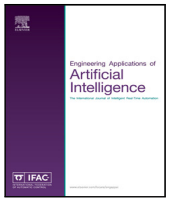
Engineering Applications of Artificial Intelligence 118 (2023) 105640

Contents lists available at ScienceDirect

Engineering Applications of Artificial Intelligence

journal homepage: www.elsevier.com/locate/engappai



Thermal images-aware guided early fusion network for cross-illumination RGB-T salient object detection

Han Wang, Kechen Song *, Liming Huang, Hongwei Wen, Yunhui Yan **

School of Mechanical Engineering & Automation, Northeastern University, Shenyang 110819, China
National Frontiers Science Center for Industrial Intelligence and Systems Optimization, Northeastern University, Shenyang 110819, China
Key Laboratory of Data Analytics and Optimization for Smart Industry (Northeastern University), Ministry of Education, China

ARTICLE INFO

Keywords:
Salient object detection
Cross-illumination
T-aware
Cross-modal fusion
Remote correction

ABSTRACT

RGB-T salient object detection (SOD) has been developed rapidly and achieved excellent results in recent years. However, some problems have not yet been solved. The current RGB-T datasets contain only a tiny amount of low-illumination data. The RGB-T SOD method trained based on these RGB-T datasets does not detect the salient objects in extremely low-illumination scenes very well. To improve the detection performance of low-illumination data, we can spend a lot of labor to label low-illumination data, but we tried a new idea to solve the problem by making full use of the properties of Thermal (T) images. Therefore, we propose a T-aware guided early fusion network for cross-illumination salient object detection. Specifically, in the training and testing stage, we use normal illumination data to train our network and then use low and extremely low-illumination data to verify the effectiveness of our method. In the early fusion stage, we propose a T-aware guided module (T-aware) for enhancing salient regions of RGB images at different illumination levels. Secondly, in the decoding stage, we use T images to guide the cross-modal fusion of RGB and T images. In addition, we propose a cross-modal fusion localization-remote correction module (CFL-RCM), which is used to deeply screen and correct redundant information generated by illumination variations. Comparative experiments on the VDT-2048 dataset validate the superior performance of our method on the cross-illumination RGB-T saliency detection. We also obtained favorable results on generalizability experiments with VT5000, VT1000, and VT821 datasets.

1. Introduction

Salient object detection mimics the human visual attention system and is used to detect and segment the most attention-grabbing regions or objects in an image. As a fundamental topic in the field of computer vision, salient object detection methods are widely used in the fields of video salient object detection (Huang et al., 2022b; Shokri et al., 2020; Kompella et al., 2021), object tracking (Liu et al., 2022b; Fiaz et al., 2019; Meinhardt et al., 2022), image segmentation (Cheng et al., 2022a; Strudel et al., 2021; Shivakumar et al., 2020) and other fields (Fu et al., 2022). In the past decade, most research has focused on RGB salient object detection and achieved excellent detection results. However, the performance of these RGB saliency detection methods degrades when detecting some images with low-illumination or complex backgrounds. With the popularity of depth cameras, depth information is pioneered to be integrated into RGB saliency detection. Because the pixel value of the depth image represents the distance from the object to the camera, it provides spatial information for saliency object detection. The introduction of depth image solves the problem of background complexity to a certain extent. In recent years, T-images have been gradually applied to the field of saliency detection and are developing rapidly because of their ability to compensate for low-illumination images.

The existing RGB-T datasets mainly include VT5000 (Tu et al., 2020a), VT1000 (Tu et al., 2019b), and VT821 (Wang et al., 2018a), which provide a large amount of RGB-T data and greatly promote the development of RGB-T SOD. However, these RGB-T datasets contain only a tiny amount of low-illumination data, and the salient features of the low-illumination data are still clearly outlined. As shown in Fig. 1, the low-illumination data of VT5000, VT1000, and VT821 account for only 10%, 5%, and 7%, respectively. The current RGB-T SOD methods mainly use the VT5000, VT1000, and VT821 to train and test. Although these RGB-T SOD methods achieve excellent detection results, the detection performance of these detection methods degrades drastically when detecting low-illumination data. Therefore, if we want to improve the performance of RGB-T SOD method to cope with low-illumination data, the most direct way is to increase the amount of

* Corresponding author at: School of Mechanical Engineering & Automation, Northeastern University, Shenyang 110819, China.
** Corresponding author.
E-mail addresses: songkc@me.neu.edu.cn (K. Song), yanyh@mail.neu.edu.cn (Y. Yan).

https://doi.org/10.1016/j.engappai.2022.105640
Received 23 August 2022; Received in revised form 3 November 2022; Accepted 10 November 2022
Available online xxxx
0952-1976/

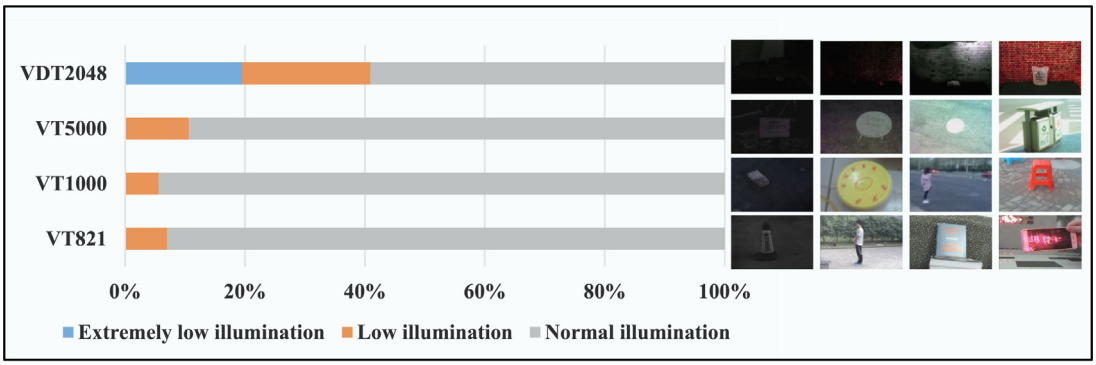


Fig. 1. The illumination percentages of the existing RGB-T datasets and the datasets we used.

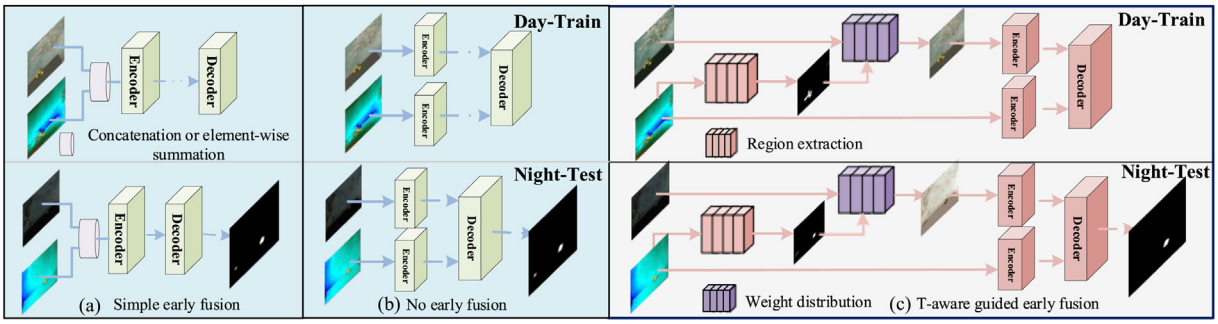


Fig. 2. Current bimodal images pre-processing architectures (a), (b), and our T-aware guided architecture (c).

low-illumination data. It is easy to obtain low-illumination data, but it is challenging to label it. As shown in Fig. 1, the recently published VDT-2048 dataset (Song et al., 2022b) contains 21% low-illumination data and 20% extremely low-illumination data, but these data are still insufficient to support the training.

The existing RGB-T SOD models can be divided into two categories: traditional methods and deep learning methods. The traditional methods are, on the one hand, top-down models, which are task-oriented models based mainly on specific high-level salience prior features and various crafted features. On the other hand, the bottom-up model, which is designed mainly for low-level features, usually uses color, texture, contrast, and borders. Traditional methods may be reliable in some specific scenarios. However, due to their lack of high-level semantic and contextual information, they will become unreliable in some variable illumination or more complex scenarios. Deep learning methods mainly deal with information fusion between bimodal, for example, multi-interaction dual decoding methods (Tu et al., 2021), unified information fusion methods for multimodal features (Gao et al., 2022) and cross-guided fusion networks (Wang et al., 2021). Deep learning methods generally perform unimodal feature extraction and then perform multimodal fusion during or after extraction to mine the information between the bimodal. However, the challenge for existing RGB-T SOD methods is to fully utilize T images in low-illumination scenes, which is difficult to solve by bimodal mid- and late-stage fusion. Therefore, our method focuses on constructing a cross-illumination SOD method using existing datasets to improve the performance of the SOD task to detect low-illumination data. Our work is similar to RGB-T and RGB-D SOD, but the existing methods do not address the following issues: (1) The state-of-the-art (SOTA) SOD methods are unsuitable for accurately detecting low-illumination data based on the existing dataset. (2) Existing SOD methods do not take full advantage of T-images to detect low-illumination data in the fusion strategy. Secondly, the existing fusion modules cannot adequately screen redundant information due to cross-illumination SOD.

Motivated by the discussions mentioned above, the main focus of this paper is to use the properties of T images to reduce the negative impact of low-illumination data on RGB-T SOD tasks. From the algorithm's perspective, we explored a new method to detect low-illumination data based on the training of normal illumination data. To this end, we propose a strategy of training with normal illumination data and testing with low illumination data to verify the robust compensability of our method for T images against RGB images in the SOD task. To support this proposal, we added early fusion and designed more robust fusion modules in the decoding phase, as discussed in detail below.

The state-of-the-art bimodal SOD methods are simply processed before the two modal images are fed into the network. As shown in Fig. 2(a), the early fusion ways of the methods proposed by Fu et al. (2020) and Huo et al. (2022b) are concatenation or element-wise summation. As shown in Fig. 2(b), other RGB-T SOD methods (Zhang et al., 2020; Tu et al., 2021; Gao et al., 2022; Zhou et al., 2022a; Guo et al., 2021; Zhang et al., 2021c; Huo et al., 2022a; Wang et al., 2021; Zhou et al., 2022b; Liu et al., 2022c; Chen et al., 2022; Tu et al., 2022; Liao et al., 2022; Xu et al., 2022; Liang et al., 2022a; Wang et al., 2022b; He et al., 2022; Zhang et al., 2022a; Ma et al., 2022; Jiang et al., 2022) do not process RGB and T images before input to the network. These methods process RGB and T images equally, and their performance decreases when a single modality is affected. To solve the problem, we constructed a T-aware guided module. Before feeding RGB images to the network, as shown in Fig. 2(c), we first extract a region of possible salient objects from the T images and use the region to enhance the RGB images. The first row of Fig. 3(a), (b), and (c) shows the RGB images under different illumination conditions. The second row shows the RGB images after the early fusion of the T-aware guided module. We can see that the T-aware guide module significantly improves the quality of RGB images in low-illumination situations. Different from Guan et al. (2018, 2019), we perform an early fusion of the RGB and T images and assign weights to the RGB and T images.

The illumination levels in the training and testing stages are different, which leads to a large amount of redundancy in the cross-modal fusion features in the testing stage. Therefore, to be suitable for scenes with changing illumination, our model focuses more on locating salient objects and screening redundant information in the decoding stage

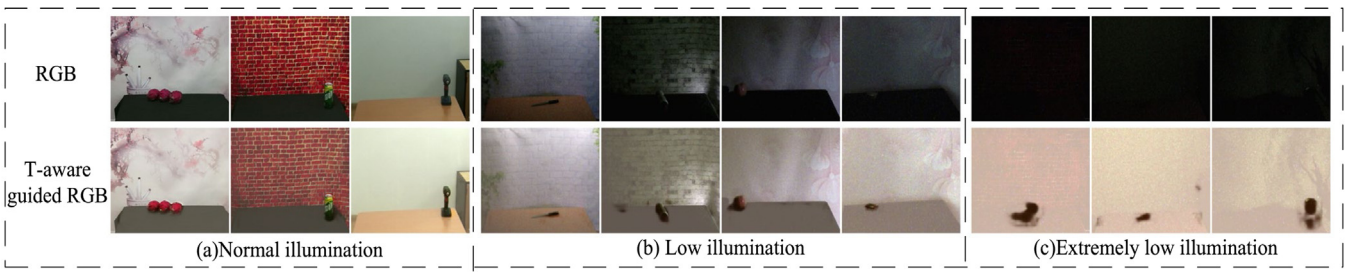


Fig. 3. Comparison of RGB images with different illumination levels before and after T-aware guided module.

compared to other methods. Different from other methods (Wang et al., 2021; Zhai et al., 2021; Wen et al., 2021), we construct a cross-modal fusion localization-remote correction module, which contains two output features to guide the decoder to generate salient maps. We first use the middle and high-level fusion features for contextual association to obtain the remote correction feature of the salient objects. Secondly, we use the remote correction feature and the high-level fusion features for deep screening operation to obtain the location information of the salient object. The decoding module will gradually recover the details of salient objects based on this location information, and the remote correction feature will further remove redundant information from the multi-scale saliency map.

In summary, the main contributions of this work can be summarized as follows:

(1) We constructed an RGB-T salient object detection model for cross-illumination with a new idea. It can take full advantage of the compensation property of T images for low-illumination RGB images. More importantly, our method can improve the detection performance on low-illumination data without expanding new low-illumination datasets.

(2) The proposed T-aware guided mechanism can effectively mine the shallow information in T images and select favorable information to complement the low-illumination RGB images.

(3) The proposed cross-modal fusion localization-remote correction module can effectively utilize mid-level and high-level features. This architecture not only provides accurate localization information for the decoder but also provides correction information in the process of salient maps from coarse to fine.

(4) We divided the VDT-2048 dataset into a training set for normal illumination and a test set for low and extremely low illumination. The latest fourteen state-of-the-art methods are compared on the VDT-2048 dataset, and the proposed method achieves superior performance. Furthermore, we also obtained favorable results on generalizability experiments in three RGB-T datasets.

2. Related work

In this section, we briefly review the existing RGB saliency object detection methods. Secondly, we provide a detailed overview of RGB-D and RGB-T salient object detection methods and present the motivation for our work.

2.1. RGB salient object detection

Salient object detection has been developed over decades, and many results have been achieved. Due to space limitations, the specific results can be viewed in the article summarized by Liu et al. (2021a). Since this year, many new methods have been proposed in the field of RGB-based Salient object detection. Song et al. (2022a) proposed a cascaded detail and backbone filling method to fill the salient subject by capturing the edges of the salient object. Wu et al. (2022) constructed an Extremely-Downsampled Network that used a shallow sampling technique to learn the global view of the entire images efficiently. Yan et al. (2022) used an unsupervised domain adaptive method based on uncertainty-aware pseudo label learning for Salient object detection, which adapts between these two domains by self-training of uncertainty perception and achieves excellent detection results. Liu et al. (2022a) constructed a pooling network (PoolNet+), which explores the potential of pooling techniques for salient object detection tasks. Zhang et al. (2022b) proposed a new progressive dual-attention residual network (PDRNet), which uses two complementary attention maps to guide residual learning, thus progressively refining the prediction in a coarse-to-fine manner. Zhuge et al. (2022) designed a new integrity cognitive network (ICONet), which defines the concept of integrity at the micro and macro levels.

2.2. RGB-D and RGB-T salient object detection

(1) RGB-D salient object detection started late compared to RGB salient object detection. However, it has been developed for several years and has achieved many results. These results can be found in the articles on RGB-D saliency detection compiled by Fan et al. (2021) and Zhou et al. (2021a). In the past two years, many new methods have emerged in the field of RGB-D salient object detection. Zhai et al. (2021) proposed a branching backbone strategy (BBSNet) that regroups multi-layer features into teacher and student features for detection. In the feature extraction stage, it adds the depth features to the RGB images for further extraction, which improves the robustness and generalization of the model. Chen et al. (2021b) used a three-dimensional convolutional neural network (RD3D) to pre-fuse RGB and depth modalities by an inflated 3D encoder, and then they created a 3D decoder enriched with inverse projection paths for decoding. Li et al. (2021) designed a Hierarchical Alternate Interaction Network (HAINet), the method that mitigates interference in the depth images and highlights salient objects in the RGB images. Ji et al. (2021) constructed a depth correction fusion strategy (DCF), which used RGB images to correct low-quality depth images, and designed a fusion module to fuse the RGB images and fixed depth images. Wen et al. (2021) proposed a dynamic selection network (DSNet), which uses a dynamic selection module (DSM) to dynamically mine cross-modal complementary information between RGB images and depth images. Zhang et al. (2021a) designed a cross-modal differential interaction network (CDINet), which models the dependency differences between two modalities based on different levels of feature representation. Liu et al. (2021b) used a Triplet Transformer Embedding Network (TriTransNet), which enhances them by learning remote dependencies across layers. Zhou et al. (2021b) proposed a specificity-preserving Salient object detection method (SPNet) to improve the performance of SOD by exploring shared information and morphology-specific attributes such as specificity. Sun et al. (2021) proposed a depth-sensitive RGB feature modeling solution (DSA2F), which achieves RGB feature enhancement and background interference reduction by capturing a depth geometric prior. Wang et al. (2022a) developed a simple yet effective network (DepthNet) to learn discriminative cross-modal features. Cheng et al. (2022b) constructed a depth-induced gap reduction network (DIGR), which is used to evaluate the depth quality and

reweight the contribution of unimodal features. Fang et al. (2022) designed a new group transformer network (GroupTransNet), which is skillful at learning long-term dependencies of cross-layer features to promote perfect feature expression. Zhu et al. (2022) developed a depth-supervised fusion transformer [DFTR], which uses depth information as supervision rather than an input. Zeng and Kwong (2022) proposed a dual Swin-transformer based mutual interactive network (DTMINet), which applies an attention-based module to enhance the features of each modality. Feng et al. (2022) designed a dual-stream depth interleaved encoder network (EDI) to extract RGB and depth information and realize their mixing simultaneously. The model obtained excellent running speed during the testing stage.

(2) In recent years, RGB-T salient object detection has developed rapidly and achieved many results due to the illumination-independent nature of T images. Here we present several deep learning-based RGB-T SOD methods, and other methods can be referred to the information compiled by Zhou et al. (2022b). Tu et al. (2020a) proposed a deep attention fusion method (ADFNet) to obtain weighted features and provided an RGB-T dataset. Zhang et al. (2020) proposed a new end-to-end network (FML) for multimodal salient object detection that takes advantage of the complementary strengths of RGB and T images. Tu et al. (2021) designed a multi-interaction double decoding method (MIDD) to mine and model multiple types of interactions. Gao et al. (2022) developed a unified information fusion network (MMNet) that can efficiently handle multimodal features. Zhou et al. (2022a) used an efficient, consistent feature fusion network (ECFFNet) to utilize the complementary information of dual-modality fully. Guo et al. (2021) proposed a two-stage fusion approach (TSFNet) to capture the features of RGB and T images fully. Zhang et al. (2021c) proposed a novel deep feature fusion network (RFF). They exploit the robustness of T images to illumination and shading. Huo et al. (2022a) built an efficient context-guided superposition refinement network (CSR-Net), which focuses on efficiency while taking into account detection performance. Wang et al. (2021) investigated a cross-guided fusion network (CGFNet) that profoundly explores the characteristics of the respective modalities through the interaction of each module. Zhou et al. (2022b) proposed an adversarial learning-assisted and perceptual importance fusion network (APNet) that can be used for salient object detection throughout the day. Zhou et applied for all-day salient object detection. Liu et al. (2022c) constructed a method to drive edge perception (SwinNet), which uses the powerful feature representation capability of Swin Transformer to guide cross-modal fusion with edge features. Chen et al. (2022) designed a cross-guided modal difference reduction network (CGMDRNet). They obtained consistent fusion by reducing the difference between RGB images and T images. Tu et al. (2022) proposed a new deep correlation network (DCNet) explores the correlation between RGB images and T images, which solves the problem that RGB and T images need to be aligned. Huo et al. (2022b) used a real-time One-Stream and Guided Refinement Network (OSRNet), which avoids the cumbersome dual-stream decoding structure by early fusion. Liao et al. (2022) constructed a cross-collaborative fusion coding network (CCFENet), which suppressed negative information between modalities by facilitating the interaction between encoders. Xu et al. (2022) proposed a strategy to process different cues in RGB and T images via the CNN feature and resultant salient map fusion. Liang et al. (2022a) proposed an end-to-end framework (MIA_DPD) which has excellent generality in RGB-D and RGB-T detection tasks. Wang et al. (2022b) designed a unidirectional RGB-T salient object detection network with intertwined driving of encoding and fusion, which makes the network more concise and effective. He et al. (2022) proposed an Enhancement and Aggregation–Feedback Network (EAF-Net) for SOD to achieve effective complementation between modalities and prevent interference from noises. Zhang et al. (2022a) designed a novel RGB-T SOD model that alleviates meaningless cross-modal fusion by leveraging a modality-aware and scale-aware feature fusion module. Ma et al. (2022) developed a novel Modal Complementary Fusion Network (MCFNet) to alleviate the contamination effect of low-quality images from both global and local perspectives. Jiang et al. (2022) proposed a novel mirror complementary Transformer network (MCNet) for RGB-T SOD to effectively extract hierarchical features of RGB and thermal images. Cong et al. (2022) used a network named TNet to solve the RGB-T SOD task and introduced a global illumination estimation module to predict the global illuminance score of the image so as to regulate the role played by the two modalities. Bi et al. (2022) used a parallel symmetric network for mining the complementary information of RGB images and T images. Liang et al. (2022b) proposed a method that can be used for RGB-T and RGB-D saliency detection tasks by exploring bimodal information.

2.3. Motivation

After the above discussion and the analysis of Table 1, most SOD methods focus on the complementarity between multimodal information and work to develop fusion solutions with more generalization and robustness. Therefore, they still have some limitations when applied to low-illumination SOD tasks. (1) The performance of SOTA methods trained on RGB-T datasets degrades sharply when detecting extremely low-illumination data. (2) In terms of fusion strategies, they mainly focus on the development of mid and late-stage fusion. However, early fusion is more effective in processing low-illumination data. (3) The SOTA methods do not develop effective fusion modules for cross-illumination strategies in the decoding phase. Although T-images provide important information for SOD, T-images are more suitable for low-illumination detection scenes.

Considering the above problems and based on the existing dataset conditions, we proposed a T images-aware guided early fusion network for cross-illumination RGB-T salient object detection.

3. Methodology

In this section, we first describe the overall architecture of our method in detail. Second, we describe the T-aware guided mechanism in particular, then we introduce the cross-modal fusion localization-remote correction module. Finally, we provide the implementation details of the method.

3.1. Architecture overview

In this paper, the overall architecture of our proposed method is shown in Fig. 4. It differs from the classical encoder–decoder structure. In the early fusion stage, we added a T-aware guided module for guiding RGB images in low-illumination scenes. Secondly, the encoder is a symmetric two-stream backbone network by VGG16. It is used to extract multi-layer features of RGB and T images. Notably, we discard the last pooling layer and the fully connected layer and keep only five convolutional blocks. They are down sampled 1, 2, 4, 8, and 16 times, and the number of channels is 64, 128, 256, 512, and 512, respectively. In the decoding stage, we designed a cross-modal fusion localization-remote correction module for effective salient object detection.

Precisely, we randomly adjust the brightness and contrast of the RGB images under normal illumination when the data are loaded. Then, these RGB images (R) are first re-assigned weights by the T-aware guided module before being input into the model. The T images (T) are directly input to the network. For the selection of the backbone network, we use a two-stream encoder VGG16 for the feature extraction network, and the extracted features for each layer are $\{f_{RGB}^{352_{64}}, f_{RGB}^{176_{128}}, f_{RGB}^{88_{256}}, f_{RGB}^{44_{512}}, f_{RGB}^{22_{512}}\} \in R_i, i \in [1,2,3,4,5]$ and $\{f_{T}^{352_{64}}, f_{T}^{176_{128}}, f_{T}^{88_{256}}, f_{T}^{44_{512}}, f_{T}^{22_{512}}\} \in T_i, i \in [1,2,3,4,5]$. After that, we build a practical decoding framework. It mainly consists of a cross-modal fusion localization-remote correction module and a decoder. The cross-modal fusion localization-remote correction module can remove the redundant information in the decoder due to cross-illumination. After processing by these modules, our method can accurately detect low and extremely low-illumination data.

Table 1
Discussion of the relevant methods used in the comparison experiments.

NO.	Year	Method	Type	Pub.	Training set	Backbone	Discussions
1	2020	JL-DCF (Fu et al., 2020)	RGB-D	CVPR	NJU2K (1.5K), NLPR (0.7K)	ResNet-101 VGG-16	Encoded using only one backbone through early fusion.
2	2020	D^3Net (Fan et al., 2021)	RGB-D	TNNLS	NJU2K (1.485K), NLPR (0.7K)	VGG-16	Trained three branch networks to output excellent detection results in the testing phase.
3	2020	BBSNet (Zhai et al., 2021)	RGB-D	ECCV	NJU2K (1.4K), NLPR (0.65K)	ResNet-50 VGG-16 VGG-19	Extracted top-level teacher features to guide the fusion of bottom-level student features.
4	2021	RD3D (Chen et al., 2021b)	RGB-D	AAAI	NJU2K (1.485K), NLPR (0.7K)	3D ResNet-50	Designed the encoder–decoder of 3D convolutional blocks for SOD tasks.
5	2021	HAINet (Li et al., 2021)	RGB-D	TIP	NJU2K (1.4K), NLPR (0.65K)	VGG-16	
6	2021	DSNet (Wen et al., 2021)	RGB-D	TIP	NJU2K (1.485K), NLPR (0.7K)	ResNet-50	Proposed different strategies to mine cross-modal complementary information between RGB images and depth images.
7	2021	CDINet (Zhang et al., 2021a)	RGB-D	ACMM	NJU2K (1.485K), NLPR (0.7K)	VGG-16	
8	2021	DCF (Ji et al., 2021)	RGB-D	CVPR	NJU2K (1.485K), NLPR (0.7K)	ResNet-50 VGG-16	Corrected low quality depth to reduce interference with the model.
9	2022	SPNet (Zhou et al., 2021b)	RGB-D	CVPR	NJU2K (1.485K), NLPR (0.7K)	ResNet-50	Explored contextual information to improve saliency detection results.
10	2022	ADFNet (Tu et al., 2020a)	RGB-T	TMM	VT5000 (2.5K)	VGG-16	Provided a deep fusion method and large-scale RGBT dataset.
11	2021	MIDD (Tu et al., 2021)	RGB-T	TIP	VT5000 (2.5K)	VGG-16	Developed an excellent fusion strategy to cross-fuse RGB features and T features in the decoding stage.
12	2021	CGFNet (Wang et al., 2021)	RGB-T	TCSVT	VT5000 (2.5K)	VGG-16	

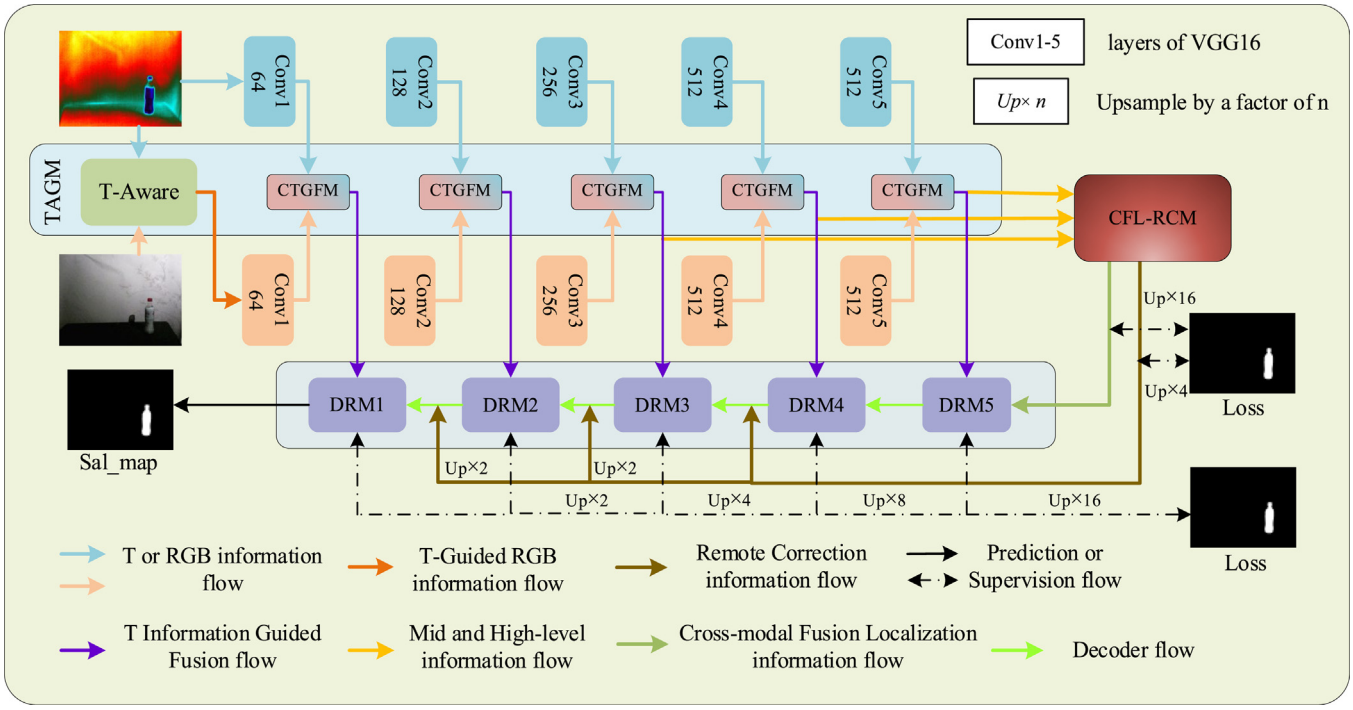


Fig. 4. The overall structure of the proposed method.

3.2. T-aware guided mechanism

(1) T-aware guided module. RGB images are rich in texture information, which plays a crucial role in recovering the details of remarkable objects during the decoding stage. However, in natural scenes, a sudden reduction or disappearance of illumination will cause a dramatic degradation in the imaging quality of RGB images. This will prevent the model from accurately extracting RGB images information. To address the drastic impact of cross-illumination on the detection performance, we tried to compensate RGB information by using T information. For this reason, we design a T-aware guided module to obtain the possible regions of salient objects in T images. This region is used to perform an initial screening of information in the RGB images. It will reduce the sensitivity of the encoder and decoder to the cross-illumination of the RGB images. Specifically, the imaging quality of T images depends mainly on the thermal factor, so it has excellent stability in coping with the environment with changing illumination. Based on the property, we apply a piece of T-aware information to RGB for guiding before the RGB images enter the encoding network. This module reduces the effect of low-illumination images on the features extracted by the backbone network. In addition, considering that the imaging quality of T images is affected by temperature, we add a confidence level of T-aware information. As shown in Fig. 5, we first input the T images to a lightweight T-aware module consisting of base convolution, and

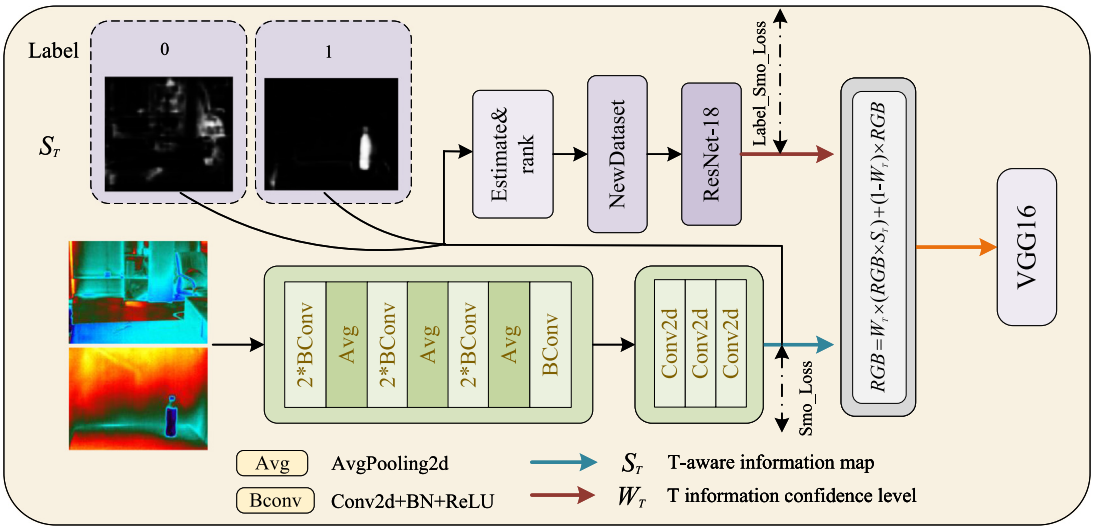


Fig. 5. T-information-aware guidance module (T-aware).

supervise the training of the T images with the ground truth. We do not overly pursue detection accuracy during the training process, aiming to obtain T-aware information with robustness. Here we use the smooth loss for supervision to ensure its robustness. Second, to obtain the confidence level of T-aware information, we first binarize this T-aware information. Then we calculate the MAE of the predicted outcome and ground truth. We then ranked each T-awareness infographic according to its MAE from smallest to largest. We take the top 30% as the high-quality T-aware information map and the bottom 30% as the inferior T-aware information map. We send the T images of these T-aware information maps to the classification network ResNet-18 for training. It is worth noting that we use label smoothing loss for supervision to prevent the classification network from being overconfident and avoid extreme training results.

Before the RGB images enter the network training, the RGB information does not rely entirely on the potentially unreliable T-aware information. Instead, we assign the T-aware information and its confidence level to the original RGB images. The specific formulas are as follows:

$$RGB = W_T \times (RGB \times S_T) + (1 - W_T) \times RGB, \tag{1}$$

where W_T represents the confidence level of T-aware information and S_T represents the T-aware information.

(2) Cross-Modal T Guided Fusion Module. This is a vital issue in the multimodal salient object detection task to mine the complementarity between cross-modal information. Considering that we are detecting salient targets under cross-illumination conditions, as shown in Fig. 6, we design a cross-modal T guided fusion module. Different from CBAM (Woo et al., 2018), the module takes advantage of the property that T images are not affected by illumination changes to explore the relevant information in RGB and T images. Meanwhile, considering the contribution of RGB information, the module not only fully integrates T information and RGB information but also highlights salient object at the spatial level.

As shown in Fig. 6, R_i and T_i respectively represent the output of RGB and T images in the ith ($i = 1...5$) convolutional block. Separately, each set of cross-modal features R_i and T_i are fed into CTGFM for cross-modal processing. Specifically, as shown in Fig. 6, for each CTGFM, we first let R_i pass the global average pooling, and then let T_i pass the global maximum pooling. Then let these two pooled features be connected and processed through a convolution block. Finally, the features output from the convolution block are mapped to 0–1. The specific formulas are as follows:

$$\begin{aligned} \Phi_i &= S(Conv(Cat(Avg(R_i), Max(T_i)))), \\ & i = [1, 2, 3, 4, 5] \end{aligned} \tag{2}$$

where Φ_i represents the weight on the spatial level of the T information master guide, S represents the sigmoid, Avg represents the global average pooling, and Max represents the global maximum pooling. In this way, after a processing method that is guided mainly by T images and retains the prominent parts of RGB images, it can provide a piece of guiding information for further mining useful features between modalities subsequently. Further, Φ_i performs element multiplication and addition operations with R_i and T_i, respectively. Then each feature goes through the global average pooling layer and then multiplies the elements with R_i and T_i separately, and finally adds them together and outputs Fus_i:

$$\begin{aligned} Fus_i = {} & S(Avg(R_i + \Phi_i \times R_i) \times (R_i + \Phi_i \times R_i)) \\ & + S(Avg(T_i + \Phi_i \times T_i) \times (T_i + \Phi_i \times T_i)) \end{aligned} \tag{3}$$

CTGFM not only fully extracts the deep-level features of both modalities but also reduces the interference of low-illumination RGB images on cross-modal feature extraction. T information is used as the primary guide and RGB information is used as the secondary guide in feature fusion, which improves the robustness of the whole model. It is noteworthy that Fus_i is unchanged from R_i and T_i before entering CTGFM in terms of channel number and spatial scale.

3.3. Cross-modal fusion localization-remote correction module

High-level features contain rich semantic information, and the effective extraction of these high-level semantic information plays an essential role in the accurately locating of salient objects. The mid-level features include both semantic and detail information, and this part of detail information also plays a vital role in the portrayal of salient objects. To make full use of the middle and high-level information of RGB and T images, we try to build a module containing both location information and remote correction information. We propose an efficient cross-modal fusion localization-remote correction module (CFL-RCM). Different from the fusion approach of Zhai et al. (2021), the module we designed not only provides information on the exact position of the salient object but also corrects the output of the latter three decoders. Specifically, the first part is the remote correction module, as shown in Fig. 7, the input of this module is the feature map Fus_i, $i = [3,4,5]$. We use a global context module GCM (Zhai et al., 2021), which has the advantages of being both effective and efficient. After Fus_i is processed by this module, we get G_Fus_i, $i = [3,4,5]$ with contextual information, and all of G_Fus_i is unified into 64 channels. Then we refine G_Fus_i from the bottom-up level by level. The difference from Zhai et al. (2021) is

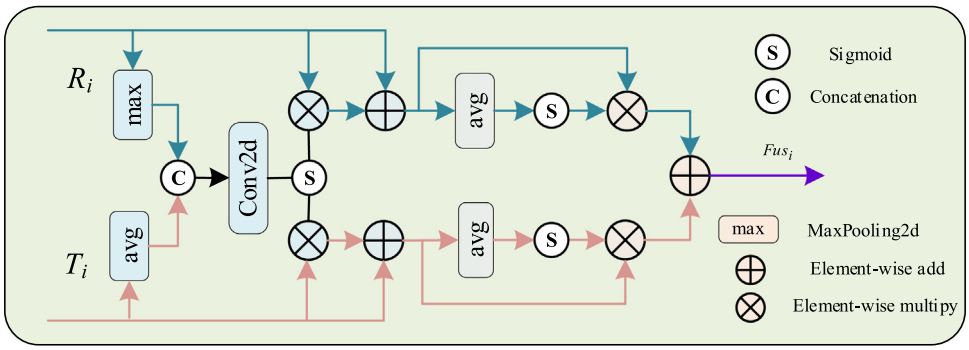


Fig. 6. Cross-modal T-information guided fusion module (CTGFM).

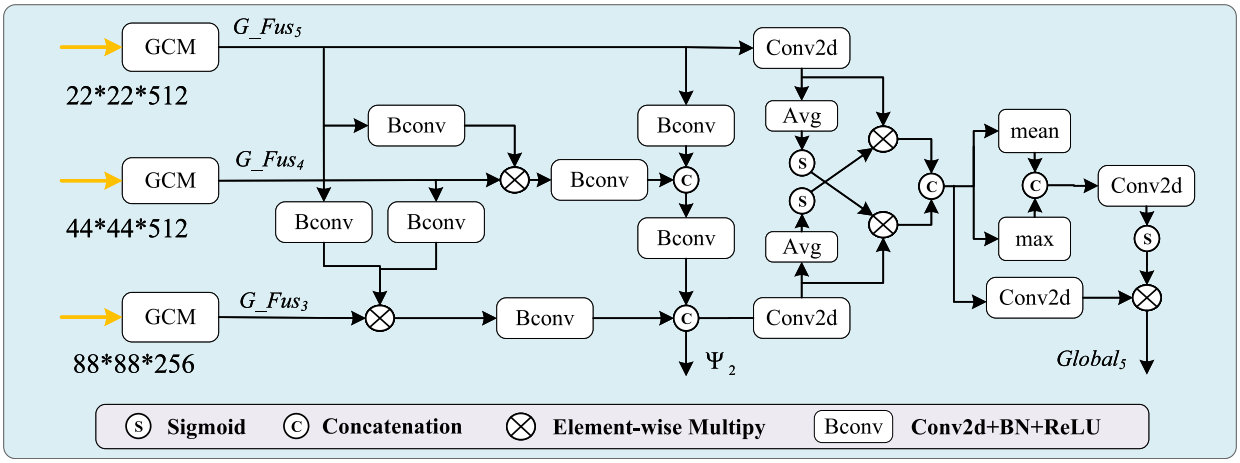


Fig. 7. Cross-modal fusion localization-remote correction module (CFL-RCM).

that we add CBR for each element multiplication and concatenation operation, as follows:

$$\Psi_1 = Cat(CBR(F_{up}(CBR(G\_Fus_5)) \times G\_Fus_4), CBR(G\_Fus_5)) \tag{4}$$

$$\Psi_2 = Cat(CBR(G\_Fus_3 \times F_{up}(CBR(G\_Fus_4) \times F_{up}(CBR(G\_Fus_5)))), CBR(\Psi_1)) \tag{5}$$

where the CBR represents the 3*3 Conv2d+BN+ReLU, F_{up} represents the up sampling, Ψ_1 represents the refined features of G_Fus_5 and G_Fus_4, and Ψ_2 represents the refined remote correction feature of G_Fus_3 and Ψ_1. The size of Ψ_2 is 88*88 and the number of channels is 64.

The second part is the cross-modal fusion localization module, as shown in Fig. 7, this part uses the integrated high-level feature G_Fus_5 and the remotely corrected feature Ψ_2 to perform the cross-modal fusion. In this way, we obtain the accurate localization information of the salient objects. Since the scale of Ψ_2 is 88*88, we have to down sample it twice, and fuse it with a 3*3 convolutional block G_Fus_5. The specific formulas are as follows:

$$\Pi_1 = Conv2d(Avg(Conv2d(Avg(\Psi_2)))) \tag{6}$$

$$\Pi_2 = Conv2d(G\_Fus_5) \tag{7}$$

$$Z = Cat(S(Avg(\Pi_2)) \times \Pi_1, S(Avg(\Pi_1)) \times \Pi_2) \tag{8}$$

$$Globel_5 = Conv2d(Z) \times S(Conv2d(Cat(Avg(Z), Max(Z)))) \tag{9}$$

where Π_1 and Π_2 represent the inputs of cross-modal fusion localization module, where Z represents the fusion feature of G_Fus_5 and Ψ_2, where $Global_5$ represents a feature of size 22*22 and channel number 1. $Global_5$ is used as the first global feature of the decoder to guide the refinement of salient object.

The following global information is based on the per-layer decoder output. This global information can guide each layer to fuse feature $\{Fus_4, Fus_3, Fus_2, Fus_1\}$ to enrich the detailed information of salient objects. In this process, detailed information is continuously added, which contains primarily favorable information and part of redundant information. As the depth of the decoding block DRM (Zhang et al., 2021b) output $Global_i$ increases from level to level, the redundant information will also continue to accumulate. The situation will lead to the deviation of the predicted results from the actual results. Therefore, to solve this problem, we use Ψ_2 to remotely correct $Global_i$, i = [0,1,2] step by step. This operation can highlight the salient features intersected by $Global_i$ and Ψ_2, and weaken the parts separated by $Global_i$ and Ψ_2. In our model, we mainly correct the global features input by the last three decoding modules. In this way the fine-tuning process is able to proceed in the expected direction.

3.4. Decoder and loss

In the step-by-step decoding stage, we handle it using $\{Fus_5, Fus_4, Fus_3, Fus_2, Fus_1\}$, $Global_5$ and Ψ_2. The specific formulas are as follows:

$$Global_i = Decoder(Fus_{i+1}, Global_{i+1} + \alpha \times \Psi_2 + \beta \times F_{up2}(\Psi_2) + \lambda \times F_{up4}(\Psi_2)), \quad i \in [0,1,2,3,4] \tag{10}$$

where $\{Global_4, Global_3, Global_2, Global_1, Global_0\}$ represents the output of each decoder layer, where $Global_0$ is our final prediction map; F_{up2} and F_{up4} represents the 2-fold and 4-fold up sampling, respectively. α, β and λ take the value of 0 in the normal case, when i = 2, α = 1,

Fig. 8. Visualization results for normal illumination, low illumination and extremely low illumination.

when i = 1, $\beta = 1$, when i = 0, $\lambda = 1$. Different from Fan et al. (2021), we add to the global features $\{Global_3, Global_2, Global_1\}$ of the last three decoding modules for corrective processing. This can further filter out the redundant information generated during the decoding process.

In addition, we perform supervised training on $\{Global_4, Global_3, Global_2, Global_1, Global_0\}$, $Global_5$ and Ψ_2. Because the five decoders are decoded step by step, and the feature map size is from small to large. Therefore, the feature map size is first expanded to 352*352 by bilinear interpolation before supervised training. Then, we obtained the auxiliary prediction map $\{G_4, G_3, G_2, G_1\} \in I^{1\times352\times352}$ and the final prediction map $G_0 \in I^{1\times352\times352}$ for the 2D convolutional output. It is worth noting that Ψ_2 and $Global_5$ are also up-sampled and 2D convolved to output the auxiliary prediction map $\{G_5, \Psi_2{}^*\} \in I^{1\times352\times352}$. Here we utilize cross-entropy *loss* and *IOU loss* to supervise them. The specific formulas are as follows:

$$\begin{cases} loss_1 = \eta \times (CE(GT, G_1)) + \gamma \times (CE(GT, G_1)) + \sum_{g_1}(CE(GT, g_1)), \\ loss_2 = \kappa \times (CE(GT, G_5)) + \lambda \times (CE(GT, \Psi_2{}^*)) \\ + \sum_{g_2}(CE(GT, g_2) + IOU(GT, g_2)), \\ g_1 \in [G_2, G_3, G_4, G_5, \Psi_2{}^*], g_2 \in [G_0, G_1, G_2, G_3, G_4] \end{cases} \tag{11}$$

where *CE* and *IOU* represent cross entropy loss and IOU *loss,* respectively. *GT* represents the ground truth of the image. Where $loss_1$ and $loss_2$ represent two different supervision strategies. Before the 45 epochs, we used the strategy of $loss_1$ to supervise. We apply different weights to the features at different levels. Here η and γ are set to 1.5 and 1.25, respectively. During the initial training, the middle and high-level information is inaccurate and the size of the middle and high-level feature maps is small. When performing bilinear interpolation up sampling, the error will be continuously amplified. These errors take up most of the weight of the overall loss and reduce the accuracy of the final prediction map. After 45 epochs, we use the strategy of $loss_2$ to supervise. Because after the model is trained by the supervision of $loss_1$, each auxiliary prediction map tends to be stable. At this time, we add *IOU loss* to supervise, which can improve the accuracy of the final prediction map. Different from Wang et al. (2021), we apply higher weights to the auxiliary prediction images G_5 and $\Psi_2{}^*$ of middle and high-level features for supervision. Here κ and λ are set to 1.5 and 2, respectively, to fine-tune the final results. In addition, the auxiliary prediction maps G_5 and $\Psi_2{}^*$ contain limited detailed information, so we do not use *IOU loss* supervision.

4. Experiments

In this section, we elaborated on the implementation details and dataset, and then described the evaluation metrics. Next, we quantitatively and qualitatively compared the proposed method with state-of-the-art RGB-T salient object detection methods to demonstrate the advantages of the proposed method for cross-illumination salient object detection. After that, we did generalizability experiments on the commonly used RGB-T dataset. In addition, we performed an ablation study to validate the role of each module in our model. Finally, we discussed some failure cases.

4.1. Experimental setup

(1) Implementation Details: Our method is based on the PyTorch framework, the device system we used is Ubuntu 18.04, and all experiments were conducted on an NVIDIA RTX2070super. Our method is trained for 65 epochs, with batch size set to 4, optimizer using SGD, initial weight decay set to 5e–4, momentum set to 0.9, and the learning rate is set to 0.001, and at 35 epochs, the learning rate decays to 1/10 of the original. The dataset and code are available at: https://github.com/VDT-2048/TAGFNet.

(2) Datasets: The existing VT5000, VT1000, and VT821 datasets contain only a tiny amount of low-illumination data, which is insufficient for detecting extremely low-illumination scenes. Therefore, we use the latest publicly available datasets VDT-2048 to validate the performance of our method. As shown in Fig. 8, we divided the VDT2048 dataset into three categories: normal illumination, low-illumination, and extremely low-illumination. Among them, **normal illumination** means that the RGB images are well illuminated, and significant objects are clearly visible. **Low-illumination** means that the RGB images have most dark parts, and the salient objects are only partially missing. **Extremely low-illumination** means that the RGB images are almost dark and the outlines of the salient objects are almost entirely lost. According to this classification standard, as shown in Fig. 9, we divided the data set into 1210 sets of normal images as the training set and 838 sets of low-illumination and extremely low-illumination images as the test set. Among them, there are 438 sets of low-illumination and 400 sets of extremely low-illumination images.

(3) Evaluation Metrics: The significance evaluation metrics can objectively describe the performance of the method and can be fairly compared with other RGB-T and RGB-D SOD methods. At present, existing SOD methods are not trained and tested using the VDT-2048 datasets. Therefore, to ensure the fairness of the comparison method, we do not use their pre-trained models for testing, we use their publicly available source code for training and testing. Specifically, all comparison experiments were conducted using a training set containing 1210 images and a test set of 838 images. The parameter settings of the comparison experiment are consistent with the original paper. We use six metrics to evaluate these methods, and the evaluation metrics include: MAE (Perazzi et al., 2012), maximum F-measure, average F-measure (Achanta et al., 2009), weighted F-measure (Margolin et al., 2014), S-measure (Fan et al., 2017), E- measure (Fan et al., 2018).

Table 2
Quantitative comparison results of different model methods, red represents the best, blue represents the second best.

Models	VDT2048					
	MAE↓	F_{max}↑	F_m↑	W_F↑	S_m↑	E_m↑
$BBSNet_{2020}$[32]	0.0053	0.8234	0.7039	0.7572	0.8407	0.8905
$JL\text{-}DCF_{2020}$[7]	0.0064	0.8169	0.6159	0.7306	0.8500	0.8159
D^3Net_{2021}[31]	0.0552	0.4128	0.2419	0.2306	0.6309	0.4995
$HAINet_{2021}$[34]	0.0078	0.7641	0.7131	0.6784	0.7960	0.8989
DCF_{2021}[35]	0.0064	0.8024	0.7691	0.6512	0.7677	0.9310
$RD3D_{2021}$[33]	0.0114	0.6999	0.4937	0.5555	0.7763	0.7254
$DSNet_{2021}$[36]	0.0076	0.7920	0.7136	0.6691	0.7806	0.8968
$CDINet_{2021}$[37]	0.0083	0.7665	0.6168	0.6830	0.8222	0.8202
$SPNet_{2022}$[39]	0.0049	0.8332	0.7628	0.7922	0.8453	0.9408
$ADFNet_{2020}$[4]	0.0297	0.67799	0.2478	0.3328	0.7227	0.4878
$MIDD_{2021}$[10]	0.0064	0.8153	0.6421	0.7051	0.8226	0.8371
$CGFNet_{2021}$[16]	0.0057	0.8201	0.6952	0.7430	0.8281	0.8847
$TriTransNet_{2021}$[38]	0.0067	0.7819	0.7554	0.6966	0.7777	0.9034
$SwinNet_{2022}$[18]	0.0054	0.8311	0.7026	0.7696	0.8466	0.884
OURS	0.0042	0.8534	0.8020	0.8210	0.8567	0.9606

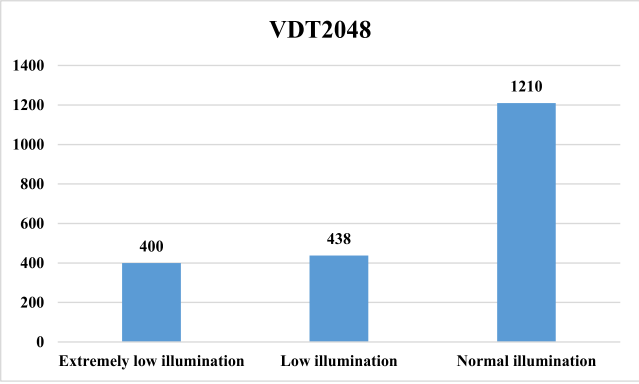


Fig. 9. Distribution of normal illumination, low illumination and extremely low illumination in the VDT-2048 dataset.

4.2. Comparison with the SOTA RGB-T methods

As shown in Table 2, we validate the robustness and generality of our proposed method in dealing with cross-illumination by comparing twelve state-of-the-art RGB-T and RGB-D methods. Among these methods, SPNet (Zhou et al., 2021b), TriTransNet (Liu et al., 2021b), CDINet (Zhang et al., 2021a), DSNet (Wen et al., 2021), DCF (Ji et al., 2021), HAINet (Li et al., 2021), RD3D (Chen et al., 2021b), D3Net (Fan et al., 2021), BBSNet (Zhai et al., 2021), and JL-DCF (Fu et al., 2020) are RGB-D saliency detection methods. The rest are RGB-T salient object detection methods, including SwinNet (Liu et al., 2022c), CGFNet (Wang et al., 2021), MIDD (Tu et al., 2021), ADFNet (Tu et al., 2020a). All of these methods are based on deep learning methods. It is worth noting that SwinNet (Liu et al., 2022c) and TriTransNet (Liu et al., 2021b) are SOTA backbone-based methods.

Quantitative metrics show that our proposed method achieves the most effective performance in coping with cross-illumination data. And it is also well demonstrated that our algorithm has greater adaptability than other algorithms. In the 838-group low-illumination test set of VDT-2048, our method has a 2.02% improvement in max_F, 3.72% improvement in average F-measure, 2.98% improvement in weighted F-measure, and 1.98% improvement in S-measure compared with other VGG16, ResNet-50, and ResNet-101 based methods. In addition, TriTransNet (Liu et al., 2021b) and SwinNet (Liu et al., 2022c) used Transformer and Swin-Transformer as the backbone, which have stronger feature extraction capability. We also achieved the best performance compared with Tri-TransNet and SwinNet.

As shown in Fig. 10 PR curves, our method covers all the compared methods. And our method achieves a pretty competitive lead under different thresholds of F-measure. As shown in Fig. 11. we provide a visualization of the results for comparison. It is evident that our proposed method can still guarantee the accurate detection of salient objects under some extreme conditions. The method can still detect salient objects and retain good edge information when dealing with challenging scenes, such as RGB images with extremely dark illumination, RGB images with a lot of noise, T images with thermal crossover, T images with thermal reflection, and T images with unclear boundaries.

4.3. Comparison with RGB-T methods in cross-illumination

Fig. 12, it shows the visualization results of our method and other methods in the process of illumination variation. Among them, rows 1 to 4 are the results under normal illumination in the VT5000 test set. We can see that our method and the other methods can detect significant objects usually. Rows 5 to 8 are the results under low-illumination in the VDT2048 test set. We can see that most models can still detect the salient objects, but the details are not well recovered. For example, CGFNet (Liao et al., 2022), HAINet (Li et al., 2021), MIDD (Tu et al., 2021), and SwinNet (Liu et al., 2022c) in the fifth row are easily affected by the shadows of the salient objects in the RGB images. Rows 9 to 12 show the results at extremely low-illumination in the VDT2048 test set. We can see that ADFNet (Tu et al., 2020a), DCF (Ji et al., 2021), and HAINet (Li et al., 2021) can no longer detect salient objects properly. The other methods are also affected by illumination, and the detection results are also degraded to different degrees.

4.4. Comparison with RGB-T methods on RGB-T datasets

To further verify that our method can adapt not only to low-illumination detection scenarios but also to normal illumination detection scenarios, we trained and tested on the commonly used RGB-T datasets. We strictly follow the mainstream RGB-T salient object detection methods for the training setup. To fully validate the effectiveness

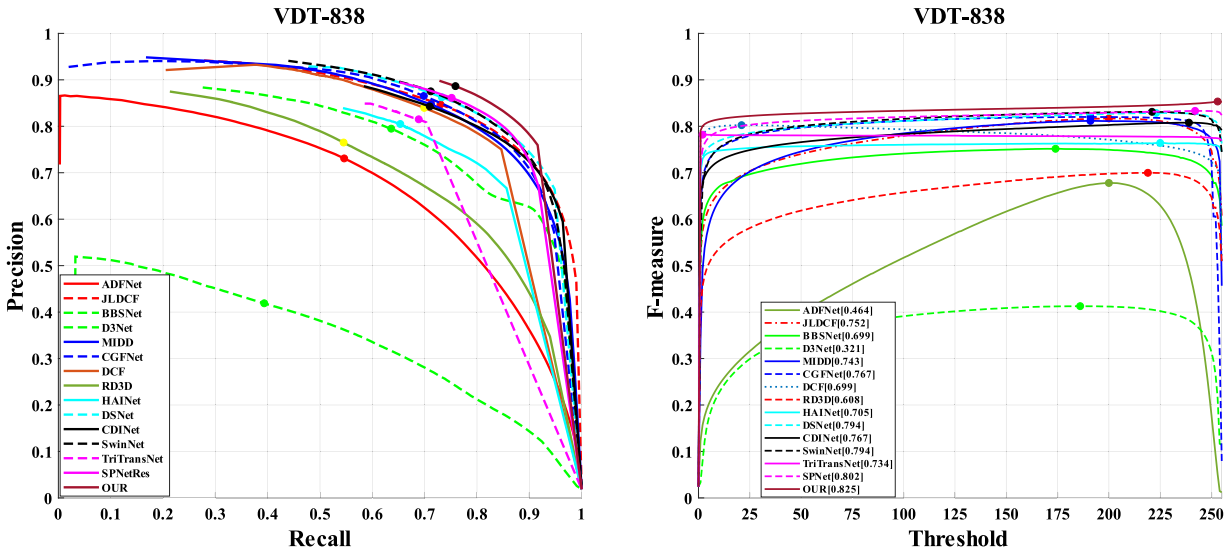


Fig. 10. Quantitative comparison results between our proposed method and the SOTA methods on the VDT-2048 dataset. The first line is Precision (vertical axis) Recall (horizontal axis) curves, and the second line shows the F-measure scores of the deep learning-based methods under different thresholds.

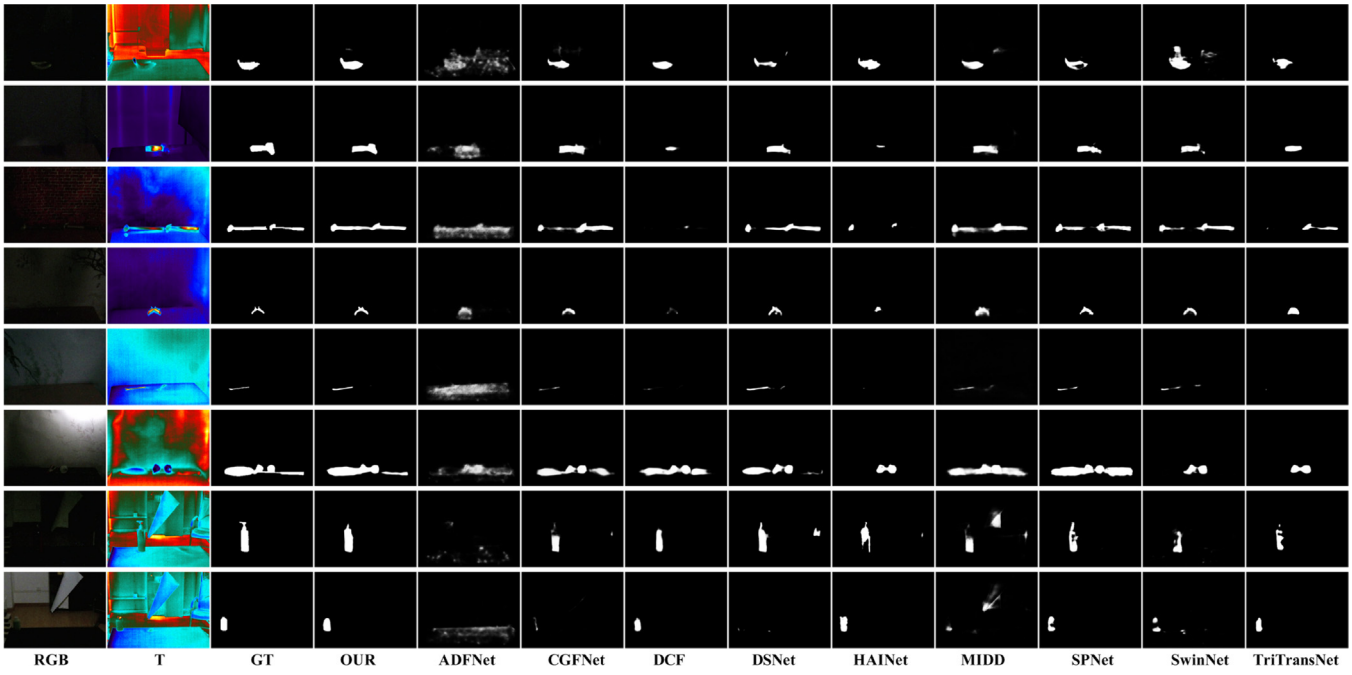


Fig. 11. Visual comparison between our method and the SOTA methods.

of our method, we tested it on three commonly used test sets, including VT821, VT1000, and VT5000.

We compared 16 the SOTA RGB-D and RGB-T salient object detection methods. JL-DCF (Fu et al., 2020), HAINet (Li et al., 2021), DCF (Ji et al., 2021) and DPANet (Chen et al., 2021a) are RGB-D saliency detection methods based on deep learning. MTMR (Wang et al., 2018b), M3S-NIR (Tu et al., 2019a), CGL (Tu et al., 2020b), LTCR (Huang et al., 2020) and MGFL (Huang et al., 2022a) are RGB-T salient object detection methods based on traditional methods, and ADFNet (Tu et al., 2020a), MIDD (Tu et al., 2021), MMNet (Gao et al., 2022), ECFFNet (Zhou et al., 2022a), TSFNet (Guo et al., 2021), CSRNet (Huo et al., 2022a), CGFNet (Wang et al., 2021) and APNet (Zhou et al., 2022b) are RGB-T salient detection methods based on deep learning methods. The specific metrics are shown in Table 3. Notably, we also plotted the corresponding RP curves and F-measure curves, as shown in Fig. 13. In summary, the results show that our method can still achieve favorable results when applied to RGB-T datasets with more normal illumination images.

4.5. Complexity analysis

Table 4 shows the difference between our method and the other methods in terms of model complexity. From train time perspective, we train on the same device with batch size set to 1 and epochs set to 30, and we can observe that our method spends less time. From the runtime perspective, our method exceeds the other methods on four datasets, where Runtime (FPS) refers to the number of images per second processed by the model during the testing stage. Different from the method proposed by Wang et al. (2021), all the methods we compared were run on the same device. Secondly, the proposed method is less complex than other methods in terms of model size, model parameters, and Flops.

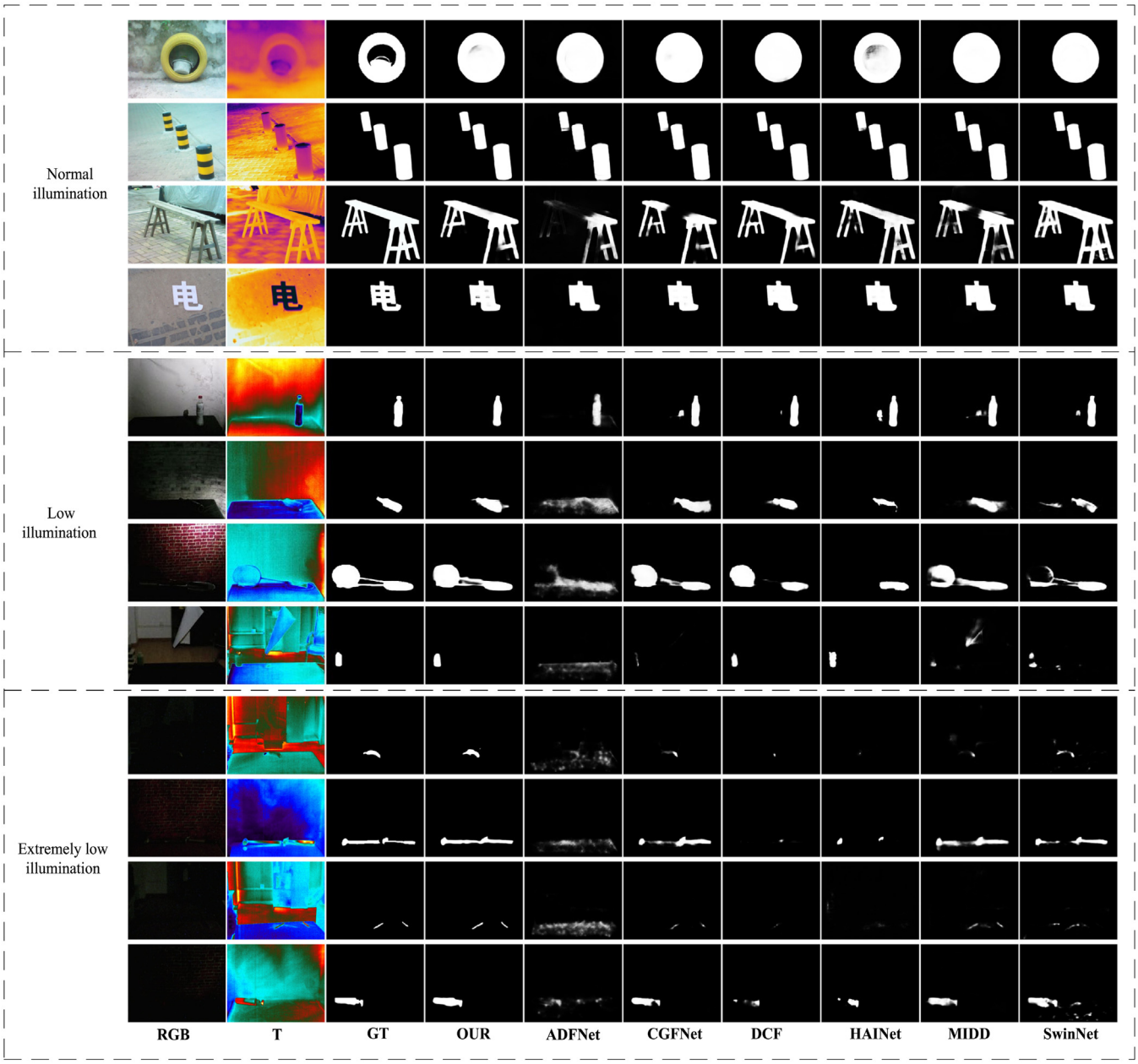


Fig. 12. Visual comparison of the proposed method with the SOTA methods at different illumination levels.

4.6. Ablation study

In this section, we mainly investigate the contribution of our proposed main module. All the ablation experiments were trained and tested based on VDT-2048, and we used six metrics for evaluation. The details are shown in Table 5.

(1) Contribution of CTGFM: As shown in the second row of Table 5, we remove the CTGFM. While we use only Element-wise summation instead for the RGB and T feature images extracted by the backbone network at all levels. The rest of the network structure remains unchanged during training and testing. Compared with the indicators in the first row of Table 5, we can see that after removing the CTGFM, each indicator has a different degree of decrease. The main purpose of this module is to improve the stability of the model, which can still maintain excellent detection results when dealing with some extremely low illumination data.

(2) Contribution of T-aware guided module: In the third row of Table 5, we remove the T-aware guided module based on (1). This part is to guide the RGB images using T information. After removing this T-aware guided module, there is no need to add alternative modules. The rest of the network structure remains unchanged during training and testing. Comparing with the metrics in the second row of Table 5, we can see that each metric decreases by about 1% after removing the T-aware guided module. As shown in the experiments, the module and CTGFM are used to solve the problems of information discrepancy arising from cross-illumination. These problems are brought about by the uneven distribution of illumination between the training and test set data.

(3) Contribution of CFL-RCM: We remove the CFL-RCM module based on (1) and (2), while we use Conv2d for the replacement, and the rest of the network structure remains unchanged. Comparing with the data in the third row of Table 5, we can see that there is a different degree of decrease in each index after removing the CFL-RCM module and a reduction of 1.61% in the S-measure. The experiments show that the contribution of the cross-modal fusion localization-remote

Table 3
Quantitative comparison results of different model methods on the RGB-T datasets, red represents the best, blue represents the second best, green represents the third best.

Models	VT5000					VT1000					VT821				
	$E_m\uparrow$	$S_m\uparrow$	$F_m\uparrow$	W_F↑	MAE↓	$E_m\uparrow$	$S_m\uparrow$	$F_m\uparrow$	W_F↑	MAE↓	$E_m\uparrow$	$S_m\uparrow$	$F_m\uparrow$	W_F↑	MAE↓
JL-DCF$_{2020}$[7]	0.865	0.860	0.743	0.733	0.050	0.907	0.916	0.837	0.838	0.028	0.846	0.848	0.732	0.711	0.063
HAINet$_{2021}$[34]	0.888	0.864	0.793	0.780	0.045	0.922	0.918	0.872	0.873	0.026	0.886	0.860	0.789	0.774	0.045
DCF$_{2021}$[35]	0.912	0.877	0.823	0.810	0.036	0.935	0.922	0.880	0.888	0.023	0.888	0.859	0.788	0.778	0.045
DPANet$_{2020}$[47]	0.836	0.813	0.688	0.640	0.070	0.881	0.881	0.781	0.767	0.051	0.735	0.736	0.574	0.548	0.157
MTMR$_{2018}$[52]	0.795	0.680	0.595	0.397	0.114	0.836	0.706	0.715	0.485	0.119	0.815	0.725	0.662	0.462	0.108
M3S-NIR$_{2019}$[53]	0.780	0.652	0.575	0.327	0.168	0.827	0.726	0.717	0.463	0.145	0.859	0.723	0.734	0.407	0.140
SGDL$_{2019}$[54]	0.824	0.750	0.672	0.558	0.089	0.856	0.787	0.764	0.652	0.090	0.847	0.765	0.730	0.583	0.085
LTCR$_{2020}$[55]	0.826	0.743	0.677	0.544	0.092	0.872	0.799	0.794	0.668	0.084	0.854	0.762	0.737	0.570	0.088
MGFL$_{2021}$[56]	0.817	0.751	0.661	0.590	0.085	0.882	0.820	0.801	0.734	0.066	0.841	0.782	0.725	0.643	0.071
ADFNet$_{2020}$[4]	0.891	0.863	0.778	0.722	0.048	0.921	0.910	0.847	0.804	0.034	0.842	0.810	0.716	0.626	0.077
MIDD$_{2021}$[10]	0.897	0.868	0.801	0.763	0.043	0.933	0.915	0.882	0.856	0.027	0.895	0.871	0.804	0.760	0.045
MMNet$_{2021}$[11]	0.896	0.850	0.796	0.783	0.045	0.926	0.905	0.874	0.876	0.032	0.919	0.883	0.843	0.832	0.033
ECFFNet$_{2021}$[12]	0.906	0.874	0.806	0.801	0.038	0.930	0.923	0.876	0.885	0.021	0.902	0.877	0.810	0.801	0.034
CSRNet$_{2021}$[15]	0.905	0.868	0.810	0.796	0.042	0.925	0.918	0.877	0.878	0.024	0.908	0.885	0.830	0.821	0.038
CGFNet$_{2021}$[16]	0.922	0.883	0.851	0.831	0.035	0.944	0.923	0.906	0.900	0.023	0.912	0.881	0.845	0.829	0.038
APNet$_{2021}$[17]	0.908	0.867	0.818	0.792	0.034	0.938	0.920	0.885	0.883	0.021	0.914	0.875	0.822	0.805	0.034
OURS	0.916	0.884	0.832	0.822	0.035	0.936	0.926	0.900	0.893	0.021	0.905	0.880	0.822	0.814	0.035

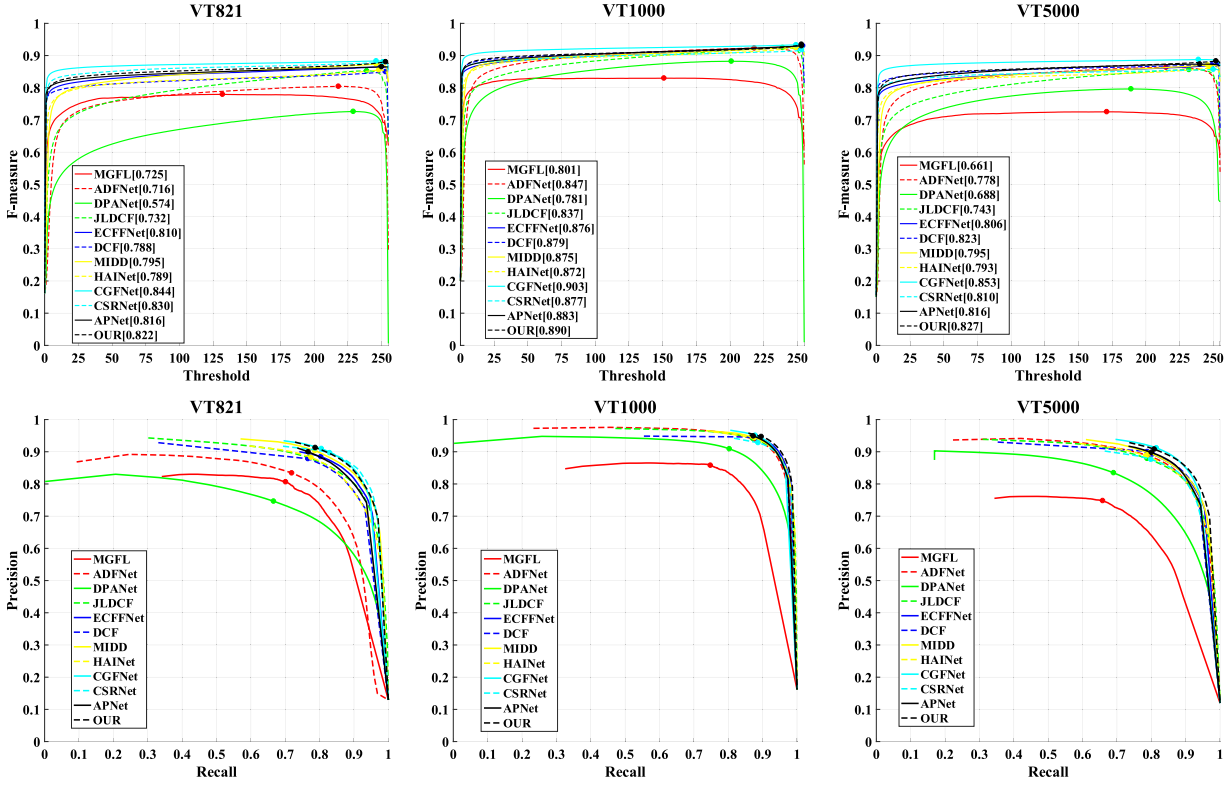


Fig. 13. Quantitative comparison results between our proposed method and the SOTA methods on three datasets. The first line is Precision (vertical axis) Recall (horizontal axis) curves, and the second line shows the F-measure scores of the deep learning-based methods under different thresholds.

correction module to the model is evident. Our model can be further improved in performance by using the efficient CFL-RCM.

4.7. Failure cases and future work

In this paper, we rethink the role of T images in RGB-T salient object detection and address the problem in RGB-T tasks from a new perspective. On the one hand, we introduce a T-aware guided module to guide the RGB images. On the other hand, we designed cross-modal fusion localization-remote correction modules for generating salient objects from coarse to fine and correcting redundant information generated in this process. Although our method shows excellent competitiveness in cross-illumination, it is still inadequate in some difficult scenes, as shown in Fig. 14. For example, in the first to third columns, if the color of the T images is similar to the background color, our detection results will be somewhat degraded. If the RGB images are dark, at that moment, neither the RGB images nor the T images can provide enough information about the salient objects. In the fourth and fifth

Table 4
Comparison of model complexity of deep learning methods running on the same device.

Models (backbone)		ADFNet (VGG16)	CGFNet (VGG16)	JL-DCF (ResNet101)	MIDD (VGG16)	SwinNet (SwinT-B)	OURS (VGG16)
Train time (h)		5.59	3.38	3.26	1.85	2.8	1.04
Runtime (fps)	VT5000	5.84	7.71	7.59	12.07	10.0	15.05
	VT1000	5.93	7.80	7.66	12.11	10.0	15.13
	VT821	5.83	7.73	7.21	11.15	9.94	13.65
	VDT2048	5.90	7.44	7.35	11.59	10.16	14.06
Model Size (MB)		317.2(56.1)	253.4(56.1)	548.1(170)	209.8(56.1)	824.3(429)	145.0(56.1)
Model Params (M)		85.2	66.38	143.4	52.43	198.7	36.18
Flops (G)		191.6	345.1	970.5	216.73	124.3	115.06

Table 5
Comparison of the different contribution metrics of the main modules.

Models				VDT838					
Backbone	CTGFM	T-aware	CFL-RCM	MAE↓	F_{max} ↑	F_m ↑	W_F↑	E_m ↑	S_m ↑
VGG16	√	√	√	0.0042	0.8534	0.8020	0.8210	0.8567	0.9606
VGG16		√	√	0.0044	0.8470	0.7959	0.8150	0.8557	0.9552
VGG16			√	0.0048	0.8390	0.7888	0.8024	0.8461	0.9532
VGG16				0.0050	0.8370	0.7759	0.7972	0.8300	0.9499

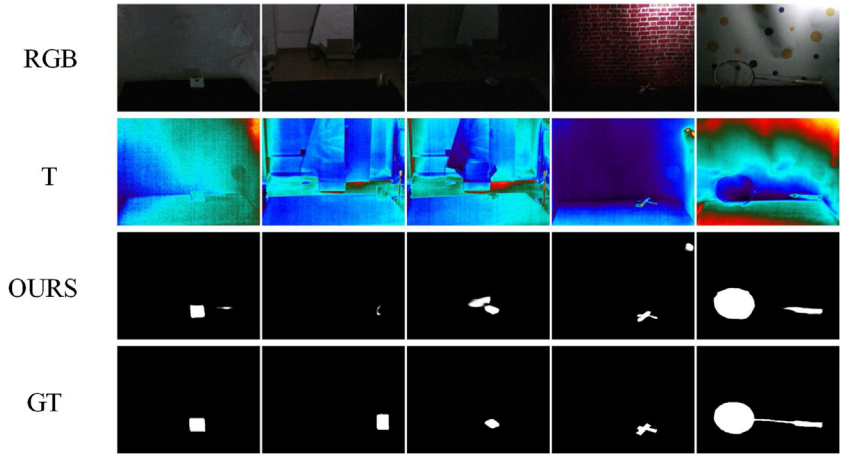


Fig. 14. Failure cases of proposed method.

columns, our method easily misdetection and missed detection due to the anomalous temperature points and some elongated structures of the T images.

For these problems, we can use some stronger feature extractors and early fusion inference mechanisms in the future. In addition, the existing public datasets for the RGB-T SOD task (such as VT821, VT1000, and VT5000 datasets) do not adequately consider the actual value of T images, which are mostly daytime data and contain only a small amount of low-illumination data. The recently proposed dataset of VDT-2048 contains most of the low-illumination data, but the scenes are limited, so it is still necessary to build a subsequent RGB-T dataset containing more low-illumination data.

5. Conclusion

In this paper, our proposed method can take full advantage of the compensation effect of T images on RGB images and achieve accurate detection of cross-illumination data. Considering that the existing RGB-T datasets only have a small amount of low-illumination data and the low-illumination data is difficult to label, in besides, other SOTA salient object detection methods do not solve this problem well. It is worth noting that to solve this problem, we do not spend a lot of labor to label low-illumination data. First of all, we developed a strategy which is to train the model using only normal illumination data and then go to test the low-illumination and extremely low-illumination data. Secondly, we propose a T-aware guided mechanism that takes full advantage of the T images to complement the low-illumination RGB images. This mechanism is applied to cross-illumination saliency detection to reduce the impact of illumination variations. It is mainly achieved by highlighting the salient regions and suppressing the background in the RGB images. In this way, our model focuses more on the information of the salient areas of the RGB images. In addition, we designed a cross-modal fusion localization-remote correction module. The cross-modal fusion localization can accurately locate the salient objects in the case of large differences in illumination between the training and test sets. The remote correction can adequately screen out favorable information and remove redundant information in response to illumination variations. The analysis of comparative and ablation experiments verifies that our method achieves the best performance. In addition, our method also achieves favorable results on the RGB-T datasets.

CRediT authorship contribution statement

Han Wang: Conceptualization, Methodology, Visualization, Experiment, Investigation, Writing – original draft, Writing – review & editing. **Kechen Song:** Conceptualization, Validation, Writing – review & editing, Project administration, Funding acquisition. **Liming Huang:** Formal analysis, Experiment, Writing – review & editing. **Hongwei Wen:** Validation, Writing – review & editing, Supervision. **Yunhui Yan:** Validation, Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (51805078), the Fundamental Research Funds for the Central Universities (N2103011), the Central Guidance on Local Science and Technology Development Fund (2022JH6/100100023), and the 111 Project (B16009).

References

Achanta, R., Hemami, S., Estrada, F., Susstrunk, S., 2009. Frequency-tuned salient region detection. In: Proc. IEEE Conf. Comput. Vis. Pattern Recognit.. pp. 1597–1604.

Bi, H., Wu, R., Liu, Z., et al., 2022. PSNet: Parallel symmetric network for RGB-T salient object detection. Eng. Appl. Artif. Intell. 511, 410–425.

Chen, Z., Cong, R., Xu, Q., Huang, Q., 2021a. 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., et al., 2021b. RGB-D salient object detection via 3D convolutional neural networks. AAAI 1063–1071.

Chen, G., Shao, F., Chai, X., et al., 2022. CGMDRNet: Cross-guided modality difference reduction network for RGB-T salient object detection. IEEE Trans. Circuits Syst. Video Technol..

Cheng, J., Tian, S., Yu, L., Liu, S., Wang, C., et al., 2022a. DDU-net: A dual dense U-structure network for medical image segmentation. Eng. Appl. Artif. Intell. 126, 109297.

Cheng, X., Zheng, X., Pei, J., Tang, H., et al., 2022b. Depth-induced gap-reducing network for RGB-D salient object detection: an interaction, guidance and refinement approach. IEEE Trans. Multimedia.

Cong, R., Zhang, K., Zhang, C., et al., 2022. Does thermal really always matter for RGB-T salient object detection? IEEE Trans. Multimed..

Fan, D.-P., Cheng, M.-M., Liu, Y., Li, T., Borji, A., 2017. Structure-measure: A new way to evaluate foreground maps. In: Proc. IEEE Int. Conf. Comput. Vis. (ICCV). 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. In: Proc. 27th Int. Joint Conf. Artif. Intell.. pp. 1–7.

Fan, D.-P., Lin, Z., Zhang, Z., Zhu, M., Cheng, M.-M., 2021. Rethinking RGB-d salient object detection: Models, data sets, and large-scale benchmarks. IEEE Trans. Neural Netw. Learn. Syst. 32 (5), 2075–2089.

Fang, X., Zhu, J., Shao, X., wang, H., 2022. GroupTransNet: Group transformer network for RGB-D salient object detection. arXiv preprint arXiv:2203.10785.

Feng, G., Meng, J., Zhang, L., Lu, H., 2022. Encoder deep interleaved network with multi-scale aggregation for RGB-D salient object detection. Pattern Recognit. 128, 108666.

Fiaz, M., et al., 2019. Handcrafted and deep trackers: Recent visual object tracking approaches and trends. ACM Comput. Surv. 52 (2), 1–44.

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: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA. pp. 3049–3059.

Fu, K., Jiang, Y., Ji, G.P., et al., 2022. Light field salient object detection: A review and benchmark. Comput. Visual Media 8, 509–534.

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.

Guan, D., Cao, Y., Yang, J., et al., 2018. Exploiting fusion architectures for multispectral pedestrian detection and segmentation. Appl. Opt. 57, 108–116.

Guan, D., Cao, Y., Yang, J., et al., 2019. Fusion of multispectral data through illumination-aware deep neural networks for pedestrian detection. Inf. Fusion 50, 148–157.

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, H., Wang, J., Li, X., et al., 2022. EAF-net: an enhancement and aggregation–feedback network for RGB-T salient object detection. Mach. Vis. Appl. 33, 1–15.

Huang, L., Song, K., Gong, A., Liu, C., Yan, Y., 2020. RGB-T saliency detection via low-rank tensor learning and unified collaborative ranking. IEEE Signal Process. Lett. 27, 1585–1589.

Huang, L., Song, K., Wang, J., Niu, M., Yan, Y., 2022a. Multi-graph fusion and learning for RGBT image saliency detection. IEEE Trans. Circuits Syst. Video Technol. 32 (3), 1366–1377.

Huang, K., Tian, C., Su, J., et al., 2022b. Transformer-based cross reference network for video salient object detection. Eng. Appl. Artif. Intell. 160, 122–127.

Huo, F., Zhu, X., Zhang, L., Liu, Q., Shu, Y., 2022a. 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., 2022b. Real-time one-stream semantic-guided refinement network for RGB-thermal salient object detection. IEEE Trans. Instrum. Meas..

Ji, W., Li, J., Yu, S., et al., 2021. Calibrated RGB-d salient object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9471–9481.

Jiang, X., Zhu, L., Hou, Y., et al., 2022. Mirror complementary transformer network for RGB-thermal salient object detection. arXiv preprint arXiv:2207.03558.

Kompella, A., et al., 2021. A semi-supervised recurrent neural network for video salient object detection. Neural Comput. Appl. 33 (6), 2065–2083.

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.

Liang, Y., Qin, G., Sun, M., Qin, J., Yan, J., Zhang, Z., 2022a. Multi-modal interactive attention and dual progressive decoding network for RGB-D/T salient object detection. Neurocomputing 490, 132–145.

Liang, Y., Qin, G., Sun, M., Qin, J., Yan, J., et al., 2022b. Multi-modal interactive attention and dual progressive decoding network for RGB-D/T salient object detection. Eng. Appl. Artif. Intell. 490, 132–145.

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..

Liu, J.-J., Hou, Q., Liu, Z.-A., Cheng, M.-M., 2022a. PoolNet+: Exploring the potential of pooling for salient object detection. IEEE Trans. Pattern Anal. Mach. Intell..

Liu, J.-J., Liu, Z.-A., Peng, P., Cheng, M.-M., 2021a. Rethinking the U-shape structure for salient object detection. IEEE Trans. Image Process. 30, 9030–9042.

Liu, Y., Pan, C., Bie, M., Li, J., 2022b. An efficient real-time target tracking algorithm using adaptive feature fusion. Eng. Appl. Artif. Intell. 85, 103505.

Liu, Z., Tan, Y., He, Q., Xiao, Y., 2022c. 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, Z., Wang, Y., Tu, Z., Xiao, Y., et al., 2021b. 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.

Ma, S., Song, K., Dong, H., et al., 2022. Modal complementary fusion network for RGB-T salient object detection. Appl. Intell. 1–18.

Margolin, R., Zelnik-Manor, L., Tal, A., 2014. How to evaluate foreground maps. In: Proc. IEEE Conf. Comput. Vis. Pattern Recognit.. pp. 248–255.

Meinhardt, T., et al., 2022. Trackformer: Multi-object tracking with transformers. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8844–8854.

Perazzi, F., Krahenbuhl, P., Pritch, Y., Hornung, A., 2012. Saliency filters: Contrast based filtering for salient region detection. In: Proc. IEEE Conf. Comput. Vis. Pattern Recognit.. pp. 733–740.

Shivakumar, et al., 2020. Pst900: Rgb-thermal calibration, dataset and segmentation network. In: IEEE International Conference on Robotics and Automation. ICRA, pp. 9441–9447.

Shokri, M., et al., 2020. Salient object detection in video using deep non-local neural networks. J. Vis. Commun. Image Represent. 68, 102769.

Song, Y., Tang, H., et al., 2022a. Disentangle saliency detection into cascaded detail modeling and body filling. ACM Trans. Multimedia Comput..

Song, K., Wang, J., et al., 2022b. A novel visible-depth-thermal image dataset of salient object detection for robotic visual perception. IEEE/ASME Trans. Mechatronics.

Strudel, R., et al., 2021. Segmenter: Transformer for semantic segmentation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 7262–7272.

Sun, P., Zhang, W., Wang, H., et al., 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.

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.

Tu, Z., Ma, Y., Li, Z., Li, C., Xu, J., Liu, Y., 2020a. Rgbt salient object detection: A large-scale dataset and benchmark. arXiv preprint arXiv:2007.03262.

Tu, Z., Xia, T., Li, C., Lu, Y., Tang, J., 2019a. M3S-NIR: Multi-modal multi-scale noise-insensitive ranking for RGB-T saliency detection. In: Proc. IEEE Conference on Multimedia Information Processing and Retrieval. pp. 141–146.

Tu, Z., Xia, T., Li, C., Wang, X., Ma, Y., Tang, J., 2019b. RGB-t image saliency detection via collaborative graph learning. IEEE Trans. Multimedia 22 (1), 160–173.

Tu, Z., Xia, T., Li, C., Wang, X., Ma, Y., Tang, J., 2020b. RGB-T image saliency detection via collaborative graph learning. IEEE Trans. Multimedia 22 (1), 160–173.

Wang, G., Li, C., Ma, Y., Zheng, A., Tang, J., Luo, B., 2018a. Rgb-t saliency detection benchmark: Dataset, baselines, analysis and a novel approach. In: Proc. Chin. Conf. Image Graph. Technol.. pp. 359–369.

Wang, G., Li, C., Ma, Y., Zheng, A., Tang, J., Luo, B., 2018b. RGB-T saliency detection benchmark: Dataset, baselines, analysis and a novel approach. In: Chinese Conference on Image and Graphics Technologies. pp. 359–369.

Wang, F., Pan, J., Xu, S., Tang, J., 2022a. 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., 2021. CGFNet: Cross-guided fusion network for RGB-T salient object detection. IEEE Trans. Circuits Syst. Video Technol. 32 (5), 2949–2961.

Wang, J., Song, K., Bao, Y., et al., 2022b. Unidirectional RGB-T salient object detection with intertwined driving of encoding and fusion. Eng. Appl. Artif. Intell. 114, 105162.

Wen, H., Yan, C., et al., 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: Proc. Eur. Conf. Comput. Vis.. pp. 3–19.

Wu, Y., Liu, Y., Zhang, L., et al., 2022. EDN: Salient object detection via extremely-downsampled network. IEEE Trans. Image Process. 31, 3125–3136.

Xu, C., Li, Q., Zhou, Q.M., Zhou, Q., et al., 2022. RGB-T salient object detection via CNN feature and result saliency map fusion. Appl. Intell. 1–20.

Yan, P., Wu, Z., Liu, M., et al., 2022. Unsupervised domain adaptive salient object detection through uncertainty-aware pseudo-label learning. arXiv preprint arXiv:2202.13170.

Zeng, C., Kwong, S., 2022. Dual swin-transformer based mutual interactive network for RGB-D salient object detection. arXiv preprint arXiv:2203.03105.

Zhai, Y., Fan, D., Yang, J., et al., 2021. Bifurcated backbone strategy for rgb-d salient object detection. IEEE Trans. Image Process. 30, 8728–8742.

Zhang, C., Cong, R., Lin, Q., et al., 2021a. Cross-modality discrepant interaction network for RGB-D salient object detection. In: Proceedings of the 29th ACM International Conference on Multimedia. pp. 2094–2102.

Zhang, Q., Huang, N., Yao, L., Zhang, D., Shan, C., Han, J., 2020. Rgb- t salient object detection via fusing multi-level cnn features. IEEE Trans. Image Process. 29, 3321–3335.

Zhang, Z., Lin, Z., Xu, J., Jin, W.-D., Lu, S.-P., Fan, D.-P., 2021b. Bilateral attention network for RGB-D salient object detection. IEEE Trans. Image Process. 30, 1949–1961.

Zhang, Q., Xi, R., Xiao, T., et al., 2022a. Enabling modality interactions for RGB-T salient object detection. Comput. Vis. Image Underst. 222, 103514.

Zhang, Q., Xiao, T., Huang, N., Zhang, D., Han, J., 2021c. Revisiting feature fusion for RGB-t salient object detection. IEEE Trans. Circuits Syst. Video Technol. 31 (5), 1804–1818.

Zhang, L., Zhang, Q., Zhao, R., 2022b. Progressive dual-attention residual network for salient object detection. IEEE Trans. Circuits Syst. Video Technol..

Zhou, T., Fan, D.P., Cheng, M.M., et al., 2021a. RGB-d salient object detection: A survey. Comput. Visual Media 7, 37–69.

Zhou, T., Fu, H., Chen, G., et al., 2021b. 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., 2022a. ECFFNet: Effective and consistent feature fusion network for RGB-T salient object detection. IEEE Trans. Circuits Syst. Video Technol. 32 (3), 1224–1235.

Zhou, W., Zhu, Y., Lei, J., Wan, J., Yu, L., 2022b. Apnet adversarial learning assistance and perceived importance fusion network for all-day RGB-t salient object detection. IEEE Trans. Emerg. Top. Comput. Intell. 6 (4), 957–968.

Zhu, H., Sun, X., Li, Y., Ma, K., Zhou, S., Zheng, Y., 2022. DFTR: Depth-supervised fusion transformer for salient object detection. arXiv preprint arXiv:2203.06429.

Zhuge, M., Fan, D.-P., Liu, N., Zhang, D., Xu, D., Shao, L., 2022. Salient object detection via integrity learning. IEEE Trans. Pattern Anal. Mach. Intell..