Available online at https://jcupt.bupt.edu.cn

The Journal of China Universities of Posts and Telecommunications

April 2025, 32(2): 44 – 55

Cross-modal attention and reinforcement for RGB-T salient object detection

Bi Hongbo[1,2,3], Sun Weihan[1], Zhang Jiayuan[1], Xia Bingjie[1], Guo Yingwei[1](✉), Zhang Cong[1]

1. School of Electrical Information Engineering, Northeast Petroleum University, Daqing 163318, China
2. Shantou University, Shantou 515063, China
3. State Key Laboratory of Public Big Data, Guizhou University, Guiyang 550025, China

Abstract

Exploring the interaction between red, green, blue (RGB) and thermal infrared modalities is critical to the success of RGB-thermal (RGB-T) salient object detection (RGB-T SOD). In this paper, a cross-modal attention and reinforcement network (CAR-Net) was proposed to explore the implicit relationship between the two modalities, which fully leverages the beneficial expression and complementary fusion of the two modalities. Specifically, CAR-Net has a cross-modal attention module (CAM) that enables efficient interaction and key information extraction through joint attention. It also includes a feature strengthener module (FSM) for improved representation using channel rank and loop methods. A large number of experiments show that the CAR-Net achieves the best performance on three publicly available datasets.

Keywords RGB-thermal (RGB-T) salient object detection (RGB-T SOD), attention, feature strengthener, multi-modal fusion

1 Introduction

Salient object detection (SOD) seeks to explore and isolate obvious objects in various scenarios. It has been implemented and popularized in plenty of computer vision tasks, such as semantic segmentation[1-2], visual tracking[3], image compression[4], and video segmentation[5]. Recently, different novel methods have been proposed to achieve rapid improvement in SOD performance, which has important significance for the above SOD tasks.

Although RGB images can provide the color, texture, contour, and other appearance information of the salient objects or regions, the RGB-based SOD[6] still has limitations in the complex environment, such as confusing shaded regions, low-light conditions, and salient objects with anfractuous boundaries. As an auxiliary tool to RGB images, depth images can provide abundant geometric and spatial layout clues of salient objects, and are conducive to improving detection performance. The existing RGB-depth (RGB-D) SOD methods[7-10] have made breakthrough progress. However, it may have side effects when the subject is close to the background and poor illumination. Thermal infrared equipment conveniently presents the thermal information of biological or other substances, and is widely used in various detection fields. Furthermore, the thermal images obtained from these thermal infrared devices provide supplementary

Special Issue: The 27th Annual Meeting of The China Association for Science and Technology

Corresponding author: Guo Yingwei, E-mail: guoyingwei8801@163.com

DOI: 10.19682/j.cnki.1005-8885.2025.0013

information for RGB images and are not susceptible to appearance and illumination changes. Therefore, it is beneficial for RGB-T SOD to effectively solve the challenges posed by various extreme environments.

Since RGB-T SOD has become a research hotspot, researchers are continually proposing new strategies to address various challenges and improve detection performance. The core of RGB-T SOD lies in effectively capturing and fusing key information from both RGB and thermal imaging modalities. Therefore, designing reasonable cross-modal cooperation methods and mitigating the negative impacts of modality differences are crucial. In earlier RGB-T SOD studies, researchers proposed to use graph learning[11] or support vector machine (SVM) regression[12] to accomplish cross-modal fusion. However, these work mainly depend on low-level hand-crafted features, which leads to deeper semantic information in the images not being captured and hinders the improvement of models, performance. Afterward, based on the powerful ability of the convolutional neural networks to capture important information and their outstanding performance in image processing, researchers tend to utilize deep learning to design RGB-T SOD models and propose different cross-modal fusion strategies[13-15]. However, many existing cross-modal fusion strategies ignore that precisely extracting information from RGB and thermal modalities is essential for effectively aggregating cross-modal clues. Insufficient feature extraction and aggregation of different modalities cause the effect of existing approaches to reach the bottleneck. In addition, realizing the full use of multi-scale features for efficient decoding is also pivotal to accurate detection. Current researches generally adopt various kinds of dilated convolution to excavate high-level semantic information or multi-scale vital salient clues, such as atrous spatial pyramid pooling (ASPP)[16] and receptive field block (RFB)[17]. Although these modules contribute to determining the location of significant regions to a certain extent, they remain in a large local receptive field, which can easily give rise to over-extraction and confuse background and salient objects.

To address these challenges mentioned above, CAR-Net was proposed in this paper. The CAR-Net mainly consists of two core components, including the CAM and the FSM. Specifically, the CAM is designed for accurate cross-modal fusion. By combining attention modules to effectively extract available semantic information and reinforce salient features in RGB and thermal images respectively. The FSM employs a semantic feature-guided strategy aimed at enhancing cross-modal saliency representations for sufficient multi-scale feature aggregation. Generally speaking, the main contributions of this paper are as follows.

1) CAM is proposed for locating and fusing RGB and thermal prominent regions efficaciously. The CAM adopts joint attention to respectively capture RGB and thermal information and realize the full utilization of cross-modal information by using a channel sorting strategy. Then, the maximum value and the average value of different modalities features are respectively taken on the channel to highlight the significance clues and aggregate multi-scale features adequately.

2) Committed to fully utilizing and aggregating multi-scale cues, FSM based on semantic feature guidance strategy is proposed. Based on the properties of different scale features, the FSM adopts the combination of channel sorting and a loop mechanism to further strengthen the extraction of significant features.

3) Training, testing, and ablation experiments on the CAR-Net are conducted. A large number of experiments show that CAR-Net has excellent performance on the three benchmark datasets, i. e., VT821, VT1000, and VT5000, and exceeds the current methods.

2 Ralated work

This section begins with an introduction to the SOD tasks. Then, the method reviews several representative researches on SOD and RGB-D SOD, mainly based on deep learning. Finally, the method discusses several RGB-T SOD frameworks.

Compared with general object detection, the challenge of SOD is to reflect the most prominent one from various regions. According to the diverse input

modal, the SOD task can be roughly classified into three categories, i. e., RGB SOD[18-19], RGB-D SOD[20-21], and RGB-T SOD[22-23]. To achieve detection, the traditional SOD models mostly concentrate on constructing manual-crafted features according to important objects' characteristics, including color, texture, shape, etc. However, the traditional SOD models achieve poor performance in complicated environments. To overcome this limitation, several structures based on deep learning are proposed. Next, some typical SOD models based on deep learning are discussed.

2.1 RGB SOD and RGB-D SOD

Benefiting from the rapid development of deep learning, RGB-based SOD deep learning methods achieve remarkable performance. Unlike existing models that focus on inter-image relationships, Yang et al.[24] introduced a multi-grained refinement module to discriminate between saliency objects and background information, and introduced inter-image relationships into the pixel-by-pixel segmentation features to enhance the discrimination of segmentation features. This multi-grained refinement is the basis for mining saliency objects. While some existing methods have shown that the additional edge supervision can facilitate SOD, edge pixels are often much less common than non-edge pixels, leading to the challenge of class imbalance. To overcome this issue, Yi et al.[19] introduced detail labels that provide additional internal details as a supplementary supervisory signal.

Depth images contain a wealth of spatial and structural detail information, which is desirable for SOD in challenging situations, i. e., in the case of cluttered backgrounds and multiple objects. However, the issue of conflicting data fusion due to modal inconsistencies between RGB image and depth image data remains to be solved in an effective way.

To overcome the above limitation, Hu et al.[25] did not integrate additional modules, such as feature enhancement and edge generation modules, to achieve higher performance. The designed network structure included only three integral parts: feature encoding, feature fusion, and feature decoding. Specifically, a dual-stream encoder is used to extract multi-modal and multi-scale features to model global information. In the feature fusion part, the attention mechanism is used to fuse multi-modal and multi-level features. In the feature decoding part, a progressive decoder is added to gradually fuse low-level features and filter the noise information to accurately predict the salient target. Kanwal et al.[26] proposed a two-way feature interaction approach to exploit the synergies of RGB and depth modalities, which effectively captures the intra-modality features at lower layers of the backbone network using a novel operation-wise shuffle channel attention module and supplements the edge guidance network. The quality of the depth map is explicitly investigated through a bottom-up edge guidance network and affluent saliency cues are exploited via a top-down global contextual aware reverse attention mechanism. Niu et al.[27] took a novel dual consistency loss function to help train the whole network, which can further force it to learn complementary features from both RGB and depth images. Zeng et al.[28] designed several decoding stages and keep decoding histories of intermediate decoding results to make full use of encoding clues. During the decoding stages, all previous decoding histories are considered progressively. In addition, the encoding features from different encoding stages are also merged as additional clues. The RGB features are emphasized with the skip convolution module. More researches that focus on RGB-D SOD can be found in Ref. [29].

2.2 RGB-T SOD

The depth map facilitates easily complementing the location cues of items inside the image and has driven the rapid progress of SOD tasks. However, in dimly light environments, the depth information can introduce noise interference instead of providing effective information support.

Wang et al.[30] conducted the first publicly relevant RGB-T SOD on dataset V T821 and designed a multi-task manifold sorting structure to segment salient targets based on multi-modal. Unlike the model of fusing multi-modal features in the early, middle, or late stages, Zhang et al.[12] proposed a multi-

interactive dual decoder to fully integrate various information, namely, cross-modal, multi-level, local details, and global context features. Gao et al.[13] employed a dual-modal detection stream mechanism to extract and integrate RGB and thermal information based on multiple stages and scale features. Jin et al.[31] designed an asymmetric feature complementary feature interaction unit to enhance RGB and thermal features, and fuse the features of the two modes in channel and spatial dimensions, while reducing the interference of thermal modes. Zhou et al.[23] used a novel bilateral fusion, multi-level coherent fusion detection method[3] for cross-modality feature extraction and refinement.

3 CAR-Net

In this section, the entire RGB-T SOD framework called CAR-Net is first outlined. Two functional modules, CAM and FSM, are described in detail next.

3.1 Overall architecture

The entire network framework of CAR-Net is displayed in Fig. 1. To elaborate sequentially, the method first adopts residual neural network (ResNet) with 2-dimensional (2D) squeeze-and-excitation network (Res2Net)[32] as the backbone to excavate the superficial features $\boldsymbol{f}_{\mathrm{rgb}}^{1}-\boldsymbol{f}_{\mathrm{rgb}}^{5}$ and $\boldsymbol{f}_{\mathrm{t}}^{1}-\boldsymbol{f}_{\mathrm{t}}^{5}$ from the input RGB images and thermal maps respectively. Then, the method design a CAM to aggregate the basic features from RGB and thermal infrared branches respectively, to realize the fusion of multi-modal features. Thus, it can get the preliminary polymerization features $\boldsymbol{f}_{\mathrm{cam}}^{1}-\boldsymbol{f}_{\mathrm{cam}}^{5}$, which can be described as

$$\boldsymbol{f}_{\mathrm{cam}}^{i}=F_{\mathrm{CAM}}(\boldsymbol{f}_{\mathrm{rgb}}^{i},\boldsymbol{f}_{\mathrm{t}}^{i});\quad i\in\{1,2,\ldots,5\} \tag{1}$$

where $F_{\mathrm{CAM}}(\cdot)$ denotes the output of CAM.

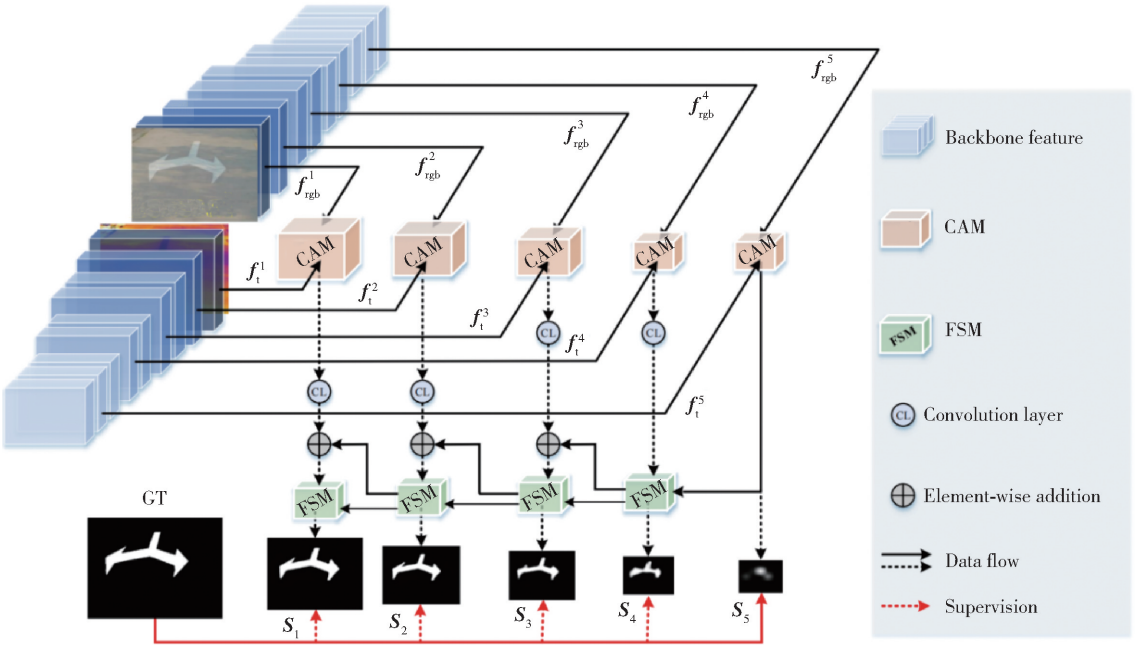


Fig. 1 Structure of CAR-Net

After that, FSM is used to mine the key critical and informative salient cues and enhance the expression of salient features. To implement it, a step-by-step decoding strategy is employed to slowly aggregate the salient features of multiple layers from high to low. Finally, generate the saliency prediction map at the first layer. Note that to roundly utilize the features of each layer, this article blends in a guidance mechanism to feedback features of the high layer to the low layer, and guides the detailed features to generate the saliency prediction map, which can be described as

$$\left.\begin{aligned}\boldsymbol{S}_i&=F_{\mathrm{FSM}}(\boldsymbol{f}_{\mathrm{cam}}^{i}+\hat{\boldsymbol{F}}_{i+1},\hat{\boldsymbol{S}}_{i+1});\quad i\in\{1,2,3,4\}\\ \boldsymbol{S}_5&=F_{\mathrm{conv}}^{3\times3}(\boldsymbol{f}_{\mathrm{cam}}^{5})\end{aligned}\right\} \tag{2}$$

where $\boldsymbol{S}_i$ means that the $\boldsymbol{f}_{\text{cam}}$ features of the ith layer are added with $\hat{\boldsymbol{F}}_{i+1}$, and then the preliminary result prediction map of the ith layer is obtained by passing through the FSM module together with the supervised feature $\hat{\boldsymbol{S}}_i$ of the ith layer ground truth (GT) map. $\hat{\boldsymbol{F}}_{i+1}$ represents the features of the $(i+1)$th layer that pass through the FSM module. $\hat{\boldsymbol{S}}_i$ represents the supervisory features of the truth map of the ith layer, and $\boldsymbol{S}_5$ represents the preliminary result prediction map of 5th layer features.

In addition, in order to get more accurate results, adopt multi-supervision to supervise the output from each layer separately. Specifically, in CAR-Net, the model collectively optimize the initial result map of each layer $\boldsymbol{S}_2 - \boldsymbol{S}_5$ and the final saliency map $\boldsymbol{S}_1$, by dening the total loss, $L(\boldsymbol{S},\boldsymbol{G})$.

$$L(\boldsymbol{S},\boldsymbol{G}) = \sum_{i=1}^{5} L_{\text{bce}}(\boldsymbol{S}_k,\boldsymbol{G}) \tag{3}$$

where the $L_{\text{bce}}(\cdot)$ represents the binary cross entropy loss and $\boldsymbol{G}$ denotes the GT saliency map. $\boldsymbol{S}$ denotes the preliminary prediction graph.

3.2 CAM

In the RGB-T SOD community, one of the essential steps is to effectively interact with multi-modal data, which is also a major concern and challenge. For this purpose, CAM is proposed to coalesce multi-modal features and further capture the valuable information, thus boosting the detection performance, shown in Fig. 2. In Fig. 2, © denotes the concatenation operation.

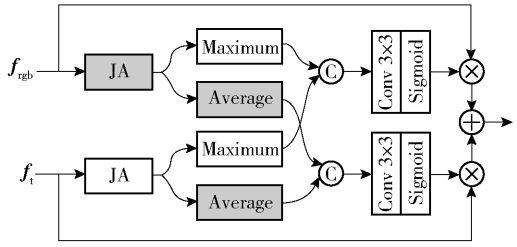


Fig. 2 Architecture of CAM

In detail, firstly, design a joint attention (JA) module composed of channel attention and position attention, as shown in Fig. 3, which mines the key clues and strengthens the expression of salient features from the two dimensions of channel and position, respectively. Specifically, on channel attention, utilize the adaptive maximum pooling operation to obtain the global features firstly. Then, cascade four convolution layers and adjust the number of channels to learn the information of channels. Noteworthy, each convolution layer (Convlayer) consists of a 3×3 convolution (Conv 3×3), a batch normalization and a retifiled linear unit (ReLU) activation function. Finally, the residual multiplication method is used to obtain the salient features.

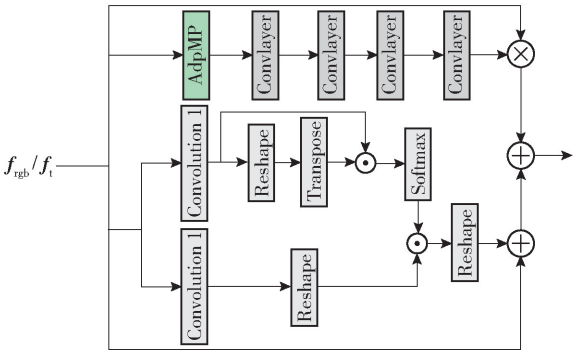


Fig. 3 Architecture of the JA module

The whole process of channel attention can be expressed by

$$F_{\text{ca}}(\boldsymbol{f}) = \boldsymbol{f} \prod_{i=1}^{4} F_{\text{CBR}}(P_{\text{M}}(\boldsymbol{f})) \tag{4}$$

where $F_{\text{ca}}(\cdot)$ denotes the channel attention operation, $F_{\text{CBR}}(\cdot)$ means a 3×3 convolution followed by batch normalization and ReLU, $P_{\text{M}}(\cdot)$ represents the adaptive maximum pooling operation. In position attention, for the input feature $\boldsymbol{f} \in \mathbb{R}^{H \times W \times C}$, including RGB feature and thermal feature. H, W, and C denote the height, width, and the number of channels of the input feature, respectively. The model first feed $\boldsymbol{f}$ into a 1×1 convolution to obtain feature maps, e. g., $\boldsymbol{f}^{\text{conv } 1\times 1}$, and capture the position information element by element, where $\boldsymbol{f}^{\text{conv } 1\times 1} \in \mathbb{R}^{H \times W \times C}$. Then reshape $\boldsymbol{f}^{\text{conv } 1\times 1}$ to $\bar{\boldsymbol{f}}^{\text{conv } 1\times 1} \in \mathbb{R}^{N \times C}$, where $N = H \times W$ means the number of pixels. Then, employ a matrix multiplication between $\boldsymbol{f}^{\text{conv } 1\times 1}$ and $\bar{\boldsymbol{f}}^{\text{conv } 1\times 1}$, and apply a softmax layer to calculate the feature map.

Moreover, reshape $\boldsymbol{f}^{\text{conv } 1\times 1}$ to $\hat{\boldsymbol{f}}^{\text{conv } 1\times 1} \in \mathbb{R}^{C \times N}$. Then

perform a matrix multiplication between $\bar{\boldsymbol{f}}^{\text{conv }1\times1}$ and $\boldsymbol{f}^{\text{conv }1\times1}$ and reshape the result to $\mathbb{R}^{H\times W\times C}$. Finally, the method use a summation operation to add it to the original features $\boldsymbol{f}$ to boost the detection performance.

The whole process of position attention can be defined as

$$\left.\begin{aligned}\boldsymbol{f}^{\text{conv }1\times1} &= F^{\text{conv }1\times1}(\boldsymbol{f})\\ \bar{\boldsymbol{f}}^{\text{conv }1\times1} &= \boldsymbol{f}^{\text{conv }1\times1} F_{\text{T}}(F_{\text{R}}(\boldsymbol{f}^{\text{conv }1\times1}))\\ F_{\text{PA}}(\boldsymbol{f}) &= \boldsymbol{f} + F_{\text{R}}(F_{\text{S}}(\bar{\boldsymbol{f}}^{\text{conv }1\times1}) F_{\text{R}}(\boldsymbol{f}^{\text{conv }1\times1}))\end{aligned}\right\} \tag{5}$$

where $F^{\text{conv }1\times1}(\cdot)$ means 1×1 convolution operation, $F_{\text{T}}(\cdot)$, $F_{\text{R}}(\cdot)$ and $F_{\text{S}}(\cdot)$ means transpose, reshape, and softmax operation respectively.

To interact the two modalities completely, the CAR-Net collects the maximum and average values of RGB and thermal features along the channel respectively, and connects the maximum and average values of the two modalities in series to make the multi-modal features preliminarily polymerized. Then, the feature weights are obtained through a 3×3 convolution and a sigmoid function, and capture the salient features by the method of residual multiplication. Finally, the CAR-Net adopts the element-wise summation to fuse the two modalities. The calculation of CAM can be described as

$$\left.\begin{aligned}\boldsymbol{f}_{\max} &= F_{\text{cat}}(F_{\text{M}}(F_{\text{JA}}(\boldsymbol{f}_{\text{rgb}})), F_{\text{M}}(F_{\text{JA}}(\boldsymbol{f}_{\text{t}})))\\ \bar{\boldsymbol{f}} &= F_{\text{cat}}(F_{\text{A}}(F_{\text{JA}}(\boldsymbol{f}_{\text{rgb}})), F_{\text{A}}(F_{\text{JA}}(\boldsymbol{f}_{\text{t}})))\\ F_{\text{CAM}} &= \sigma(F^{3\times3}_{\text{conv}}(\boldsymbol{f}_{\max})\boldsymbol{f}_{\text{rgb}}) + \sigma(F^{3\times3}_{\text{conv}}(\bar{\boldsymbol{f}})\boldsymbol{f}_{\text{t}})\end{aligned}\right\} \tag{6}$$

where $F_{\text{JA}}(\cdot)$ indicates JA module and $F_{\text{cat}}(\cdot)$ means concatenation operation, $F_{\text{A}}(\cdot)$ and $F_{\text{M}}(\cdot)$ denote taking the average and maximum values in the channel, $F_{\text{CAM}}(\cdot)$ denotes the final output of the CAM, $\sigma(\cdot)$ expresses the sigmoid activation function.

3.3 FSM

As we all know, different CNN layers capture features with different scales. The deeper layer contains rich semantic information while the lower layer includes abundant structure cues. However, in the process of feature fusion, the semantic features are gradually diluted, especially when combining shallow features with noise information. To address it, a FSM was designed, shown in Fig. 4, which mainly contains a channel rank operation and adopts a circular way to enhance the salient features.

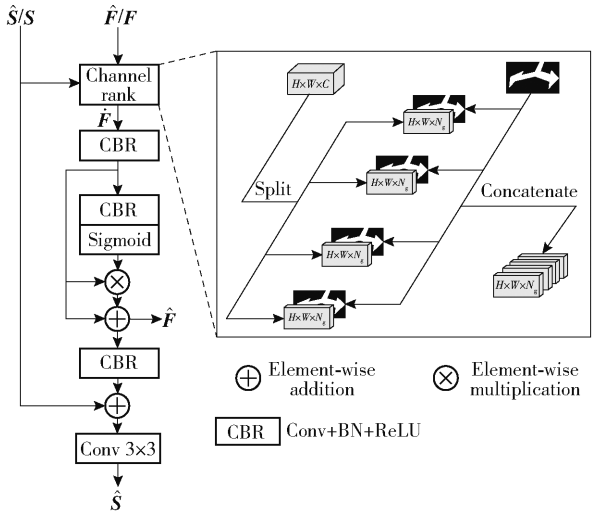


Fig. 4 Architecture of FSM

Specifically, given two inputs in the channel rank operation, e. g., a convolutional feature $\boldsymbol{F}$ with N_{C} channels and prediction map $\boldsymbol{S}$. The FSM first splits $\boldsymbol{F}$ into N_{g} groups. Then $\boldsymbol{S}$ is regarded as a guidance feature map to concatenate with each slice feature. After concatenation, a feature with $N_{\text{c}} + N_{\text{g}}$ channels can be obtained, which can be formulated as

$$\left.\begin{aligned}F_{\text{split}} &= \{\boldsymbol{F}_1, \boldsymbol{F}_2, \ldots, \boldsymbol{F}_{N_{\text{g}}}\}\\ F_{\text{rank}} &= F_{\text{cat}}(\boldsymbol{F}_1, \boldsymbol{S}, \boldsymbol{F}_2, \boldsymbol{S}, \ldots, \boldsymbol{F}_{N_{\text{g}}}, \boldsymbol{S})\end{aligned}\right\} \tag{7}$$

where the F_{split} and F_{rank} mean splitting channels and reranking channels respectively.

After the channel rank operation, feed the concatenated feature to a convolution layer for guided learning and reduce the number of channels from $N_{\text{c}} + N_{\text{g}}$ to N_{c}, which contains a 3×3 convolution, a batch normalization, and a ReLU activation function. The feature representation is then enhanced using a convolutional layer followed by a sigmoid activation function, and employs residual to complement salient features. Then, through a convolution layer, it can get the first reinforced feature map $\hat{\boldsymbol{F}}$, which can be

calculated as

$$\left.\begin{aligned}\dot{\boldsymbol{F}} &= F_{\mathrm{CR}}(\boldsymbol{F})\\ \hat{\boldsymbol{F}} &= F_{\mathrm{CBR}}(\dot{\boldsymbol{F}})\sigma(F_{\mathrm{CBR}}(F_{\mathrm{CBR}}(\dot{\boldsymbol{F}})))\end{aligned}\right\} \tag{8}$$

where the $F_{\mathrm{CR}}(\cdot)$ represents the channel rank operation. Another 3×3 convolutional layer is further applied to generate the saliency prediction map $\hat{\boldsymbol{S}}$, which can be defined as

$$\hat{\boldsymbol{S}} = F_{\mathrm{conv}}^{3\times3}(F_{\mathrm{CBR}}(\hat{\boldsymbol{F}}) + \boldsymbol{S}) \tag{9}$$

where the $F_{\mathrm{conv}}^{3\times3}(\cdot)$ expresses the 3×3 convolutional. Note that, to reduce the computational complexity, this paper set the number of channels as 64. Moreover, the model adopt a cyclic strategy to strengthen the salient features repeatedly. In each cycle, the number of groups N_{g} is different, and N_{g} has the following corresponding relationship with the nth cycle.

$$N_{\mathrm{g}} = \frac{64}{2^n};\quad n\in\{0,1,\ldots,7\} \tag{10}$$

where n means the nth circulate.

4 Experiment

4.1 Datasets and evaluation metric

The CAR-Net is experimented on three publicly available RGB-T SOD benchmark datasets VT821[30], VT1000[32], and VT5000[33]. VT821 is the first RGB-T SOD relevant dataset containing 821 image pairs collected from approximately 60 environmental scenes. Compared with VT821, which uses the manual alignment of RGB and thermal images, VT1000 leverages RGB and thermal cameras to match RGB and thermal images highly. VT5000 is the largest RGB-T SOD dataset up to now, consisting of 5 000 image pairs covering 11 challenging scenes.

In this paper, four commonly used evaluation indicators are used to demonstrate the effectiveness of CAR-Net, including structure measure (S_{m})[34], frequency measure (F_{m})[35], enhanced-alignment measure (E_{m})[36], mean absolute error (ε)[37]. It is important to note that max F_{m} and max E_{m} refer to maximum values. Structure measure is leveraged to calculate the region and object similarity between predicted maps and inputs' corresponding ground truths. F_{m} is to reflect the frequency tuning of object areas. E_{m} is based on cognitive vision studies, jointly matching features at the pixel level and capturing statistical features at the image level. For S_{m}, max F_{m}, and max E_{m}, the larger the data value, the better the performance, in contrast, the smaller the value of ε, the better the performance.

4.2 Implementation detail

This experiment is conducted based on PyTorch and a single NVIDIA GTX1080Ti GPU. The training and testing images are uniformly resized to 224×224. In this work, Res2Net[32] is used as a backbone network. The basic parameters of this experiment are set, for example, as more learning rates can increase the learning accuracy, the batch size is equal to 10 and the learning rate is equal to 1×10^{-4}.

4.3 Comparison with state-of-the-art

To demonstrate the effectiveness of CAR-Net, it was compared to 16 high-grade methods, including two-stage fusion network (TSFNet)[38], adversarial learning assistance and perceived importance fusion network (APNet)[39], multi-interactive dual-decoder (MIDD) for RGB-T SOD[14], modal complementary fusion network (MCFNet)[40], multi-modal interactive attention and dual progressive decoding (MIA-DPD)[22] network, cross-guided fusion network (CGFNet)[15], efficient context-guided stacked refinement network (CSRNet)[41], effective and consistent feature fusion network (ECFFNet)[23], multi-graph fusion and learning for RGB-T image saliency detection (MGFL)[42], multi-stage and multi-scale fusion network (MMNet)[13], RGB-T image saliency detection via collaborative graph learning (SDGL)[33], real-time one-stream semantic-guided refinement network (OSRNet)[43], cross modal view-mixed transformer for bi-modal SOD (CAVER)[44], multitype fusion and enhancement network (MFENet)[45], patch-to-pixel attention-aware

transformer network (PATNet)[46], multi-scale fusion and edge-supervised network (MSEDNet)[47]. For a fair comparison, all predicted maps for these frameworks were either provided by the authors or generated by models retrained with open-source code.

1) Quantitative evaluation

Table 1 lists the quantitative evaluation results of differenst methods on three datasets. Compared with other recent methods, the CAR-Net achieves outstanding results in RGB-T SOD tasks with different scenarios. In particular, on the VT5000 dataset, the best results were achieved for all four reference indicators. Specifically, compared with the 2nd best MFENet, CAR-Net increases S_m by 0.91%, max E_m by 0.64%, and max F_m by 1.51%, and decreases ε by 3.13%. Compared with the 3rd best C2FNet, the CAR-Net increases max E_m by 0.86%, and max F_m by 0.92% and decreases ε by 3.13%.

Table 1 Quantitative results of different methods on three public RGB-T datasets

Method	S_m			ε			max E_m			max F_m		
	VT821	VT1000	VT5000	VT821	VT1000	VT5000	VT821	VT1000	VT5000	VT821	VT1000	VT5000
TSFNet	0.847	0.912	0.851	0.049	0.027	0.049	0.878	0.946	0.893	0.790	0.899	0.807
APNet	0.867	0.921	0.875	0.034	0.021	0.035	0.909	0.955	0.920	0.825	0.913	0.848
MIDD	0.871	0.915	0.867	0.045	0.027	0.043	0.918	0.957	0.920	0.851	0.91	0.849
MCFNet	0.891	0.932	0.887	0.029	0.019	0.033	0.928	0.967	0.931	0.863	0.929	0.867
MIA-DPD	0.844	0.924	0.878	0.070	0.025	0.040	0.902	0.964	0.929	0.831	0.928	0.865
CGFNet	0.881	0.923	0.883	0.038	0.023	0.035	0.920	0.959	0.926	0.866	0.923	0.689
CSRNet	0.885	0.918	0.867	0.038	0.024	0.042	0.923	0.953	0.914	0.858	0.908	0.836
ECFFNet	0.877	0.923	0.874	0.034	0.021	0.038	0.910	0.959	0.921	0.834	0.917	0.847
MGFL	0.873	0.914	0.862	0.040	0.027	0.043	0.921	0.953	0.914	0.848	0.906	0.836
MMNet	0.873	0.914	0.862	0.040	0.027	0.043	0.921	0.953	0.914	0.848	0.906	0.836
SDGL	0.765	0.787	0.750	0.085	0.090	0.089	0.839	0.859	0.829	0.735	0.770	0.695
OSRNet	0.875	0.926	0.883	0.043	0.022	0.033	0.916	0.965	0.933	0.839	0.924	0.862
CAVER	0.783	0.870	0.801	0.070	0.041	0.060	0.871	0.937	0.894	0.720	0.845	0.742
MFENet	0.884	0.929	0.883	0.030	0.019	0.033	0.927	0.965	0.933	0.854	0.924	0.862
MSEDNet	0.849	0.914	0.878	0.070	0.032	0.041	0.891	0.960	0.930	0.825	0.920	0.864
PATNet	0.822	0.876	0.876	0.069	0.054	0.054	0.877	0.922	0.922	0.779	0.862	0.862
CAR-Net	0.887	0.929	0.891	0.037	0.020	0.032	0.929	0.968	0.939	0.868	0.930	0.875

2) Qualitative evaluation

In Fig. 5, examples of different types of visualizations are listed centrally, further demonstrating the effectiveness of CAR-Net by comparison with other competing networks. It can be seen that the saliency map obtained by CAR-Net is much closer to reality. The existing RGB-T method cannot accurately detect important areas due to the confusion of the environmental scene. In contrast, the CAR-Net can achieve better performance, has good prospects and complete structure.

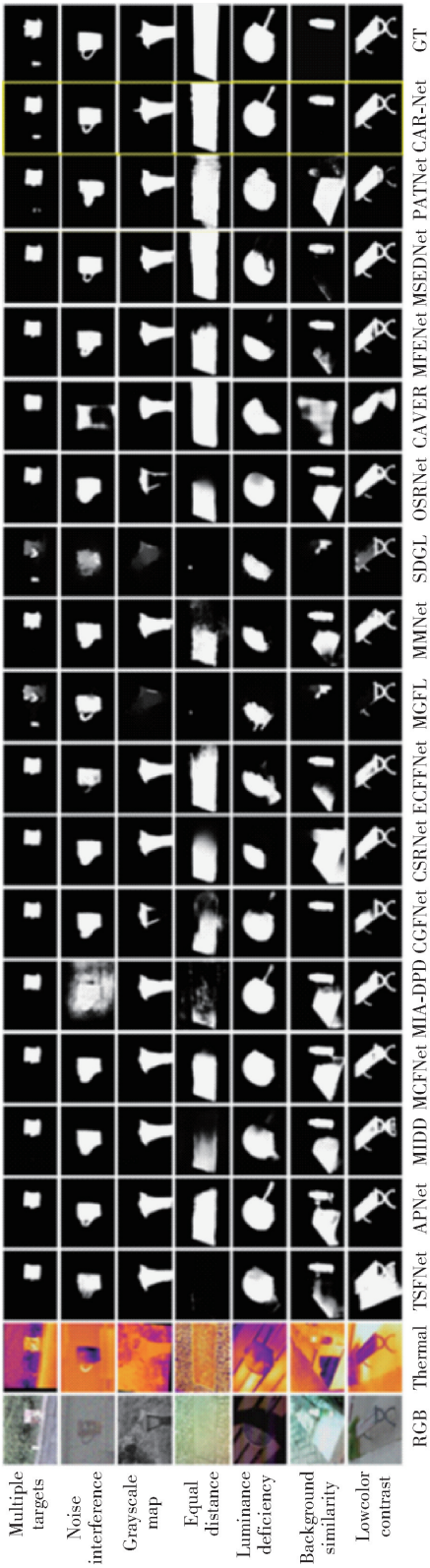


Fig.5 Visual instances of the CAR-Net contrast to other models

In this paper, the size and efficiency of the model are evaluated using floating point operations per second (FLOPs), the number of parameters, and fram per sencond (FPS) metrics, and the evaluation results are compared with some other methods. As shown in Table 2, it can be seen from the evaluation results that the size and efficiency of the algorithm are at a high level compared with other models. Therefore, CAR-Net further strengthen the efficiency of the model and reduce the complexity and parameters of the model in the future.

Table 2 Comparison of FLOPs, the number of parameters, and efficiency

Method	FLOPs/10^9	Number of parametes	Efficiency/FPS
MIDD	114.6	200.0	16.0
CGFNet	233.6	253.4	10.6
MMNet	26.9	251.7	21.0
CAVER	30.3	64.0	11.7
MSEDNet	38.0	369.9	19.7
PATNet	74.7	339.3	16.0
CAR-Net	14.4	59.2	11.8

4.4 Ablation study

To further validate the effectiveness of each CAR-Net's component, several ablation experiments are designed and the results are listed in Table 3. For the backbone (B) module, all the extra modules of the CAR-Net are removed (i. e., CAM and FSM) and only kept the GT supervision for the five layers in backbone to ensure that the model accurately extracts salient features.

1) Effectiveness of CAM

From Table 3, "B + CAM" consistently acquires encouraging results. By inserting CAM, it can select necessary cross-modal information, and key bodies are highlighted from complicated backgrounds. However, without a deeper analysis of the unconformity between multi-modal features, it may introduce inconsistent noises between cross-modalities during direct fusion. Therefore, key cues are extracted from the two dimensions of channel and position and the expression of salient features was enhanced respectively, to integrate effective salient features from various modal information. As listed in the 2nd row of Table 3, the

model with CAM improved detection performance significantly. Compared with the backbone module, the performance gains over the baseline model are 0.2% – 0.6%, 0.001 – 0.007, 0.5% – 1.0%, 0.5% – 1.5% for the metrics S_m, ε, max E_m, and max F_m, respectively, on these three challenging datasets.

Table 3 Ablation results of the CAR-Net

Modle combination	S_m			ε			max E_m			max F_m		
	VT821	VT1000	VT5000	VT821	VT1000	VT5000	VT821	VT1000	VT5000	VT821	VT1000	VT5000
B	0.856	0.920	0.863	0.046	0.025	0.042	0.890	0.953	0.906	0.815	0.908	0.827
B + CAM	0.862	0.920	0.861	0.039	0.023	0.041	0.900	0.958	0.911	0.830	0.913	0.832
B + FSM	0.880	0.920	0.891	0.037	0.020	0.032	0.910	0.965	0.936	0.851	0.926	0.872
B + CAM + FSM	0.887	0.920	0.891	0.037	0.020	0.032	0.920	0.968	0.939	0.868	0.930	0.875

2) Effectiveness of FSM

One of the core detection mechanisms is that highlighting multi-scale features can further benefit accurate SOD. To give evidence for this claim, the FSM is applied to the CAR-Net framework for evaluation. From Table 3, the model with FSM achieves impressive accuracy performance. In addition, employing FSM can help to capture and highlight more details, as shown in Table 4. Furthermore, the S_m score indicates that prominent regions take on more unabridged structures by embedding FSM. It further shows its effectiveness of multi-level aggregating features.

Table 4 Visual maps for different combinations of each module

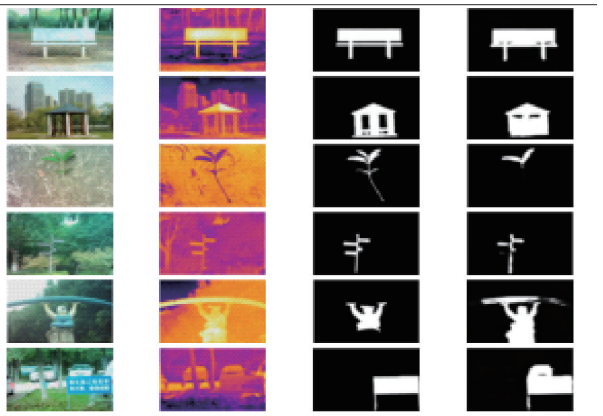
Input		Output				
RGB	Thermal	B	B+CAM	B+FSM	B+CAM+FSM	GT

4.5 Limitation and failure case

Although the CAR-Net is effective for RGB-T SOD, it also has some limitations. From Table 5, it can be observed that when the detected object is small in shape or has serious noise pollution, the detection performs under expectations. In the future, more attention will be paid to the influence of noise and local details, and more effective RGB-T SOD schemes will be explored. Although there have been some failures in scenarios where CAR-Net is highly similar to the background or extremely complex. In the future, all scenarios will be considered and a more applicable model will be proposed.

Table 5 Visualization maps in failure cases

Input		Output	
RGB	Thermal	GT	CAR-Net



5 Conclusions

CAR-Net consists of two modules, CAM and FSM. CAM based on the joint attention mechanism explores the key features of salience from the two directions of channel and position, and then achieves the multi-modal feature fusion in the channel maximum and channel average. Besides, FSM uses the methods of channel splitting and channel ranking, and combines the cycle strategy to fulfill multiple enhancements of salient features. The experimental data of CAR-Net proves that the CAR-Net outperforms other 16 existing methods.

Acknowledgements

The work was supported by the National Natural Science Foundation of China (62471124), the Heilongjiang Province Natural Science Foundation (LH2022F005).

References

[1] LAI B S, GONG X J. Saliency guided dictionary learning for weakly-supervised image parsing. Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR'16), 2016, Jun 27 – 30, Las Vegas, NV, USA. Piscataway, NJ, USA: IEEE, 2016: 3630 – 3639.

[2] AKSOY Y, OH T H, PARIS S, et al. Semantic soft segmentation. ACM Transactions on Graphics, 2018, 37(4): Article 72.

[3] SHEN L Q, LIU Z, ZHANG Z Y. A novel H.264 rate control algorithm with consideration of visual attention. Multimedia Tools and Applications, 2013, 63(3): 709 – 727.

[4] TU Z Z, XIA T, LI C L, et al. M3S-NIR: multi-modal multi-scale noise-insensitive ranking for RGB-T saliency detection. Proceedings of the 2019 IEEE Conference on Multimedia Information Processing and Retrieval (MIPR'19), 2019, Mar 28 – 30, San Jose, CA, USA. Piscataway, NJ, USA: IEEE, 2019: 141 – 146.

[5] WANG W G, SHEN J B, PORIKLI F. Saliency-aware geodesic video object segmentation. Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR'15), 2015, Jun 7 – 12, Boston, MA, USA. Piscataway, NJ, USA: IEEE, 2015: 3395 – 3402.

[6] HOU Q B, CHENG M M, HU X W, et al. Deeply supervised salient object detection with short connections. Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR'17), 2017, Jul 21 – 26, Honolulu, HI, USA. Piscataway, NJ, USA: IEEE, 2017: 3203 – 3212.

[7] LIU D, ZHANG K, CHEN Z Z. Attentive cross-modal fusion network for RGB-D saliency detection. IEEE Transactions on Multimedia, 2020, 23: 967 – 981.

[8] ZHANG J, FAN D P, DAI Y C, et al. Uncertainty inspired RGB-D saliency detection. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2022, 44(9): 5761 – 5779.

[9] XIE C X, XIA C Q, MA M C, et al. Pyramid grafting network for one-stage high resolution saliency detection. Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR'22), 2022, Jun 18 – 24, New Orleans, LA, USA. Piscataway, NJ, USA: IEEE, 2022: 11717 – 11726.

[10] JIN W D, XU J, HAN Q, et al. CDNet: complementary depth network for RGB-D salient object detection. IEEE Transactions on Image Processing, 2021, 30: 3376 – 3390.

[11] LIU Y, HAN J G, ZHANG Q, et al. Salient object detection via two-stage graphs. IEEE Transactions on Circuits and Systems for Video Technology, 2018, 29(4): 1023 – 1037.

[12] ZHANG L H, ZHANG D D, SUN J Y, et al. Salient object detection by local and global manifold regularized SVM mode. Neurocomputing, 2019, 340: 42 – 54.

[13] GAO W, LIAO G B, MA S W, et al. Unified information fusion network for multi-modal RGB-D and RGB-T salient object detection. IEEE Transactions on Circuits and Systems for Video Technology, 2022, 32(4): 2091 – 2106.

[14] TU Z, LI Z, LI C, et al. Multi-interactive dual-decoder for RGB-thermal salient object detection. IEEE Transactions on Image Processing, 2021, 30: 5678 – 5691.

[15] WANG J, SONG K C, BAO Y Q, et al. CGFNet: cross-guided fusion network for RGB-T salient object detection. IEEE Transactions on Circuits and Systems for Video Technology, 2022, 32(5): 2949 – 2961.

[16] ZHAO T, WU X Q. Pyramid feature attention network for saliency detection. Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR'19), 2019, Jun 15 – 20, Long Beach, CA, USA. Piscataway, NJ, USA: IEEE, 2019: 3085 – 3094.

[17] LIU S T, HUANG D, WANG Y H. Receptive field block net for accurate and fast object detection. Proceedings of the 15th European Conference on Computer Vision (ECCV'18): Part XI, 2018, Sep 8 – 14, Munich, Germany. LNCS 11211. Cham, Switzerland: Springer Nature Switzerland AG, 2018: 404 – 419.

[18] YAN P X, WU Z Y, LIU M M, et al. Unsupervised domain adaptive salient object detection through uncertainty-aware pseudo-label learning. Proceedings of the 36th AAAI Conference on Artificial Intelligence (AAAI'22), 2022, Feb 22 – Mar 1, Online. Palo Alto, CA, USA: AAAI, 2022: 3000 – 3008.

[19] YI Y G, ZHANG N Y, ZHOU W, et al. GPONet: a two-stream gated progressive optimization network for salient object detection. Pattern Recognition, 2024, 150: Article 110330.

[20] SUN F M, REN P, YIN B W, et al. CATNet: a cascaded and aggregated transformer network for RGB-D salient object detection. IEEE Transactions on Multimedia, 2023, 26: 2249 – 2262.

[21] GAO L N, LIU B, FU P, et al. TSVT: token sparsification vision transformer for robust RGB-D salient object detection. Pattern Recognition, 2024, 148: Article 110190.

[22] LIANG Y H, QIN G H, SUN M H, et al. Multi-modal interactive attention and dual progressive decoding network for RGB-D/T salient object detection. Neurocomputing, 2022, 490: 132 – 145.

[23] ZHOU W J, GUO Q L, LEI J S, et al. ECFFNet: effective and consistent feature fusion network for RGB-T salient object

detection. IEEE Transactions on Circuits and Systems for Video Technology, 2022, 32(3): 1224 – 1235.

[24] YANG M, LIU Z Y, WU Y, et al. Salient object detection via multi-grained refinement polygon topology positive feedback. Expert Systems with Applications, 2024, 250: Article 123903.

[25] HU X H, SUN F M, SUN J, et al. Cross-modal fusion and progressive decoding network for RGB-D salient object detection. International Journal of Computer Vision 2024, 132(8): 3067 – 3085.

[26] KANWAL S, TAJ I A. CVIT-Net: a conformer driven RGB-D salient object detector with operation-wise attention learning. Expert Systems with Applications, 2023, 225: Article 120075.

[27] NIU Y, ZHOU S P, DONG Y H, et al. Bidirectional feature learning network for RGB-D salient object detection. Pattern Recognition, 2024, 150: Article 110304.

[28] ZENG C, KWONG S, IP H. Dual swin-transformer based mutual interactive network for RGB-D salient object detection. Neurocomputing, 2023, 559: Article 126779.

[29] FAN D P, LIN Z, ZHANG Z, et al. Rethinking RGB-D salient object detection: models, data sets, and large-scale benchmarks. IEEE Transactions on Neural Networks and Learning Systems, 2021, 32(5): 2075 – 2089.

[30] WANG G Z, LI C L, MA Y P, et al. RGB-T saliency detection benchmark: dataset, baselines, analysis and a novel approach. Proceedings of the 13th Conference on Image and Graphics Technologies and Applications (IGTA'18), 2018, Apr 8 – 10, Beijing, China. CCIS 875. Berlin, Germany: Springer, 2018: 359 – 369.

[31] JIN D Z, SHAO F, XIE Z X, et al. CAFCNet: cross-modality asymmetric feature complement network for RGB-T salient object detection. Expert Systems with Applications, 2024, 247: Article 123222.

[32] GAO S H, CHENG M M, ZHAO K, et al. Res2NET: a new multi-scale backbone architecture. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2021, 43(2): 652 – 662.

[33] TU Z Z, XIA T, LI C L, et al. RGB-T image saliency detection via collaborative graph learning. IEEE Transactions on Multimedia, 2020, 22(1): 160 – 173.

[34] TU Z Z, MA Y, LI Z, et al. RGBT salient object detection: a large-scale dataset and benchmark. IEEE Transactions on Multimedia, 2023, 25: 4163 – 4176.

[35] FAN D P, CHENG M M, LIU Y, et al. Structure-measure: a new way to evaluate foreground maps. Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV'17), 2017, Oct 22 – 29, Venice, Italy. Piscataway, NJ, USA: IEEE, 2017: 4548 – 4557.

[36] ACHANTA R, HEMAMI S, ESTRADA F, et al. Frequency-tuned salient region detection. Proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognition, 2009, Jun 20 – 25, Miami, FL, USA. Piscataway, NJ, USA: IEEE, 2009: 1597 – 1604.

[37] PERAZZI F, KRÄHENBÜHL P, PRITCH Y, et al. Saliency filters: contrast based filtering for salient region detection. Proceedings of the 2012 IEEE Conference on Computer Vision and Pattern Recognition, 2012, Jun 16 – 21, Providence, RI, USA. Piscataway, NJ, USA: IEEE, 2012: 733 – 740.

[38] GUO Q L, ZHOU W J, LEI J, et al. TSFNet: two-stage fusion network for RGB-T salient object detection. IEEE Signal Processing Letters, 2021, 28: 1655 – 1659.

[39] ZHOU W J, ZHU Y, LEI J S, et al. APNet: adversarial learning assistance and perceived importance fusion network for all-day RGB-T salient object detection. IEEE Transactions on Emerging Topics in Computational Intelligence, 2022, 6(4): 957 – 968.

[40] MA S, SONG K C, DONG H W, et al. Modal complementary fusion network for RGB-T salient object detection. Applied Intelligence, 2022, 53(8): 9038 – 9055.

[41] HUO F G, ZHU X G, ZHANG L, et al. Efficient context-guided stacked refinement network for RGB-T salient object detection. IEEE Transactions on Circuits and Systems for Video Technology, 2022, 32(5): 3111 – 3124.

[42] HUANG L M, SONG K C, WANG J, et al. Multi-graph fusion and learning for RGBT image saliency detection. IEEE Transactions on Circuits and Systems for Video Technology, 2022, 32(3): 1366 – 1377.

[43] HUO F S, ZHU X G, ZHANG Q, et al. Real-time one-stream semantic-guided refinement network for RGB-thermal salient object detection. IEEE Transactions on Instrumentation and Measurement, 2022, 71: DOI: 10.1109/TIM.2022.3185323.

[44] PANG Y W, ZHAO X Q, ZHANG L H, et al. CAVER: cross-modal view-mixed transformer for bi-modal salient object detection. IEEE Transactions on Image Processing, 2023, 32: 892 – 904.

[45] WU J Y, ZHOU W J, QIAN X H, et al. MFENet: multitype fusion and enhancement network for detecting salient objects in RGB-T images. Digital Signal Processing, 2023, 133: Article 103827.

[46] JIANG M F, MA J H, CHEN J T, et al. PATNet: patch-to-pixel attention-aware transformer network for RGB-D and RGB-T salient object detection. Knowledge-Based Systems, 2024, 291: Article 111597.

[47] PENG D G, ZHOU W Y, PAN J Z, et al. MSEDNet: multi-scale fusion and edge-supervised network for RGB-T salient object detection. Neural Networks, 2024, 171: 410 – 422.

(Editor: Wang Xuying)