion of multi-modal features at different levels.In contrast to the aforementioned

methods, we design a relatively simple yet highly efficient interactive fusion module. This module achieves bidirectional guidance between RGB and depth features. By combining global difference and local structural attention, it effectively highlights salient regions while preserving fine detail boundaries, thus realizing efficient cross-modal fusion.

2.3. Transformers for RGB-D SOD

In recent years, Transformers have been successfully applied to numerous computer vision tasks. In the field of RGB-D SOD, some researchers have employed Transformers as the backbone for feature extraction. For instance, Liu et al. [31] utilized the Swin Transformer to extract dual-stream features, which were then fed into spatial alignment and channel recalibration modules for enhancement. Chen et al. [5] leveraged PVTv2 to extract features from both modalities and conducted thorough exploration into the fusion of multi-modal features and the interaction of cross-level features. Furthermore, some researchers utilize Transformers for feature enhancement, while others employ them for cross-modal interaction. Pang et al. [33] designed a novel top-down, Transformer-based information pathway to integrate RGB and depth features, thereby achieving cross-modal interaction. Wu et al. [50] proposed a cross-modal interactive parallel Transformer module that effectively captures long-range multi-modal interactions and generates more comprehensive fused features. Chen et al. [7] proposed a heterogeneous cross-modal attention mechanism based on self- and cross-attention, which achieves complementary global contextual information and enhances cross-modal local semantic information. In our work, we use the PVT encoder as the backbone network to extract features from the RGB and depth modalities, cooperating with our proposed modules to achieve excellent performance.

3. Method

3.1. Overview

Figure 1 illustrates the architecture of our proposed GMFNet. It consists of a dual-stream PVT encoder, a Global-Local Interaction Module (GLM), and a Multi-scale Feature Refinement Module (MRM). First, RGB and depth images are input into the dual-stream PVT to extract multi-level features. These features are then paired and fed into the GLM module to generate cross-modal RGB-D interactive features. Finally, the MRM module refines these features to progressively generate the final salient object prediction map.

3.2. Encoder

Traditional Convolutional Neural Networks (CNNs) are inherently limited by their fixed-size local receptive fields, making it difficult to model long-range dependencies and fully capture global contextual information within images. To overcome this limitation, this paper adopts the Pyramid Vision Transformer (PVT) as the backbone encoder, leveraging its global perception capability and advantages in multi-scale hierarchical feature representation. Specifically, for RGB images, which contain rich color and texture information, we employ the more powerful PVTv2-B2 as the backbone network. For depth maps, which primarily reflect spatial structure, we select the lightweight PVTv2-B1 to reduce computational overhead while ensuring adequate feature representation capability. At the input stage, to maintain channel consistency, the single-channel depth map is replicated into a three-channel feature map. After passing through the dual-stream encoder, we obtain four-stage RGB features f_r^i and depth features f_d^i(where i=1,2,3,4), providing multi-scale representations for subsequent cross-modal fusion and decoding.

3.3. Global-Local Interaction Module

RGB features contain rich texture information, enabling the capture of fine details and textures in natural scenes, while depth features focus on spatial positional information. Therefore, effectively fusing valuable and complementary information from different modalities is crucial. To this end, we design a GLM module, which aims to achieve deep fusion of RGB and depth features through a global-local collaborative interaction mechanism, thereby enhancing the robustness and discriminability of cross-modal representations. The structure is illustrated in Figure 2. The main procedure is as follows: for given features f_r^i (i=1,2,3,4) and f_d^i (i=1,2,3,4), we first employ 1×1 convolutions with Sigmoid activation to generate gating weights separately. To further enhance feature fusion, we adopt a mutual modulation strategy, where RGB features guide depth features and depth features guide RGB features. The enhanced RGB and depth features are concatenated and passed through a channel attention module to model inter-channel dependencies across modalities, amplifying discriminative channel responses, yielding the result denoted as F_c^i (i=1,2,3,4). This process is formulated as:

$$F_c^i = \mathrm{CA}\left(Cat\left(\left(\mathrm{CRC1}(f_r^i) \otimes f_d^i\right), \left(\mathrm{CRC1}(f_d^i) \otimes f_r^i\right)\right)\right) \tag{1}$$

where $\mathrm{CRC1}(\cdot)$ denotes a module consisting of two 1×1 convolutional layers with a ReLU activation in between and a Sigmoid function at the end, $\mathrm{CA}(\cdot)$ denotes the channel attention mechanism, $\otimes$ denotes element-wise multiplication, and $\mathrm{Cat}(\cdot)$ denotes concatenation along the channel dimension.

To further model global context and local structural information, the features F_c^i $(i = 1, 2, 3, 4)$ are processed through global and local spatial attention branches. The global branch captures large-scale contextual differences via average pooling and max pooling to obtain a global

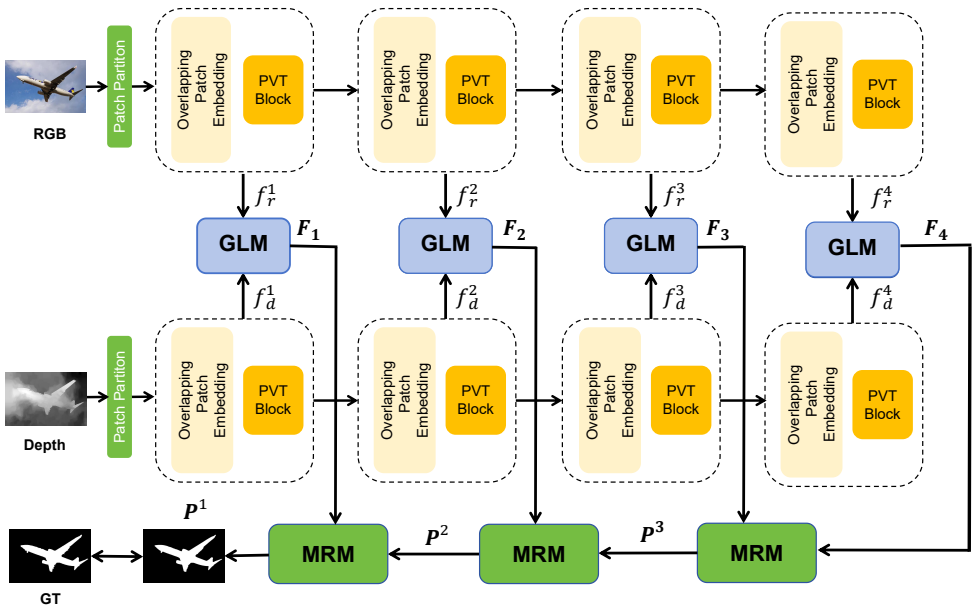

Figure 1. Overall architecture of GMFNet, comprising a dual-stream PVT encoder, GLM, and MRM. Specifically, RGB and depth images are first input into the encoder. Subsequently, the encoder extracts four hierarchical features for each modality: f_r^i and f_d^i (where $i = 1, 2, 3, 4$). These features, f_r^i and f_d^i, are then input into the GLM module to obtain the fused features F_i ($i = 1, 2, 3, 4$). Finally, under the guidance of the MRM, the final prediction map is inferred from the highest-level features.

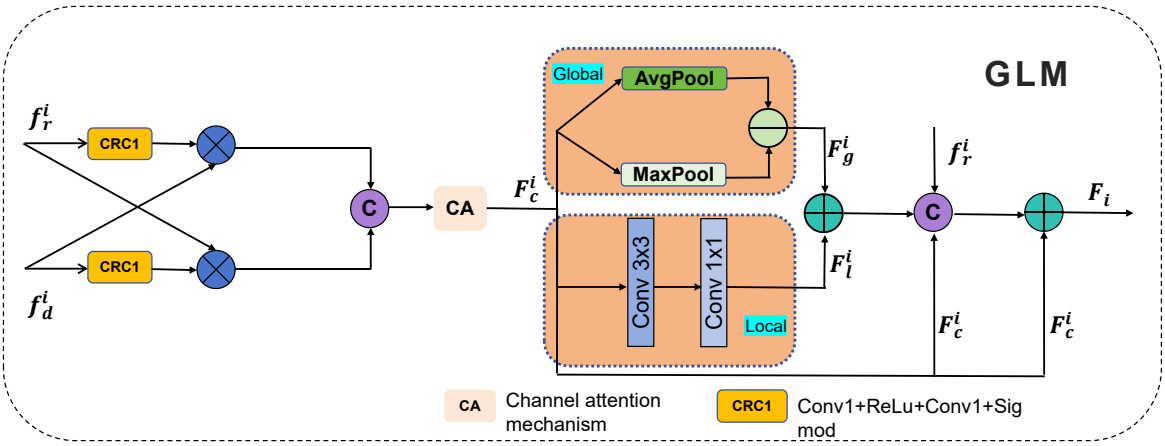

Figure 2. Global-Local Interaction Module (GLM)

guidance signal, while the local branch employs dilated convolutions to extract local structural patterns and preserve detailed boundary information. The global attention F_g^i ($i = 1, 2, 3, 4$) and local attention F_l^i ($i = 1, 2, 3, 4$) are then added together to form a balanced spatial attention map. Finally, the original RGB features f_r^i ($i = 1, 2, 3, 4$), the cross-modal interactive features F_c^i ($i = 1, 2, 3, 4$), and the spatial attention map are concatenated and fused through a convolutional layer. Residual connections are applied to maintain training stability and enhance the robustness of feature representations.This process is formulated as:

$$\begin{cases} F_g^i = \sqrt{\sum_{c=1}^{C} \left(Max(F_c^i) - Avg(F_c^i)\right)^2}, \\ F_l^i = \mathrm{Conv}_{1\times1}(\mathrm{Conv}_{3\mathrm{x}3}(F_c^i)), \\ F_i = Cat(F_c^i, f_r^i, \sigma(F_l^i \oplus F_g^i)) \oplus \mathrm{F}_c^i \end{cases} \tag{2}$$

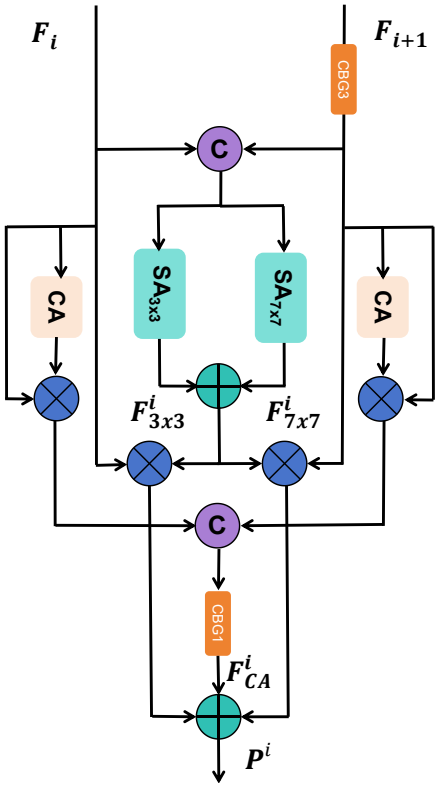

Figure 3. Multi-scale Feature Refinement Module (MRM)

where $\mathrm{Max}(\cdot)$ and $\mathrm{Avg}(\cdot)$denote max pooling and average pooling,$\sigma(\cdot)$ represents the Sigmoid activation operation.

3.4. Multi-scale Feature Refinement Module

To effectively fuse shallow detail features and deep semantic features from the encoder output, we propose a Multi-scale Feature Refinement Module (MRM) to enhance the fused representations. MRM employs multi-scale spatial attention to capture complementary spatial cues, followed by cross-channel interaction for feature enhancement, and integrates an optimization convolution to obtain superior feature representations. The detailed structure of MRM is shown in Figure 3. Specifically, the high-level features F_{i+1} from deeper layers are first processed through a residual convolution block consisting of a 3×3 convolution, batch normalization (BN), and GELU activation, which performs nonlinear transformation and feature enhancement. The resulting features are then combined with shallow features F_i and fed into two spatial attention modules with different receptive fields (using 3×3 and 7×7 convolution kernels, respectively) to capture local details and global contextual information. The two attention maps, $F^i_{3\times3}$ and $F^i_{7\times7}$, are added together and normalized by a Sigmoid activation to generate a unified spatial weight map $\alpha \in [0,1]^{H\times W}$. Two learnable scalar parameters, w_1 and w_2, are then introduced to weight and fuse the two branches, yielding F^i_{SA}:

$$\begin{cases} \mathbf{F}^i_{3\times3} = \mathrm{SA}_{3\times3}(Cat\left(\mathrm{CBG3}(F_{i+1}), F_i\right)), \\ \mathbf{F}^i_{7\times7} = \mathrm{SA}_{7\times7}(Cat\left(\mathrm{CBG3}(F_{i+1}), F_i\right)), \\ \boldsymbol{\alpha} = \sigma\left(\mathbf{F}^i_{3\times3} \oplus \mathbf{F}^i_{7\times7}\right), \\ \mathbf{F}^i_{SA} = \boldsymbol{\alpha} \otimes (w_1 \cdot F_i \oplus w_2 \cdot \mathrm{CBG3}(F_{i+1})), \end{cases} \tag{3}$$

where $\mathrm{CBG3}(\cdot)$ denotes a 3×3 convolution followed by batch normalization and a GELU activation function. Meanwhile, along the channel dimension, channel attention modules are applied to F_i and F_{i+1} separately to enhance the responses of key channels. The enhanced features are then concatenated and passed through a CBG1 structure implemented with a 1×1 convolution to enable cross-channel interaction, producing the channel-fused features $\mathbf{F}^i_{\mathrm{CA}}$, which further exploits cross-modal complementarity. Finally, the outputs of the spatial pathway $\mathbf{F}^i_{\mathrm{SA}}$ and the channel pathway $\mathbf{F}^i_{\mathrm{CA}}$ are added together for the final fusion and output:

$$\begin{cases} \mathbf{F}^i_{CA} = \mathrm{CBG1}(Cat\left(\mathrm{CA}(F_{i+1}) \otimes F_{i+1}, \mathrm{CA}(F_i) \otimes F_i\right)), \\ \mathbf{P}^i = \mathbf{F}^i_{CA} \oplus \mathbf{F}^i_{\mathrm{SA}}. \end{cases} \tag{4}$$

where $\mathrm{CBG1}(\cdot)$ denotes a 1×1 convolution followed by batch normalization and a GELU activation function.

3.5. Loss function

During the training phase, a hybrid loss composed of Binary Cross-Entropy (BCE) loss and Intersection over Union (IoU) loss is employed to effectively optimize the network. The overall loss function L of the model is defined as:

$$L = \sum_{i=1}^{i} \left[L^i_{BCE}\left(P^i, G\right) + L^i_{IoU}\left(P^i, G\right)\right], \tag{5}$$

where P denotes the predicted map and G represents the ground truth map.

4. Experiments

4.1. Datasets

To validate the effectiveness of the proposed model, we conducted experiments on five challenging RGB-D datasets, including DUT [36], NJU2K [26], NLPR [34], SIP [16], and LFSD [29]. DUT consists of 1,200 images captured from real-world scenes using a Lytro camera. NJU2K is a mixed dataset composed of samples collected from both the Internet and daily life. NLPR contains 1,000 images from various scenes. SIP is composed of 1,000 high-resolution images of salient objects, characterized by single or multiple salient objects in diverse scenarios. LFSD includes 100 images containing multiple small objects with

Table 1. Performance comparison on RGB-D SOD datasets (Full Methods).The best two results are shown in red and blue font.

Dataset		Existing Methods																Ours
		CIR TIP 2022	Swin TCSVT 2022	SPSN ECCV 2022	C2DF TMM 2023	EGA INS 2023	CAVER TIP 2023	MITF TCSVT 2023	CAT TMM 2023	DMG NN 2024	FCF TCSVT 2024	TPCL TMM 2024	STA Neurocom 2024	CPNet IJCV 2024	MAG KBS 2024	ADI Neurocom 2025	GAI Neurocom 2025	
DUT	$M\downarrow$	0.035	0.024	0.040	0.025	0.032	0.028	0.025	0.019	0.076	0.034	0.020	_	0.019	0.021	0.018	0.019	0.018
	$F_\beta\uparrow$	0.925	0.948	0.899	0.941	0.937	0.939	0.934	0.953	0.853	0.923	0.951	_	0.954	0.953	0.954	0.958	0.958
	$E_\xi\uparrow$	0.953	0.967	0.937	0.962	0.951	0.962	0.960	0.973	0.888	0.953	0.971	_	0.971	0.969	0.974	0.970	0.974
	$S_m\uparrow$	0.920	0.940	0.900	0.933	0.920	0.931	0.937	0.947	0.838	0.918	0.944	_	0.948	0.944	0.946	0.958	0.949
NJU2K	$M\downarrow$	0.040	0.028	0.031	0.038	0.035	0.032	0.030	0.027	0.035	0.034	0.028	0.034	0.025	0.027	0.027	0.027	0.026
	$F_\beta\uparrow$	0.917	0.935	0.918	0.909	0.923	0.923	0.923	0.930	0.927	0.923	0.927	0.927	0.934	0.935	0.938	0.938	0.936
	$E_\xi\uparrow$	0.946	0.961	0.949	0.942	0.946	0.954	0.957	0.956	0.930	0.953	0.955	0.954	0.960	0.962	0.965	0.958	0.961
	$S_m\uparrow$	0.914	0.931	0.918	0.908	0.914	0.921	0.926	0.932	0.913	0.918	0.926	0.924	0.933	0.928	0.931	0.930	0.932
NLPR	$M\downarrow$	0.026	0.018	0.022	0.021	0.021	0.021	0.018	0.018	0.022	0.024	0.018	0.021	0.017	0.017	0.017	0.018	0.017
	$F_\beta\uparrow$	0.903	0.927	0.910	0.917	0.925	0.918	0.928	0.926	0.927	0.911	0.920	0.926	0.929	0.931	0.930	0.934	0.931
	$E_\xi\uparrow$	0.953	0.968	0.958	0.961	0.967	0.962	0.968	0.967	0.956	0.960	0.965	0.968	0.969	0.969	0.970	0.964	0.970
	$S_m\uparrow$	0.921	0.938	0.923	0.927	0.933	0.929	0.933	0.937	0.924	0.924	0.932	0.934	0.939	0.938	0.937	0.936	0.936
SIP	$M\downarrow$	0.066	0.040	0.042	0.052	0.049	0.042	0.040	0.038	0.049	_	0.040	0.040	0.039	0.036	0.033	0.032	0.032
	$F_\beta\uparrow$	0.863	0.912	0.899	0.860	0.891	0.904	0.913	0.917	0.904	_	0.907	0.919	0.915	0.924	0.924	0.933	0.931
	$E_\xi\uparrow$	0.903	0.939	0.934	0.917	0.924	0.932	0.940	0.941	0.924	_	0.937	0.943	0.938	0.947	0.951	0.941	0.953
	$S_m\uparrow$	0.862	0.900	0.892	0.871	0.883	0.894	0.890	0.904	0.878	_	0.889	0.902	0.898	0.907	0.909	0.914	0.915
LFSD	$M\downarrow$	0.066	0.065	0.086	0.065	0.069	0.056	0.063	0.054	0.067	0.061	0.052	0.069	0.053	0.053	0.053	0.053	0.049
	$F_\beta\uparrow$	0.881	0.871	0.810	0.863	0.865	0.886	0.876	0.878	0.873	0.876	0.883	0.869	0.881	0.891	0.888	0.897	0.895
	$E_\xi\uparrow$	0.907	0.900	0.864	0.883	0.899	0.918	0.911	0.922	0.907	0.913	0.924	0.905	0.917	0.923	0.925	0.918	0.927
	$S_m\uparrow$	0.875	0.873	0.816	0.863	0.861	0.883	0.874	0.886	0.859	0.875	0.886	0.867	0.882	0.888	0.892	0.888	0.892

Table 2. Ablation study of the components of GMFNet. (a) Symmetric baseline (B1+B1). (b) Our baseline (B1+B2). (c) Our baseline combined with GLM only. (d) Our baseline combined with MRM only. (e)Symmetric baseline combined with GLM and MRM. (f) Our GMFNet. The best results are highlighted in red.

No.	B1+B1	B1+B2	GLM	MRM	SIP				LFSD				DUT			
					$E_\xi\uparrow$	$S_m\uparrow$	$F_\beta\uparrow$	$\mathcal{M}\downarrow$	$E_\xi\uparrow$	$S_m\uparrow$	$F_\beta\uparrow$	$\mathcal{M}\downarrow$	$E_\xi\uparrow$	$S_m\uparrow$	$F_\beta\uparrow$	$\mathcal{M}\downarrow$
a	✓				0.928	0.872	0.871	0.055	0.907	0.867	0.864	0.067	0.962	0.919	0.919	0.031
b		✓			0.932	0.880	0.885	0.050	0.917	0.881	0.876	0.059	0.968	0.933	0.940	0.025
c		✓	✓		0.942	0.905	0.924	0.035	0.918	0.885	0.889	0.053	0.970	0.943	0.952	0.021
d		✓		✓	0.945	0.909	0.923	0.036	0.924	0.890	0.889	0.051	0.971	0.947	0.955	0.019
e	✓		✓	✓	0.947	0.905	0.921	0.036	0.914	0.877	0.878	0.055	0.965	0.936	0.945	0.024
f		✓	✓	✓	0.953	0.915	0.931	0.032	0.927	0.892	0.895	0.049	0.974	0.949	0.958	0.018

Table 3. Ablation study of the GLM module: (a) Removing the global branch; (b) Removing the local branch; (c) Removing the channel attention (CA) mechanism; (d) Our GMFNet. The best results are highlighted in red.

No.	Methods	SIP				LFSD				DUT			
		$E_\xi\uparrow$	$S_m\uparrow$	$F_\beta\uparrow$	$\mathcal{M}\downarrow$	$E_\xi\uparrow$	$S_m\uparrow$	$F_\beta\uparrow$	$\mathcal{M}\downarrow$	$E_\xi\uparrow$	$S_m\uparrow$	$F_\beta\uparrow$	$\mathcal{M}\downarrow$
a	w/o globa	0.945	0.907	0.924	0.036	0.927	0.890	0.890	0.049	0.970	0.945	0.953	0.020
b	w/o local	0.952	0.914	0.930	0.033	0.922	0.888	0.881	0.050	0.969	0.948	0.956	0.019
c	w/o CA	0.944	0.905	0.924	0.037	0.923	0.889	0.886	0.051	0.971	0.946	0.956	0.019
d	GMFNet	0.953	0.915	0.931	0.032	0.927	0.892	0.895	0.049	0.974	0.949	0.958	0.018

Table 4. (a) Removing the spatial attention (SA) mechanism ; (b) Using only the left part of MRM,removing the right part that contains channel attention and spatial attention; (c) Using only the right part of MRM,removing the left part that contains channel attention and spatial attention; (d) Our GMFNet. The best results are highlighted in red.

No.	Methods	SIP				LFSD				DUT			
		$E_\xi \uparrow$	$S_m \uparrow$	$F_\beta \uparrow$	$\mathcal{M} \downarrow$	$E_\xi \uparrow$	$S_m \uparrow$	$F_\beta \uparrow$	$\mathcal{M} \downarrow$	$E_\xi \uparrow$	$S_m \uparrow$	$F_\beta \uparrow$	$\mathcal{M} \downarrow$
a	w/o SA	0.948	0.910	0.924	0.034	0.922	0.889	0.886	0.051	0.974	0.948	0.956	0.019
b	+MRM(left)	0.947	0.910	0.925	0.035	0.924	0.889	0.885	0.052	0.972	0.945	0.952	0.020
c	+MRM(right)	0.947	0.907	0.925	0.035	0.921	0.885	0.887	0.052	0.971	0.944	0.954	0.021
d	GMFNet	0.953	0.915	0.931	0.032	0.927	0.892	0.895	0.049	0.974	0.949	0.958	0.018

complex backgrounds. Following the setup in Hu et al. [21], we selected 800 samples from DUT, 1,485 samples from NJU2K, and 700 samples from NLPR for training.

4.2. Evaluation Metrics

We adopt four widely used evaluation metrics to assess our model, namely E-measure (E_ξ) [15], S-measure (S_m) [14], F-measure (F_β) [1], and Mean Absolute Error (MAE) [35]. Specifically, E-measure (E_ξ) evaluates the alignment of both local pixel-level and global image-level information. S-measure (S_m) measures the structural similarity of saliency maps in terms of region-aware and object-aware spatial structures. F-measure (F_β) is the weighted harmonic mean of precision and recall, which reflects overall performance. MAE calculates the average absolute difference between the predicted saliency map and the ground truth map at the pixel level.

4.3. Implementation Details

We trained our model on an NVIDIA RTX 4090D GPU. During both training and testing, the input image resolution was resized to 384×384. To prevent overfitting during training, we applied several data augmentation strategies to the training images, including random horizontal flipping, random cropping, random rotation, and normalization. The network was optimized using the Adam optimizer. We trained for 300 epochs with a batch size of 20, an initial learning rate of 5e-5, and a learning rate decay factor of 0.1 applied every 150 epochs.

4.4. Comparison with State-of-the-Art Methods

To validate the superiority of GMFNet, we compared it against 16 competing methods, including CIRNet [31], SwinNet [31], SPSN [27], C2DF [54], EGANet [49], CAVER [33], MITF [5], CATNet [39], DMGNet [40], FCFNet [55], TPCL [50], STANet [38], CPNet [21], MAGNet [59], ADINet [43], and GAINet [24]. The results of all comparison models were either obtained from their published papers or reproduced by running the released code.

4.4.1 Quantitative Evaluation

We conducted quantitative evaluations on five benchmark datasets using four evaluation metrics. The results presented in Table 1 demonstrate that GMFNet outperforms the vast majority of competing methods in terms of overall performance. All four evaluation metrics proposed in this study achieve values close to the optimal, confirming the effectiveness of the proposed method.

4.4.2 Qualitative Evaluation

Figure 4 presents a visual comparison between the proposed method and existing approaches. Experimental results demonstrate that, under several complex scenarios, the saliency maps generated by GMFNet outperform those produced by other models, highlighting the competitiveness of the proposed method. For instance, GMFNet exhibits superior performance in handling object occlusion (Rows 1–2), multiple objects (Rows 3–5), detection of small targets (Rows 6–7), and low-quality depth images (Row 8). These results underscore the effectiveness and reliability of the proposed approach.

4.5. Ablation Study

To evaluate the effectiveness of each module in GMFNet, we conduct ablation studies on the DUT, LFSD, and SIP datasets. The detailed results of these ablation experiments are presented in the following sections.

4.5.1 Effectiveness of Modules

To verify the effectiveness of the proposed modules, we conducted corresponding ablation experiments. Starting from the baseline model, we progressively integrated different modules, including the GLM module and the MRM module. Six model variants were designed: (a) Symmetric baseline (B1+B1); (b) Our baseline (B1+B2); (c) Our baseline combined with GLM only; (d) Our baseline combined with MRM only; (e) Symmetric baseline combined with GLM and MRM; (f) Our baseline with both

Table 5. Comparison of modeling complexity between the proposed method and several state-of-the-art (SOTA) methods.

Method	SwinNet	SPSN	MITF	CAVER	CATNet	TPCL	CPNet	ADI	STA	Ours
FLOPs(G)	124.7	100.3	24.2	44.5	344.1	274.3	258.6	50.3	24.5	23.1
Params(M)	198.7	37.0	127.5	55.7	262.7	112.6	216.5	131.1	85.0	44.1

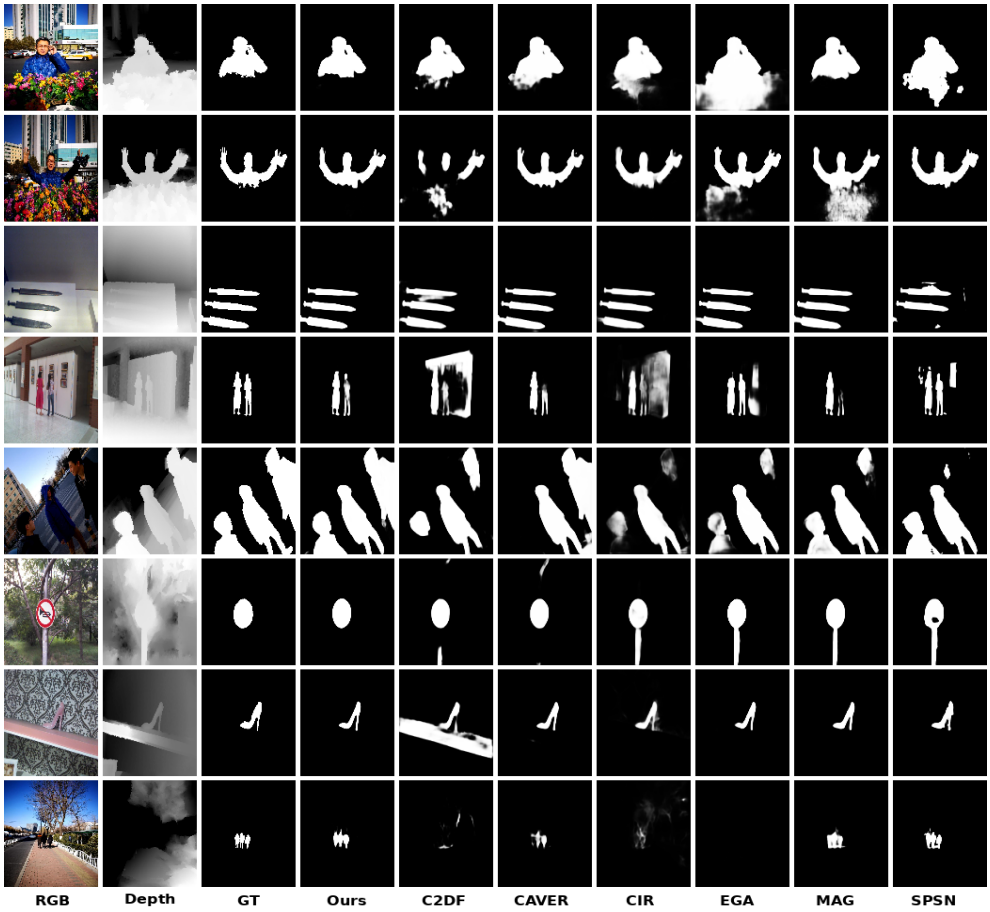

Figure 4. The results of a qualitative evaluation for the detection of RGB-D significant models in various challenging scenarios.

Table 6. Quantitative comparison of Mean absolute error (M), F-measure (F_β), E-measure (E_ξ) and S-measure (S_α) on three benchmark RGB-T datasets.

Datasets	**Metric**	CGMDR	APNet	ECFFNet	OSR	CGF	LSNet	Scribble	TNet	CMBDIF	CAVER	WaveNet	CAFCNet	**Ours**
VT5000	$M\downarrow$	0.032	0.037	0.037	0.042	0.035	0.039	0.036	0.035	0.032	0.035	0.029	0.027	0.025
	$F_\beta\uparrow$	0.877	0.832	0.850	0.833	0.853	0.834	0.842	0.867	0.876	0.857	0.878	0.889	0.901
	$E_\xi\uparrow$	0.939	0.919	0.922	0.917	0.922	0.923	0.932	0.936	0.937	0.935	0.948	0.947	0.953
	$S_\alpha\uparrow$	0.896	0.876	0.876	0.876	0.883	0.878	0.877	0.895	0.886	0.893	0.911	0.899	0.911
VT1000	$M\downarrow$	0.020	0.024	0.021	0.025	0.023	0.025	0.023	0.024	0.019	0.021	0.018	0.017	0.016
	$F_\beta\uparrow$	0.927	0.905	0.919	0.911	0.906	0.913	0.913	0.921	0.925	0.925	0.936	0.937	0.938
	$E_\xi\uparrow$	0.966	0.954	0.959	0.959	0.944	0.962	0.963	0.965	0.967	0.968	0.974	0.975	0.974
	$S_\alpha\uparrow$	0.931	0.922	0.924	0.926	0.923	0.926	0.925	0.929	0.927	0.936	0.942	0.935	0.938
VT821	$M\downarrow$	0.035	0.033	0.034	0.042	0.036	0.032	0.027	0.029	0.032	0.032	0.026	0.028	0.024
	$F_\beta\uparrow$	0.872	0.824	0.835	0.835	0.845	0.842	0.875	0.885	0.863	0.874	0.883	0.874	0.898
	$E_\xi\uparrow$	0.932	0.908	0.911	0.913	0.912	0.920	0.940	0.936	0.927	0.933	0.941	0.934	0.949
	$S_\alpha\uparrow$	0.894	0.868	0.877	0.875	0.881	0.879	0.895	0.899	0.882	0.891	0.908	0.891	0.911

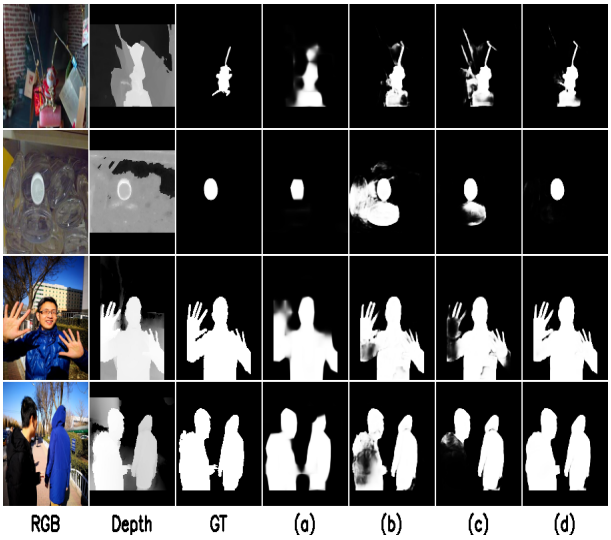

Figure 5. Visual comparison of ablation studies with different models.(a) Our baseline(B1+B2) . (b) Our baseline combined with GLM only. (c) Our baseline combined with MRM only. (d) Our GMFNet.

GLM and MRM modules, namely the proposed GMFNet. The experimental results demonstrate that both GLM and MRM significantly improve network performance, and their combination yields even greater performance gains. As shown in Table 2 , case (f) outperforms case (c), indicating that the GLM module emphasizes global-local complementarity but lacks further refinement of boundaries and details. In addition, case (f) achieves better accuracy than case (d), since using only the MRM module cannot fully exploit the complementarity between RGB and depth information. Meanwhile, the corresponding visualization results shown in Figure 5.

4.5.2 Effectiveness of GLM

To validate the effectiveness of the proposed GLM module, we conducted the following experiments: (a) removing the global attention branch from GLM; (b) removing the local attention branch from GLM; (c) removing the channel attention (CA) mechanism from GLM; and (d) Our GMFNet. As shown in Table 3, based on the results of case (d), we conclude that the GLM module effectively leverages the complementary properties of RGB and depth information, achieving feature fusion between the two modalities. The comparative results show that fully exploiting and interacting with RGB and depth information leads to superior detection performance. We further observe that effective extraction of global and local features, as well as their efficient fusion, has a significant impact on saliency detection performance.

4.5.3 Effectiveness of MRM

To validate the effectiveness of the MRM module, we conducted the following experiments: (a) removing the spatial attention (SA) mechanism from MRM; (b) using only the left part of MRM,removing the right part that contains channel attention and spatial attention; (c) using only the right part of MRM,removing the left part that contains channel attention and spatial attention; and (d) Our GMFNet. As shown in Table 4, the MAE values of cases (a), (b), and (c) are significantly higher than that of case (d) across the three datasets. The inferior performance of case (a) can be attributed to the crucial role of spatial attention in localizing salient regions, since its absence makes it difficult for the model to accurately separate foreground from background. The results of cases (b) and (c) indicate that relying solely on a single branch is insufficient to adequately model both the spatial and semantic information of cross-modal features. Case (d) achieves the best performance, further demonstrating the effectiveness of our MRM module and strategy.

4.5.4 Complexity Comparison

We systematically calculated the FLOPs and parameters of several representative classical models, with the results presented in Table 5. Compared with these models, our method achieves lower FLOPs and parameter counts while maintaining superior performance.

4.6. Discussion

4.6.1 Failure cases

Although GMFNet demonstrates effectiveness and efficiency, achieving optimal performance in certain challenging scenarios remains a formidable task, and failure cases still occur occasionally, as illustrated in Figure 6. In the first row, the low-quality depth map prevents the model from accurately distinguishing the salient object from the background. In the second and third rows, the salient objects are more than half occluded, leading to a significant degradation in prediction performance. In the fourth row, the salient object in the RGB image is interfered with by foreground elements, and the depth map is also highly misleading. Under these conditions, our proposed method partially fails to accurately segment the salient objects. Nevertheless, many state-of-the-art (SOTA) methods encounter similar difficulties.

4.6.2 Extension to RGB-T SOD

We also evaluated GMFNet on the RGB-T SOD task to verify its generalization capability. We retrained GMFNet on the RGB-T SOD dataset, following the same training protocol as Wang et al. [46]. The VT5000 dataset [41], which

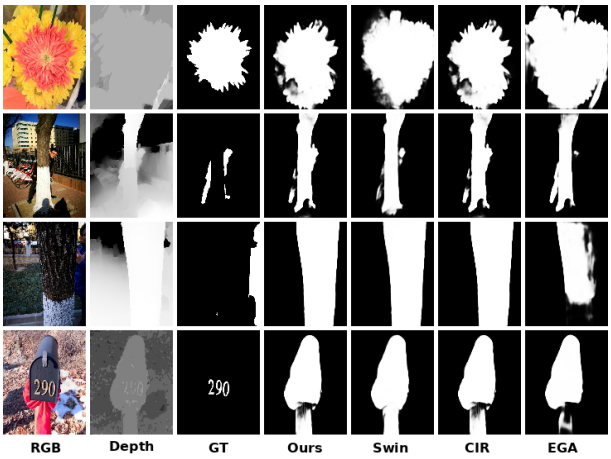

Figure 6. Failure cases of our model and comparison with state-of-the-art (SOTA) methods.

contains 5,000 image pairs, was split evenly—half for training and half for testing. The remaining datasets, VT821 [45] with 821 image pairs and VT1000 [42] with 1,000 image pairs, were used solely for testing. According to the quantitative results in Table 6, GMFNet achieves remarkable performance compared with recent state-of-the-art methods—including CGMDR [4], APNet [62], ECFFNet [60], OSR [22], CGF [46], LSNet [63], Scribble [28], TNet [13], CMDBIF [51], CAVER [33], WaveNet [61], and CAFC-Net [25]– demonstrating its strong generalization ability.

5. Conclusion

The proposed Global-Local Multi-scale Fusion Network (GMFNet) aims to address the insufficient cross-modal feature fusion and inadequate utilization of multi-level information in RGB-D salient object detection. Specifically, a dual-stream PVT encoder extracts multi-level features from both RGB and depth modalities; the Global-Local Interaction Module (GLM) achieves deep cross-modal fusion through cross-modal gating and global-local collaborative attention, thereby enhancing the discriminability and robustness of cross-modal representations; the Multi-scale Refinement Module (MRM) effectively integrates high-level semantic features with low-level detail features by leveraging multi-scale spatial attention and channel interaction mechanisms. Experimental results demonstrate that GMFNet outperforms 16 state-of-the-art methods across five public RGB-D datasets . Future work may further explore lightweight design, weakly supervised learning, and cross-modal temporal modeling to extend the approach to broader application scenarios such as video saliency detection.

References

[1] R. Achanta, S. Hemami, F. Estrada, and S. Susstrunk. Frequency-tuned salient region detection. In *2009 IEEE conference on computer vision and pattern recognition*, pages 1597–1604. IEEE, 2009. 7

[2] A. Borji, M.-M. Cheng, Q. Hou, H. Jiang, and J. Li. Salient object detection: A survey. *Computational visual media*, 5(2):117–150, 2019. 1

[3] A. Borji, M.-M. Cheng, H. Jiang, and J. Li. Salient object detection: A benchmark. *IEEE transactions on image processing*, 24(12):5706–5722, 2015. 1

[4] G. Chen, F. Shao, X. Chai, H. Chen, Q. Jiang, X. Meng, and Y.-S. Ho. Cgmdrnet: Cross-guided modality difference reduction network for rgb-t salient object detection. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(9):6308–6323, 2022. 10

[5] 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. *IEEE Transactions on Circuits and Systems for Video Technology*, 33(4):1787–1801, 2022. 3, 7

[6] H. Chen and Y. Li. Progressively complementarity-aware fusion network for rgb-d salient object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3051–3060, 2018. 2

[7] H. Chen, F. Shen, D. Ding, Y. Deng, and C. Li. Disentangled cross-modal transformer for rgb-d salient object detection and beyond. *IEEE Transactions on Image Processing*, 33:1699–1709, 2024. 3

[8] Q. Chen, Z. Liu, Y. Zhang, K. Fu, Q. Zhao, and H. Du. Rgb-d salient object detection via 3d convolutional neural networks. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pages 1063–1071, 2021. 2

[9] T. Chen, J. Xiao, X. Hu, G. Zhang, and S. Wang. Adaptive fusion network for rgb-d salient object detection. *Neurocomputing*, 522:152–164, 2023. 2

[10] Z. Chen, R. Cong, Q. Xu, and Q. Huang. Dpanet: Depth potentiality-aware gated attention network for rgb-d salient object detection. *IEEE Transactions on Image Processing*, 30:7012–7024, 2020. 2

[11] R. Cong, J. Lei, H. Fu, Q. Huang, X. Cao, and N. Ling. Hscs: Hierarchical sparsity based co-saliency detection for rgbd images. *IEEE Transactions on Multimedia*, 21(7):1660–1671, 2018. 1

[12] R. Cong, Q. Lin, C. Zhang, C. Li, X. Cao, Q. Huang, and Y. Zhao. Cir-net: Cross-modality interaction and refinement for rgb-d salient object detection. *IEEE Transactions on Image Processing*, 31:6800–6815, 2022. 1, 2

[13] 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 Transactions on Multimedia*, 25:6971–6982, 2022. 10

[14] D.-P. Fan, M.-M. Cheng, Y. Liu, T. Li, and A. Borji. Structure-measure: A new way to evaluate foreground maps. In *Proceedings of the IEEE international conference on computer vision*, pages 4548–4557, 2017. 7

[15] D.-P. Fan, C. Gong, Y. Cao, B. Ren, M.-M. Cheng, and A. Borji. Enhanced-alignment measure for binary foreground map evaluation. *arXiv preprint arXiv:1805.10421*, 2018. 7

[16] D.-P. Fan, Z. Lin, Z. Zhang, M. Zhu, and M.-M. Cheng. Rethinking rgb-d salient object detection: Models, data sets, and large-scale benchmarks. *IEEE Transactions on neural networks and learning systems*, 32(5):2075–2089, 2020. 5

[17] D.-P. Fan, W. Wang, M.-M. Cheng, and J. Shen. Shifting more attention to video salient object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8554–8564, 2019. 1

[18] D.-P. Fan, Y. Zhai, A. Borji, J. Yang, and L. Shao. Bbs-net: Rgb-d salient object detection with a bifurcated backbone strategy network. In *European conference on computer vision*, pages 275–292. Springer, 2020. 1

[19] G. Feng, J. Meng, L. Zhang, and H. Lu. Encoder deep interleaved network with multi-scale aggregation for rgb-d salient object detection. *Pattern Recognition*, 128:108666, 2022. 2

[20] 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. *IEEE transactions on pattern analysis and machine intelligence*, 44(9):5541–5559, 2021. 2

[21] X. Hu, F. Sun, J. Sun, F. Wang, and H. Li. Cross-modal fusion and progressive decoding network for rgb-d salient object detection. *International Journal of Computer Vision*, 132(8):3067–3085, 2024. 7

[22] F. Huo, X. Zhu, Q. Zhang, Z. Liu, and W. Yu. Real-time one-stream semantic-guided refinement network for rgb-thermal salient object detection. *IEEE Transactions on Instrumentation and Measurement*, 71:1–12, 2022. 10

[23] W. Ji, J. Li, M. Zhang, Y. Piao, and H. Lu. Accurate rgb-d salient object detection via collaborative learning. In *European conference on computer vision*, pages 52–69. Springer, 2020. 2

[24] Z. Jiang, L. Yu, Y. Han, J. Li, and F. Niu. Global-aware interaction network for rgb-d salient object detection. *Neurocomputing*, 621:129204, 2025. 7

[25] D. Jin, F. Shao, Z. Xie, B. Mu, H. Chen, and Q. Jiang. Cafc-net: Cross-modality asymmetric feature complement network for rgb-t salient object detection. *Expert Systems with Applications*, 247:123222, 2024. 10

[26] R. Ju, L. Ge, W. Geng, T. Ren, and G. Wu. Depth saliency based on anisotropic center-surround difference. In *2014 IEEE international conference on image processing (ICIP)*, pages 1115–1119. IEEE, 2014. 5

[27] M. Lee, C. Park, S. Cho, and S. Lee. Spsn: Superpixel prototype sampling network for rgb-d salient object detection. In *European conference on computer vision*, pages 630–647. Springer, 2022. 7

[28] L. Li, J. Han, N. Liu, S. Khan, H. Cholakkal, R. M. Anwer, and F. S. Khan. Robust perception and precise segmentation for scribble-supervised rgb-d saliency detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(1):479–496, 2023. 10

[29] N. Li, J. Ye, Y. Ji, H. Ling, and J. Yu. Saliency detection on light field. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2806–2813, 2014. 5

[30] Y. Liang, G. Qin, M. Sun, J. Qin, J. Yan, and Z. Zhang. Multi-modal interactive attention and dual progressive decoding network for rgb-d/t salient object detection. *Neurocomputing*, 490:132–145, 2022. 2

[31] Z. Liu, Y. Tan, Q. He, and Y. Xiao. Swinnet: Swin transformer drives edge-aware rgb-d and rgb-t salient object detection. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(7):4486–4497, 2021. 3, 7

[32] Y. Pang, L. Zhang, X. Zhao, and H. Lu. Hierarchical dynamic filtering network for rgb-d salient object detection. In *European conference on computer vision*, pages 235–252. Springer, 2020. 1, 2

[33] Y. Pang, X. Zhao, L. Zhang, and H. Lu. Caver: Cross-modal view-mixed transformer for bi-modal salient object detection. *IEEE Transactions on Image Processing*, 32:892–904, 2023. 3, 7, 10

[34] H. Peng, B. Li, W. Xiong, W. Hu, and R. Ji. Rgbd salient object detection: A benchmark and algorithms. In *European conference on computer vision*, pages 92–109. Springer, 2014. 1, 5

[35] F. Perazzi, P. Krähenbühl, Y. Pritch, and A. Hornung. Saliency filters: Contrast based filtering for salient region detection. In *2012 IEEE conference on computer vision and pattern recognition*, pages 733–740. IEEE, 2012. 7

[36] Y. Piao, W. Ji, J. Li, M. Zhang, and H. Lu. Depth-induced multi-scale recurrent attention network for saliency detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 7254–7263, 2019. 5

[37] L. Qu, S. He, J. Zhang, J. Tian, Y. Tang, and Q. Yang. Rgbd salient object detection via deep fusion. *IEEE transactions on image processing*, 26(5):2274–2285, 2017. 2

[38] P. Song, W. Li, P. Zhong, J. Zhang, P. Konuisz, F. Duan, and N. Barnes. Synergizing triple attention with depth quality for rgb-d salient object detection. *Neurocomputing*, 589:127672, 2024. 7

[39] 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 Transactions on Multimedia*, 26:2249–2262, 2023. 7

[40] Y. Tang and M. Li. Dmgnet: Depth mask guiding network for rgb-d salient object detection. *Neural Networks*, 180:106751, 2024. 7

[41] Z. Tu, Y. Ma, Z. Li, C. Li, J. Xu, and Y. Liu. Rgbt salient object detection: A large-scale dataset and benchmark. *IEEE Transactions on Multimedia*, 25:4163–4176, 2022. 9

[42] Z. Tu, T. Xia, C. Li, X. Wang, Y. Ma, and J. Tang. Rgb-t image saliency detection via collaborative graph learning. *IEEE Transactions on Multimedia*, 22(1):160–173, 2019. 10

[43] F. Wang, Y. Li, L. Wang, and P. Zheng. Asymmetric deep interaction network for rgb-d salient object detection. *Expert Systems with Applications*, 266:126083, 2025. 7

[44] F. Wang, S. Yin, J. T. Mbelwa, and F. Sun. Context and saliency aware correlation filter for visual tracking. *Multimedia Tools and Applications*, 81(19):27879–27893, 2022. 1

[45] 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 *Chinese Conference on Image and Graphics Technologies*, pages 359–369. Springer, 2018. 10

[46] J. Wang, K. Song, Y. Bao, L. Huang, and Y. Yan. Cgfnet: Cross-guided fusion network for rgb-t salient object detection. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(5):2949–2961, 2021. 9, 10

[47] W. Wang, G. Sun, and L. Van Gool. Looking beyond single images for weakly supervised semantic segmentation learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(3):1635–1649, 2022. 1

[48] W. Wang, E. Xie, X. Li, D.-P. Fan, K. Song, D. Liang, T. Lu, P. Luo, and L. Shao. Pvt v2: Improved baselines with pyramid vision transformer. *Computational visual media*, 8(3):415–424, 2022. 2

[49] L. Wei and G. Zong. Ega-net: Edge feature enhancement and global information attention network for rgb-d salient object detection. *Information Sciences*, 626:223–248, 2023. 7

[50] J. Wu, F. Hao, W. Liang, and J. Xu. Transformer fusion and pixel-level contrastive learning for rgb-d salient object detection. *IEEE Transactions on Multimedia*, 26:1011–1026, 2023. 3, 7

[51] 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 Transactions on Circuits and Systems for Video Technology*, 33(8):4149–4163, 2023. 10

[52] B. Yin, X. Zhang, Z. Li, L. Liu, M. Cheng, and Q. Hou. Dformer: Rethinking rgbd representation learning for semantic segmentation. arxiv 2023. *arXiv preprint arXiv:2309.09668*. 2

[53] J. Zhang, D.-P. Fan, Y. Dai, S. Anwar, F. Saleh, S. Aliakbarian, and N. Barnes. Uncertainty inspired rgb-d saliency detection. *IEEE transactions on pattern analysis and machine intelligence*, 44(9):5761–5779, 2021. 1

[54] M. Zhang, S. Yao, B. Hu, Y. Piao, and W. Ji. C{2}{2} dfnet: Criss-cross dynamic filter network for rgb-d salient object detection. *IEEE Transactions on Multimedia*, 25:5142–5154, 2022. 7

[55] Q. Zhang, Q. Qin, Y. Yang, Q. Jiao, and J. Han. Feature calibrating and fusing network for rgb-d salient object detection. *IEEE Transactions on Circuits and Systems for Video Technology*, 34(3):1493–1507, 2023. 7

[56] W. Zhang, G.-P. Ji, Z. Wang, K. Fu, and Q. Zhao. Depth quality-inspired feature manipulation for efficient rgb-d salient object detection. In *Proceedings of the 29th ACM international conference on multimedia*, pages 731–740, 2021. 2

[57] X. Zhao, Y. Pang, L. Zhang, H. Lu, and X. Ruan. Self-supervised pretraining for rgb-d salient object detection. In *Proceedings of the AAAI conference on artificial intelligence*, volume 36, pages 3463–3471, 2022. 2

[58] X. Zhao, L. Zhang, Y. Pang, H. Lu, and L. Zhang. A single stream network for robust and real-time rgb-d salient object detection. In *European conference on computer vision*, pages 646–662. Springer, 2020. 1

[59] 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. *Knowledge-Based Systems*, 299:112126, 2024. 7

[60] W. Zhou, Q. Guo, J. Lei, L. Yu, and J.-N. Hwang. Ecffnet: Effective and consistent feature fusion network for rgb-t salient object detection. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(3):1224–1235, 2021. 10

[61] 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 Transactions on Image Processing*, 32:3027–3039, 2023. 10

[62] W. Zhou, Y. Zhu, J. Lei, J. Wan, and L. Yu. Apnet: Adversarial learning assistance and perceived importance fusion network for all-day rgb-t salient object detection. *IEEE Transactions on Emerging Topics in Computational Intelligence*, 6(4):957–968, 2021. 10

[63] W. Zhou, Y. Zhu, J. Lei, R. Yang, and L. Yu. Lsnet: Lightweight spatial boosting network for detecting salient objects in rgb-thermal images. *IEEE Transactions on Image Processing*, 32:1329–1340, 2023. 10