Hindawi
Advances in Multimedia
Volume 2023, Article ID 9921988, 14 pages
https://doi.org/10.1155/2023/9921988

Research Article

Coordinate Attention Filtering Depth-Feature Guide Cross-Modal Fusion RGB-Depth Salient Object Detection

Lingbing Meng,[1] **Mengya Yuan**,[1] **Xuehan Shi**,[1] **Qingqing Liu**,[1] **Le Zhange**,[1] **Jinhua Wu**,[1] **Ping Dai**,[1] **and Fei Cheng**[1,2]

[1]*School of Anhui Institute of Information Technology, Wuhu 241199, China*
[2]*School of Hangzhou Dianzi University, Hangzhou 310018, China*

Correspondence should be addressed to Fei Cheng; 1776041825@qq.com

Received 24 October 2022; Revised 14 February 2023; Accepted 28 April 2023; Published 8 May 2023

Academic Editor: Zhongxu Hu

Existing RGB + depth (RGB-D) salient object detection methods mainly focus on better integrating the cross-modal features of RGB images and depth maps. Many methods use the same feature interaction module to fuse RGB and depth maps, which ignores the inherent properties of different modalities. In contrast to previous methods, this paper proposes a novel RGB-D salient object detection method that uses a depth-feature guide cross-modal fusion module based on the properties of RGB and depth maps. First, a depth-feature guide cross-modal fusion module is designed using coordinate attention to utilize the simple data representation capability of depth maps effectively. Second, a dense decoder guidance module is proposed to recover the spatial details of salient objects. Furthermore, a context-aware content module is proposed to extract rich context information, which can predict multiple objects more completely. Experimental results on six benchmark public datasets demonstrate that, compared with 15 mainstream convolutional neural network detection methods, the saliency map edge contours detected by the proposed model have better continuity and the spatial structure details are clearer. Perfect results are achieved on four quantitative evaluation metrics. Furthermore, the effectiveness of the three proposed modules is verified through ablation experiments.

1. Introduction

Salient object detection (SOD) [1–5] aims to locate the most attractive objects in natural scene images and has been widely used in various computer vision tasks, such as image resolution [6], object detection [7], learning-based compression [8], and image quality assessment [9]. In recent years, benefiting from the rapid development of convolutional neural networks (CNNs), SOD has achieved great success. However, when dealing with some challenging scenarios, such as when the contrast between the object and background is low or there are multiple objects in the image, many models have difficulty predicting the objects clearly and completely. Microsoft Kinect sensors and Huawei mobile phones are widely used tools that can capture depth maps easily. Compared with previous models that only used RGB images for training, models with depth maps as auxiliary information can achieve improved detection performance, which has resulted in the development of various RGB + depth (RGB-D) SOD algorithms [10–14]. However, because RGB images and depth maps contain different modal information, it remains challenging to achieve cross-modal feature fusion effectively, which significantly impacts the robustness of the model. Although many previous methods [15–18] have explored cross-modal feature fusion, its application remains limited due to (1) the effects of the RGB image background and (2) the effects of illumination on the RGB image. Regarding (1), RGB images provide rich color information, but the detection accuracy is seriously disturbed by color information. For example, as illustrated in the first row of Figure 1, the consistency of the salient object color and background color causes the model to generate incorrect detection results. The detected object (a chair) is extremely similar in color to the background. A small part of the chair is detected by the 3DCNN [3] and LDCM [15], whereas the rest of the chair is swallowed by the

background. Regarding (2), as shown in the second row of Figure 1, because the images are affected by illumination, the background area is high in brightness, whereas the object area is low in brightness, so 3DCNN and LDCM misjudge the background area as the object area, and the detected area is blurred. Furthermore, although many methods predict the complete object area, such as the carts generated by the 3DCNN and LDCM in the third row of Figure 1, the edge spatial structure details of the salient objects are lost through the upsampling convolution. Although encoder feature maps are usually introduced into the decoder feature map through skip connections to recover the spatial details of salient objects more effectively and the ground truth map is used to supervise the loss of the decoder stage in every layer, it remains impossible to generate more complete detailed features. There are multiple objects in the image, as shown in the fourth row of Figure 1, and the saliency maps predicted by both the 3DCNN and LDCM lose the object and generate only a single object. The combination of the above shows that the detection performance of the model is affected by the color and illumination of RGB images, edge spatial structure details, and number of salient objects.

As a remedy for the aforementioned problems, an RGB-D SOD network is proposed that uses the depth-feature guide cross-modal fusion module with coordinate attention filtering. First, coordinate attention is used to filter invalid information from the depth map and to strengthen the expressive ability of salient objects, which can guide the model to learn more advanced semantic features. It can also better locate the position of salient objects while significantly suppressing the background information interference of RGB images. Second, a dense decode guidance (DDG) module is proposed, which can not only provide a more comprehensive semantic guidance for the encoder features of skip connections but also compensate for the loss of high-level semantic information in the decoder stages, thereby better recovering the structural details of salient objects. Finally, to remove the variation in the number of objects, a context-aware content (CAC) module is designed that aims to explore rich contextual feature information effectively and efficiently as well as to extract the most discriminative salient features. Three encoder-decoder U-nets are jointly trained in an end-to-end manner.

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

(i) To suppress the effects of RGB image color and illumination for model detection, a coordinate attention filtering depth-feature guide cross-modal fusion module is proposed that uses coordinate attention filtering to enhance the feature representation of salient objects in the depth map such that the generated attention map can guide the model to highlight the locations and contour features of objects more prominently.

(ii) A dense decoder guidance module is designed to compensate effectively for the loss of high-level semantic features in the decoder process to restore the edge structural detail features of the salient objects better.

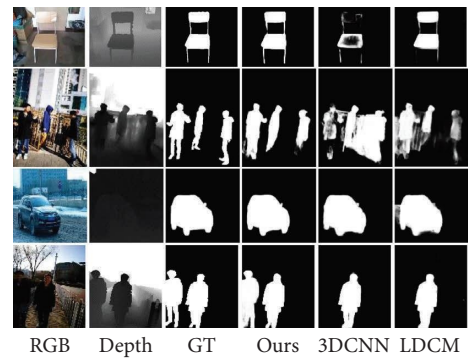

Figure 1: Saliency detection results generated by different models.

(iii) A context-aware content module is designed that can effectively capture rich contextual feature information, which is used to improve the feature capability to enhance the model performance in detecting multiobject scenes.

(iv) Comprehensive experiments on six benchmark datasets with four evaluation metrics demonstrate that, compared to 15 other models, the proposed model has superior detection performance, and the generated saliency map has a better visual effect.

2. Related Work

2.1. Salient Object Detection. With the development and popular application of deep learning, an increasing number of studies [19–22] have utilized deep learning to detect salient objects. Zhao et al. [19] developed a lightweight and real-time model that directly uses the depth map to guide early and middle fusion between an RGB image and the depth map. Sun et al. [20] introduced a depth-sensitive attention module to enhance RGB features effectively, which can utilize the depth geometry feature to reduce background distraction.

Multilevel feature aggregation and cross-modal feature fusion strategies [23–26] are widely used in models to improve detection performance. Wang et al. [23] proposed cross-modality consistency of correlation for RGB-D SOD. Zhang et al. [24] designed a cross-modality discrete interaction network that includes an RGB-induced detail enhancement module and depth-induced semantic enhancement of different layers. Zhou et al. [25] proposed a crossflow and cross-scale adaptive fusion network to detect salient objects in RGB-D images. Other methods have also achieved good results, such as uncertainty learning [27], collaborative learning [28], saliency prior [21], graph neural networks [29], edge detection [30], and transformers [31, 32].

2.2. Attention Mechanisms. Attention mechanisms have been widely used in computer vision tasks, such as visual tracking [33], image classification [34], video question answering [35], person reidentification [36], and image segmentation [37]. Zhang et al. [38] developed a selection

attention mechanism to fuse multimodal information. Chen et al. [4] introduced the channel-wise attention mechanism to achieve a selectively cross-modal cross-level combination. Because the attention mechanism has a strong feature selection ability, its application is well suited to RGB-D SOD [39–41].

Previous methods have directly added or multiplied RGB and depth features when fusing them. The elaborate fusion module also treats RGB features and depth features equally, and only fused features are employed layer by layer in the decoding stage. In this paper, inspired by the above methods, the inherent characteristics of RGB images and depth maps are rethought; moreover, it is argued that the advantages and disadvantages of the inherent characteristics of each modality should be considered in cross-modal feature interaction rather than being treated equally. According to observations, the performance of SOD is greatly affected by the background information in the collected RGB images. Therefore, the performance is reduced by extracting background noise from the RGB features with the network. The objects in the depth map are not disturbed by color; therefore, a CFD module is proposed that uses coordinate attention filtering such that the depth features can effectively suppress the interference of background information, which improves the expressive ability of salient objects. In addition, the three-branch decode structure is adopted in this paper to preserve the original RGB features and depth features for decoding to achieve effective utilization of multimodal features and improve the detection accuracy of the model.

3. Proposed Method

The proposed RGB-D SOD network is shown in Figure 2. In the feature extraction stage, one ordinary convolution is used to reduce the image resolution quickly, and the four residual blocks of the ResNet-50 architecture are used as the subsequent feature extractor, which uses two identical backbone branches to extract the features of the RGB image and the depth map. These extracted features are denoted as F_I^R and F_I^D, respectively, where $I \in \{1, 2, ..., 5\}$ represents the level of feature layers. At the low levels (first and second layers), the RGB and depth feature maps are added to generate a fusion branch feature map. Next, the CFD module is embedded into the higher levels (third, fourth, and fifth layers), and the fusion branch feature map is represented by F_I^{ORD}. For the decoder stage, DDG and CAC modules are designed. Finally, the RGB, depth, and fusion branch streams are designed as three encoder-decoder architectures with the same structure for joint end-to-end training. The final saliency map is generated by the fusion branch stream.

3.1. Depth-Feature Guide Cross-Modal Fusion Module with Coordinate Attention Filtering. RGB images contain rich colors and appearances. Compared with RGB images, depth maps discard complex color information and can intuitively describe the shapes and positions of objects, which means the feature expression ability of objects is provided more directly and effectively. At the low level of the encoder, the detailed features of the object are learned by the model, including the clear boundary, texture, and spatial structure, but these also contain significant background noise. At the high level of the encoder, the features learned by the model contain more semantic information. The high-level semantic features of the depth map are relatively simple; therefore, they can be used to guide the fusion of cross modalities. However, some images exist in which the collected depth maps are of lower quality. Therefore, a CFD module is designed and then embedded in the high levels of the network to make better use of the depth map features. The noise in the depth map is filtered by the coordinate attention, which largely suppresses the nonsalient region features in the RGB image, thereby helping the model locate and identify the salient regions more accurately. The structure is shown in Figure 3.

Specifically, the RGB feature map (F_I^R) and the depth feature map (F_I^D) are fed into a convolutional layer with a kernel size of 3×3 and stride of 1, which are aggregated to generate the feature map (F_I^{RD}) as follows:

$$F_I^{RD} = \mathrm{Cov}\left(F_I^R\right) + \mathrm{Cov}\left(F_I^D\right), \tag{1}$$

where Cov represents the convolutional layer.

Coordinate attention is used to filter the noise of the depth map to utilize the feature information of the depth map more effectively. The coordinate attention module is shown in Figure 4. Specifically, pool kernels of size (H, 1) and (1, W) are selected to encode each channel along the horizontal and vertical coordinate directions for the input depth map, respectively, which correspond to X Avg Pool and Y Avg Pool. Thus, the output features of channel c with height h and width w can be expressed as follows:

$$\begin{aligned} Z_C^h &= \frac{1}{W} \sum_{0 \le p < W} F_I^D(h, p), \\ Z_C^w &= \frac{1}{H} \sum_{0 \le q < H} F_I^D(q, w). \end{aligned} \tag{2}$$

The aforementioned transformations aggregate features in two different directions. Two types of transformations enable the coordinate vector to capture long-distance dependencies in one spatial direction and preserve precise location information in the other spatial direction, which helps the network locate salient objects more accurately.

The coordinate vector is used to generate feature information with a global receptive field and an accurate position to generate coordinate attention maps. The specific operation of generating attention maps is described next.

First, the two feature vectors (Z_C^h and Z_C^w) are concatenated with a 1×1 convolutional layer and then divided into two separate feature maps (Z_H and Z_W) along the spatial dimension. Next, two 1×1 convolutions are used to transform the feature maps Z_H and Z_W to have the same number of channels as the input depth map. The two attention maps are generated using the sigmoid function, expressed as follows:

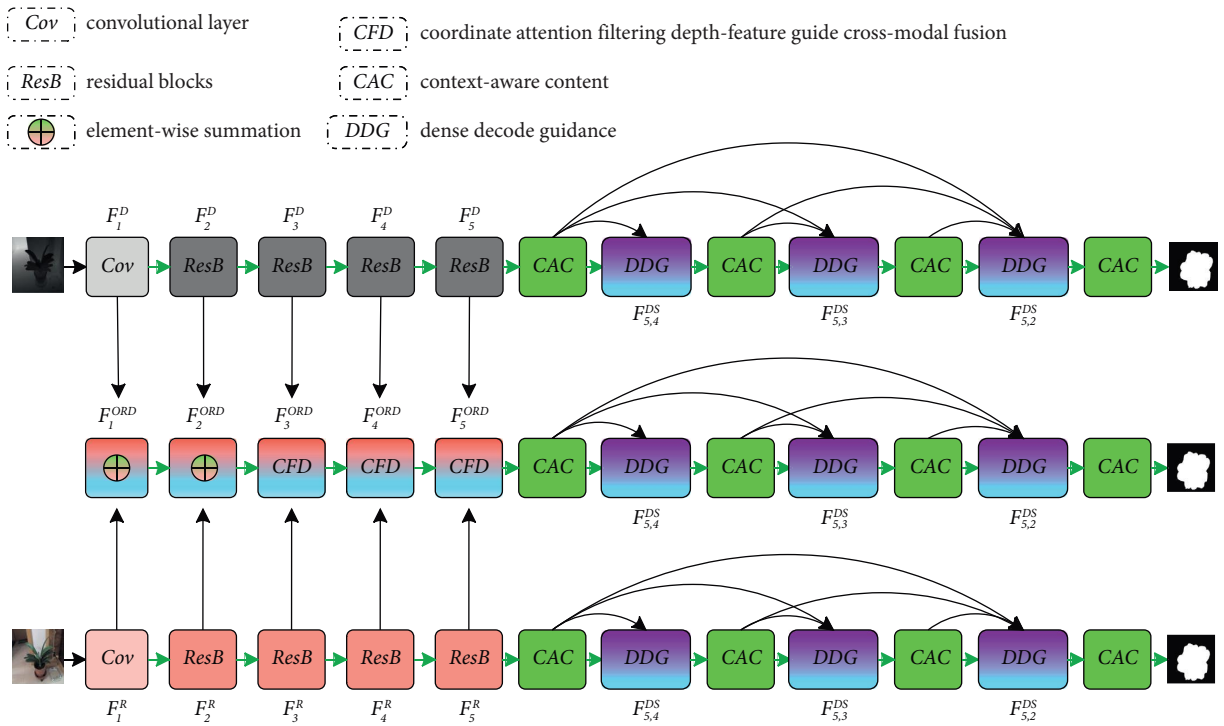

Figure 2: The framework of the coordinate attention filtering depth-feature guide cross-modal fusion RGB-D SOD.

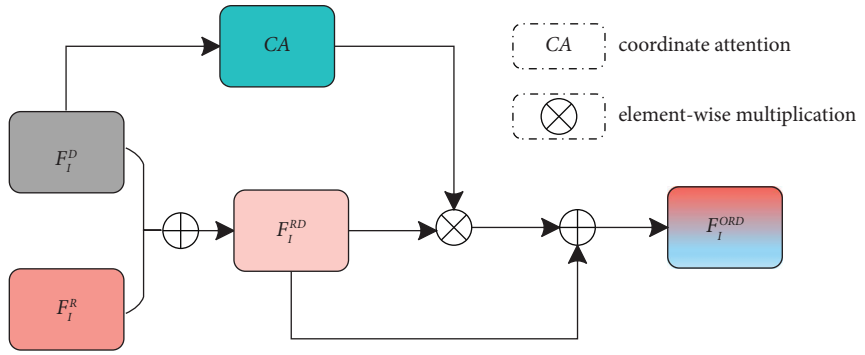

Figure 3: The coordinate attention filtering depth-feature guide cross-modal fusion module.

$$\begin{aligned} Z &= \mathrm{Cov}\left(\mathrm{Concat}\left(Z_C^H, Z_C^W\right)\right), \\ Z_H, Z_W &= \mathrm{Split}\left(Z\right), \\ G_H &= \sigma\left(\mathrm{Cov}\left(Z_H\right)\right), \\ G_W &= \sigma\left(\mathrm{Cov}\left(Z_W\right)\right). \end{aligned} \tag{3}$$

Finally, the two attention maps are multiplied together and added to F_I^{RD} to obtain the enhanced feature map:

$$F_I^{\mathrm{ORD}} = F_I^{RD} \times G_H + F_I^{RD} \times G_W + F_I^{RD}. \tag{4}$$

The CFD module can not only suppress the effects of RGB image color and illumination but also effectively capture the relationship among feature map channels, which guides the effective information interaction among cross-modal features to improve SOD performance.

3.2. Context-Aware Content Module. In the decoder stage, existing methods directly use upsampling convolution to generate the final saliency map. However, for the multiobject case, the same convolutional layer cannot extract distinguishable features, causing the entire object to be lost. Therefore, a CAC module is designed, which aims to explore rich contextual information effectively and efficiently as well as to deal with the changes caused by inconsistent numbers of salient objects more effectively.

The CAC module is shown in Figure 5. Four 3×3 depth-wise convolutions are used, with dilate convolution rates of 1, 3, 5, and 7 to enlarge the receptive field, capturing multiscale features comprehensively. Meanwhile, the number of channels and sizes of all feature maps are kept the same. Subsequently, the input feature map and four feature maps are added as follows to output more discriminative salient features:

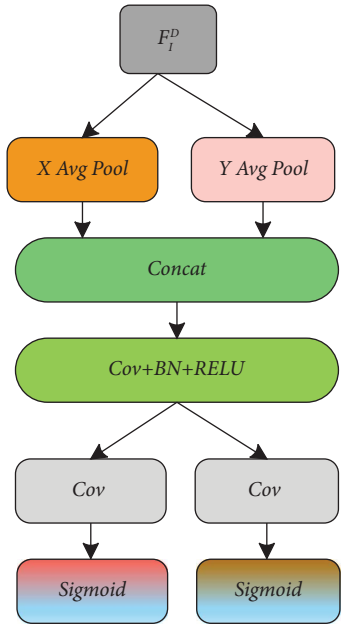

Figure 4: The structure of the coordinate attention module.

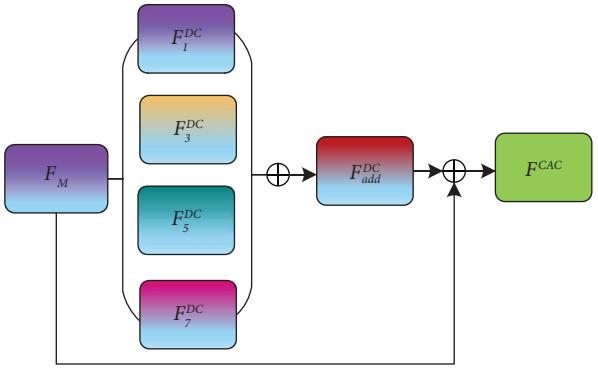

Figure 5: The context-aware content module.

$$
\begin{aligned}
F_1^{DC} &= D\mathrm{Cov}\left(F_M, r = 1\right),\\
F_3^{DC} &= D\mathrm{Cov}\left(F_M, r = 3\right),\\
F_5^{DC} &= D\mathrm{Cov}\left(F_M, r = 5\right),\\
F_5^{DC} &= D\mathrm{Cov}\left(F_M, r = 7\right),\\
F_{\mathrm{add}}^{DC} &= \mathrm{Cov}\left(F_1^{DC} + F_3^{DC} + F_5^{DC} + F_7^{DC}\right),\\
F^{\mathrm{CAC}} &= F_{\mathrm{add}}^{DC} + F_M.
\end{aligned}
\tag{5}
$$

Here, *DCov* and r represent depth-wise convolutions and the dilate factor, respectively; F_M represents the input feature map of the CAC module, and the j-th CAC is designated as F_j^{CAC} and $j \in \{4, 3, 2, 1\}$.

In this way, the CAC module can obtain multiscale information and bring powerful feature representation, which is beneficial for producing multisalient objects and results with high performance.

3.3. Dense Decode Guidance Module. Herein, the intrinsic properties of the depth map are demonstrated as guiding the learning of cross-modal interaction features during the feature encoder stage, whereas the decoder is committed to learning features related to saliency regions and predicting the saliency maps of same size as the ground truth map. The encoder features are introduced into the decoder stage by skip connections, as is common in SOD models. The attention module, which is applied between the encoder and the decoder, is also a popular methodology. However, these methods only establish relationships between the encoder and decoder features of same size, ignoring the effects of different levels of features. As high-level features provide rich semantic information that can provide semantic guidance for each layer of the decoder and compensate for the loss of semantic information in layer-by-layer upsampling, a DDG module is designed to enhance and refine the saliency maps generated by each layer, which better restore the edge structural detail features of the salient objects.

The DDG module considers the RGB branch flow as an example (the other two branch flows adopt the same strategy). First, the encoder feature map (F_I^R) is fed into a 3×3 convolution kernel, and the feature map is output with 256 channels. Similarly, the decoder feature map is

adjusted by convolution operation and upsampling interpolation to obtain a feature map with the same size and same number of channels as the encoder feature map. Finally, the decoder feature map of each layer is multiplied by the encoder feature map and concatenated to be sent to the CAC module. The entire process can be formulated as follows:

$$\begin{aligned} F_{5,4}^{DS} &= \mathrm{Conat}\left(\mathrm{CAC}\left(\mathrm{up}\left(F_5^R\right)\right), F_4^R\right), \\ F_{5,3}^{DS} &= \mathrm{Concat}\begin{pmatrix} \mathrm{CAC}\left(\mathrm{up}\left(F_{5,4}^{DS}\right)\right) \times F_3^R \\ \mathrm{CAC}\left(\mathrm{up}_{\times 2}\left(F_5^R\right)\right) \times F_3^R \end{pmatrix}, \\ F_{5,2}^{DS} &= \mathrm{Concat}\begin{pmatrix} \mathrm{CAC}\left(\mathrm{up}\left(F_{5,3}^{DS}\right)\right) \times F_2^R, \\ \mathrm{CAC}\left(\mathrm{up}_{\times 2}\left(F_{5,4}^{DS}\right)\right) \times F_2^R \\ \mathrm{CAC}\left(\mathrm{up}_{\times 4}\left(F_5^R\right)\right) \times F_2^R \end{pmatrix}, \end{aligned} \tag{6}$$

where up() represents bilinear interpolation, and the subscript numbers represent the upsampling times.

3.4. Loss Function. The binary cross entropy (BCE) and intersection over union (IoU) loss functions are often used to optimize SOD models.

The BCE loss function can be expressed as follows:

$$l^{\mathrm{bce}} = -\frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} \left| G_{i,j} \log\left(P_{i,j}\right) + \left(1 - G_{i,j}\right) \log\left(1 - P_{i,j}\right) \right|. \tag{7}$$

Moreover, the IoU loss function is defined as follows:

$$l^{\mathrm{iou}} = 1 - \frac{\sum_{\mathbf{i}=1}^{H} \sum_{j=1}^{W} P_{i,j} \cdot G_{i,j}}{\sum_{\mathbf{i}=1}^{H} \sum_{j=1}^{W} \left(P_{i,j} + G_{i,j} - P_{i,j} \cdot G_{i,j}\right)}, \tag{8}$$

where H and W represent the width and height of the image, respectively. The subscripts i and j represent the pixel value coordinates. Additionally, P and G represent the predicted saliency map and the ground truth map, respectively.

BCE and IoU are combined for the optimization loss function of the proposed model:

$$l = l^{\mathrm{bce}} + l^{\mathrm{iou}}. \tag{9}$$

The auxiliary loss function is used to optimize the model in the decoding stage and to prevent gradient vanishing during the training process. Specifically, a 3×3 convolutional layer is applied to the feature map of each layer in the decoder stage to convert the input feature map with 256 channels into a feature map with 1 channel. Simultaneously, the feature map is bilinearly interpolated to the same scale as the ground truth map, and the sigmoid function is used to normalize the generated saliency map.

Next, the loss functions of the three branch streams are as follows:

$$\begin{aligned} l^{\mathrm{RGB}} &= l_F^{4R} + 0.8 \times l_F^{3R} + 0.6 \times l_F^{2R} + 0.4 \times l_F^{1R}, \\ l^{\mathrm{de\,pth}} &= l_F^{4d} + 0.8 \times l_F^{3d} + 0.6 \times l_F^{2d} + 0.4 \times l_F^{1d}, \\ l^{Rd} &= l_F^{4Rd} + 0.8 \times l_F^{3Rd} + 0.6 \times l_F^{2Rd} + 0.4 \times l_F^{1Rd.} \end{aligned} \tag{10}$$

Therefore, the total loss function of the model is as follows:

$$l^{\mathrm{total}} = l^{\mathrm{RGB}} + l^{\mathrm{de\,pth}} + l^{Rd}, \tag{11}$$

where l^{RGB}, l^{Depth}, and l^{Rd} represent the loss functions of the RGB, deep, and fusion branch streams, respectively. l_F^{iR}, l_F^{id}, and l_F^{iRd} represent the CAC feature map of the i-th layer in the RGB, deep, and fusion branch streams, respectively.

4. Experiments and Results

4.1. Dataset. To verify the effectiveness of the proposed model, experiments were performed on six public datasets: NJU2K [42], DES [43], NLPR [44], SSD [45], DUT-RGBD [46], and SIP [47]. NJU2K contains 1985 image pairs collected from the Internet and 3D movies. The DES (RGBD135) dataset contains 135 RGB-D image pairs from seven indoor locations. The NLPR dataset consists of 1000 image pairs collected by Kinect from 11 different scenes, including more than 400 kinds of common objects. The SIP dataset contains 1000 image pairs collected by smartphones with camera resolutions of 992×744. The SSD dataset contains 80 images extracted from three stereoscopic movies for which the depth maps are generated by the depth estimation method. The DUT-RGBD dataset includes 1200 indoor and outdoor complex scenes, of which 800 and 400 image pairs are used for training and testing, respectively.

4.2. Evaluation Metrics. In this paper, the maximum F-measure ($F_\beta^{\max}$) [48], maximum E-measure ($E_\varphi^{\max}$) [49], S-measure (S_α) [50], and the mean absolute error (M) [51] are used as evaluation metrics. F_β is proposed to consider the importance of precision and recall in a comprehensive manner. Its calculation formula is as follows:

$$F_\beta = \frac{\left(1 + \beta^2\right) \times \mathrm{precision} \times \mathrm{recall}}{\beta^2 \times \mathrm{precision} + \mathrm{recall}}, \tag{12}$$

where $\beta^2 = 0.3$, and the maximum F-measure is denoted as $F_\beta^{\max}$.

M is the average of the absolute errors between the predicted saliency map and the ground truth map:

$$M = \frac{1}{H \times W} \sum_{\mathbf{i}=1}^{H} \sum_{\mathbf{j}=1}^{W} |P(i,j) - G(i,j)|. \tag{13}$$

S_α calculates the structural similarity between object-aware and region-aware:

$$S_\alpha = \alpha S_O + (1 - \alpha) S_r, \tag{14}$$

where S_O and S_r represent object and region awareness, respectively. Typically, α is set to 0.5. The larger the value of S_α, the more similar are the saliency and ground truth maps in their spatial structures.

E_φ calculates the local pixel-level and global image-level errors and is defined as follows:

$$E_\varphi = \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} \phi FM(i, j), \tag{15}$$

where ϕFM denotes the enhanced alignment matrix.

4.3. Experimental Details. The network in this study was implemented using the deep learning framework PyTorch, and the model was executed on a machine with an Nvidia RTX 3090 GPU. There are 1985 image pairs in the NJU2K dataset, of which 1485 and 500 images are used for training and testing, respectively. There are 1000 image pairs in the NLPR dataset, of which 700 and 300 images are used for training and testing, respectively. In particular, when the DUT-RGBD dataset is tested, an additional 800 DUT-RGBD image pairs are supplemented for training. The Adam optimizer is used to optimize the model, and the batch size is set to 10. The initial learning rate is set to 0.0001 and updated every two iterations with a decay rate of 0.9. All training and testing images are resized to 352×352. To prevent the model from overfitting, the optimal model is selected based on the validation dataset (800 image pairs), and the model saves the best result in 126 epochs of training, taking approximately 10 h. The proposed model does not require any preprocessing or postprocessing.

4.4. Experimental Comparison. The proposed method was compared with 15 state-of-the-art CNN-based RGB-D models: SSP [52], EENet [53], LDCM [15], DSN [54], 3DCNN [3], CDNet [14], CMDI [24], CCAF [25], DSAM [20], BiANet [10], DQFM [55], DCF [56], JLDCF [57], and ICNet [5]. For a fair comparison, all saliency maps were directly obtained from the original author or generated from the train model provided by the original author.

The results of various SOD methods on the six datasets are listed in Table 1. According to the experimental results, the proposed method notably outperforms the other methods in multiple metrics. Compared with the other methods, on the SIP and NLPR datasets, the proposed method is superior in all metrics. For example, compared with the second-best model (SSP) on the SIP dataset, $F_\beta^{\max}$, $E_\varphi^{\max}$, S_α, and M are improved by 0.000, 0.004, 0.001, and 0.001, respectively. The proposed model is also compared with the 3DCNN model on the NLPR dataset, with $F_\beta^{\max}$, $E_\varphi^{\max}$, S_α, and M improving by 0.002, 0.002, 0.004, and 0.002, respectively. On the DUT-RGBD dataset, the proposed and 3DCNN methods both added an additional 800 image pairs for training, giving the same M, but the proposed method outperforms the 3DCNN in terms of $F_\beta^{\max}$ and $E_\varphi^{\max}$. In addition, for the proposed method on the SSD dataset, except for M, the other three metrics are far lower than those of the DSN method, which is caused by the low quality of the depth map. Because the proposed model relies on the quality of the depth map, the detection performance on the SSD dataset is relatively weak. However, a comprehensive analysis of all datasets and evaluation metrics demonstrates that the proposed detection method is better than the other methods.

The precision-recall (PR) and F-measure curves are illustrated in Figure 6. Note that the proposed model achieves both better precision and recall than the other models. Some visual saliency map results for the proposed and nine other methods are shown in Figure 7. Next, several specific challenging cases are summarized. When the background information is similar to the color of the object (first, fifth, and sixth rows), many models only detect a portion of salient objects; in contrast, the proposed model performs well and can detect salient objects clearly and completely. Additionally, for scenes with extremely low brightness (seventh, eighth, and ninth rows), which is a very challenging situation, the shadow between the legs of the person in the eighth row is not detected by the other methods, but the proposed method can detect the complete object in low-light scenes. This finding demonstrates that the CFD module can use depth features to differentiate the object region clearly from a similar background, whereas the objects detected by other methods are submerged into the background.

Low-contrast and multiobject scenarios are also shown in the bottom three rows, in which the other methods wrongly miss objects when dealing with such cases. For example, there are two objects in the penultimate row of the image, but many methods can detect one person only, whereas the proposed method detects two objects completely. This finding shows that the CAC module can effectively capture rich contextual feature information and improve the detection performance in multiobject scenarios. From the displayed visualization results, the saliency map generated by the proposed method verifiably has a finer spatial structure, which indicates that the DDG module effectively makes the salient object more uniform and clearer. On the whole, the objects detected by the proposed method are more complete, the texture is clearer, and the boundary contour is more prominent. The proposed model gives better results visually, and the generated saliency map is closer to the ground truth map.

4.5. Ablation Experiment. Ablation experiments were mainly conducted to prove the effectiveness of each module, and the experimental results on the NLPR and SIP datasets are listed in Table 2.

4.6. Effectiveness of CFD. In the feature encoder stage, an add operation is used instead of the CFD module to concatenate the RGB and depth modalities. Specifically, for the feature maps of the two modalities of F_I^R and F_I^D, the enhanced feature map is F_I^{RD} ($F_I^{RD} = F_I^R + F_I^D$), which is denoted as w/o CFD in Table 2. Considering the experimental results on the SIP and NLPR datasets, the proposed CFD reduces M by 0.002 and improves S_α by 0.003 and 0004, respectively. This finding proves that the model detection performance can be

Table 1: Benchmark results of 16 deep learning models on six public RGB-D saliency detection datasets using four common evaluation metrics.

Method	DES				DUT-RGBD				NJU2K				NLPR				SIP				SSD			
	$F_\beta^{\max}$	$E_\xi^{\max}$	S_α	M	$F_\beta^{\max}$	$E_\xi^{\max}$	S_α	M	$F_\beta^{\max}$	$E_\xi^{\max}$	S_α	M	$F_\beta^{\max}$	$E_\xi^{\max}$	S_α	M	$F_\beta^{\max}$	$E_\xi^{\max}$	S_α	M	$F_\beta^{\max}$	$E_\xi^{\max}$	S_α	M
Ours	*0.944*	0.973	0.933	**0.017**	**0.947**	**0.960**	0.928	**0.029**	*0.929*	0.950	*0.919*	**0.032**	**0.932**	**0.967**	**0.934**	**0.020**	**0.908**	**0.931**	**0.889**	**0.045**	*0.882*	*0.914*	0.870	**0.044**
SSP	0.943	**0.978**	0.936	**0.017**	*0.946*	0.957	*0.931*	**0.029**	0.922	0.942	0.909	0.039	0.922	0.961	0.922	0.025	**0.908**	*0.927*	*0.888*	*0.046*	—	—	—	—
DSN	0.939	0.967	0.928	0.021	—	—	—	—	**0.930**	**0.952**	**0.921**	*0.034*	0.927	0.958	0.926	0.024	0.902	0.917	0.876	0.052	**0.895**	**0.922**	**0.885**	*0.045*
3DCNN	0.941	0.972	0.935	0.019	0.939	*0.958*	**0.932**	**0.029**	0.923	0.947	0.915	0.037	*0.927*	*0.965*	0.930	*0.022*	0.906	0.924	0.885	0.048	0.873	0.904	0.872	0.048
CDNet	0.944	0.973	**0.940**	0.019	0.845	0.872	0.830	0.071	0.892	0.927	0.885	0.048	*0.927*	0.964	*0.931*	0.024	0.896	0.914	0.872	0.056	0.831	0.900	0.844	0.060
CMDI	0.943	0.970	0.937	0.020	0.945	0.957	0.927	**0.029**	0.928	*0.951*	*0.919*	0.035	0.923	0.960	0.927	0.024	0.904	0.915	0.875	0.054	0.868	0.899	0.853	0.056
LDCM	—	—	—	—	0.939	0.956	0.928	0.034	0.922	0.947	0.909	0.046	0.913	0.956	0.922	0.029	0.889	0.911	0.870	0.062	0.880	0.921	0.882	0.054
CCAF	*0.944*	*0.974*	*0.938*	*0.018*	0.926	0.942	0.904	0.037	0.921	0.944	0.910	0.038	0.918	0.957	0.922	0.027	0.900	0.917	0.877	0.054	0.790	0.861	0.786	0.075
DSAM	0.930	0.954	0.917	0.023	0.940	0.956	0.921	*0.030*	0.917	0.938	0.904	0.040	0.916	0.952	0.919	0.024	0.891	0.912	0.862	0.057	0.877	0.913	*0.877*	0.048
BiANet	0.939	0.968	0.930	0.021	—	—	—	—	0.930	0.948	0.916	0.036	0.924	0.962	0.926	0.023	0.904	0.926	0.887	0.047	0.872	0.911	0.863	0.048
HAIN	**0.945**	0.973	0.935	*0.018*	0.932	0.943	0.909	0.038	0.925	0.944	0.912	0.038	0.924	0.960	0.924	0.024	*0.907*	0.922	0.880	0.053	0.858	0.903	0.857	0.052
DQFM	0.930	0.973	*0.938*	0.019	0.784	0.844	0.792	0.090	0.898	0.946	0.908	0.043	0.916	0.961	0.925	0.024	0.904	0.926	0.885	0.049	0.810	0.877	0.815	0.072
EENet	0.933	0.970	0.929	0.021	*0.946*	0.948	0.895	0.044	0.914	0.950	0.910	0.038	0.916	0.961	0.920	0.025	0.898	0.916	0.873	0.053	0.865	0.914	0.862	0.052
DCF	0.909	0.951	0.905	0.024	*0.946*	0.898	0.837	0.070	0.923	0.950	0.912	0.036	0.918	0.963	0.924	*0.022*	0.899	0.922	0.876	0.052	0.867	0.909	0.864	0.050
JLDCF	0.934	0.968	0.931	0.020	0.872	0.898	0.837	0.070	0.911	0.937	0.893	0.045	0.925	0.963	0.925	*0.022*	0.904	0.924	0.880	0.049	0.797	0.870	0.808	0.086
ICNet	0.925	0.960	0.920	0.027	0.874	0.898	0.837	0.070	0.901	0.924	0.894	0.052	0.919	0.952	0.923	0.028	0.873	0.903	0.854	0.069	0.856	0.902	0.848	0.064

Note. Bold represents the best result, and italic represents the second-best result.

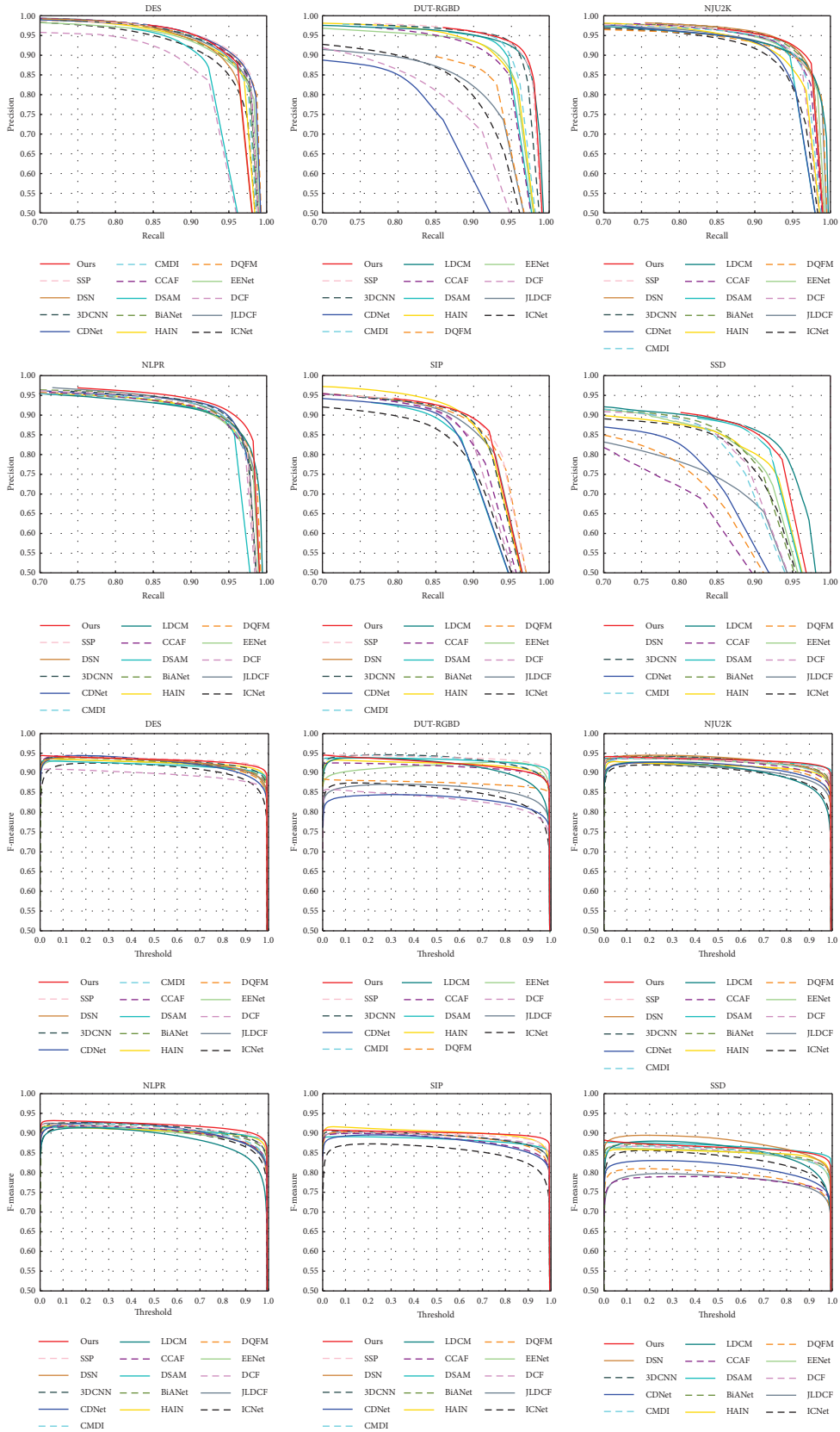

Figure 6: PR and F-measure curves of different models on six datasets.

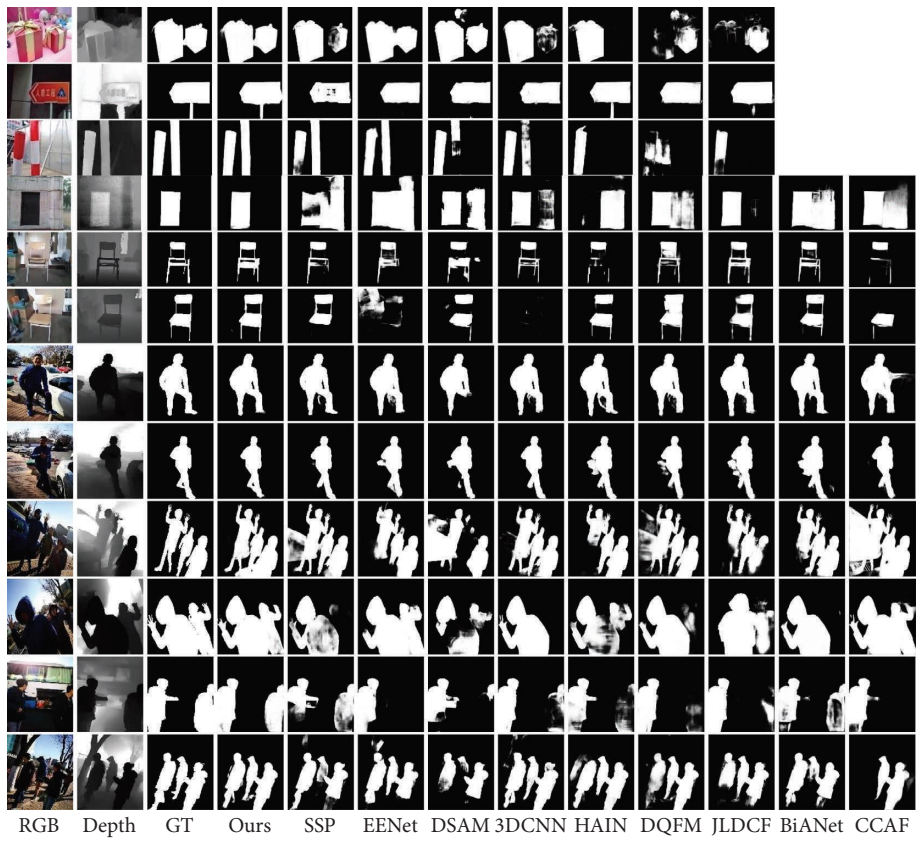

Figure 7: Comparison of saliency maps of our model with those of other RGB-D SOD models.

Table 2: Ablation results of different components on the SIP and NLPR datasets.

Model	SIP				NLPR			
	$F_\beta^{\max}$	$E_\xi^{\max}$	S_α	M	$F_\beta^{\max}$	$E_\xi^{\max}$	S_α	M
Ours	0.908	0.931	0.889	0.045	0.932	0.967	0.934	0.020
w/o CFD	0.901	0.927	0.886	0.047	0.929	0.963	0.930	0.022
w/o DDG	0.903	0.927	0.887	0.046	0.928	0.964	0.928	0.021
w/o CAC	0.904	0.930	0.886	0.046	0.928	0.965	0.931	0.021
w/o D	0.905	0.927	0.886	0.046	0.923	0.960	0.928	0.024
w/o R	0.904	0.928	0.886	0.047	0.927	0.961	0.927	0.022
w/o RD	0.899	0.920	0.869	0.053	0.924	0.961	0.926	0.023

improved when using the CFD module instead of simply adding feature maps. Some visual results are shown in Figure 8. Without the CFD module, the models predict the background information as salient objects of varying degrees for illumination effects (first row), salient objects consistent with background information (second row), and complex background (third row). The model that uses the CFD module can accurately predict the salient objects, which effectively suppresses the influence of background information and accurately generates the salient objects.

4.7. Effectiveness of DDG. In addition, in the feature decoder stage, the DDG module is deleted, like in the U-net method, and only the encoder and decoder feature maps of the same scale are concatenated. This map is referred to as w/o DDG in Table 2. The DDG module improves $E_\varphi^{\max}$ by 0.004 and 0.003 on the SIP and NLPR datasets, respectively. The visualization results are shown in Figure 9. Considering the saliency map, note that without the help of the DDG module, although the salient object can be accurately detected, the spatial structure is not sufficiently clear. With the help of the DDG module, the model generates clearer salient objects with more detailed spatial structures.

4.8. Effectiveness of CAC. To verify the effectiveness of the CAC module, the CAC module is replaced with a 3×3 convolutional layer, which is denoted as w/o CAC in Table 2. On the SIP dataset, $F_\beta^{\max}$, $E_\varphi^{\max}$, and S_α increase by 0.004, 0.001, and 0.003, respectively; moreover, M decreases by 0.001. The four metrics also have different degrees of improvement on the NLPR dataset. The visual saliency map for comparison is shown in Figure 10. For multiobject scenes,

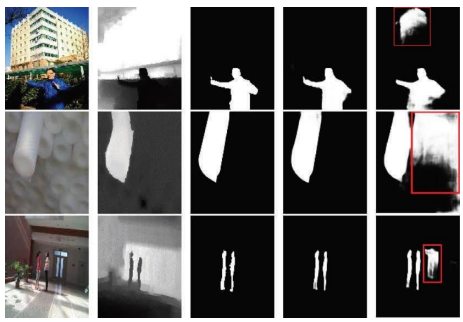

Figure 8: Visual result comparison between our full model and our model without the CFD module. The red boxes indicate the differences between the two models.

Figure 9: Visual result comparison between our full model and our model without the DDG module. The red boxes indicate the differences between the two models.

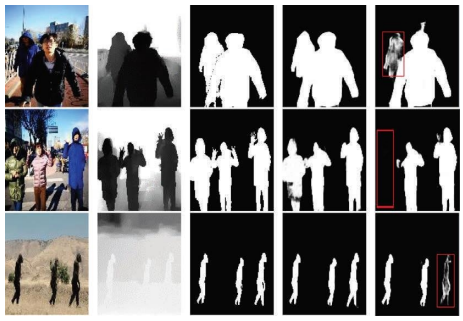

Figure 10: Visual result comparison between our full model and our model without the CAC module. The red boxes indicate the differences between the two models.

the saliency map generated by the model without the CAC module either has missing objects or the objects are considerably blurred and incomplete. However, the salient objects generated by the proposed method are more complete, which shows that the CAC module can effectively explore rich contextual information.

4.9. Effectiveness of the Three Branch Streams. The effectiveness of jointly training the three branch streams is also verified in the decoder stage. First, the RGB and depth branch streams are removed from the decoder, which only keeps the fusion branch stream for training, denoted as w/o RD. The experimental results show that when the RGB and depth branch flows are removed for training, the detection performance is greatly reduced, which indicates that the detection performance is considerably degraded when only fusion branch flow is used. For example, on the SIP dataset, compared with the full method, $E_{\varphi}^{\max}$ and S_{α} are reduced by 0.011 and 0.020, respectively. Second, the depth and RGB branch stream are removed separately, leaving only the remaining two branch streams, denoted as w/o R and w/o D, respectively. Note that whether the model with the deep or RGB branch streams is removed, the detection metrics are lower than those of the full model, which indicates that training three branch streams together produces the best results.

4.10. Effectiveness of Our Model on the Three Datasets. Additionally, image pairs were specifically collected for low-illumination scenes, complex backgrounds, and multiobject scenes. There are 122 image pairs for the low-light scenes, all collected from the SIP dataset, defined as the low-illumination (LI) dataset. A total of 255 image pairs with complex backgrounds were collected from the NLPR dataset, called the complex background (CB) dataset. The multiobject (MO) dataset was collected from the NLPR and SIP datasets and contains 38 and 327 image pairs from NLPR and SIP, respectively. The experimental results are listed in Table 3. Compared with the other methods, the proposed method showcases better detection performances in these three scenarios, far ahead of other methods in terms of M, which further verifies the effectiveness of each proposed module. All models find it more difficult to detect salient objects effectively in multiobject scenes, which confirms the need to improve the performance of multiobject detection in RGB-D SOD.

4.11. Failure Cases and Analyses. As mentioned above, the results of the quantitative and qualitative evaluations demonstrate the superiority and effectiveness of the proposed method. However, the proposed method still has limitations in some cases. Some detection failures of the saliency maps are shown in Figure 11. It can be seen that the quality of the depth maps is very low, which not only makes it difficult to characterize the salient objects but also causes a lot of noise information. We can see that although the object location is correctly predicted in the first and third rows, redundant and erroneous object regions are generated due to being affected by the noise of the depth map. As can be seen from the second row, the locations of salient objects in the RGB image are not obvious, and the depth map makes

Table 3: M and S_α metrics of different models on three datasets.

Method	LI		MO		CB	
	M	S_α	M	S_α	M	S_α
Ours	0.047	0.886	0.068	0.844	0.020	0.939
SSP	0.057	0.874	0.069	0.847	0.026	0.926
DSN	0.048	0.884	0.082	0.812	0.023	0.931
3DCNN	0.054	0.875	0.074	0.836	0.023	0.933
CDNet	0.062	0.862	0.092	0.804	0.025	0.934
CMDI	0.064	0.865	0.087	0.812	0.024	0.932
LDCM	0.068	0.866	0.089	0.821	0.028	0.930
CCAF	0.062	0.864	0.085	0.825	0.027	0.925
DSAM	0.071	0.848	0.095	0.789	0.023	0.926
BiANet	0.055	0.873	0.073	0.839	0.023	0.931
HAIN	0.051	0.886	0.087	0.817	0.024	0.928
DQFM	0.051	0.882	0.072	0.837	0.024	0.931
EENet	0.054	0.877	0.081	0.818	0.026	0.924
DCF	0.060	0.866	0.083	0.820	0.021	0.930
JLDCF	0.058	0.868	0.081	0.819	0.021	0.931
ICNet	0.077	0.845	0.108	0.786	0.026	0.931

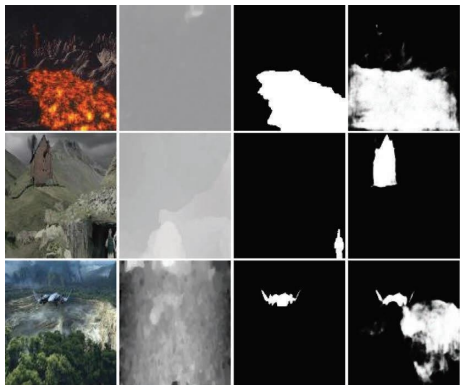

Figure 11: False cases of RGB-D SOD.

it difficult to provide effective saliency features, which leads the model to misclassify the prominent background area as the salient area. In summary, the proposed method is not effective in generating objects with low-quality depth maps. Now that the attention map generated by the depth map is used in the feature encoder stage to guide the generation of cross-modal features, low-quality depth maps can interfere with the generation of valid cross-modal saliency features, which can cause the model to produce incorrect object regions. To address the problem of low-quality depth maps, a depth map quality score can be used to determine the proportion of depth maps in the model, and the detection performance can be further improved by preprocessing.

5. Conclusion

In this paper, a novel depth-feature guide cross-modal fusion method for RGB-D SOD is proposed. Unlike most previous works, which mostly focused on learning to fuse cross modalities, the proposed model is based on depth maps of inherent simplicity, which guide the learning of shared modal information to improve the detection performance. In addition, the proposed DDG module can effectively recover the spatial detail structure features of salient objects, and the CAC module achieves effective multiobject detection by extracting rich contextual information. Quantitative and qualitative evaluations on six challenging benchmark datasets demonstrate that the proposed model outperforms the existing RGB-D SOD methods.

Data Availability

The data used to support the findings of this study are available from the corresponding author upon request. All images used in this study are publicly available and have been approved by the publisher, and the datasets are available at https://github.com/lartpang/awesome-segmentation-saliency-dataset.

Conflicts of Interest

The authors declare that there are no conflicts of interest.

Authors' Contributions

Lingbing Meng wrote the original draft and was involved in the methodology. Mengya Yuan wrote the original draft. Xuehan Shi wrote the review, reviewed the manuscript, and acquired funding. Qingqing Liu wrote the review, reviewed the manuscript, acquired funding, and investigated the study. Le Zhang wrote the review and reviewed the manuscript. Jinhua Wu acquired funding, was responsible for software, and validated the study. Ping Dai wrote the review and reviewed the manuscript. Fei Cheng acquired funding, curated the data, and supervised the study.

Acknowledgments

This work was supported by the General Project of the Natural Science Foundation of Anhui Province, China (2008085MF201); the General Project of Anhui Philosophy and Social Sciences Planning, China (AHSKY2021D142);

the Natural Science Research Project of Anhui University (KJ2020a0824, 2022AH051887, and 2022AH051894); the Advanced Talent Scientific Research Project of Anhui Institute of Information Technology (rckj2021A002); the Support Program for Outstanding Young Talents in Colleges and Universities (gxyq2022147); and the Scientific and Technological Innovation 2030 Major Project (2020AAA0103600).

References

[1] R. Cong, Y. Zhang, L. Fang et al., "RRNet: relational reasoning network with parallel multiscale attention for salient object detection in optical remote sensing images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–11, 2022.

[2] H. Mei, Y. Liu, Z. Wei et al., "Exploring dense context for salient object detection," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 32, no. 3, pp. 1378–1389, 2022.

[3] Q. Chen, Z. Liu, Y. Zhang, K. Fu, Q. Zhao, and H. Du, "RGB-D salient object detection via 3D convolutional neural networks," in *Proceedings of the AAAI conference on artificial intelligence*, pp. 1–9, California CA USA, May 2021.

[4] Z. Li, C. Lang, L. Liang et al., "Dense attentive feature enhancement for salient object detection," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 32, no. 12, pp. 8128–8141.

[5] G. Li, Z. Liu, and H. Ling, "ICNet: information conversion network for RGB-D based salient object detection," *IEEE Transactions on Image Processing*, vol. 29, pp. 4873–4884, 2020.

[6] M. Jalwana, N. Akhtar, M. Bennamoun, and A. Mian, "CAMERAS: enhanced resolution and sanity preserving class activation mapping for image saliency," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 16327–16336, Nashville, TN, USA, June 2021.

[7] Y. Pang, L. Ye, X. Li, and J. Pan, "Incremental learning with saliency map for moving object detection," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 28, no. 3, pp. 640–651, 2018.

[8] D. Mishra, S. Singh, R. Singh, and D. Kedia, "Multi-scale network (MsSG-CNN) for joint image and saliency map learning-based compression," *Neurocomputing*, vol. 460, no. 14, pp. 95–105, 2021.

[9] K. Gu, S. Wang, H. Yang et al., "Saliency-Guided quality assessment of screen content images," *IEEE Transactions on Multimedia*, vol. 18, no. 6, pp. 1098–1110, 2016.

[10] Z. Zhang, Z. Lin, J. Xu, W.-D. Jin, S.-P. Lu, and D.-P. Fan, "Bilateral attention network for RGB-D salient object detection," *IEEE Transactions on Image Processing*, vol. 30, pp. 1949–1961, 2021.

[11] J. Zhang, D. Fan, Y. Dai et al., "RGB-D saliency detection via cascaded mutual information minimization," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4338–4347, Montreal, BC, Canada, October 2021.

[12] K. Fu, D. Fan, G. Ji, Q. Zhao, J. Shen, and C. Zhu, "Siamese network for RGB-D salient object detection and beyond," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 9, pp. 5541–5559, 2021.

[13] K. Yi, J. Zhu, F. Guo, and J. Xu, "Cross-stage multi-scale interaction network for RGB-D salient object detection," *IEEE Signal Processing Letters*, vol. 29, pp. 2402–2406, 2022.

[14] W. Jin, J. Xu, Q. Han, Y. Zhang, and M. Cheng, "CDNet: complementary depth network for RGB-D salient object detection," *IEEE Transactions on Image Processing*, vol. 30, pp. 3376–3390, 2021.

[15] F. Wang, J. Pan, S. Xu, and J. Tang, "Learning discriminative cross-modality features for RGB-D saliency detection," *IEEE Transactions on Image Processing*, vol. 31, pp. 1285–1297, 2022.

[16] Z. Huang, H. Chen, T. Zhou, Y. Yang, and B. Liu, "Multi-level cross-modal interaction network for RGB-D salient object detection," *Neurocomputing*, vol. 452, no. 10, pp. 200–211, 2021.

[17] H. Chen, Y. Li, and D. Su, "Multi-modal fusion network with multi-scale multi-path and cross-modal interactions for RGB-D salient object detection," *Pattern Recognition*, vol. 86, pp. 376–385, 2019.

[18] J. Han, H. Chen, N. Liu, C. Yan, and X. Li, "CNNs-based RGB-D saliency detection via cross-view transfer and multiview fusion," *IEEE Transactions on Cybernetics*, vol. 48, no. 11, pp. 3171–3183, 2018.

[19] X. Zhao, L. Zhang, Y. Pang, H. Lu, and L. Zhang, "A single stream network for robust and real-time RGB-D salient object detection," in *Proceedings of the ECCV 2020: 16th European Conference Computer Vision*, pp. 646–662, Glasgow, UK, August 2020.

[20] S. Sun, C. Feng, S. Tong, Y. Zhao, N. Chen, and M. Zhu, "Evaluation of advanced phosphorus removal from slaughterhouse wastewater using industrial waste-based adsorbents," *Water Science and Technology: A Journal of the International Association on Water Pollution Research*, vol. 83, no. 6, pp. 1407–1417, 2021.

[21] N. Huang, Y. Yang, D. Zhang, Q. Zhang, and J. Han, "Employing bilinear fusion and saliency prior information for RGB-D salient object detection," *IEEE Transactions on Multimedia*, vol. 24, pp. 1651–1664, 2022.

[22] Z. Gao, C. Xu, H. Zhang, S. Li, and V. H. C. de Albuquerque, "Trustful Internet of surveillance things based on deeply represented visual Co-saliency detection," *IEEE Internet of Things Journal*, vol. 7, no. 5, pp. 4092–4100, 2020.

[23] N. Wang and X. Gong, "Adaptive fusion for RGB-D salient object detection," *IEEE Access*, vol. 7, pp. 55277–55284, 2019.

[24] C. Zhang, R. Cong, Q. Lin et al., "Cross-modality discrepant interaction network for RGB-D salient object detection," in *Proceedings of the 29th ACM international conference on multimedia*, pp. 2094–2102, Chengdu, China, October 2021.

[25] W. Zhou, Y. Zhu, J. Lei, J. Wan, and L. Yu, "CCAFNet: crossflow and cross-scale adaptive fusion network for detecting salient objects in RGB-D images," *IEEE Transactions on Multimedia*, vol. 24, pp. 2192–2204, 2022.

[26] N. Huang, Y. Liu, Q. Zhang, and J. Han, "Joint cross-modal and unimodal features for RGB-D salient object detection," *IEEE Transactions on Multimedia*, vol. 23, pp. 2428–2441, 2021.

[27] J. Zhang, D.-P. Fan, Y. Dai et al., "Uncertainty inspired RGB-D saliency detection via conditional variational autoencoders," *Proc. CVPR*, vol. 44, pp. 8579–8588, 2020.

[28] W. Ji, J. Li, M. Zhang, Y. Piao, and H. Lu, "Accurate RGB-D salient object detection via collaborative learning," in *Proceedings of the ECCV 2020: 16th European Conference Computer Vision*, pp. 52–69, Glasgow, UK, August 2020.

[29] H. Luo, C. B. Hill, G. Zhou, X. Q. Zhang, C. Li, and S. Lyu, "Genome-wide association mapping reveals novel genes associated with coleoptile length in a worldwide collection of barley," *BMC Plant Biology*, vol. 20, no. 1, pp. 346–364, 2020.

[30] B. V. Lad, M. F. Hashmi, and A. G. Keskar, "Boundary preserved salient object detection using guided filter based hybridization approach of transformation and spatial domain analysis," *IEEE Access*, vol. 10, pp. 67230–67246, 2022.

[31] Z. Liu, Y. Wang, Z. Tu, Y. Xiao, and B. Tang, "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, Chengdu China, October 2021.

[32] N. Liu, N. Zhang, K. Wan, and L. Shao, "J. Han,visual saliency transformer," in *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 4702–4712, Montreal, BC, Canada, October 2021.

[33] H. Ge, S. Wang, C. Huang, and Y. An, "A visual tracking algorithm combining parallel network and dual attention-aware mechanism," *IEEE Access*, vol. 11, pp. 15831–15844, 2023.

[34] J. Bai, Z. Wen, Z. Xiao et al., "Hyperspectral image classification based on multibranch attention transformer networks," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–17, August 2022.

[35] S. Xiao, Y. Li, Y. Ye et al., "Hierarchical temporal fusion of multi-grained attention features for video question answering," *Neural Processing Letters*, vol. 52, no. 2, pp. 993–1003, 2020.

[36] J. Yang, C. Zhang, Y. Tang, and Z. Li, "PAFM: pose-drive attention fusion mechanism for occluded person re-identification," *Neural Computing & Applications*, vol. 34, no. 10, pp. 8241–8252, 2022.

[37] Z. Chen, Y. Shang, A. Python, Y. Cai, J. Yin, and Db-BlendMask, "DB-BlendMask: decomposed attention and balanced BlendMask for instance segmentation of high-resolution remote sensing images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–15, 2022.

[38] M. Zhang, W. Ren, Y. Piao, Z. Rong, and H. Lu, "Select, supplement and focus for RGB-D saliency detection," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 3469–3478, Seattle, WA, USA, June 2020.

[39] W. Ji, G. Yan, J. Li et al., "DMRA: depth-induced multi-scale recurrent attention network for RGB-D saliency detection," *IEEE Transactions on Image Processing*, vol. 31, pp. 2321–2336, 2022.

[40] C. Li, R. Cong, S. Kwong et al., "ASIF-Net: attention steered interweave fusion network for RGB-D salient object detection," *IEEE Transactions on Cybernetics*, vol. 51, no. 1, pp. 88–100, Jan. 2021.

[41] Z. Liu, Q. Duan, S. Shi, and P. Zhao, "Multi-level progressive parallel attention guided salient object detection for RGB-D images," *The Visual Computer*, vol. 37, no. 3, pp. 529–540, 2021.

[42] R. Ju, L. Ge, W. Geng, T. Ren, and G. Wu, "Depth saliency based on anisotropic center-surround difference," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 1115–1119, Paris, France, October 2014.

[43] Y. Cheng, H. Fu, X. Wei, J. Xiao, and X. Cao, "Depth enhanced saliency detection method," in *Proceedings of the international conference on internet multimedia computing and service*, pp. 23–27, Xiamen, China, July 2014.

[44] H. Peng, B. Li, W. Xiong, W. Hu, and R. Ji, "RGBD salient object detection: a benchmark and algorithms," in *Proceedings of the 13th European Conference Computer Vision--ECCV 2014*, pp. 92–109, Zurich, Switzerland, September 2014.

[45] G. Li and C. Zhu, "A three-pathway psychobiological framework of salient object detection using stereoscopic technology," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV) Workshops*, pp. 3018–3014, Carolina WB, USA, October 2017.

[46] Y. Piao, W. Ji, J. Li, M. Zhang, and H. Lu, "Depth-induced multi-scale recurrent attention network for saliency detection," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 7253–7262, Seoul, Korea(South), November 2019.

[47] D.-P. Fan, Z. Lin, Z. Zhang, M. Zhu, and M.-M. Cheng, "Rethinking RGB-D salient object detection: models, data sets, and large-scale benchmarks," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 32, no. 5, pp. 2075–2089, 2021.

[48] R. Achanta, S. Hemami, F. Estrada, and S. Susstrunk, "Frequency-tuned salient region detection," in *Proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1597–1604, Miami, FL, USA, June 2009.

[49] D.-P. Fan, M.-M. Cheng, Y. Liu, T. Li, and A. Borji, "Structure-measure: a new way to evaluate foreground maps," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pp. 4548–4557, Carolina WB, USA, October 2017.

[50] D.-P. Fan, C. Gong, Y. Cao, B. Ren, M.-M. Cheng, and A. Borji, "Enhanced-alignment measure for binary foreground map evaluation," 2018, https://arxiv.org/abs/1805.10421.

[51] F. Perazzi, P. Krähenbühl, Y. Pritch, and A. Hornung, "Contrast based filtering for salient region detection," in *Proceedings of the 2012 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 733–740, Providence, RI, USA, June 2012.

[52] X. Zhao, Y. Pang, L. Zhang, H. Lu, and X. Ruan, "Self-supervised pretraining for rgb-d salient object detection," in *Proceedings of the AAAI Conference on Artificial Intelligence*, Columbia, Canada, June 2022.

[53] Y.-H. Wu, Y. Liu, J. Xu, J.-W. Bian, Y.-C. Gu, and M.-M. Cheng, "Mobilesal: extremely efficient rgb-d salient object detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, 2022.

[54] H. Wen, C. Yan, X. Zhou et al., "Dynamic selective network for RGB-D salient object detection," *IEEE Transactions on Image Processing*, vol. 30, pp. 9179–9192, 2021.

[55] W. Zhang, G. Ji, Z. Wang, K. Fu, and Q. Zhao, "Depth quality-inspired feature manipulation for efficient RGB-D salient object detection," in *Proceedings of the 29th ACM international conference on multimedia*, pp. 731–740, Chengdu China, October 2021.

[56] W. Ji, J. Li, S. Yu et al., "Calibrated RGB-D saliency object detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9471–9481, Nashville, TN, USA, June 2021.

[57] K. Fu, D. Fan, G. Ji, and Q. Zhao, "Joint learning and densely-cooperative fusion framework for RGB-D salient object detection," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 3049–3059, Seattle, WA, USA, June 2020.