Journal of King Saud University – Computer and Information Sciences 35 (2023) 101702

HOSTED BY

Contents lists available at ScienceDirect

Journal of King Saud University – Computer and Information Sciences

journal homepage: www.sciencedirect.com

Lightweight multi-level feature difference fusion network for RGB-D-T salient object detection

Kechen Song [a,b,c], Han Wang [a,b,c], Ying Zhao [a,b,c], Liming Huang [a,b,c], Hongwen Dong [d], Yunhui Yan [a,b,c,*]

[a] School of Mechanical Engineering & Automation, Northeastern University, Shenyang 110819, China
[b] National Frontiers Science Center for Industrial Intelligence and Systems Optimization, Northeastern University, Shenyang 110819, China
[c] Key Laboratory of Data Analytics and Optimization for Smart Industry (Northeastern University), Ministry of Education, China
[d] Shanghai Radio Equipment Research Institute, Shanghai 200090, China

ARTICLE INFO

Article history:
Received 12 May 2023
Revised 3 July 2023
Accepted 3 August 2023
Available online 9 August 2023

Keywords:
Salient object detection
RGB-depth-thermal images
Lightweight network
Cross-modal feature fusion

ABSTRACT

In recent years, bimodal salient object detection has developed rapidly. In view of the advanced performance of their robustness to extreme situations such as background similarity and illumination variation, researchers began to focus on RGB-Depth-Thermal salient object detection (RGB-D-T SOD). However, most existing bimodal methods usually need expensive computational costs to complete accurate prediction, and this situation is even more serious for three-modal methods, which undoubtedly limits their applicability. To solve this problem, we are the first to propose a lightweight multi-level feature difference fusion network (MFDF) for real-time RGB-D-T SOD. In view of the depth modality contains less useful information, we design an asymmetric three-stream encoder based on MobileNetV2. Due to the differences in semantics and details between high and low level features, using the same module without discrimination will lead to a large number of redundant parameters. On the contrary, in the coding stage, we introduce a cross-modal enhancement module (CME) and a cross-modal fusion module (CMF) to fuse low-level and high-level features respectively. In order to reduce redundant parameters, we design a low-level feature decoding module (LFD) and a multi-scale high-level feature fusion module (MHFF). A great deal of experiments proves that the proposed MFDF has more advantages than the 17 state-of-the-art methods. On the efficiency side, MFDF has a faster speed (124 FPS when the image size is 320 × 320) and much fewer parameters (8.9 M).

1. Introduction

Salient object detection (SOD) aims to make computers imitate human beings and automatically detect and segment the most prominent objects in input images. As one of the most popular tasks in computer vision, SOD has been widely used in various downstream vision directions, such as image-based quality assessment (Risnandar, , 2022), semantic segmentation (Mohakud and Dash, 2022; Zeng et al., 2022; Liu et al., 2023; Janneh et al., 2023; Chen and Zhao, 2023), object detection (Sharma and Mir, 2022; Kesav and Jibukumar, 2022; Arulprakash and Aruldoss, 2022; Chen et al., 2022), object tracking (Xia et al., 2022) and so on.

In recent years, SOD has made a long-term advance with the culmination of deep learning techniques. According to the number of input image data sources, the existing research fields of SOD based on deep learning can be divided into single-modality salient object detection (RGB SOD), RGB-Depth salient object detection (RGB-D SOD), RGB-Thermal infrared salient object detection (RGB-T SOD) and RGB-D-T SOD.

Mainstream SOD networks are designed to process RGB images generated by visible cameras and have made significant progress. However, when dealing with some difficult scenes, such as similar foreground and cluttered background, it is impossible to rely solely on RGB images. Recent work has achieved remarkable results by inputting RGB images together with a network of correspondingly aligned depth images. Because the grayscale value of each pixel in a depth image can be used to characterize the proximity of a point in the scene to the camera, it possesses the ability to distinguish prominent objects from the similar and complex background surrounding them. However, depth information is easily disturbed in practical applications, and it contains limited information in the face of weak light, uneven illumination and small objects. In contrast, thermal infrared images can simulate the spatial distribution of surface temperature of objects, and have low sensitivity to illumination changes. Therefore, even in the above unfavorable

* Corresponding author.
E-mail address: songkc@me.neu.edu.cn (Y. Yan).

https://doi.org/10.1016/j.jksuci.2023.101702
1319-1578/

environment, thermal infrared images can capture objects clearly and supplement rich contour information for RGB images. Therefore cross-modal complementary information from depth and thermal infrared images brings new vitality and progress to RGB SOD from different angles.

A hierarchical weighted suppress interference (HWSI) method (Song et al., 2022) is the pioneering work for RGB-D-T SOD, which exchanges the information of three modalities and makes a new VDT SOD benchmark dataset (VDT-2048).

Although the above methods have obtained excellent performance, two critical issues have not been well settled. First of all, they mainly focus on exploring the effective complementarity of cross-modal information, while ignoring the essential differences between high and low-level features. In the encoding and decoding stage, only the same module is used for the fusion interaction of high and low-level features, resulting in redundant parameters and difficulty in achieving further performance breakthroughs, as shown in Fig. 1(a). Second, as shown in Fig. 2, these frameworks obtain higher precision at the expense of memory size and running speed, which limits their application in real life.

To track the above issues, we develop a multi-level feature difference fusion network (MFDF) which can reduce the deployment period of RGB-D-T SOD. In view of the differences in semantic level and spatial resolution between high-low level features, we use different modules for high-level and low-level features in the encoding and decoding stages, as shown in Fig. 1(d).

The main contributions of this paper can be summarized as follows:

1) We propose a novel lightweight network named MFDF for real-time RGB-D-T SOD. Considering the amount of information contained in modalities due to different imaging mechanisms, we design an asymmetric backbone network based on MobileNetV2 to reduce parameter redundancy.

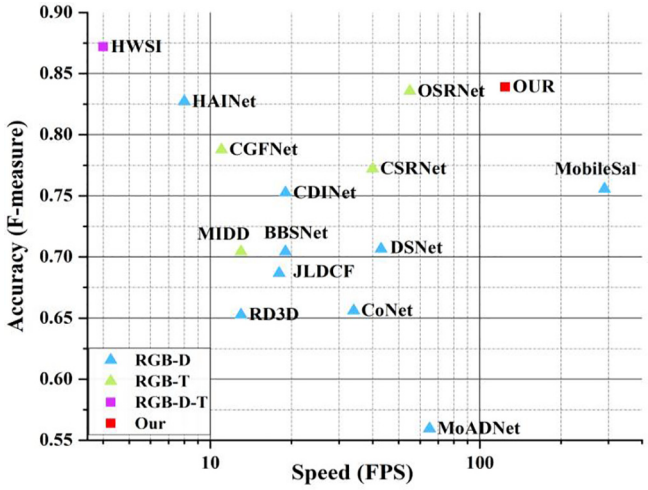

Fig. 2. Comparison with some advanced methods on the VDT-2048 dataset. Our method shows very competitive accuracy and a much faster speed.

2) We devise a simple and effective CMF to do a preliminary cross-modal fusion for high level features, and introduce the CME to reduce the diversity between modalities and improve the effect of cross-modal fusion.
3) In the aspect of high-level feature fusion, the MHFF is presented to improve the model inference speed by reducing the use of large dilated convolution kernels. In low-level feature decoding, we design the LFD to enlarge the receptive field of low-level features.
4) The MFDF is far superior to the 17 SOTA methods in the F-measure, and E-measure of reference benchmark VDT-2048. For efficiency analysis, our network has only 8.9 M

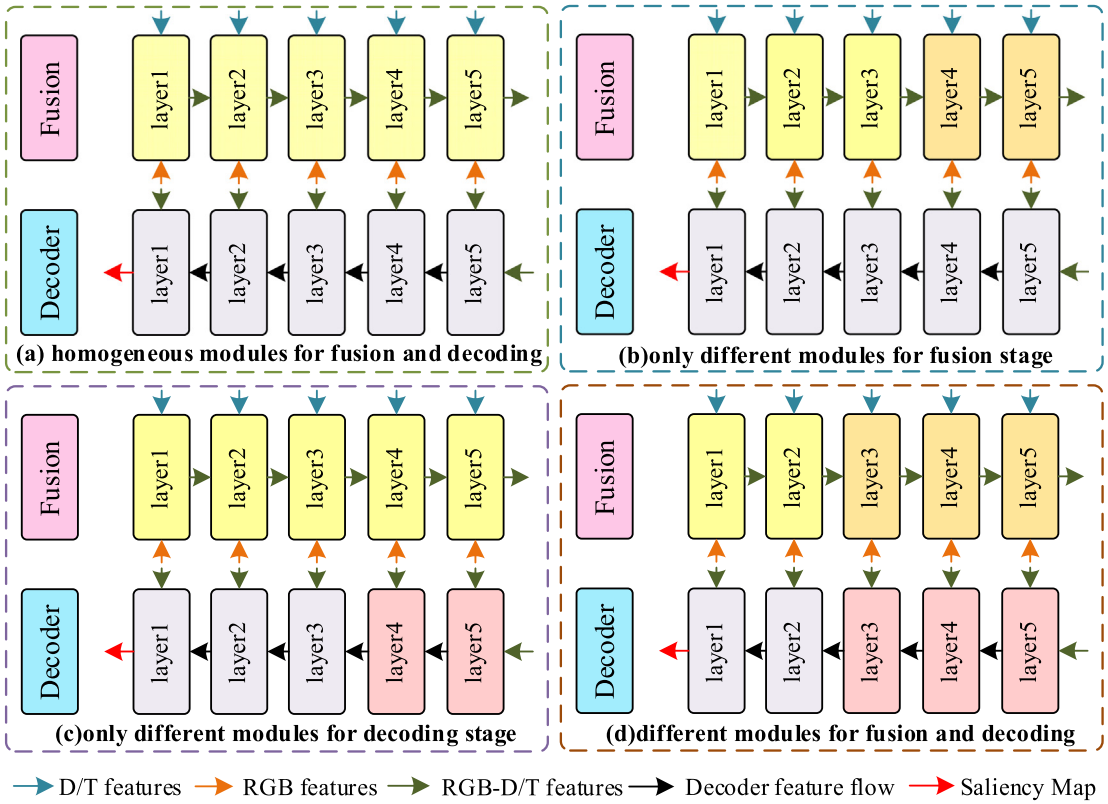

Fig. 1. Existing mainstream fusion and decoding architectures (a), (b) and (c), and our proposed architecture (d).

parameters, achieving 124 FPS processing speed with the input image size 320 × 320 on a single NVIDIA RTX2070 super.

2. Related work

2.1. RGB salient object detection

Due to the limitation of computer resources, the early traditional RGB SOD algorithm can only detect targets based on artificial features and various prior knowledge, like foreground consistency (Zhang et al., 2017), histograms (Lu et al., 2014), boundary prior (Zhu et al., 2014), center prior (Cheng et al., 2015) and color prior (Zhou et al., 2021). However, traditional methods cannot describe the object structure in complex scenes and lack generalization. There is no doubt that this weakness limits the development of SOD.

Recently, due to the strong feature representation ability of deep CNN, the RGB SOD methods have gone out of the restrictions of traditional methods and made a significant breakthrough. Zhang *et al.* (Zhang et al., 2017) proposed a generic aggregating multi-level variable feature framework (Amulet), which first integrates features at each resolution to predict salient map. Liu *et al.* (Liu et al., 2018) presented the pixel-wise contextual attention network (PiCANet) to improve the salient detection performance by integrating global context and multi-scale local context. Hou *et al.* (Hou et al., 2017) introduced a series of short connections into the HED architecture so that the output layer of the network can accurately capture the prominent objects and their boundaries in the picture. Liu *et al.* (Liu et al., 2020) designed a unified end-to-end training architecture to deal with the three tasks simultaneously. Tang *et al.* (Tang et al., 2021) used a high-resolution refinement network (HRRN) to regress and predict the pixels in the uncertain region. Wu *et al.* (Wu et al., 2022) adopted an extremely-downsampled network (EDN) to learn image features from a global perspective, so as to achieve accurate target location.

However, researchers are no longer satisfied with improving the accuracy blindly, and begin to consider the deployment of SOD applications in practice. Li *et al.* (Liu et al., 2021) designed the first lightweight SOD network called HVPNet, which carries out hierarchical perceptual learning on the premise of achieving a balance between accuracy and efficiency. Liu *et al.* (Liu et al., 2021) presented a very lightweight stereoscopically attentive multi-scale network (SAMNet), which adaptively fuses features of various scales using stereo attention mechanism. For purpose of detecting accurate targets quickly, a position prior attention network (PPA-Net) presented by Zhang *et al.* (Zhang et al., 2022) brings in a prior position to increase the confidence of objects in the central area.

2.2. RGB-D salient object detection

Although the SOD algorithm relying on CNN have made remarkable achievements at present, it is difficult to distinguish background from foreground and reflect the logical relationship between objects only by RGB images when faced with some realistic complex scenes, such as prominent objects and background with similar appearance, multiple objects, transparent objects, one object with different colors, etc. With the emergence of a large number of cheap and portable RGB-D camera recently, depth images containing more position and contour information are introduced into SOD to improve the detection results of objects. Subsequently, a new research direction of RGB-D SOD is formed.

RGB-D SOD methods depending on deep CNN can be roughly classified into two categories according to the importance of depth information. The first one is to fuse RGB and D modalities as equally important to explore modal differences.

Fu *et al.* (Fu et al., 2020) constructed a very flexible joint learning and dependent-cooperative fusion (JL-DCF) framework for cross-modal complementarity. Zhang *et al.* (Zhang et al., 2021) raised a bi-directional transfer-and-selection net (BTS-Net), which introduces bidirectional interaction between RGB and depth in the feature encoding phase. Aiming at the potential noise of the original depth map, Ji *et al.* (Ji et al., 2021) tried to calibrate the original depth directly through a depth calibration and fusion (DCF) framework, which significantly improves the performance. Chen *et al.* (Chen et al., 2021) established a depth potentiality-aware gated attention network (DPANet) to explicitly model the confidence response of depth maps and reduce pollution. A dynamic selective network (DSNet) presented by Wen *et al.* (Wen et al., 2021) seeks the possibility of consistently fusing cross-modal, cross-level and multi-scale cues through dynamic selection. Li *et al.* (Li et al., 2021) constructed a hierarchical alternate interactions network (HAINet), specifically, firstly using RGB features to filter unfavorable information in depth features, and then the process is repeated in reverse. Chen *et al.* (Chen et al., 2021) used the aggregation ability of 3D convolution (RD3D) for the first time to complete the pre-fusion in encoder stage and the deep fusion in decoder stage. X. Fang *et al.* (Fang et al., 2022) introduced transformer, and developed the group transformer network (GroupTransNet) to learn the long-range dependencies at the least cost and obtain perfect feature expression. A novel specification-preserving network (SPNet) proposed by Zhao *et al.* (Zhou et al., 2021) learns shared representation while preserving the unique features of each bimodal modality. Wang *et al.* (Wang et al., 2022) adopted remote cross-modal correlation and local depth correlation to refine features from different angles to ensure the fine structure of the target.

Because high-level features emphasize semantic information and low-level features highlight spatial information, Zhou *et al.* (Zhou et al., 2022) raised a crossflow and cross-scale adaptive fusion network (CCAFNet) which adopts different adaptive fusion strategies for different levels of bimodal features in the encoding stage to promote SOD. Zhang *et al.* (Zhang et al., 2021) used the differences between modalities to realize the mutual supplement of different levels of features in the encoding stage. The above process establishes a cross-modal differential interaction network (CDI-Net), as shown in Fig. 1(b).

The second is to let the auxiliary D modality information to seek and strengthen RGB main features, and accomplish cross-modal complementary fusion under the guidance of main modality. Ji et al. (Ji et al., 2020) proposed a novel collaborative learning framework (CoNet) that utilizes saliency knowledge and edge in a mutually beneficial way to improve detector performance. Chen *et al.* (Chen and Fu, 2020) used depth prediction to reduce mutual degradation between modalities, thus making up for missing parts and wrong predictions. Liu *et al.* (Liu et al., 2020) provided the selective self-mutual attention (S2MA) module to integrate self-attention and others' attention to accurately propagate context and filter out unreliable modal information. A depth distiller (A2dele) presented by Piao *et al.* (Piao et al., 2020) explores the use of network prediction and attention to adaptively transfer pixel-level depth knowledge to prediction of RGB streams. In order to suppress the interference in low-level cross-modal features, Zhao *et al.* (Zhai et al., 2021) presented a bifurcated backbone strategy network (BBS-Net) to mine deep and rich information by cascade refinement mechanism. According to the different features of high and low levels in the decoding stage, Liu *et al.* (Liu et al., 2021) designed a triplet transformer embedding network (TriTransNet), which firstly strengthens the features of the upper three levels, and then combines them with the features of the lower two levels to realize prediction.

Benefiting from the geometric structure clues in depth images, RGB-D SOD also shows amazing results in some extreme scenes. However, there is no gainsaying that using separate feature extraction networks for different modality brings huge parameters, which may lead to high computation and expensive storage costs. This also limits the actual deployment process of the current RGB-D model, especially in mobile devices. In order to solve this limitation, Chen *et al.* (Chen and Fu, 2020) constructed a lightweight depth backbone to reduce redundancy. By embedding A2dele, Piao et al. (Piao et al., 2020) implemented a lightweight architecture that does not use depth data during testing. The collaborative learning strategy allows CoNet (Ji et al., 2020) without no additional deep networks and deep input during testing, thus making it lighter and more versatile. Jin *et al.* (Jin et al., 2022) constructed mobile asymmetric dual-stream networks (MoADNet), which uses inverted bottleneck structure and atrous pyramid to speed up reasoning and compensate for information loss in lightweight backbone. Wu *et al.* (Wu et al., 2022) raised an implicit depth restoration (IDR) technology used only in training to enhancement the feature presentation of the mobile network (MobileSal) for an efficient RGB-D SOD. By revisiting the feature fusion period, Huang et al. (Huang et al., 2021) firstly presented a middle-level fusion way for real-time SOD. An efficient lightweight RGB-D SOD model (DFM-Net) designed by Zhang *et al.* (Zhang et al., 2021) can dynamically filter the depth features according to the depth quality.

2.3. RGB-T salient object detection

Although depth images contain abundant geometric structure and 3D information, they can be combined with visual cues to promote object detection and location. However, under complex conditions such as weak light, dark or uneven light, the information contained in the RGB and depth images may not be sufficient for accurate detection. The thermal infrared imager can capture the infrared radiation emitted by targets to create images, allowing it to operate effectively during nighttime or under adverse weather conditions. Therefore, RGB-T SOD algorithm for fusing visible and thermal infrared images is gradually emerging. For detailed methods, please refer to the review written by Song *et al.* (Song et al., 2023). Unlike RGB-D SOD, where depth modality may be in auxiliary position, RGB and thermal modality are in the same position. In extreme environments, thermal images are used to supplement details such as contours and boundaries, while RGB images are used to provide complementary color, texture and semantic information.

Tu *et al.* (Tu et al., 2020) proposed a novel attention-based deep fusion network (ADFNet) and constructed a large-scale, high-diversity benchmark VT5000. Zhou *et al.* (Zhou et al., 2022) introduced generation countermeasure learning into RGB-T SOD and improved generation of an adversarial learning assistance and perceived impact fusion network (APNet). Aiming at the insufficiency of cross-modal information fusion, Wang *et al.* (Wang et al., 2022) raised a novel cross-guided fusion network (CGFNet) to seek and fuse the information of different modalities. An effective and consistent feature fusion network (ECFFNet) presented by Zhou *et al.* (Zhou et al., 2022) fuses feature of corresponding levels and performs bilateral fusion of background information. Tu *et al.* (Tu et al., 2021) raised a multi-interactive dual-decoder (MIDD) to upsample the two modalities separately to prevent the information of the two modalities from influencing each other excessively in the interaction process. Liu *et al.* (Liu et al., 2022) developed the SwinNet driven by Swin Transformer, which relies on attention mechanism to bridge the modal gap, with edge information guiding the decoder to highlight the contour. A two-stage fusion network (TSFNet) presented by Guo *et al.* (Guo et al., 2021) fuses the intersection and union information of local regions and the background features. Gao *et al.* (Gao et al., 2022) developed a novel multi-stage and multi-scale fusion network (MMNet) to explore complementarity. Liao *et al.* (Liao et al., 2022) constructed a cross-collaborative fusion-encoder network (CCFENet) to reduce the negative response of contaminated data and encourage efficiency and accurate SOD. A novel deep correlation network (DCNet) proposed by Tu *et al.* (Tu et al., 2022) solves a new weak alignment-free RGBT SOD task at spatial, feature and semantic levels. Taking account into indoor complex illumination changes, Song *et al.* (xxxx) build a VI-RGBT1500 dataset and a multi-graph affinity interactive (MGAI) network. Chen *et al.* (Chen et al., 2022) established a cross-guided modality difference reduction network (CGMDRNet) to reduce modal differences and become more different fusion features. And according to the diverse properties of high-low layer features, they took different treatment measures in the decoding stage, as shown in Fig. 1(c).

However, the above RGB-T SOD frameworks need high computing cost and memory consumption while obtaining high precision, which is not suitable for running on resource-limited devices. To solve this problem, Huo *et al.* (Huo et al., 2022) created a real-time one-stream semantic-guided refinement network (OSRNet), which uses an easy and useful early fusion to avoid the cumbersome two-stream encoder structure. Huo *et al.* (Huo et al., 2022) used the lightweight backbone and designed highly effective decoders, named context-guided stacked refinement network (CSRNet).

2.4. RGB-D-T salient object detection

For purpose of better coping with the complex interference environment, Song *et al.* (Song et al., 2022) proposed a method named HWSI and provided a new VDT-2048 dataset by combining the advantages of RGB, Thermal, and Depth modalities for the first time. However, HWSI adopts pairwise fusion for the three modalities, and then synthesizes the final decoder features. The stacking and homogenization of multiple identical modules make the complexity and running speed of the model not dominant.

Most available methods of RGB-D, RGB-T, RGB-D-T mentioned above mainly focus on the design of fusion module, ignoring the problem that there are differences between high and low level features in terms of semantic level and spatial resolution. Therefore, they only use the same module to process the encoding and decoding stages of fusion interaction of features at different levels, resulting in very redundant parameters and limiting the performance improvement. There are only a few exceptions, such as CCAFNet (Zhou et al., 2022) and CDINet (Zhang et al., 2021), which distinguish the encoding process, while TriTransNet (Liu et al., 2021) and CGMDRNet (Chen et al., 2022) are different from the decoding process. But they are not put forward for the purpose of lightweight.

Unlike the above bimodal and trimodal algorithms, the proposed MFDF is the first one to adopt different processing means for the interaction of the three modalities in the encoding and decoding stages in response to the difference between high and low level features, as shown in Fig. 1(d). It solves the above limitations by ensuring the high performance of the benchmark dataset VDT-2048, fewer model parameters and competitive reasoning speed for RGB-D-T SOD. Our network has obtained accurate real-time salient target detection results in extensive experiments.

3. Methodology

In this part, we present the overall look of our net. Specifically, low-level features, they are firstly enhanced by the CME, and then we utilize the CMF to fuse them, while high-level features are directly aggregated by the CMF. Secondly, high-level semantic

information *RTD5* of trimodal images is obtained by high-level cross-modal fusion between *RT5* and *D5* for salient prediction. After that, the fused features are sent to DAS to restore the depth map, which is used to play an auxiliary role in supervision. Finally, the LFD and MHFF are developed for reducing unnecessary parameters.

3.1. Architecture overview

Fig. 3 paints the overall architecture of the proposed MFDF, where we utilize three branches, RGB, depth, and thermal infrared, to extract trimodal information respectively.

3.1.1. RGB branch and T branch

We use MobileNetV2 (Sandler et al., 2018) as the feature extraction backbone of our method, and remove the GAP layer and the final FC layer in order to apply the salient object detection task. Each layer of the RGB and T branches is a convolutional layer with a step size of 2, and the number of channels output from each layer is 16, 24, 32, 96, 320. For subsequent convenient representation, we denote the RGB branch's five layers of the RGB branch are represented as R1, R2 in the lower layer and R3, R4, and R5 in the higher layer. Moreover, the five layers of the T branch are represented as T1, T2 in the lower layer and T3, T4, and T5 in the higher layer.

3.1.2. Depth branch

It is basically the same as RGB and T branches. Since depth images have less texture and semantic information compared to RGB and T images, we use a simplified version of MobileNetV2 (Sandler et al., 2018) to extract depth features. This approach can reduce unnecessary computational complexity. For the depth branch is a convolutional layer with a step size of 2 per layer and the number of channels output per layer is 16, 24, 32, 96, 320. For easy representation, the five layers of features of the depth branch are denoted as D1, D2, D3, D4, and D5.

As shown in Fig. 3, we first fuse the five layers of features extracted from RGB and T branches respectively. The lower-level features have more texture information, so we first enhance them using the CME and then fuse them employing the CMF, while the higher-level features are fused directly using the CMF. For ease of representation, RGB and T features fused by the CMF are denoted as *RT1*, *RT2*, *RT3*, *RT4*, *RT5*. Secondly, the high-level cross-modal fusion is performed using *RT5* and *D5* to obtain the high-level semantic information *RTD5* of the trimodal images for salient prediction. Then, *RT1*, *RT2*, *RT3*, *RT4* and *RTD5* features are input to the DAS to reconstruct the depth map for aiding supervised RGB and T branches to enhance their feature learning capability. Finally, for the prediction part of the salient map, we raise the LFD and MHFF for the characteristics of high and low-level features to cut down the number of unnecessary parameters. The final output of the low-level feature decoding module is the salient map. The details are described in the next parts.

3.2. Cross-Modal enhancement and fusion modules

1) **Cross-modal enhancement module (CME).** There are five main layers in the feature extraction branch of RGB and T. As in previous studies (Zhou et al., 2022; Zhang et al., 2021), texture information is mainly concentrated in the first two layers of feature extraction, and semantic information is mainly concentrated in the last three layers. Therefore, in the first two layers of extracted features, we believe that the texture information of the images is affected by different modalities with large variability, while in the last three layers of extracted features, the semantic information accounts for the major part and the modal variability of the features is small. To solve the problem of large modal differences in texture information of shallow features, we develop a cross-modal enhancement module to decrease the modal differences between RGB and T images, and then boost the effect of cross-modal fusion.

As shown in Fig. 4, the structure of the CME is shown. Where *Ri* and *Ti* represent the low-level features of the input RGB and T, and we explicitly model the differences between the two modalities by element-wise subtraction RT_i^{dif}:

$$RT_i^{dif} = Ri - Ti, i \in [1,2] \tag{1}$$

The variability of the RGB and T modalities is later reduced by the following operations:

$$R\_i = CBR(CBR(Ri) \times CBR(RT_i^{dif}) + CBR(Ri)), i \in [1,2] \tag{2}$$

$$T\_i = CBR(CBR(Ti) \times CBR(-RT_i^{dif}) + CBR(Ti)), i \in [1,2] \tag{3}$$

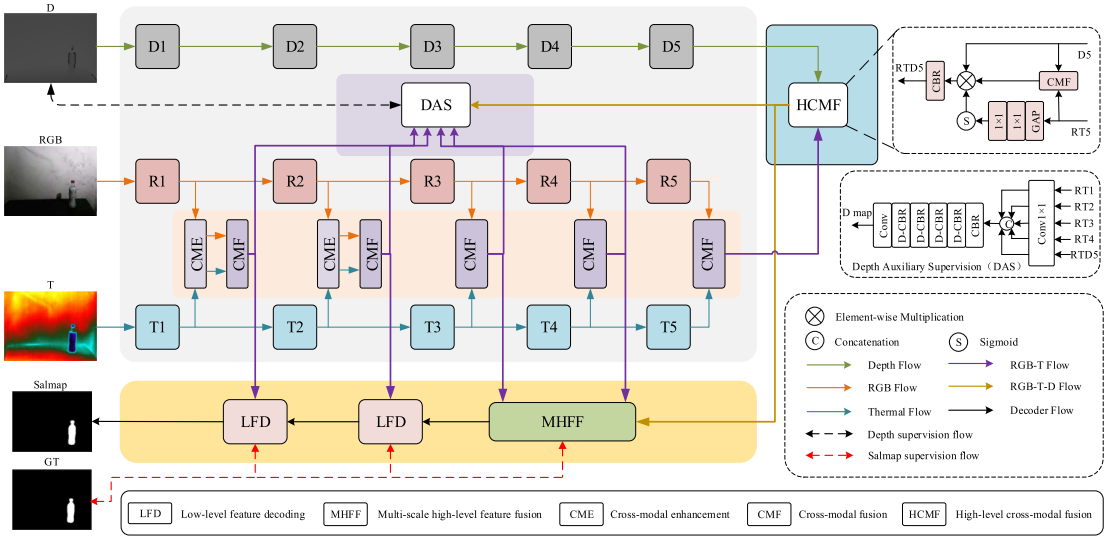

Fig. 3. The overall structure of the proposed method.

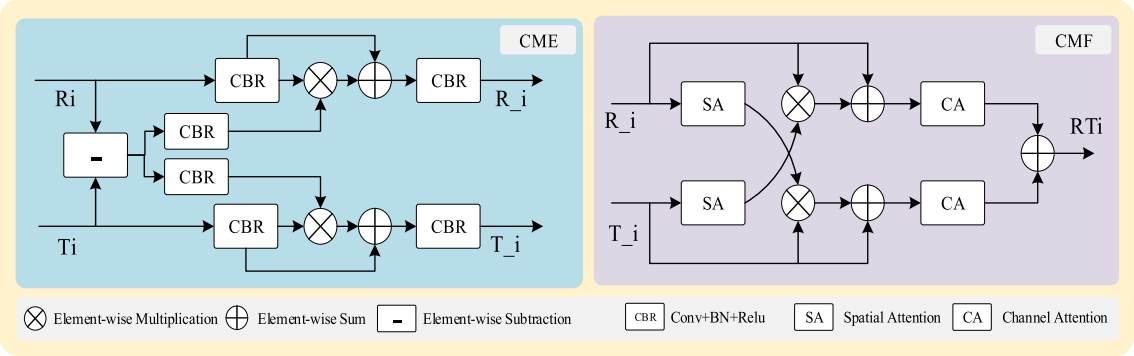

Fig. 4. The architecture of Cross-modal enhancement and fusion modules.

where *R_i* and *T_i* stand for the enhanced *Ri* and *Ti*, respectively, and CBR represents the convolution layer (Conv2d), normalization layer (BN) and activation function layer (Relu). For the discrepancy information RT_i^{dif}, the multiplication operation with RGB features is performed to enhance the T information weight in RGB, and the original information of RGB is retained by residual connection.

2) **Cross-modal fusion module (CMF).** Cross-modal interactive fusion is a crucial step in predicting multimodal saliency maps, which can be accurately predicted by fusing information from different modalities. At present, the favorable information of the two modal images cannot be fully fused by simply multiplying, adding and connecting operations (Wang et al., 2022). Therefore, we build a cross-modal fusion module to aggregate the features of different modalities simply and effectively. As shown in Fig. 4, in the low-level feature fusion stage, the input of the CMF is *R_i* and *T_i* after feature strengthening. In the high-level feature fusion stage, the input of the CMF is directly provided by the feature extraction backbone. The detailed operation process is shown in Fig. 3. The formula is expressed as follows:

$$RTi = CA((SA(R\_i) \times T\_i) + T\_i) + CA((SA(T\_i) \times R\_i) + R\_i), i \in [1,2] \tag{4}$$

$$RTi = CA((SA(Ri) \times Ti) + Ti) + CA((SA(Ti) \times Ri) + Ri), i \in [3,4,5] \tag{5}$$

where CA and SA (Woo et al., 2018) stand for the channel attention mechanism and spatial attention mechanism, respectively. *RTi* represents the features after cross-modal fusion of *Ri* and *Ti*. By using the cross-parallel operation of CA and SA and by multiplying and adding the features, the background redundant information in RGB and T images can be removed and the differentiated information of both can be retained. The gain of this module is also shown in the ablation experiment section. The final features after cross-modal fusion can be directly sent to the final decoding section for predicting the salient maps.

3) **High-level cross-modal fusion module (HCMF).** The depth map mainly provides the spatial structure information of objects. When the complex texture of RGB image or temperature intersection of T image is encountered in detection, it is helpful to distinguish the target and background information in the region. As previously studied (Sun et al., 2021), it is crucial to apply the depth map to the salient object detection task in a rational way. Here, since the texture information and imaging quality of the depth map are in most cases inferior to those of RGB and T images, but the depth map has an advantage in distinguishing foreground and background information, and to save computational cost, we fuse only the high-level features of the depth information. High-level depth features are introduced to compensate for the lack of semantic information when RGB and T images encounter challenging scenes. And in order to efficiently and effectively fuse the high-level trimodal features, as shown in Fig. 3, we raise the HCMF, as illustrated by Eqs. (6).

$$RTD5 = CBR(CMF(D5, RT5) \times D5 \times \partial(GAP(RT5))) \tag{6}$$

where CMF stands for the cross-modal fusion module, ∂ represents the Sigmoid activation function, and GAP denotes for global adaptive pooling. We first perform cross-modal fusion with *D5* and *RT5* to obtain the initial fusion features. To retain the advantage of depth features in distinguishing foreground and background, we then multiply with *D5*. To retain the high-level semantic features of RGB and T, we then multiply with *RT5* processed by GAP and Sigmoid layers. Finally, the final trimodal high-level semantic features are output by CBR and used to guide the multi-scale high-level feature fusion.

3.3. Multi-scale high-level feature fusion (MHFF) module and low-level feature decoding (LFD) module

In the process of feature extraction, multi-scale features are lost step by step and semantic information in contrast increases as the number of layers deepens. Therefore, rational utilization of semantic and texture information of features at different scales is crucial in the salient object detection task. Nevertheless, most detection methods (Wen et al., 2021; Zhou et al., 2022) available only use the same decoder in the decoding phase, directly from the top to low level mapping to decode the final output significant map. Although these methods have some compatibility in dealing with multiscale features with different distributed texture and semantic information, this compatibility is achieved through the redundant design of the decoder. Therefore, in order to save computational cost and ensure the accuracy of saliency detection, we exploit two different decoders in the decoding phase.

1) **Multi-scale high-level feature fusion module (MHFF).** High-level features mainly contain semantic information, and have many channels and low resolution, while middle-level features have a balanced position of channel number and resolution. In order to reduce unnecessary computation, we need to obtain a coarse salient feature to guide the decoding of low-level features. We present a lightweight MHFF module, as shown in Fig. 5.

Unlike existing multi-scale feature decoding modules (Zhai et al., 2021), we use depthwise separable convolutions instead of ordinary convolutions to reduce unnecessary parameters. Simultaneously, due to the low resolution of high-level features, using large convolution kernel cannot effectively improve the feature receptive field, but will slow down the inference speed. Therefore, we utilize different dilated convolution rates (Yu and Koltun, 2015) to accelerate the inference speed of the net. As shown, D-GCM represents a depthwise separable global context module, where D-GCM-1 uses dilated rates of 1, 3, 5, D-GCM-2 using 1,3,7, and D-GCM-3 using 1, 3, 5, 7. The number of characteristic channels of *RT3*, *RT4*, and *RTD5* treated by D-GCM is unified as 32, which we represent *G_RT3*, *G_RT4*, and *G_RTD5*. The specific formulas of multi-scale high-level feature fusion module are expressed as follows:

$$\Psi_1 = CAT\left(F_{up}(DCBR(G\_RTD5)), DCBR\left(F_{up}(DCBR(G\_RTD5)) \times G\_RT4\right)\right) \tag{7}$$

$$\Psi_2 = CBR\left(CAT\left(DCBR\left(RT3 \times DCBR(F_{up2}(RT4)) \times DCBR(F_{up4}(RT5))\right), DCBR(\Psi_1)\right)\right) \tag{8}$$

where Ψ_1 represents the intermediate result of the feature fusion of the fifth and fourth layers. Ψ_2 is the result of the third layer feature and a fusion with scale 44 × 44 and number of channels 32. CAT indicates the concatenation operation and $F_{up\times n}$ represents n times upsampling of the features.

2) **Low-level feature decoding module (LFD).** Different from the high-level features, the low-level features mainly contain texture information, with a small number of channels, high resolution, and the detail accuracy of the final salient map is closely related to the low-level features. For this purpose, we design the LFD module, as shown in Fig. 6.

Unlike other methods, we do not use the original features at the end of the decoding operation, but also the features output by a U-Net structure. It is worth noting that using a large convolution kernel on the lower-level features, and the effect on the inference speed is not obvious, so we use multiple dilated convolution kernels of different sizes on the LFD to improve the detection accuracy. This way can reduce the interference of the background information in the original feature brought about by the direct residual connection. The specific formulas are expressed as follows:

$$\Phi_i = CAT(RTi, CF_{up2\times i}(\Psi_2)), i \in [1,2] \tag{9}$$

$$F_{mapi} = GAP(\Phi_i) \times CBR(Conv(BR\left(\sum_{3}^{j} DConv_{d=j}(Conv(\Phi_i))\right)) + U_3(\Phi_i)), \quad i \in [1,2], j \in [3,6,9,12] \tag{10}$$

where $CF_{up2\times i}$ represents the mapping function of Ψ_2, consisting of CBR + F_{up}. And Φ_i indicates the features after the fusion of middle-level feature maps and low-level features and serves as input to the LFD. $DConv_d$ represents a depthwise separable convolution with a dilated convolution rate. U3 represents the U-Net structure function of the three-layer residual connections. F_{mapi} represents salient features, which are finally output through a 1 × 1 convolution.

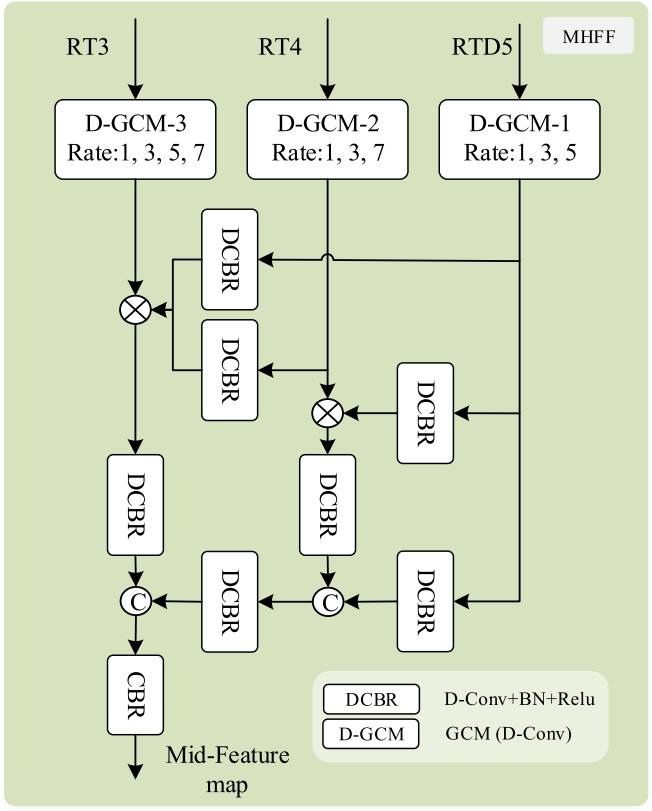

Fig. 5. Details of Multi-scale high-level feature fusion module.

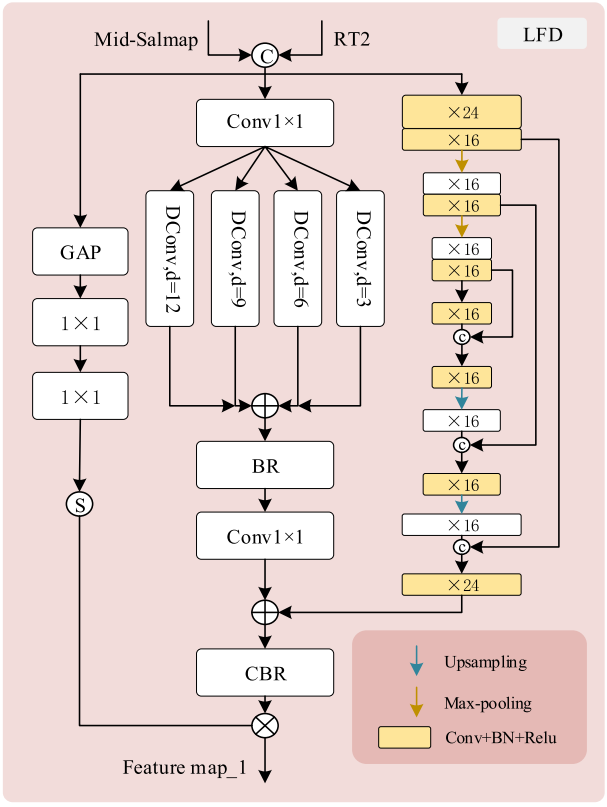

Fig. 6. Details of low-level feature decoding module.

3.4. Supervised and loss functions

1) **Deep auxiliary supervision.** Lightweight feature extraction backbone such as MobileNet (Sandler et al., 2018) is not as powerful as large numbers of parameters such as ResNet (He et al., 2016), VGG (Simonyan and Zisserman, 2014). To increase the accuracy of multi-modality salient object detection, we tried to improve the feature extraction ability of the lightweight feature extraction backbone. We found that the imaging of the depth map does not have the texture information of RGB and T pictures, but is presented with a spatial smoothing degree. This character enables the depth map to suppress a part of the background information, and the suppressed background information is basically consistent with the background information of RGB and T images. Therefore, we consider that the depth map can be used as an auxiliary supervision during the training stage to supervise the backbones to improve their inhibitory ability to the background features. For this purpose, we design the DAS module, as shown in Fig. 3. The DAS uses *RT1*, *RT2*, *RT3*, *RT4*, *RTD 5* as input, through 1×1 convolution normalized to 256 channels, then sampled to the scale size of *RT4*, and connected together through a CBR and four DCBR operations, finally through a 1×1 convolution output prediction depth map, and the depth map as supervision to feedback, as illustrated by Eqs. (11) to (12).

$$D_{map} = Conv_{1\times1}(DCBR^4(CBR(Conv_{1\times1}(RT1, RT2, RT3, RT4, RDT5)))) \tag{11}$$

$$loss_D = 1 - SSIM(D, D_{map}) \tag{12}$$

where $DCBR^4$ represents four sets of DCBR operations. SSIM is used to calculate the structural similarity of two features and $loss_D$ represents the loss of depth versus predicted depth maps for auxiliary supervision. In the inference stage, this part will not occupy the calculated amount.

2) **Salient map supervision.** We use the output of these three decoders, F_{map1}, F_{map2}, Ψ_2 to supervise our approach. These three features are converted into a single channel by 1×1 convolution, and the predicted salient map is output by sigmoid and bilinear interpolation up-sampling. For convenience of representation, we note as $F_{sali}, i \in [1, 2, 3]$, where F_{sal1} is the final predicted salient map. We supervise our model using the binary cross-entropy loss and the dice loss (Milletari et al., 2016). The specific loss calculation formula is as follows:

$$loss = \sum_{1}^{i} (\lambda_i \times BCE(GT, F_{sali}) + \beta_i \times Dice(GT, F_{sali})), i \in [1, 2, 3] \tag{13}$$

where the *loss* stands for the main loss of the model. The λ_i and β_i represent the loss equilibrium coefficient, and we set the default parameters 1.

4. Experiments

In this part, we explain the experimental setup. Then we compare the proposed method with the most advanced SOD methods. Meanwhile, we conduct a comprehensive ablation experimental analysis of the proposed module, a complexity analysis, and finally we analyze the reason of failure examples.

4.1. Experimental setup

1) *Implementation Details:* The equipment system we use is Ubuntu18.04, the model training and testing framework is PyTorch, and all the experiments are calculated on NVIDIA RTX2070 super. Our model is trained for 65 iterations with an initial learning rate of 0.0001. The batch size is 4. Using Adam as the optimizer of our model, momentum, weight decay coefficient, β_1, and β_2 is set to 0.9, 0.0001, 0.9 and 0.99, respectively. During data loading, we use data enhancement operations such as horizontal flipping and random cropping, and finally adjusted the multimodal images to 320×320. The dataset and code are available at: https://github.com/VDT-2048/MFDF.
2) *Datasets:* We use the publicly available VDT-2048 dataset. We use the data completed by HWSI (Song et al., 2022) as our training and test sets, so as to ensure the fairness of the contrast experiments.
3) *Evaluation Metrics:* The accuracy evaluation index of salient object detection can objectively measure the accuracy of different methods to detect salient objects. Similar to recent work, we mainly use the following metrics, absolute average error (MAE) (Perazzi et al., 2012), the lower the better; F-measure (Fm) (Achanta et al., 2009), which considers precision and recall; weighted F-measure (W_F) (Margolin et al., 2014), which extends the base quantity to non-binary values and determines the weight error according to their location and neighborhood; E-measure (Em) (Fan et al., 2018), which is an enhanced alignment method, considering local pixels and image mean; S-measure (Sm) (Fan et al., 2017), which evaluates spatial structure similarity by combining regional perceived structural similarity Sr and object perceived structural similarity. Model complexity evaluation index can objectively measure the requirements of different methods for computing hardware. The Frames Per Second (FPS), which is the number of pictures per second of the model computing device in the inference stage; the Model Size refers to the space size of the computing device occupied by the pre-training weight; the amount of model parameters (Model Params), which refers to the number of parameters contained in the model; the Floating Point Operation Per Second (FLOPs), which includes multiplication and addition, only related to the model, and can be used to measure its complexity.

4.2. Comparison with the SOTA bimodal/trimodal methods

Currently, only individual method for salient object detection uses the VDT-2048 dataset. To ensure the comprehensiveness of the comparative experiment, we choose 16 state-of-the-art bimodal SOD methods relying on deep learning in the past few years. Meanwhile, for purpose of justice of the experiment, those methods that do not use VDT-2048 dataset, we do not use the training model provided by them, we use their open-source code to retrain and test, and the experimental parameters are set according to the original paper.

As shown in Table 1, these methods prove the effect of our proposed method on detection. Among these methods, DPANet (Chen et al., 2021), JL-DCF (Fu et al., 2020), BBS-Net (Zhai et al., 2021), CoNet (Ji et al., 2020), RD3D (Chen et al., 2021), HAINet (Li et al., 2021), CDINet (Zhang et al., 2021), DSNet (Wen et al., 2021), MobileSal (Wu et al., 2022), and MoADNet (Jin et al., 2022) are the RGB-D SOD methods. The RGB-T methods include ADFNet (Tu et al., 2020), MIED (Tu et al., 2020), MIDD (Tu et al., 2021), CGFNet (Wang et al., 2022), CSRNet (Huo et al., 2022), and OSRNet (Huo et al., 2022). HWSI (Song et al., 2022) is the RGB-D-T method.

Table 1
Quantitative comparison results for different methods, where red represents the best, blue represents the sub best, and green represents the third best. (Chen et al., 2021; Fu et al., 2020; Zhai et al., 2021; Ji et al., 2020; Chen et al., 2021; Li et al., 2021; Zhang et al., 2021; Wen et al., 2021; Wu et al., 2022; Jin et al., 2022; Tu et al., 2020; Tu et al., 2020; Tu et al., 2021; Wang et al., 2022; Huo et al., 2022; Huo et al., 2022; Song et al., 2022).

Models	Precision Metrics					Complexity Metrics			
	MAE↓	F_m ↑	W_F↑	E_m ↑	S_m ↑	FPS↑	MS(MB)↓	MP(M)↓	FLOPs(G)↓
DPANet$_{2021}$[31]	0.0045	0.6995	0.7826	0.8758	0.8889	24	370.2	92.4	58.9
JL-DCF$_{2020}$[28]	0.0056	0.6868	0.7227	0.8686	0.8543	18	574.8	143.5	861.2
BBS-Net$_{2021}$[44]	0.0051	0.7044	0.8168	0.8812	0.9123	19	199.8	49.8	31.1
CoNet$_{2020}$[40]	0.0078	0.6561	0.6785	0.8762	0.8080	34	167.6	42.5	20.3
RD3D$_{2021}$[34]	0.0044	0.6527	0.7904	0.8393	0.9091	13	189.7	46.9	50.7
HAINet$_{2021}$[33]	0.0035	**0.8272**	**0.8534**	**0.9648**	0.9097	8	239.7	59.8	181.4
CDINet$_{2021}$[39]	0.0038	0.7525	0.8267	0.9211	0.9093	19	217.5	54.4	150.1
DSNet$_{2021}$[32]	0.0060	0.7065	0.6766	0.8806	0.8051	43	692.8	172.4	117.3
MobileSal$_{2022}$[47]	0.0043	0.7557	0.7890	0.9156	0.8778	**290**	**41.7**	**6.5**	**1.6**
MoADNet$_{2022}$[46]	0.0152	0.5596	0.5506	0.8050	0.7615	**65**	**20.6**	**5.0**	**2.0**
ADFNet$_{2020}$[51]	0.0091	0.4658	0.6355	0.6860	0.8668	5	332.6	83.1	191.5
MIED$_{2020}$[76]	0.0054	0.5784	0.7501	0.7828	0.8889	3	216.4	54.0	198.9
MIDD$_{2021}$[55]	0.0040	0.7043	0.8168	0.8804	0.9136	13	209.8	52.4	66.4
CGFNet$_{2021}$[53]	**0.0036**	0.7879	0.8498	0.9359	**0.9205**	11	265.8	66.4	345.2
CSRNet$_{2022}$[64]	0.0058	0.7722	0.7773	0.9450	0.8615	40	**4.7**	**0.98**	**4.2**
OSRNet$_{2022}$[63]	**0.0030**	**0.8357**	**0.8817**	**0.9658**	**0.9286**	55	62.7	15.6	42.3
OUR	**0.0031**	**0.8392**	**0.8665**	**0.9687**	**0.9191**	124	56.4	8.9	**4.2**
HWSI$_{2022}$[12]	0.003	0.872	0.897	0.981	0.932	4	403.4	100.7	357.7

Quantitative metrics show that our proposed method is comparable to the most advanced method in bimodal SOD task in accuracy, and even some metrics have reached the current optimal. It is notable that the accuracy of our method is comparable to that of OSRNet (Huo et al., 2022), but in terms of complexity, the FPS of our method is about twice that, the Model Size (MS) is only about half that, and the FLOPs is one tenth of that. Therefore, under the condition of certain hardware requirements, our method is easier to implement than OSRNet (Huo et al., 2022). At the same time, we can see MobileSal (Wu et al., 2022), MoADNet (Jin et al., 2022) and CSRNet (Huo et al., 2022) in the table. These methods belong to lightweight methods, which perform well in complexity index, but their accuracy index is average. In practical application, these methods have low robustness and generalization, and will be analyzed in visualization results. Finally, the HWSI belongs to the trimodal salient target detection method. This method makes full use of the complementarity of trimodal images in model design, and has excellent detection performance. However, its model complexity exceeds that of most bimodal salient target detection methods which is too large. Its practical application difficulty is even greater than that of general bimodal methods. The results in Fig. 2 also show more intuitively that MFDF is an attractive multispectral salient object detector in terms of accuracy and speed. To sum up, our method has the advantage of providing enough information by using the trimodal image, and from the perspective of lightweight model, the detection accuracy is comparable to that of the most advanced bimodal method. Meanwhile, our model is lighter than it, which has the possibility of practical application.

In order to show the advantages of our proposed method more directly, we compare all the comparison methods in two new metrics, max E-measure scores (E_{max}) and max F-measure scores (F_{max}), and show them through the histogram in Fig. 7. The results also prove that our method is competitive with other methods.

4.3. Analysis of the visual results

As shown in Fig. 8, the salient objects in the first line of pictures have multiple bar profiles, and it can be seen that our method can detect the profiles more accurately than the others. The second and third rows are the detection of small objects, and it can be seen that our method is hardly disturbed by other information. In the fourth row, the color of the salient object is similar to the background color, our method can be accurately detected, and the salient object profile detected by other methods has different degrees of deletion. The fifth, sixth and seventh rows are to detect salient under low illumination conditions. It can be seen that other methods are affected by illumination, and there are problems such as partial missing structure of salient objects or unclear boundary between multiple objects. The visualization results verify that our method maintains good detection accuracy and has good robustness in some challenge scenarios. As shown in Fig. 9 PR curves and F-measure curves, our method shows excellent performance compared with other methods, and it is noteworthy that our method also gains a fairly competitive lead at different F-measure thresholds.

As shown in Fig. 10, we conduct a comprehensive performance comparison of our method with others. In Fig. 10(a), we compare the model parameter size and E-measure accuracy. For clarity of the experimental results, we remove the coordinates of RD3D (Chen et al., 2021), which falls between BBS-Net (Zhai et al., 2021) and MIDD (Tu et al., 2021), and only keep the coordinates of CDINet (Zhang et al., 2021) for the comparison, as the results of CDINet and HAINet (Li et al., 2021) were similar. We can see that although the model parameter size of CSRNet (Huo et al., 2022), MobileSal (Wu et al., 2022), and MoADNet (Jin et al., 2022) is low, their detection accuracy is poor. In contrast, our proposed method has a model parameter size of only 8.9 MB and high

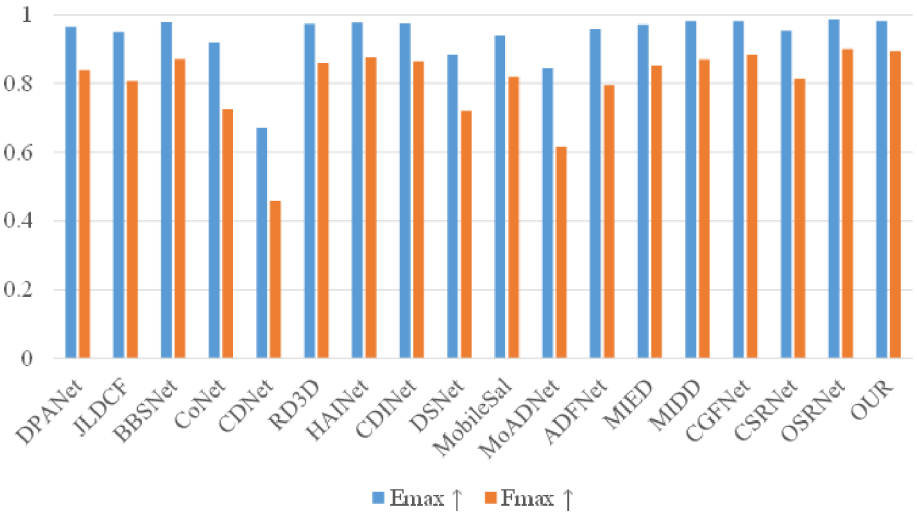

Fig. 7. Comparison of visualization results between our MFDF and other methods on E_{max} and F_{max}.

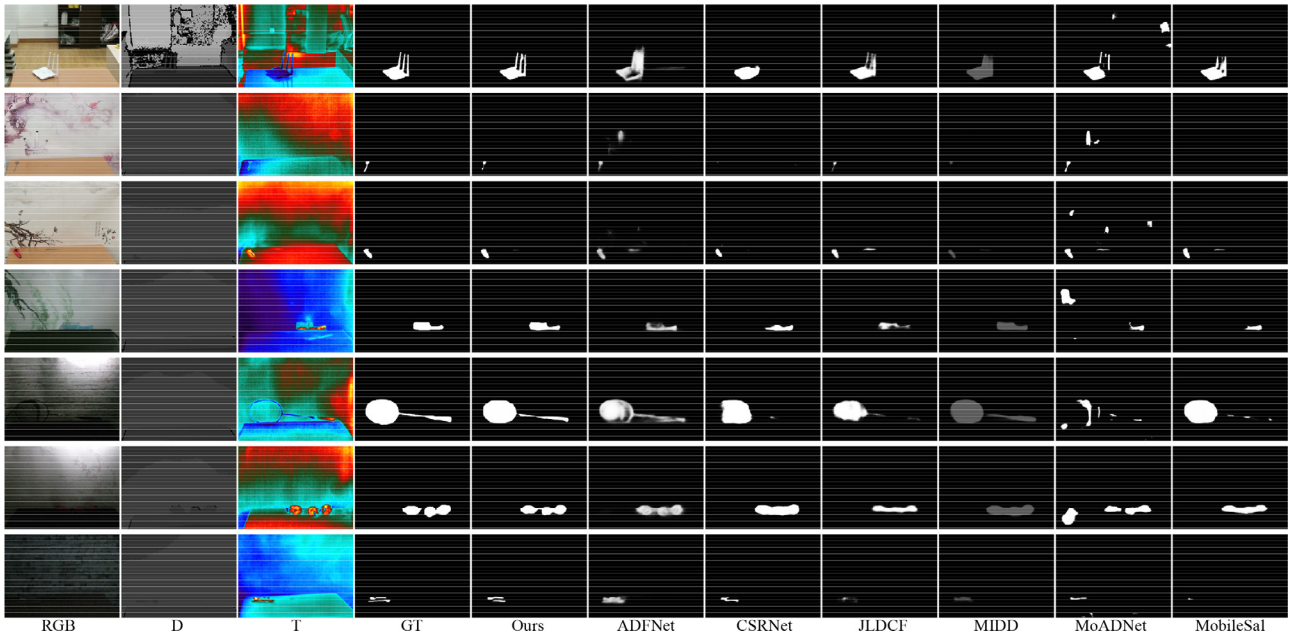

Fig. 8. Comparison of the visualization results of our MFDF and other methods.

detection accuracy. In Fig. 10(b), we compare the model FLOPs and S-measure accuracy. Here, we remove CoNet (Ji et al., 2020), MIDD (Tu et al., 2021), and ADFNet (Tu et al., 2020) for clarity, as they were close to BBS-Net. We can see that MobileSal and MoADNet have a low computational cost, but their detection accuracy is also low. CSRNet has a similar computational cost as our method, but we have a much higher detection accuracy. In summary, our MFDF achieves a good balance between detection accuracy and model complexity.

4.4. Ablation study

4.4.1. The effect analysis of different module

In this part, we study the contribution of the main modules we used. All of the ablation experiments are trained and tested on the VDT-2048 dataset. We use five evaluation metrics. The specific data are shown in Table 2.

1) Contribution of the CMF module: As shown in the second row of Table 2, we add a cross-modal fusion module on the basis of backbone, and other parameter settings remain unchanged from the proposed method. We can see that the F-measure a maximum increase of 4.4%. Other indexes have also improved to varying degrees. It is worth noting that some indexes will decline slightly, which may be caused by the low stability of the network. Secondly, it is possible that the CMF mainly uses the spatial and channel attention mechanism in the fusion stage, and the information processing is not perfect. As far as the overall index is concerned, this module contributes to improve the detection performance.
2) Contribution of the DAS module: As shown in the third row of Table 2, we add a deep auxiliary supervision module on the basis of 1), and other parameter settings remain unchanged with the proposed method. We can see that most

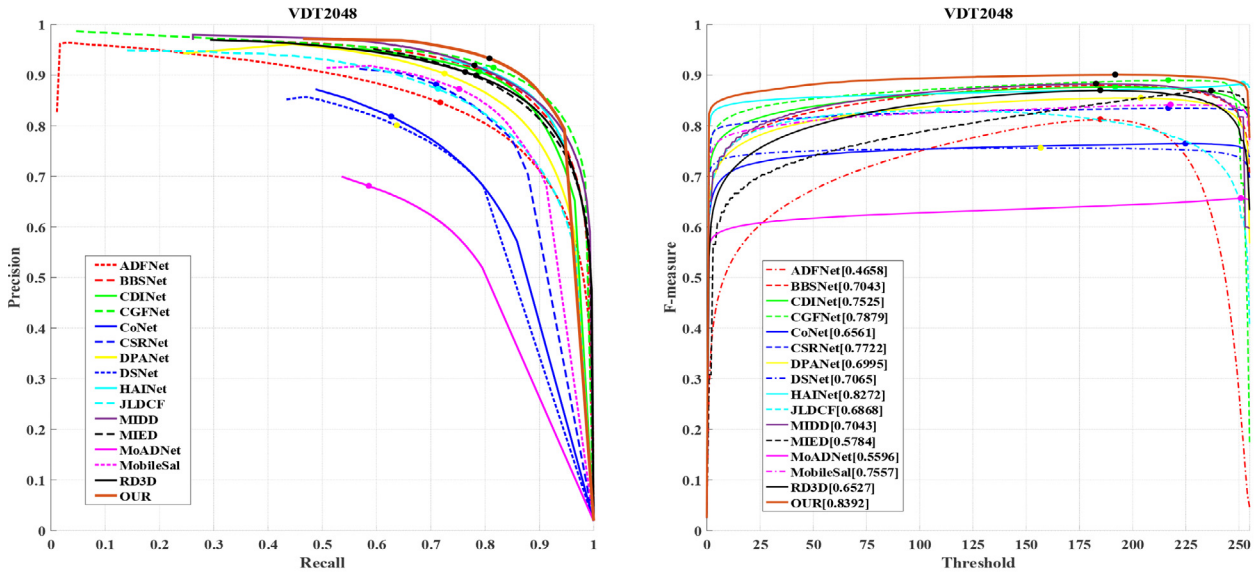

Fig. 9. Quantitative comparison results between our proposed method and the SOTA methods on the VDT-2048 dataset. The first line is Precision (vertical axis) Recall (horizontal axis) curves, and the second line shows the F-measure scores of the deep learning-based methods under different thresholds.

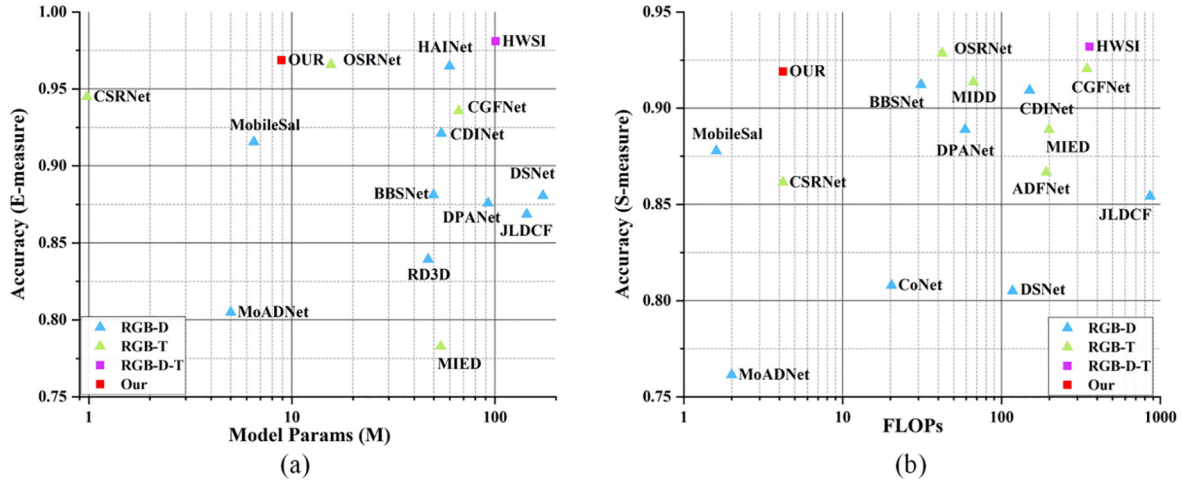

Fig. 10. Comprehensive performance comparison. The larger the E-measure and S-measure, the better. The smaller the Model Params and FLOPs, the better.

Table 2
Comparison of the different contributions of the main modules.

Model	MAE↓	F_m ↑	W_F↑	E_m ↑	S_m ↑
–	0.0044	0.7675	0.8100	0.9382	0.8907
CMF	0.0041	0.8115	0.8147	0.9589	0.8868
CMF-DAS	0.0039	0.8136	0.8120	0.9614	0.8899
CMF-DAS-HCMF	0.0038	0.8104	0.8314	0.9605	0.8984
CMF-DAS-HCMF-MHFF	0.0036	0.8205	0.8468	0.9640	0.9079
CMF-DAS-HCMF-MHFF-LFD	0.0032	0.8291	0.8619	0.9644	0.9170
CMF-DAS-HCMF-MHFF-LFD-CME	0.0031	0.8392	0.8665	0.9687	0.9191

indexes have been improved to some extent. This module mainly improves the feature extraction ability of the feature extraction backbone and improves the stability of our method.

3) Contribution of the HCMF module: As shown in the fourth row of Table 2, we add a high-level cross-modal fusion module on the basis of 2), with other parameter settings and the proposed method remaining unchanged. We can see that the index W_F has a maximum 1.96% increase, the S-measure is a 0.85% increase, and some indexes have a slight decrease. The reason is that this module mainly provides semantic information when RGB and T images encounter challenge scenes at the same time, and has certain robustness in the detection of some hard scenes. If the depth information is unreliable, it will affect the already unreliable RGB and T information.

4) Contribution of the MHFF module: As shown in the fifth row of Table 2, we add a multi-scale high-level feature fusion module on the basis of 3), keeping other parameter settings and the proposed method. We can see that all indexes have improved, among which the W_F increased by 1.54%, the F-measure increased by 0.99%, and the S-measure increased by 0.95%. This module mainly carries out multi-scale fusion of high-level features, and outputs a middle-level salient feature to guide the subsequent low-level decoding. This fusion method can reduce the parameter redundancy of step-by-step decoding, and plays a crucial part in improving the detection performance and reducing the model complexity.
5) Contribution of the LFD module: As shown in the sixth line of Table 2, we add a low-level feature decoding module on the basis of 4), and other parameter settings remain unchanged with the proposed method. We can see that all indexes have improved, with the F-measure increasing by 0.86%, the W_F by 1.51%, and the S- measure by 0.91%. This module mainly aims at the characteristics of low-level features with rich texture information. By using U-Net structure instead of residual structure to remove part of the background information of the low-layer features, and using multiple empty convolution cores to expand the feature receptive field, so as to reduce the interference of the background information on the final salient map.
6) Contribution of the CME module: As shown in the last row of Table 2, we add the CME based on 5), and other parameter settings remain unchanged from the proposed method. We can see an improvement in all indexes, with F-measure increased by 1.01%. This module is mainly before the CMF. The reason why reducing the difference between modalities is not used in high-level is that the difference is mainly obvious in low-level features. By doing so, it can avoid cross-modal fusion in the wrong direction because the features of a certain modality deviate from the truth map extremely. This module can improve our method to deal with the fusion problem caused by poor imaging quality of one of the modalities.

4.4.2. The influence analysis of different factors

1) Effectiveness analysis of U-Net structure in LFD module. As shown in Table 3, the first row indicates that we change the original U-Net structure into a simple residual connection, and other structural parameters remain unchanged. The experimental results show that the complexity of the model is only slightly reduced, but the detection results are greatly reduced, in which Fm is reduced by 3.03%, and other accuracy indexes are also reduced to varying degrees, so this part of the structure is effective.
2) Different Dilation Rates used in LFD module. As shown in Table 4, the numbers to the right of the name represent the parallel features at different spatial rates. We can see that using different dilation rates in LFD module has little effect on model complexity, but the detection accuracy is improved to varying degrees, and the Fm of the proposed method is improved by 1.12%.
3) Analysis of depthwise separable convolution in the MHFF module. As shown in Table 5, we replace the depthwise separable convolution with ordinary convolution in the MHFF module of our method while keeping other structural parameters unchanged. From the experimental results, it can be observed that the model complexity increases significantly, and only some detection accuracy indicators improve slightly. Therefore, it is feasible to replace vanilla convolution with depthwise separable convolution.
4) Analysis of MHFF module with convolution kernels of different sizes. As shown in Table 6, the number to the right of the name represents the parallel convolution kernel size we use in the high-level feature. In the first row, we all use convolution kernels of size 1 and 3 in the three layers of high-level features. The complexity of the model decreases, but the detection results also decrease obviously. In the second row, we all use kernels of size 1, 3, 5 and 7 in the three layers of high-level features. We can see that the complexity of the model has increased and the detection accuracy indicator has partially increased. In the third, the size distribution of convolution kernels used in the three levels of high-level features is opposite to the proposed method. The complexity indicator of the model partially increases, because the dimension of the high-level features is high and the scale is small, and the convolution kernel used has many parameters. However, because the scale is small, the computation is small, and there is a balance relationship here. We can also see that the accuracy indicator here has declined. To sum up, the different high-level feature convolution kernel distribution strategies used in the proposed method are better.
5) Analysis of homogeneous module in cross-modal fusion stage. As shown in Table 7, the first row indicates that we use the CME module for each layer of features, while keeping other structure constant. We can see that the model complexity increases significantly, and the detection accuracy does not reach the optimal level. The second row indicates that we do not use the CME for each layer of features, while keeping other structure constant. The model complexity decreases slightly, and the detection accuracy does not reach the optimal level. In conclusion, our proposed method only uses CME for low-level feature cross-modal fusion stages, which has little impact on model complexity but improves detection accuracy to varying degrees, with Fm increasing by 1.01%.
6) Analysis of homogeneous module in decoding stage. As shown in Table 8, the first row indicates that we use LFD to decode all layers in the decoding stage, while keeping other network structure unchanged. We can see that the model complexity decreases, but the detection accuracy decreases to varying degrees, with Fm decreasing by 1.67%. The second row indicates that we use MHFF to decode all layers in the decoding stage, while keeping other network structure unchanged. The model complexity increases significantly, and the detection accuracy of the model decreases by a large margin, with Fm decreasing by 7.86%. In summary, the proposed method's differential treatment of high and

Table 3
Effectiveness analysis of U-Net structure in LFD module.

Models	Precision Metrics					Complexity Metrics			
	MAE↓	F_m ↑	W_F↑	E_m ↑	S_m ↑	FPS↑	MS(MB)↓	MP(M)↓	FLOPs(G)↓
LFD-U ×	0.0033	0.8089	0.8572	0.9566	0.9148	**130**	**55.1**	**8.8**	**3.8**
OUR	**0.0031**	**0.8392**	**0.8665**	**0.9687**	**0.9191**	124	56.4	8.9	4.2

Table 4
Different Dilation Rates used in LFD module.

Models	Precision Metrics					Complexity Metrics			
	MAE↓	F_m ↑	W_F↑	E_m ↑	S_m ↑	FPS↑	MS(MB)↓	MP(M)↓	FLOPs(G)↓
LFD-3	0.0035	0.8155	0.8536	0.9567	0.9118	**128**	56.4	8.9	**4.1**
LFD-3,6	0.0033	0.8157	0.8562	0.9590	0.9147	125	56.4	8.9	4.1
LFD-3,9	0.0033	0.8280	0.8552	0.9624	0.9112	125	56.4	8.9	4.1
OUR	**0.0031**	**0.8392**	**0.8665**	**0.9687**	**0.9191**	124	**56.4**	**8.9**	4.2

Table 5
Analysis of depthwise separable convolution replacement in MHFF module.

Models	Precision Metrics					Complexity Metrics			
	MAE↓	F_m ↑	W_F↑	E_m ↑	S_m ↑	FPS↑	MS(MB)↓	MP(M)↓	FLOPs(G)↓
MHFF-Dwise ×	0.0032	0.8096	0.8623	0.9553	**0.9211**	95	62.4	9.5	5.6
OUR	**0.0031**	**0.8392**	**0.8665**	**0.9687**	0.9191	**124**	**56.4**	**8.9**	**4.2**

Table 6
Analysis of MHFF module with convolution kernels of different sizes.

Models	Precision Metrics					Complexity Metrics			
	MAE↓	F_m ↑	W_F↑	E_m ↑	S_m ↑	FPS↑	MS(MB)↓	MP(M)↓	FLOPs(G)↓
MHFF-1,3	0.0033	0.8306	0.8618	0.9650	0.9138	**129**	**53.7**	**8.2**	**3.6**
MHFF-1,3,5,7	0.0033	0.8355	0.8627	**0.9698**	0.9130	112	57.9	9.3	4.3
MHFF- opposite	0.0033	0.8259	0.8538	0.9644	0.9099	117	56.6	8.9	4.0
OUR	**0.0031**	**0.8392**	**0.8665**	0.9687	**0.9191**	124	56.4	8.9	4.2

Table 7
Analysis of homogeneous module in cross-modal fusion stage.

Models	Precision Metrics					Complexity Metrics			
	MAE↓	F_m ↑	W_F↑	E_m ↑	S_m ↑	FPS↑	MS(MB)↓	MP(M)↓	FLOPs(G)↓
CME (all-layers)	0.0033	0.8267	0.8607	0.9637	0.9165	94	73.6	13.2	4.8
CME ×	0.0032	0.8291	0.8619	0.9644	0.9170	132	56.3	8.9	3.8
OUR	**0.0031**	**0.8392**	**0.8665**	**0.9687**	**0.9191**	**124**	**56.4**	**8.9**	**4.2**

Table 8
Analysis of homogeneous module in decoding stage.

Models	Precision Metrics					Complexity Metrics			
	MAE↓	F_m ↑	W_F↑	E_m ↑	S_m ↑	FPS↑	MS(MB)↓	MP(M)↓	FLOPs(G)↓
LFD (all-layers)	0.0031	0.8225	0.8585	0.9598	0.9150	138	53.2	8.3	3.0
MHFF(all-layers)	0.0038	0.7606	0.8369	0.9248	0.9141	40	61.9	10.1	25.5
OUR	**0.0031**	**0.8392**	**0.8665**	**0.9687**	**0.9191**	**124**	**56.4**	**8.9**	**4.2**

low-level features is an effective decoding strategy that achieves advantages in both model complexity and detection accuracy.

4.5. Failure examples analysis and future direction

In this section, we mainly analyze the failure of our method, and put forward relevant solutions that may be used in future schemes. As shown in Fig. 11, in the first row, when the RGB image is extremely dim, the salient in the T image resembles the background color, which leads to the failure of our method to detect salient. This situation may require a more powerful backbone for feature extraction.

In the second row, the salient object in both RGB and T images is very obvious. However, the color information in RGB images crosses and thermal crossover occurs in T images, which eventually leads to the incomplete detection of the salient map at the junction of the two modalities of color. In this case, it may be necessary to choose the one with high confidence for the local information of different modalities in the cross-modal fusion stage. In the third row, the salient objects in RGB, D and T images are not obvious, so the available effective information is not enough to support the existing methods to detect. In this case, it is generally necessary to improve the imaging quality of the equipment or improve the imaging conditions of the scene to be detected. The fine pole part of the fourth line of badminton racket is not detected. This

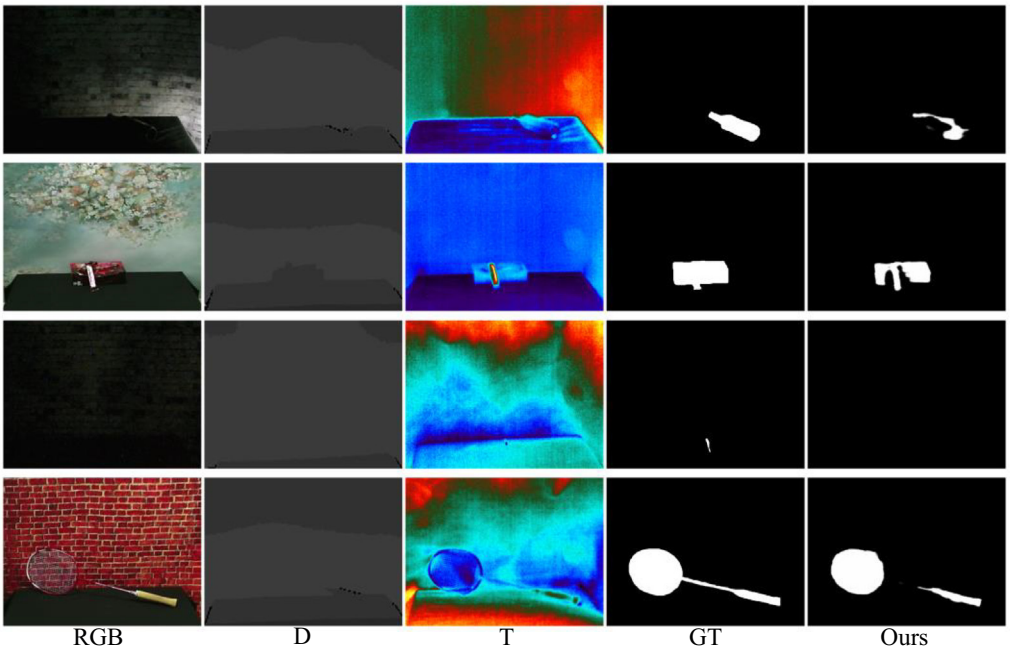

Fig. 11. Failure cases of the proposed MFDF. The first row shows the RGB image is extremely dim. The second row presents thermal crossover phenomenon exists in the T images. The third row illustrates the salient object of RGB, D, and T images is not obvious.

situation may be that part of the color of the bar in RGB and T images is similar to the background, which makes it difficult for our method to extract features. Secondly, it may be that we extract this part of features and are filtered out in the fusion decoding process. To solve this problem, it may be necessary to design a module for detecting slender objects in the future.

5. Conclusion

In this paper, we propose a lightweight multi-level feature difference fusion network for real-time RGB-D-T SOD. Firstly, we leverage the distinct information present in different modal images. We design an asymmetric network structure with MobileNetV2 as the backbone to effectively utilize the information from each modality while reducing unnecessary parameters. Secondly, in the cross-modal fusion stage, we employ the CME module to minimize the discrepancy between modalities prior to fusing low-level features. This helps to prevent incorrect low-level texture information from one modality from influencing the detection results. Next, we employ multi-scale high-level features for level-by-level refinement and optimize the model's inference speed by reducing the usage of large null convolution kernels. For low-level feature decoding, we adopt multiple dilated convolution kernels to expand the perceptual field of low-level features. Additionally, we use a U-shaped structure instead of a residual structure to filter out background information. Our comparative experiments demonstrate the excellent performance of our method in terms of detection accuracy and complexity. Furthermore, our analysis of different factors shows that the employed module structure and model architecture are effective. In the future, we will continue to optimize our method and aim to apply it to terminal devices.

Declaration of Competing Interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: [Kechen Song reports financial support was provided by National Natural Science Foundation of China.].

References

Achanta, R., Hemami, S., Estrada, F. and Susstrunk, S., 2009. Frequency-tuned salient region detection. In: 2009 IEEE Conference on Computer Vision and Pattern Recognition, pp. 1597-1604.

Arulprakash, E., Aruldoss, M., 2022. A study on generic object detection with emphasis on future research directions. J. King Saud Univ.–Comput. Information Scientist 34 (9), 7347–7365.

Chen, S., Fu, Y., 2020. Progressively guided alternate refinement network for RGB-D salient object detection. In: Computer Vision–ECCV 2020: 16th European Conference, pp. 520-538.

Chen, Z., Cong, R., Xu, Q., Huang, Q., 2021. DPANet: depth potentiality-aware gated attention network for RGB-D salient object detection. IEEE Trans. Image Process. 30, 7012–7024.

Chen, Q., Liu, Z., Zhang, Y., Fu, K., Zhao, Q., Du, H., 2021. RGB-D salient object detection via 3D convolutional neural networks. In: AAAI, pp. 1063-1071.

Chen, S.H., Lai, Y.W., Kuo, C.L., Lo, C.Y., Lin, Y.S., Lin, Y.R., Kang, C.H., Tsai, C.C., 2022. A surface defect detection system for golden diamond pineapple based on CycleGAN and YOLOv4. J. King Saud Univ.–Comput. Inf. Sci. 34 (10), 8041–8053.

Chen, G., Shao, F., Chai, X., Chen, H., Jiang, Q., Meng, X., Ho, Y.S., 2022. CGMDRNet: cross-guided modality difference reduction network for RGB-T salient object detection. IEEE Trans. Circuits Syst. Video Technol. 32 (9), 6308–6323.

Chen, H., Zhao, J., 2023. 3D mesh classification and panoramic image segmentation using spherical vector networks with rotation-equivariant self-attention mechanism. J. King Saud Univ.–Comput. Inf. Sci. 35, (5) 101546.

Cheng, M.-M., Mitra, N.J., Huang, X., Torr, P.H.S., Hu, S.-M., 2015. Global contrast based salient region detection. IEEE Trans. Pattern Anal. Mach. Intell. 37 (3), 569–582.

Fan, D.P., Cheng, M.M., Liu, Y., Li, T., Borji, A., 2017. Structure-measure: a new way to evaluate foreground maps. In: Proceedings of the IEEE international conference on computer vision, pp. 4548-4557.

Fan, D.P., Gong, C., Cao, Y., Ren, B., Cheng, M.M., Borji, A., 2018. Enhanced-alignment measure for binary foreground map evaluation, arXiv preprint arXiv:1805.10421.
Fang, X., Zhu, J., Shao, X., Wang, H., 2022. GroupTransNet: Group Transformer Network for RGB-D Salient Object Detection, arXiv preprint arXiv:2203.10785.
Fu, K., Fan, D.P., Ji, G.P., Zhao, Q., 2020, JL-DCF: Joint Learning and Densely-Cooperative Fusion Framework for RGB-D Salient Object Detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 3049-3059.
Gao, W., Liao, G., Ma, S., Li, G., Liang, Y., Lin, W., 2022. Unified information fusion network for multi-modal RGB-D and RGB-T salient object detection. IEEE Trans. Circuits Syst. Video Technol. 32 (4), 2091–2106.
Guo, Q., Zhou, W., Lei, J., Yu, L., 2021. TSFNet: two-stage fusion network for RGB-T salient object detection. IEEE Signal Process Lett. 28, 1655–1659.
He, K., Zhang, X., Ren, S., Sun, J., 2016, Deep residual learning for image recognition. In: Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), 2016, pp. 770–778.
Hou, Q., Cheng, M.M., Hu, X., Borji, A., Tu, Z., Torr, P.H., 2017, Deeply supervised salient object detection with short connections, in: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 3203-3212.
Huang, N., Jiao, Q., Zhang, Q., Han, J., 2022, Middle-level fusion for lightweight RGB-D salient object detection, 2021, arXiv preprint arXiv:2104.11543.
Huo, F., Zhu, X., Zhang, L., Liu, Q., Shu, Y., 2022. Efficient context-guided stacked refinement network for RGB-T salient object detection. IEEE Trans. Circuits Syst. Video Technol. 32 (5), 3111–3124.
Huo, F., Zhu, X., Zhang, Q., Liu, Z., Yu, W., 2022. Real-Time one-stream semantic-guided refinement network for RGB-thermal salient object detection. IEEE Trans. Instrum. Meas. 71, 1–12.
Janneh, L.L., Zhang, Y., Cui, Z., Yang, Y., 2023. Multi-Level feature Re-weighted fusion for the semantic segmentation of crops and weeds. J. King Saud Univ.–Comput. Inf. Sci. 101545
Ji, W., Li, J., Zhang, M., Piao, Y., Lu, H., 2020, Accurate RGB-D salient object detection via collaborative learning. In: Computer Vision–ECCV 2020: 16th European Conference, 2020, pp. 52-69.
Ji, W., Li, J., Yu, S., Zhang, M., Piao, Y., Yao, S., Bi, Q., Ma, K., Zheng, Y., Lu, H., Cheng, L., 2021., Calibrated RGB-D salient object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 9471-9481.
Jin, X., Yi, K., Xu, J., 2022. MoADNet: mobile asymmetric dual-stream networks for real-time and lightweight RGB-D salient object detection. IEEE Trans. Circuits Syst. Video Technol. 32 (11), 7632–7645.
Kesav, N., Jibukumar, M.G., 2022. Efficient and low complex architecture for detection and classification of brain tumor using RCNN with two channel CNN. J. King Saud Univ.–Comput. Inf. Sci. 34 (8), 6229–6242.
Li, G., Liu, Z., Chen, M., Bai, Z., Lin, W., Ling, H., 2021. Hierarchical alternate interaction network for RGB-D salient object detection. IEEE Trans. Image Process. 30, 3528–3542.
Liao, G., Gao, W., Li, G., Wang, J., Kwong, S., 2022. Cross-collaborative fusion-encoder network for robust RGB-thermal salient object detection. IEEE Trans. Circuits Syst. Video Technol. 32 (11), 7646–7661.
Liu, N., Han, J., Yang, M.H., 2018. Picanet: Learning pixel-wise contextual attention for saliency detection. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 3089-3098.
Liu, N., Zhang, N., Han, J., 2020, Learning selective self-mutual attention for RGB-D saliency detection. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 13756-13765.
Liu, Z., Wang, Y., Tu, Z., Xiao, Y., Tang, B., 2021, TriTransNet: RGB-D salient object detection with a triplet transformer embedding network. In: Proceedings of the 29th ACM International Conference on Multimedia, pp. 4481-4490.
Liu, Y., Gu, Y.-C., Zhang, X.-Y., Wang, W., Cheng, M.-M., 2021. Lightweight salient object detection via hierarchical visual perception learning. IEEE Trans. Cybern. 51 (9), 4439–4449.
Liu, J.-J., Hou, Q., Cheng, M.-M., 2020. Dynamic feature integration for simultaneous detection of salient object, edge, and skeleton. IEEE Trans. Image Process. 29, 8652–8667.
Liu, X., Hou, S., Liu, S., Ding, W., Zhang, Y., 2023. Attention-based multimodal glioma segmentation with multi-attention layers for small-intensity dissimilarity. J. King Saud. Univ.–Comput. Inf. Sci. 35 (4), 183–195.
Liu, Z., Tan, Y., He, Q., Xiao, Y., 2022. SwinNet: Swin transformer drives edge-aware RGB-D and RGB-T salient object detection. IEEE Trans. Circuits Syst. Video Technol. 32 (7), 4486–4497.
Liu, Y., Zhang, X.-Y., Bian, J.-W., Zhang, L., Cheng, M.-M., 2021. SAMNet: stereoscopically attentive multi-scale network for lightweight salient object detection. IEEE Trans. Image Process. 30, 3804–3814.
Lu, S., Tan, C., Lim, J.-H., 2014. Robust and efficient saliency modeling from image co-occurrence histograms. IEEE Trans. Pattern Anal. Mach. Intell. 36 (1), 195–201.
Margolin, R., Zelnik-Manor, L., Tal, A., 2014, How to evaluate foreground maps? In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 248-255.
Milletari, F., Navab, N., Ahmadi, S.A., 2016, V-Net: Fully convolutional neural networks for volumetric medical image segmentation. In: Proc. 4th Int. Conf. 3D Vis., pp. 565–571.
Mohakud, R., Dash, R., 2022. Skin cancer image segmentation utilizing a novel EN-GWO based hyper-parameter optimized FCEDN. J. King Saud Univ.–Comput. Inf. Sci. 34 (10), 9889–9904.
Perazzi, F., Krähenbühl, P., Pritch, Y., Hornung, A., 2012, Saliency filters: contrast based filtering for salient region detection. In: 2012 IEEE Conference on Computer Vision and Pattern Recognition, pp. 733-740.
Piao, Y., Rong, Z., Zhang, M., Ren, W., Lu, H., 2020. A2dele: Adaptive and attentive depth distiller for efficient RGB-D salient object detection. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 9060-9069.
Risnandar, 2022. DeSa COVID-19: Deep salient COVID-19 image-based quality assessment. J. King Saud Univ.–Comput. Inf. Sci. 34 (10), 9501–9512.
Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., Chen, L.C., 2018, MobileNetV2: Inverted residuals and linear bottlenecks. In: Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2018, pp. 4510–4520.
Sharma, V., Mir, R.N., 2022. Saliency guided faster-RCNN (SGFr-RCNN) model for object detection and recognition. J. King Saud Univ.–Comput. Inf. Sci. 34 (5), 1687–1699.
Simonyan, K., Zisserman, A., 2014, Very deep convolutional networks for large-scale image recognition, arXiv preprint arXiv:1409.1556.
K. Song, L. Huang, A. Gong, Y. Yan, Multiple graph affinity interactive network and a variable illumination dataset for RGBT image salient object detection. In: IEEE Transactions on Circuits and Systems for Video Technology, doi: 10.1109/TCSVT.2022.3233131.
Song, K., Wang, J., Bao, Y., Huang, L., Yan, Y., 2022. A novel visible-depth-thermal image dataset of salient object detection for robotic visual perception. IEEE/ASME Trans. Mechatron. https://doi.org/10.1109/TMECH.2022.3215909.
Song, K., Zhao, Y., Huang, L., Yan, Y., Meng, Q., 2023. RGB-T image analysis technology and application: a survey. Eng. Appl. Artif. Intel. 120, 105919.
Sun, P., Zhang, W., Wang, H., Li, S., Li, X., 2021, Deep RGB-D saliency detection with depth-sensitive attention and automatic multi-modal fusion. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 1407-1417.
Tang, L., Li, B., Zhong, Y., Ding, S., Song, M., 2021, Disentangled high quality salient object detection. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 3580-3590.
Tu, Z., Ma, Y., Li, Z., Li, C., Xu, J., Liu, Y., 2022, Rgbt salient object detection: A large-scale dataset and benchmark, arXiv preprint arXiv:2007.03262.
Tu, Z., Li, Z., Li, C., Lang, Y., Tang, J., 2020, Multi-interactive encoder-decoder network for RGBT salient object detection, arXiv preprint arXiv:2005.02315.
Tu, Z., Li, Z., Li, C., Lang, Y., Tang, J., 2021. Multi-Interactive dual-decoder for RGB-thermal salient object detection. IEEE Trans. Image Process. 30, 5678–5691.
Tu, Z., Li, Z., Li, C., Tang, J., 2022. Weakly alignment-free RGBT salient object detection with deep correlation network. IEEE Trans. Image Process. 31, 3752–3764.
Wang, F., Pan, J., Xu, S., Tang, J., 2022. Learning discriminative cross-modality features for RGB-D saliency detection. IEEE Trans. Image Process. 31, 1285–1297.
Wang, J., Song, K., Bao, Y., Huang, L., Yan, Y., 2022. CGFNet: cross-guided fusion network for RGB-T salient object detection. IEEE Trans. Circuits Syst. Video Technol. 32 (5), 2949–2961.
Wen, H., Yan, C., Zhou, X., Cong, R., Sun, Y., Zheng, B., Ding, G., 2021. Dynamic selective network for RGB-D salient object detection. IEEE Trans. Image Process. 30, 9179–9192.
Woo, S., Park, J., Lee, J.Y., Kweon, I.S., 2018, Cbam: Convolutional block attention module. In: Proceedings of the European conference on computer vision (ECCV), 2018, pp. 3-19.
Wu, Y.-H., Liu, Y., Xu, J., Bian, J.-W., Gu, Y.-C., Cheng, M.-M., 2022. MobileSal: extremely efficient RGB-D salient object detection. IEEE Trans. Pattern Anal. Mach. Intell. 44 (12), 10261–10269.
Wu, Y.-H., Liu, Y., Zhang, L., Cheng, M.-M., Ren, B., 2022. EDN: salient object detection via extremely-downsampled network. IEEE Trans. Image Process. 31, 3125–3136.
Xia, R., Chen, Y., Ren, B., 2022. Improved anti-occlusion object tracking algorithm using Unscented Rauch-Tung-Striebel smoother and kernel correlation filter. J. King Saud Univ.–Comput. Inf. Sci. 34 (8), 6008–6018.
Yu, F., Koltun, V., 2015. Multi-scale context aggregation by dilated convolutions, arXiv preprint arXiv:1511.07122.
Zeng, J., Ouyang, H., Liu, M., Leng, L.U., Fu, X., 2022. Multi-scale YOLACT for instance segmentation. J. King Saud Univ.–Comput. Inf. Sci. 34 (10), 9419–9427.
Zhai, Y., Fan, D.P., Yang, J., Borji, A., Shao, L., Han, J., Wang, L., 2021. Bifurcated backbone strategy for rgb-d salient object detection. IEEE Trans. Image Process. 30, 8728–8742.
Zhang, J., Ehinger, K.A., Wei, H., Zhang, K., Yang, J., 2017. A novel graph-based optimization framework for salient object detection. Pattern Recogn. 64, 39–50.
Zhang, P., Wang, D., Lu, H., Wang, H., Ruan, X., 2017, Amulet: Aggregating multi-level convolutional features for salient object detection. In: Proceedings of the IEEE international conference on computer vision, pp. 202-211.
Zhang, W., Ji, G.P., Wang, Z., Fu, K., Zhao, Q., 2021. Depth quality-inspired feature manipulation for efficient RGB-D salient object detection. In: Proceedings of the 29th ACM international conference on multimedia, pp. 731-740.
Zhang, W., Jiang, Y., Fu, K., Zhao, Q., 2021, BTS-Net: Bi-Directional Transfer-And-Selection Network for RGB-D Salient Object Detection. In: 2021 IEEE International Conference on Multimedia and Expo (ICME), 2021, pp. 1-6.
Zhang, C., Cong, R., Lin, Q., Ma, L., Li, F., Zhao, Y., Kwong, S., 2021. Cross-modality discrepant interaction network for RGB-D salient object detection. In: Proceedings of the 29th ACM International Conference on Multimedia, 2021, pp. 2094-2102.

Zhang, J., Liang, Q., Shi, Y., 2022. Accurate and efficient salient object detection via position prior attention. Image Vis. Comput. 124, 104508.
Zhou, T., Fu, H., Chen, G., Zhou, Y., Fan, D.P., Shao, L., 2021. Specificity-preserving rgb-d saliency detection. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 4681-4691.
Zhou, W., Guo, Q., Lei, J., Yu, L., Hwang, J.-N., 2021. IRFR-Net: Interactive recursive feature-reshaping network for detecting salient objects in RGB-D images. IEEE Trans. Neural Networks Learn. Syst. https://doi.org/10.1109/TNNLS.2021.3105484.
Zhou, W., Zhu, Y., Lei, J., Wan, J., Yu, L., 2022. APNet adversarial learning assistance and perceived importance fusion network for all-day RGB-T salient object detection. IEEE Trans. Emerging Topics Computational Intell. 6 (4), 957–968.
Zhou, W., Zhu, Y., Lei, J., Wan, J., Yu, L., 2022. CCAFNet: crossflow and cross-scale adaptive fusion network for detecting salient objects in RGB-D images. IEEE Trans. Multimedia 24, 2192–2204.
Zhou, W., Guo, Q., Lei, J., Yu, L., Hwang, J.-N., 2022. ECFFNet: effective and consistent feature fusion network for RGB-T salient object detection. IEEE Trans. Circuits Syst. Video Technol. 32 (3), 1224–1235.
Zhu, W., Liang, S., Wei, Y., Sun, J., 2014, Saliency optimization from robust background detection. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 2814-2821.