IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 15 October 2023, accepted 11 December 2023, date of publication 18 December 2023,
date of current version 28 December 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3344644

RESEARCH ARTICLE

RGB-D Salient Object Detection Method Based on Multi-Modal Fusion and Contour Guidance

YANBIN PENG, MINGKUN FENG, AND ZHIJUN ZHENG
School of Information and Electronic Engineering, Zhejiang University of Science and Technology, Hangzhou 310023, China

Corresponding author: Yanbin Peng (pyb2010@126.com)

This work was supported in part by the Basic Public Welfare Research Program of Zhejiang Province under Grant LGF22F020017 and Grant GG21F010013, and in part by the Natural Science Foundation of Zhejiang Province under Grant Y21F020030.

ABSTRACT Salient object detection is a critical task in the field of computer vision. However, existing detection methods still face certain challenges, such as the inability to effectively integrate multimodal features and the blurring of detection result boundaries. To address these issues, this paper proposes a novel RGB-D salient object detection method that combines multimodal feature fusion and contour-guiding techniques. Initially, we employ ResNet50 as the backbone network, and by removing its final pooling layer and fully connected layer, we construct a fully convolutional network specifically for feature extraction from RGB images and depth images. Subsequently, we leverage channel attention mechanism and spatial attention mechanism separately to optimize the RGB image features and depth image features. Following this, we design an interactive feature fusion module to blend the optimized features, thereby obtaining the multimodal fusion features. Furthermore, based on the localization ability of high-level fusion features, we constrain the low-level fusion features, eliminate non-salient objects, and generate salient object contour features. Eventually, we use this contour feature to guide the recognition process of salient objects, resulting in salient objects with clear boundaries. Our approach has been validated across seven RGB-D salient object detection datasets. The experimental results indicate an improvement of 0.21% $\sim$ 1.84% and 0.32% $\sim$ 1.25% respectively in maxF and S metrics, compared to the best competing methods (CMINet, CIR-Net, and CPFP).

INDEX TERMS RGB-D salient object detection, multimodal feature fusion, contour guidance, attention mechanism, fully convolutional network.

I. INTRODUCTION

Salient Object Detection (SOD) holds a significant place in the field of computer vision, with its primary objective being to identify and emphasize the most attention-grabbing objects within images [1], [2], [3], [4]. SOD has been widely adopted in numerous practical application domains, such as image and video editing [5], [6], object tracking [7], [8], human-computer interaction [9], autonomous driving [10], [11], robotic navigation [12], [13], and medical image analysis [14], [15].

Existing salient object detection methods [18], [19], [20], [21], [22], [23], [24], [25], [26] primarily rely on deep learning technologies, especially Convolutional Neural Networks (CNNs). However, consecutive convolution and pooling operations significantly reduce the size of deep feature maps, which, although able to accurately locate salient objects, cannot finely segment them. Methods based on U-Net [64], [65] connect encoders and decoders to form a U-shaped structure, gradually recovering resolution and learning local detailed features through upsampling of high-level feature maps and direct inter-layer connections of the encoder and decoder. Nevertheless, due to the information loss during the upsampling process, the obtained salient object contours are blurred. Therefore, introducing salient object contour information and using it to guide the segmentation of salient objects can yield clear salient object boundaries. Currently, various contour-enhanced RGB image salient object detection methods [51], [52], [53], [54], [55], [56], [57], [58], [59], [60] exist, however, in many challenging

The associate editor coordinating the review of this manuscript and approving it for publication was Gustavo Olague.

scenarios, such as complex backgrounds, low contrast, and similarity between foreground and background, salient object detection still faces significant challenges. Fortunately, with the widespread use of depth sensors, depth maps can provide geometric and spatial information, becoming a vital supplement to RGB image information. Therefore, how to combine RGB image and depth map information, and introduce salient object contour guiding technology on this basis, becomes key to solving the problem. At present, some research on RGB-D salient object detection based on multimodal and contour guidance has been carried out [61], [62], [63]. Jiang et al. [61] and Zhang et al. [62] extract contour features from RGB images, but do not consider depth image information. Liu et al. [63] extract the contour prior information of salient objects from depth images but do not consider RGB image information. Since RGB images and depth images have complementary fusion characteristics, considering the features of both modalities can extract more accurate contour features. Therefore, we attempt to perform cross-modal contour feature extraction and further guide the contour features into salient object detection.

In summary, this paper proposes an RGB-D salient object detection method based on multimodal fusion and contour guidance (MFCG-Net), which integrates multimodal feature fusion and contour-guided techniques. Firstly, we extract features from the RGB and depth images, and apply an attention mechanism for feature optimization. Subsequently, we design an interactive feature fusion module to blend the optimized features, thereby generating the fused features across modalities. Then, we leverage the saliency localization capability of the top-layer fused features to guide the lower-layer fused features, eliminating background noise and producing salient object contour features. Finally, we use these contour features to guide the identification process of the salient objects, resulting in salient objects with clear boundaries. In conclusion, we have made three main contributions in this study:

1) In order to make effective use of features from different modalities, we have devised an interactive enhancement fusion module. This module, through interactive feature learning, fully explores and utilizes the correlation between different modalities, thus achieving mutual reinforcement of features across modalities. Concurrently, we implement optimized feature fusion via element-wise multiplication and convolution smoothing. Building on this foundation, the fused features are sequentially propagated to the next layer, thereby realizing cross-level feature integration.

2) In order to fully explore the contour information of salient objects, we have designed a unique contour feature extraction module. Within this module, we directly convey the spatial position information of top-level fusion features to the low-level fusion features, thus constraining the contours of non-salient objects and accurately extracting the contour features of salient objects. Upon collecting contour features and modality fusion features, we utilize the contour features to guide the modality fusion features. Through the complementary information between the two, we achieve the extraction of salient objects with clear boundaries and prominent features.

3) We validated our approach on seven RGB-D salient object detection datasets. The experimental results reveal that our method outperforms existing techniques in terms of both accuracy and effectiveness. This strongly indicates that our approach can effectively utilize the information from RGB and depth images, while fully exploiting the contour information of the salient objects, thereby enhancing the performance of salient object detection.

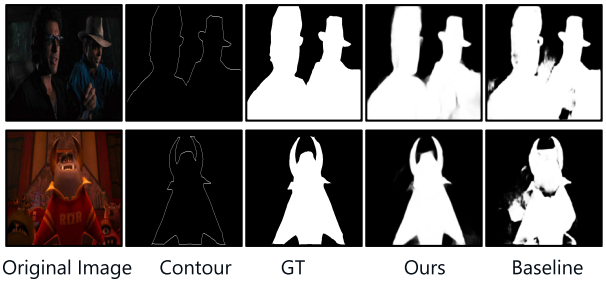


FIGURE 1. Visual effects of the method we proposed. Once we model the salient object contour information and guide the generation of salient maps via these contour features, non-salient objects are effectively eliminated. Simultaneously, the edge contours of salient objects appear more distinctly and strikingly.

II. RELATED WORKS

A. SALIENT OBJECT DETECTION

The development of salient object detection (SOD) has evolved from manually extracting features to utilizing Convolutional Neural Networks (CNN) and Fully Convolutional Networks (FCN). Early methods primarily relied on manually extracted features [16], [17]. However, manual feature extraction often requires expert experience and domain knowledge, implying that the feature selection process might be influenced by individual preferences. If inappropriate features are selected, it could negatively impact the performance of the model. Furthermore, manual features often can only handle relatively simple patterns. For complex patterns and non-linear relationships, manually extracted features may not effectively capture them. With the advancement of deep learning, the emergence of CNN and FCN has automated the feature extraction process and facilitated processing of more complex image patterns, significantly improving the efficiency and performance of salient object detection. Therefore, methods based on CNN and FCN gradually replaced those based on manual feature extraction.

B. MULTI-MODAL FUSION-BASED SALIENT OBJECT DETECTION

Significant progress has been made in the research of using Convolutional Neural Networks (CNN) for RGB-D salient object detection [18], [19], [20], [21], [22], [23], [24], [25], [26], [27], [28], [29], [30]. For instance,

Fan et al. [19] utilized two parallel CNNs to independently process RGB and depth images, and then organically fused the outputs of these two networks. They utilized the semantic localization information provided by high-level features to effectively eliminate the interference introduced by low-level features, thus generating the final saliency map. This method ingeniously utilizes the complementarity of RGB and depth information. Han et al. [20] proposed an innovative RGB-D salient object detection method called Layered Interactive Attention Network (LIANet), which mainly consists of feature encoding, layered fusion mechanism, and feature decoding. In the feature encoding stage, they introduced a concise and efficient attention module (SAM), which defined an energy function that considers the weights of channel and spatial dimensions, enabling the network to learn more discriminative neurons without adding extra parameters. They then designed a layered interactive fusion module (LIFM), which significantly enhances the interaction between RGB features and depth features, simultaneously eliminating the interference in the depth map, accurately highlighting the features of salient objects. Finally, through feature decoding and a mixed loss function, they further optimized and trained the model. Zhao et al. [21] designed a deep model that can utilize global and local context information to capture the saliency of objects. To better initialize the training of the deep neural network, they explored several different pre-training strategies and designed a pre-training program tailored to the specific task, allowing the multi-context model to adapt to the saliency detection task, thus improving the performance of saliency detection. Huang et al. [23] proposed a multi-modal feature interaction module, which captures their cross-modal complementary information by jointly using several simple linear fusion strategies and bi-linear fusion strategies. They then proposed a fusion module guided by saliency prior information to utilize the multi-level supplementary information of fusion cross-modal features at different levels. Finally, they designed a saliency refinement and prediction module to more effectively utilize the extracted multi-level cross-modal information for RGB-D saliency detection, which can more comprehensively use the information of RGB-D images. Endo and Premachandra [24] proposed a bathing accident monitoring system using a depth sensor. The system aims to prevent accidental drowning caused by loss of consciousness during bathing. To protect the privacy of bathers, no RGB images were captured using a 2D camera. Matsumura and Premachandra [25] provided an accident prevention method using deep learning techniques to notify visually impaired individuals about the presence of height and steps when approaching stairs. The study employed deep learning on generated 3D point cloud data to detect stairs. A preprocessing stage was proposed to reduce the weight of the point cloud data for application in deep learning-based training. We investigated a method for extracting salient object contour information from cross-modal fused features and applying it to improve RGB-D salient object detection.

C. CONTOUR-GUIDED SALIENT OBJECT DETECTION

To preserve important structural information in salient object detection, an increasing number of networks incorporate contour information to improve the effectiveness of RGB salient object detection. Chen et al. [51] proposed a contour-aware salient object segmentation model that achieved better segmentation results through Contour Loss and global attention modules. Wang et al. [52] introduced a deep model called Focal-BG for salient object detection and segmentation. By jointly learning the segmentation mask and boundary detection of salient objects, the model accurately captures the shape details, especially near the object boundaries. Guan et al. [53] proposed an edge-aware convolutional neural network for salient object detection. By combining global contextual information and low-level edge features, and utilizing pyramid pooling modules and auxiliary side-output supervision, this algorithm generates more distinct edge-aware features and effectively utilizes multi-scale global information. Zhuge et al. [54] proposed a boundary-guided feature aggregation network for salient object detection. This network utilizes multi-level convolutional features and edge information for salient object detection, achieving precise localization through feature extraction, boundary prediction, and feature fusion. Wu et al. [55] proposed a method called SCRAN for edge-aware salient object detection. By bidirectionally propagating information and iteratively improving features at each layer, this method optimizes the accuracy of both salient object detection and edge detection. Zhou et al. [56] introduced SE2Net, a Siamese edge-enhanced network for salient object detection. SE2Net employs a multi-stage Siamese network architecture to aggregate low-level and high-level features and simultaneously estimate the edge and region saliency maps. By enhancing edge response and suppressing background false alarms, SE2Net improves the accuracy and semantics of salient objects. Additionally, the paper also presents edge-guided inference algorithms that further improve salient masks along the edges. Su et al. [57] proposed a boundary-aware salient object detection method that addresses the selectivity-invariance dilemma by combining boundary localization and internal perception. Continuous dilated modules are used to enhance feature extraction capabilities, and transitional compensation flow is introduced to address the problem of transition regions. Lin et al. [58] presented a boundary-aware salient object detection method that improves detection performance by introducing a cyclic bidirectional guided refinement network (RTGRNet). This method leverages the complementarity between saliency and boundary features and progressively refines these features iteratively. Specifically, RTGRNet consists of two streams of guided refinement modules (TGRMs), each composed of a guidance block and saliency/boundary feature streams. The refined features from the previous TGRM are used to improve the performance of the saliency and boundary feature streams in the current TGRM. Zhang et al. [59] proposed a novel deep learning approach to address three major challenges in visual saliency

detection: complex scenes, multiple salient objects, and salient objects at different scales. By introducing a fully convolutional neural network, combining handcrafted and deep learning features, and utilizing contextual information, saliency detection is performed. Wang et al. [60] introduced a new method called PAGE-Net for detecting salient objects in images using convolutional neural networks (CNNs). This method enhances saliency representation and improves the localization and segmentation of salient objects through a pyramid attention structure and salient edge detection module.

III. PROPOSED METHOD

A. OVERVIEW

In the field of RGB salient object detection, contour information has been thoroughly investigated [51], [52], [53], [54], [55], [56], [57], [58], [59], [60]. However, comparatively, research into RGB-D salient object detection is still in its infancy. Existing methodologies either focus on extracting contour features from RGB images [61], [62] or concentrate on mining contour information from depth images [63], without fully utilizing cross-modal information for contour feature extraction. Considering the complementary nature of RGB image features and depth image features, the comprehensive use of both modal information can enhance the performance of salient object detection. Therefore, we carry out cross-modal contour feature extraction and further guide the extracted contour features to the phase of salient object detection, aiming to augment the performance of the latter.

The overall architecture of our proposed RGB-D salient object detection method (MFCG-Net), which is based on multi-modal fusion and contour guidance, is illustrated in Fig. 2. In this method, we adopt ResNet50 as the backbone network architecture, and by removing its final pooling layer and fully connected layer, we construct a fully convolutional network. This network independently performs feature extraction from both the RGB and depth images. The extracted features from the RGB image are denoted as $\mathbf{F}_{\mathrm{R}}^{i}$, while the features from the depth image are represented as $\mathbf{F}_{\mathrm{D}}^{i}$, where i is a natural number ranging from 1 to 5.

1) Attention Optimization Mechanism: As shown in Fig. 2, we use the Channel Attention mechanism to optimize the RGB image features, resulting in the optimized features $\mathbf{F}_{\mathrm{RC}}^{i}$. We also utilize the Spatial Attention mechanism to optimize the features of the depth image, leading to the optimized features $\mathbf{F}_{\mathrm{DS}}^{i}$.

2) Interactive Modal Feature Fusion: We have designed an interactive feature fusion module to combine the two sets of optimized features, resulting in the fused modal features $\mathbf{F}_{\mathrm{F}}^{i}$.

3) Contour-Guided Salient Object Detection: We utilize the positioning capability of the high-level fused features to constrain the low-level fused features, thus eliminating non-salient objects and forming the contour features of the salient objects. Subsequently, guided by these contour features, we conduct the detection process of the salient objects, ultimately obtaining salient objects with distinct boundaries.

B. ATTENTION OPTIMIZATION MECHANISM

In this study, the ResNet50 network is used to extract features from the input RGB images and depth images, respectively. On this basis, we apply two different attention mechanisms to optimize the features of RGB images and depth images. The attention mechanism provides an efficient strategy for deep learning models, which focuses on and understands key parts of the input data, thereby enhancing the performance of the model, improving the interpretability of the model, and opening up the possibility for the implementation of more complex tasks.

Currently, there are three commonly used attention mechanisms in the field of deep learning: channel attention, spatial attention, and self-attention. In this paper, according to the inherent attributes of different features, we have chosen the corresponding attention mechanism for optimization. As the RGB image is a color three-channel image with rich channel data, we choose to handle it with channel attention mechanism. As for the depth image, it is a single-channel image, mainly showing the spatial position of the salient object, with relatively less channel information, therefore we choose to handle it with spatial attention mechanism.

The implementation process of channel attention is shown in Fig. 3, which is described as follows:

$$\mathbf{C}_{\max} = \mathrm{FC}(\sigma(\mathrm{FC}(\mathrm{P}_{\max}(\mathbf{F}_{\mathrm{R}}^{i})))) \tag{1}$$

$$\mathbf{C}_{\mathrm{mean}} = \mathrm{FC}(\sigma(\mathrm{FC}(\mathrm{P}_{\mathrm{mean}}(\mathbf{F}_{\mathrm{R}}^{i})))) \tag{2}$$

$$\mathbf{F}_{\mathrm{RC}}^{i} = \delta(\mathbf{C}_{\max} \oplus \mathbf{C}_{\mathrm{mean}}) * \mathbf{F}_{\mathrm{R}}^{i} \quad i \in \{1, 2, 3,4,5\} \tag{3}$$

where $\mathrm{P}_{\max}$ denotes the maximum pooling operation performed on each feature map, while $\mathrm{P}_{\mathrm{mean}}$ signifies the average pooling operation also applied to each feature map. FC stands for the fully connected layer. The symbol σ represents the Relu activation function, the symbol δ represents the Sigmoid activation function, and the symbol $*$ indicates the element-wise multiplication operation carried out in spatial expansion.

The spatial attention mechanism, as illustrated in Fig. 4, can be described as follows:

$$\mathbf{F}_{\mathrm{DS}}^{i} = \delta(\mathrm{Conv}_{3\times3}(\ \mathrm{concat}(\mathrm{Q}_{\max}(\mathbf{F}_{\mathrm{D}}^{i}),\mathrm{Q}_{mean}(\mathbf{F}_{\mathrm{D}}^{i})))) \odot \mathbf{F}_{\mathrm{D}}^{i} \quad i \in \{1, 2, 3,4,5\} \tag{4}$$

where $\mathrm{Q}_{\max}$ denotes the maximum pooling operation carried out along the channel, while $\mathrm{Q}_{\mathrm{mean}}$ signifies the average pooling operation also performed along the channel. $\mathrm{Conv}_{3\times3}$ represents a convolution operation with a 3×3 kernel, and concat() denotes a concatenation operation. The symbol $\odot$ stands for an element-wise multiplication operation executed in channel expansion.

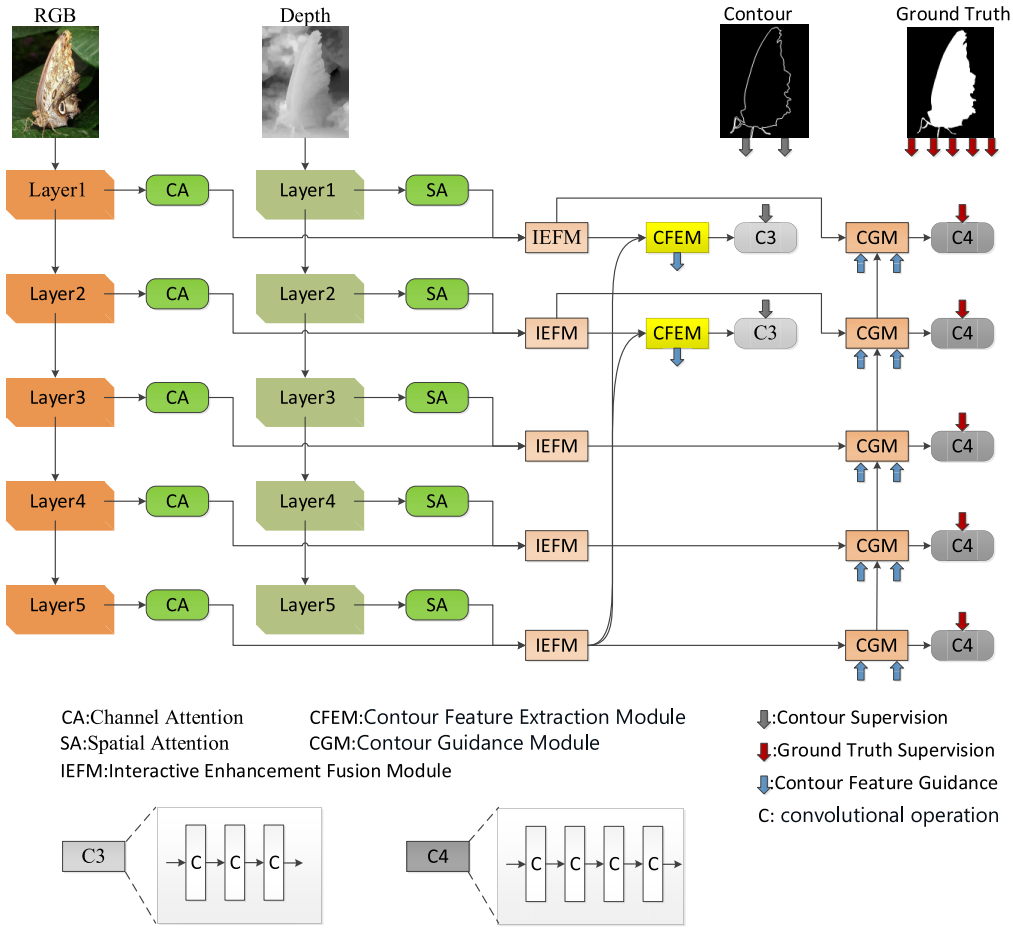


FIGURE 2. Overall structure of MFCG-Net.

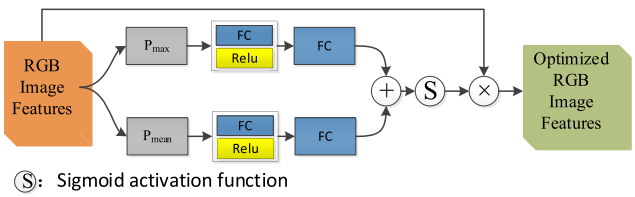


FIGURE 3. Channel attention optimization mechanism.

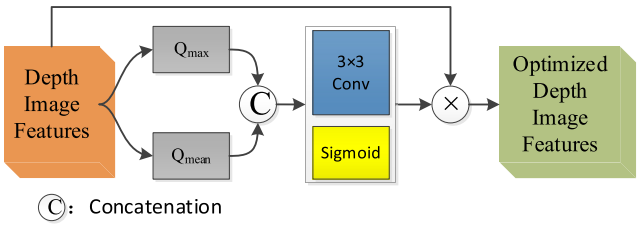


FIGURE 4. Spatial attention optimization mechanism.

C. INTERACTIVE ENHANCEMENT FUSION MODULE

To effectively achieve cross-modal feature fusion, we have designed an Interactive Enhancement Fusion Module (IEFM). The operating process of IEFM consists of two stages: initially, interactive modal feature enhancement is conducted; thereafter, modal feature fusion takes place.

During the first stage, we adopt an interactive enhancement strategy. Both modal features undergo a convolution operation with a 3×3 kernel, followed by a batch normalization operation, and then activation via a Sigmoid activation function. The resulting feature maps enhance the other modal feature through an element-wise multiplication method. Subsequently, another convolution operation with a 3×3 kernel is conducted, and a residual connection with the original modal feature is established. Ultimately, we obtain the interactive enhanced modal features $\mathrm{F}^{i}_{\mathrm{IR}}$ and $\mathrm{F}^{i}_{\mathrm{ID}}$. The specific process can be described as follows:

$$\mathbf{F}^{i}_{\mathrm{IR}} = \mathrm{BRConv}_{3\times3}(\delta(\mathrm{BN}(\mathrm{Conv}_{3\times3}(\mathbf{F}^{i}_{\mathrm{DS}}))) \otimes \mathbf{F}^{i}_{\mathrm{RC}}) \oplus \mathbf{F}^{i}_{\mathrm{RC}} \tag{5}$$

$$\mathbf{F}^{i}_{\mathrm{ID}} = \mathrm{BRConv}_{3\times3}(\delta(\mathrm{BN}(\mathrm{Conv}_{3\times3}(\mathbf{F}^{i}_{\mathrm{RC}}))) \otimes \mathbf{F}^{i}_{\mathrm{DS}}) \oplus \mathbf{F}^{i}_{\mathrm{DS}} \tag{6}$$

where BN stands for batch normalization operation, $\otimes$ represents element-wise multiplication operation, and $\oplus$ denotes element-wise addition operation. $\mathrm{BRConv}_{3\times3}$ refers to the convolution operation with a 3×3 kernel, which

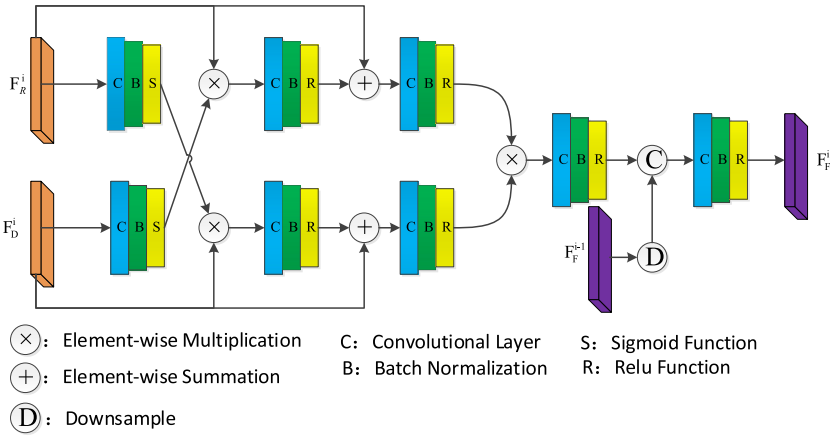


FIGURE 5. Interactive enhancement fusion module.

is followed by a batch normalization operation and a Relu activation function

In the second stage, we carry out an element-wise multiplication operation on the two interactively enhanced modal features, followed by a smoothing process using a convolution operation with a 3 × 3 kernel. Subsequently, the result is concatenated with the fused feature from the previous layer. The concatenated result goes through another convolution operation with a 3 × 3 kernel to obtain the fused feature of the current layer. This process can be described as follows:

$$\mathbf{F}_{\mathrm{F}}^{i} = \mathrm{BRConv}_{3\times3}(\mathrm{concat}(\mathrm{BRConv}_{3\times3} \times (\mathbf{F}_{\mathrm{IR}}^{i} \otimes \mathbf{F}_{\mathrm{ID}}^{i}),\mathrm{Down}(\mathbf{F}_{\mathrm{F}}^{i-1}))) \quad (7)$$

where Down() signifies the process of downsampling the feature map to half of its original size.

It should be noted that when i equals 1, in the absence of fused features from the previous layer, we only need to compute the features of the RGB image and depth map of the current layer. This process can be described as follows:

$$\mathbf{F}_{\mathrm{F}}^{1} = \mathrm{BRConv}_{3\times3}(\mathrm{BRConv}_{3\times3}(\mathbf{F}_{\mathrm{IR}}^{1} \otimes \mathbf{F}_{\mathrm{ID}}^{1})) \quad (8)$$

D. CONTOUR FEATURE EXTRACTION MODULE

In this module, our aim is to model the contour information of salient objects and extract contour features from it. The backbone network for salient object detection usually contains multi-layer features, where the low-level features can capture the detail information of the object, such as its shape and edges, while the high-level features can capture semantic information, such as the category and location of the salient object. Therefore, we have carefully designed a method to extract the edge contour information of salient objects from low-level fusion features $\mathbf{F}_{\mathrm{F}}^{1}$ and $\mathbf{F}_{\mathrm{F}}^{2}$. At the same time, we also need semantic location information from high-level fusion features. Given that the receptive field of the top-level fusion feature $\mathbf{F}_{\mathrm{F}}^{5}$ is the largest, its localization capability is also the most accurate. For this reason, we have designed a top-level guided localization mechanism, which directly propagates location information of top-level fusion features to low-level fusion features, constraining the outline of non-salient objects and thereby obtaining the contour features of salient objects. The calculation process of contour features can be described as follows:

$$\mathrm{CloneSize}(\mathbf{F}_{\mathrm{F}}^{5}, \mathbf{F}_{\mathrm{F}}^{i}) = UP(\mathrm{BRConv}_{1\times1}(\mathbf{F}_{\mathrm{F}}^{5}),\mathbf{F}_{\mathrm{F}}^{i}) \quad (9)$$

$$\mathbf{F}_{\mathrm{CF}}^{\mathrm{i}} = \mathbf{F}_{\mathrm{F}}^{i} \oplus \mathrm{CloneSize}(\mathbf{F}_{\mathrm{F}}^{5}, \mathbf{F}_{\mathrm{F}}^{i}) \quad i \in \{1, 2\} \quad (10)$$

where $\mathrm{F}_{\mathrm{CF}}^{\mathrm{i}}$ represents the contour feature of the i-th layer. UP(A,B) denotes the operation of upsampling feature map A to the same size as feature map B. $\mathrm{BRConv}_{1\times1}$ refers to a convolution operation with a kernel size of 1 × 1, followed by a batch normalization operation and a ReLU activation function. The purpose of this convolution operation is to change the number of channels in the features. CloneSize(A,B) indicates the operation of adjusting the size of feature map A to match the size of feature map B.

After obtaining the contour features, we perform three convolution operations on them. The first two operations use a 3 × 3 convolution kernel, which primarily serves to enhance the contour features. The third operation employs a 1 × 1 convolution kernel, the purpose of which is to adjust the number of channels in the feature map to 1, thus obtaining the predicted contour map. We use the real contour map to supervise the generation of the predicted contour map. The supervision method utilizes a cross-entropy loss function. The specific process can be described as follows:

$$\mathbf{CF}^{i} = \mathrm{BRConv}_{1\times1}(\mathrm{BRConv}_{3\times3}(\mathrm{BRConv}_{3\times3}(\mathbf{F}_{\mathrm{CF}}^{i}))) \quad (11)$$

$$\mathrm{L}(\mathbf{CF}^{i}, \mathbf{GTC}) = -\sum_{k\in U+} \log \mathrm{p}(\mathrm{y}_k = 1) - \sum_{k\in U-} \log \mathrm{p}(\mathrm{y}_k = 0) \quad i \in \{1, 2\} \quad (12)$$

where $\mathbf{CF}^{i}$ represents the i-th predicted contour map, while **GTC** stands for the real contour map. U+ denotes the set of

contour pixel points in the real contour map, and U- signifies the set of background pixel points in the real contour map. $\log p(y_k = 1)$ represents the probability that pixel point k in the predicted contour map is predicted as a contour, and $\log p(y_k = 0)$ signifies the probability that pixel point k in the predicted contour map is predicted as a background.

E. LAYER-BY-LAYER CONTOUR GUIDANCE MODULE

After obtaining the contour features and modality fusion features, our goal is to use the contour features to guide the modality fusion features, and to capture salient objects with clear boundaries by utilizing the complementary information between the two sets of features. During the decoding phase, in order to fully utilize the multi-resolution modality fusion features, we adopt a network structure similar to UNet, integrating the modality fusion features layer by layer from the higher layers to the lower ones. To avoid dilution of the contour information during the integration process, we design a layer-by-layer contour guidance module. In each layer of the decoding process, we incorporate the contour features into the decoding features, thus making the predicted position of the salient objects more accurate, the edges clearer, and the detail of segmentation better. The entire process can be described as follows:

$$\mathbf{DF}^5 = \text{CloneSize}((\text{CloneSize}(\mathbf{F}_F^5, \mathbf{F}_{CF}^2) \oplus \mathbf{F}_{CF}^2), \mathbf{F}_{CF}^1) \oplus \mathbf{F}_{CF}^1 \tag{13}$$

$$\mathbf{UF}^i = \text{CloneSize}(\mathbf{DF}^{i+1}, \mathbf{F}_F^i) \oplus \mathbf{F}_F^i \tag{14}$$

$$\mathbf{DF}^i = \text{CloneSize}((\text{CloneSize}(\mathbf{UF}^i, \mathbf{F}_{CF}^2) \oplus \mathbf{F}_{CF}^2), \mathbf{F}_{CF}^1) \oplus \mathbf{F}_{CF}^1 \quad i \in \{1, 2, 3,4\} \tag{15}$$

Herein, DF^i represents the decoding features at the i-th layer.

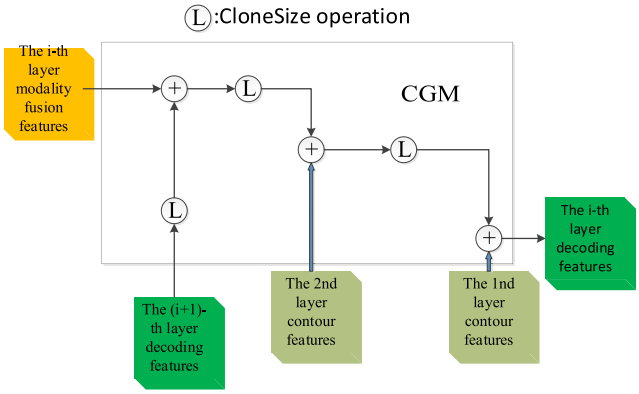


FIGURE 6. Contour guidance module.

After obtaining the decoding features, we perform four convolution operations on each layer of decoding features. Specifically, for the first and second layers, we first carry out three convolution operations with a 3×3 convolution kernel. For the third, fourth, and fifth layers, we initially conduct three convolution operations with a 5×5 convolution kernel. Finally, for all five layers of decoding features, we each use a 1×1 convolution kernel for a single convolution operation to adjust the number of channels to 1, thereby obtaining the predicted saliency map. Subsequently, we merge these predicted saliency maps to obtain the final fused saliency map. This process can be described as follows:

$$\mathbf{PF}^i = \text{BRConv}_{1\times1}(\text{BRConv}_{3\times3}(\text{BRConv}_{3\times3}(\text{BRConv}_{3\times3}(\mathbf{DF}^i)))) \quad i \in \{1, 2\} \tag{16}$$

$$\mathbf{PF}^j = \text{BRConv}_{1\times1}(\text{BRConv}_{5\times5}(\text{BRConv}_{5\times5}(\text{BRConv}_{5\times5}(\mathbf{DF}^j)))) \quad j \in \{3, 4, 5\} \tag{17}$$

$$\mathbf{S} = \sum_{i=1}^{5} \alpha_i \mathbf{PF}^i \tag{18}$$

Herein, $\mathbf{PF}^k$ represents the k-th predicted saliency map, and S denotes the fused saliency map, which is the final prediction result. The real saliency map is used to supervise the predicted saliency map and the fused saliency map. For the supervision method, we use the cross-entropy loss function. This process can be described as follows:

$$L(\mathbf{PF}^i, \mathbf{GT}) = -\sum_{m\in V+} \log p(y_m = 1) - \sum_{m\in V-} \log p(y_m = 0) \quad i \in \{1, 2, 3,4,5\} \tag{19}$$

$$L(\mathbf{S}, \mathbf{GT}) = -\sum_{m\in V+} \log p(y_m = 1) - \sum_{m\in V-} \log p(y_m = 0) \tag{20}$$

Herein, GT represents the real saliency map, V+ denotes the set of pixel points of the salient object in the real saliency map, and V- represents the set of pixel points of the background in the real saliency map. $\log p(y_m = 1)$ indicates the probability of pixel point m in the predicted saliency map or the fused saliency map being predicted as a salient object, while $\log p(y_m = 0)$ indicates the probability of pixel point m in the predicted saliency map or the fused saliency map being predicted as background. The total loss function can be expressed as follows:

$$L_{ALL} = L(\mathbf{S}, \mathbf{GT}) + \sum_{i=1}^{2} L(\mathbf{CF}^i, \mathbf{GTC}) + \sum_{j=1}^{5} L(\mathbf{PF}^j, \mathbf{GT}) \tag{21}$$

IV. EXPERIMENTS

In subsection A, we introduce some key details of the experiments. Subsequently, in subsections B and C, we provide detailed descriptions of the datasets used and the evaluation metrics respectively. Further, in subsection D, we conduct both quantitative and qualitative comparisons between the newly proposed method and other advanced methods. Lastly, in subsection E, we present the results of ablation experiments conducted on seven different datasets.

A. EXPERIMENTAL DETAILS

We utilize ResNet50 [31] as the backbone network and train our model on the NJU2K [32] and NLPR [33] datasets.

Our model, implemented using PyTorch, was executed on an NVIDIA RTX 3090 GPU for all experimental processes. The hyperparameters were set as follows: learning rate = 1e-4, gradient clipping threshold = 0.5, and batch size = 10. We trained the model for 200 epochs, reducing the learning rate by a tenth every 50 epochs. When processing input images, we first resized them to 352 × 352, then applied random flipping, random cropping, random rotation, and color enhancement. In the inference process, we obtained two predicted contour maps and five predicted saliency maps. By fusing the five predicted saliency maps, we produced the final saliency map.

B. DATASETS

In this study, we conducted an in-depth evaluation of MFCG-Net, covering seven widely used RGB-D datasets, specifically NJU2K [32], NLPR [33], STERE [34], SSD [35], SIP [36], DES [37], and LFSD [38]. The STERE dataset contains 1,000 pairs of RGB-D images, primarily focusing on the representation of outdoor scenes. The NJU2K dataset includes 1,985 pairs of RGB images and their corresponding depth images, providing a rich array of objects and complex environmental scenarios for the study. The NLPR dataset, as a representative dataset, encompasses 1,000 stereo images, with depth images obtained under diversified lighting conditions and capture scenes. The LFSD is a challenging dataset that contains 100 color images with complex backgrounds and foregrounds. The SSD dataset is relatively small, containing just 80 images selected from stereoscopic movies, these images cover various aspects of movie scenes, including characters, animals, buildings, and other foreground elements. SIP is a newly released dataset that includes 929 high-definition RGB-D images of individuals, making it highly suitable for research in the field of human detection. The DES dataset contains 135 images, most of which display relatively simple foreground objects and visual scenes, and the depth maps in this dataset are of excellent quality.

C. EVALUATION METRICS

In order to quantitatively evaluate the data, we employed a series of methods, including Precision-Recall (P-R) curves, F-measure [39], E-measure [40], S-measure [41], and Mean Absolute Error (MAE). We set multiple thresholds to process the saliency maps, and by comparing the binarized saliency maps after processing with the ground truth salient maps, we obtained a set of precision-recall values, which were then used to plot the Precision-Recall curves.

The F-measure is a comprehensive evaluation metric, defined as the weighted harmonic mean of precision and recall, with the specific definition as follows:

$$F_\beta = \frac{(1+\beta^2)P \times R}{\beta^2 \times P + R} \tag{22}$$

Herein, β^2 is set to 0.3, where P represents precision and R represents recall. In this experiment, we only report the maximum F-measure (maxF) value.

The definition of E-measure comprehensively considers both the values of local pixels and the average value of the entire image, which aligns with the philosophy of cognitive visual research. The specific definition is as follows:

$$E_m = \frac{l}{\mathrm{W} \times \mathrm{H}} \sum_{i=1}^{\mathrm{W}} \sum_{j=1}^{\mathrm{H}} \psi(i,j) \tag{23}$$

Herein, W and H respectively represent the width and height of the saliency map, while ψ represents the enhanced alignment matrix. In this experiment, we only report the maximum E-measure (maxE) value.

The S-measure is an evaluation method used to assess the structural similarity between the saliency map and the ground truth salient map. The specific definition is as follows:

$$S_\alpha = \alpha \times S_\mathrm{o} + (1-\alpha) S_\mathrm{r} \tag{24}$$

Herein, the weight coefficient α is set to 0.5, indicating that object S_o and region S_r contribute equally to the structural similarity.

MAE is used to calculate the Mean Absolute Error between the saliency map S and the ground truth salient map GT, with the specific formula expressed as follows:

$$\mathrm{MAE} = \frac{1}{\mathrm{W} \times \mathrm{H}} \sum_{i=1}^{\mathrm{W}} \sum_{j=1}^{\mathrm{H}} |\mathbf{S}(i,j) - \mathbf{GT}(i,j)| \tag{25}$$

Herein, W and H respectively represent the width and height of the saliency map.

D. COMPARISON WITH STATE-OF-THE-ART METHODS

To fully demonstrate the effectiveness of our proposed MFCG-Net, we compare it with nine existing RGB-D based salient object detection (SOD) methods, including DCMF [42], CIR-Net [43], CFPF [44], DMRA [45], UCNet [46], HDFNet [47], CMINet [48], JLDCF [49], and HINet [50]. To ensure a fair comparison, we use the saliency maps provided by the authors. If no such maps are provided, we generate them using the source code and model files provided by the authors.

1) QUANTITATIVE COMPARISON

Table 1 presents the quantitative evaluation results for four assessment metrics. It is clear from the table that our proposed MFCG-Net method excels in all four metrics, surpassing most of the cutting-edge methods. In datasets such as NJU2K, NLPR, LFSD, SIP, SSD, and STERE, our method outperforms all compared algorithms in terms of performance. This can be primarily attributed to our carefully designed multimodal fusion strategy and contour-guided technique. By optimizing multimodal features through different attention mechanisms, and then integrating them in

TABLE 1. Comparison of evaluation results on four assessment metrics - MAE, max F-measure (maxF), max E-measure (maxE), and S-measure (S) - across seven datasets. The arrow ↑ indicates that a higher value is better, while ↓ signifies that a lower value is preferable. The best performance in each row is highlighted in bold.

datasets	assessment metrics	Comparison Methods									
		HDFNet	JL-DCF	UCNet	CPFP	DMRA	CIR-Net	DCMF	HINet	CMINet	Ours
DES	MAE ↓	0.022	0.021	0.019	0.036	0.031	0.029	0.023	0.022	**0.016**	0.018
	maxF ↑	0.921	0.918	0.930	0.851	0.889	0.892	0.924	0.922	0.939	**0.941**
	maxE ↑	0.970	0.957	0.976	0.932	0.941	0.941	0.968	0.967	**0.979**	0.974
	S ↑	0.926	0.928	0.934	0.875	0.903	0.907	0.932	0.927	0.940	**0.943**
LFSD	MAE ↓	0.077	0.081	0.067	0.086	0.074	0.068	0.095	0.076	0.063	**0.062**
	maxF ↑	0.862	0.854	0.863	0.824	0.858	0.882	0.815	0.847	0.874	**0.890**
	maxE ↑	0.896	0.887	0.905	0.870	0.905	0.909	0.877	0.889	0.913	**0.925**
	S ↑	0.854	0.849	0.864	0.831	0.845	0.876	0.827	0.852	0.879	**0.890**
NJU2K	MAE ↓	0.039	0.039	0.035	0.028	0.049	0.035	0.036	0.039	0.031	**0.026**
	maxF ↑	0.910	0.915	0.910	0.937	0.892	0.928	0.925	0.914	0.934	**0.949**
	maxE ↑	0.944	0.951	0.949	0.962	0.937	0.955	0.958	0.945	0.957	**0.977**
	S ↑	0.908	0.913	0.911	0.930	0.889	0.925	0.925	0.915	0.933	**0.944**
NLPR	MAE ↓	0.023	0.022	0.025	0.034	0.030	0.028	0.029	0.026	0.021	**0.020**
	maxF ↑	0.917	0.918	0.903	0.870	0.875	0.907	0.906	0.906	0.922	**0.933**
	maxE ↑	0.963	0.965	0.956	0.922	0.942	0.955	0.954	0.957	0.963	**0.971**
	S ↑	0.923	0.931	0.920	0.890	0.898	0.921	0.922	0.922	0.932	**0.939**
SIP	MAE ↓	0.048	0.049	0.051	0.062	0.082	0.069	0.062	0.066	0.044	**0.042**
	maxF ↑	0.894	0.894	0.879	0.855	0.835	0.866	0.872	0.855	0.910	**0.912**
	maxE ↑	0.930	0.931	0.919	0.906	0.883	0.905	0.911	0.899	0.939	**0.945**
	S ↑	0.886	0.885	0.875	0.853	0.816	0.862	0.870	0.856	0.899	**0.904**
SSD	MAE ↓	0.046	0.052	0.049	0.081	0.057	0.052	0.073	0.049	0.051	**0.043**
	maxF ↑	0.870	0.839	0.849	0.792	0.849	0.855	0.811	0.852	0.860	**0.886**
	maxE ↑	0.925	0.909	0.921	0.869	0.911	0.912	0.897	0.916	0.903	**0.930**
	S ↑	0.880	0.864	0.869	0.812	0.855	0.873	0.838	0.865	0.873	**0.887**
STERE	MAE ↓	0.042	0.044	0.039	0.050	0.064	0.046	0.043	0.049	0.036	**0.034**
	maxF ↑	0.900	0.895	0.899	0.878	0.852	0.897	0.906	0.883	0.916	**0.927**
	maxE ↑	0.943	0.942	0.944	0.919	0.917	0.939	0.946	0.933	0.951	**0.962**
	S ↑	0.900	0.900	0.903	0.881	0.838	0.901	0.910	0.892	0.918	**0.925**

a complementary fusion, we have successfully extracted discriminative information from RGB images and depth images, forming a fused feature. Following this, the contour-guided technique effectively eliminates non-salient objects and enhances the clarity of the boundaries of salient objects. As for the DES dataset, apart from the MAE and maxE metrics, our method surpasses all other methods in the remaining two metrics, maintaining a leading position. Fig. 7 shows the comparison results of the PR curves for different methods, which also reflects the superior performance of our method across seven datasets, thereby confirming that our method outperforms all compared methods.

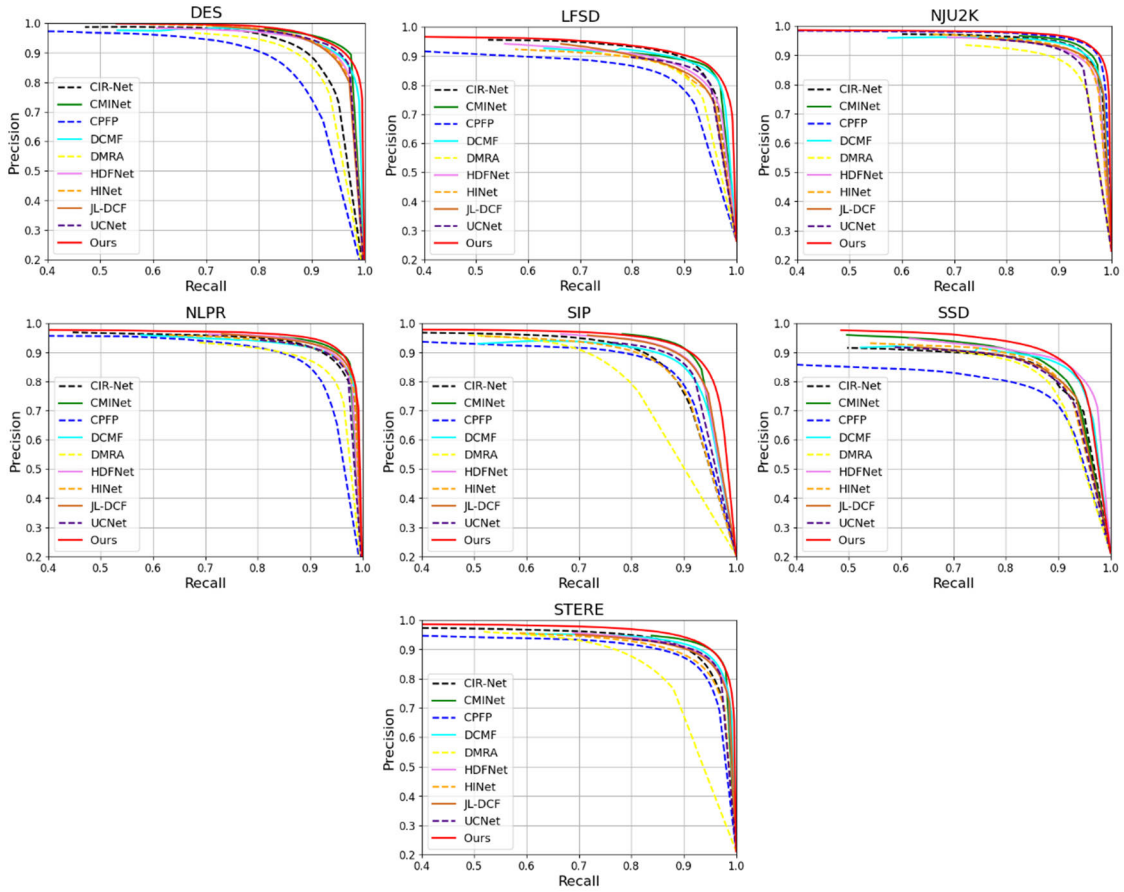


FIGURE 7. Comparison of P-R curves for different methods on seven RGB-D datasets. Our MFCG-Net method is represented by the red solid line.

2) QUALITATIVE COMPARISON

Fig. 9 displays several typical scenarios, including those where the foreground and background are similar (rows 1 and 2), the background is complex (rows 3 and 4), the quality of the depth image is poor (rows 5 and 6), multiple objects coexist (rows 7 and 8), and small objects are present (rows 9 and 10). From a visual perspective, our proposed method displays exceptional performance. Our method is capable of more precisely locating salient objects and generating more accurate saliency maps. Thus, both qualitative and quantitative evaluations sufficiently demonstrate the effectiveness of our proposed method.

In the process of salient object detection, the absence of guidance from contour features often results in the generated saliency maps having blurred boundaries. The multi-modal fusion module is capable of achieving complementary integration of RGB image features and depth image features, thereby assisting in the elimination of background interference in complex scenarios and enhancing the precision of saliency detection. Figure 8 displays the segmentation results of predicted saliency maps in three complex scenarios, with only boundary guidance (column 4) and only multi-modal fusion (column 5) being applied, respectively. By examining Figure 9, we can observe that, when only applying boundary guidance, the predicted saliency map still retains background noise, while when only employing the multi-modal fusion method, although most of the background noise is successfully removed, the boundaries of the salient objects appear blurred. Therefore, only by integrating the advantages of both techniques can we generate accurate and clear saliency maps.

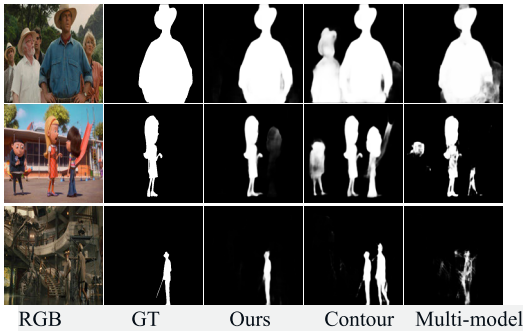


FIGURE 8. Visual comparison of applying only boundary guidance and applying only multi-modal fusion.

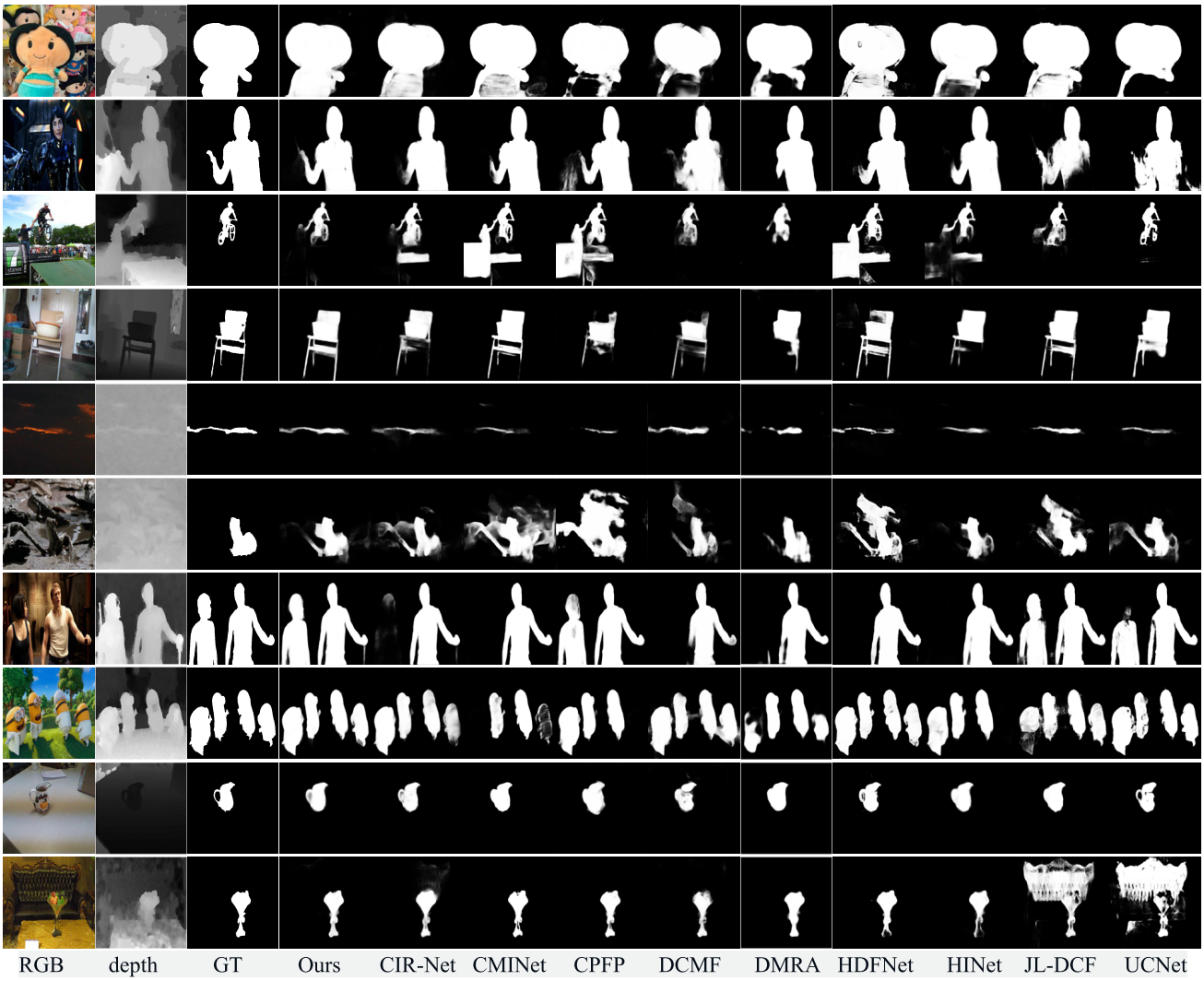


FIGURE 9. Visual comparison between MFCG-Net and the state-of-the-art RGB-D models.

E. ABLATION STUDY

The ablation analysis shown in Table 2 clearly reveals the effectiveness of each module. Herein, CAM stands for Channel Attention Module, SAM denotes Spatial Attention Module, IEFM represents Interactive Enhancement Fusion Module, CFEM signifies Contour Feature Extraction Module, and CGM refers to the Contour Guidance Module. The terms "without CAM", "without SAM", "without IEFM", and "without CFEM&CGM" represent the resultant models after removing the corresponding modules from the MFCG-Net model. By comparing the data in the fifth and seventh columns, it is evident that the incorporation of the IEFM module significantly improves the model's performance. Similarly, comparing data from the sixth and seventh columns shows that the inclusion of the CFEM&CGM module notably enhances the model's performance. Furthermore, the CAM and SAM modules also contribute to the model's performance boost. These results underscore the importance of the four sets of modules: the CAM module optimizes RGB image features by introducing channel attention, the SAM module enhances depth image features by introducing spatial attention, the IEFM module realizes complementary fusion of different modal features, and the CFEM&CGM module implements contour guidance to exclude non-salient objects and refine the boundaries of salient objects. These four functional modules all significantly improve the model's performance. The final column of data demonstrates that the MFCG-Net model, which integrates all four sets of modules, achieves the best results.

In order to reduce computational load, we replaced the backbone network in this paper, consisting of two ResNet50 networks, with a Siamese network that includes a single ResNet50. Table 3 presents a comparison of different backbone networks used. Experimental data demonstrates that our proposed method outperforms Siamese network-based method in terms of performance.

TABLE 2. Comparison of ablation study results.

datasets	assessment metrics	Without CAM	Without SAM	Without IEFM	Without CFEM&CGM	MFCG-Net
DES	MAE↓	0.021	0.020	0.025	0.030	0.018
	maxF↑	0.920	0.923	0.912	0.903	0.941
	maxE↑	0.964	0.964	0.956	0.941	0.974
	S↑	0.928	0.930	0.924	0.915	0.943
LFSD	MAE↓	0.072	0.077	0.081	0.088	0.062
	maxF↑	0.861	0.858	0.841	0.844	0.890
	maxE↑	0.900	0.895	0.888	0.880	0.925
	S↑	0.863	0.851	0.843	0.838	0.890
NJU2K	MAE↓	0.037	0.038	0.042	0.045	0.026
	maxF↑	0.917	0.919	0.908	0.913	0.949
	maxE↑	0.945	0.950	0.942	0.938	0.977
	S↑	0.915	0.916	0.910	0.907	0.944
NLPR	MAE↓	0.024	0.026	0.029	0.032	0.020
	maxF↑	0.914	0.907	0.901	0.893	0.933
	maxE↑	0.960	0.955	0.951	0.944	0.971
	S↑	0.927	0.922	0.919	0.912	0.939
SIP	MAE↓	0.054	0.060	0.062	0.069	0.042
	maxF↑	0.883	0.879	0.874	0.866	0.912
	maxE↑	0.919	0.913	0.911	0.904	0.945
	S↑	0.878	0.872	0.863	0.858	0.904
SSD	MAE↓	0.049	0.048	0.059	0.067	0.043
	maxF↑	0.851	0.855	0.819	0.805	0.886
	maxE↑	0.910	0.916	0.878	0.859	0.930
	S↑	0.871	0.874	0.852	0.832	0.887
STERE	MAE↓	0.043	0.051	0.053	0.062	0.034
	maxF↑	0.897	0.876	0.878	0.859	0.927
	maxE↑	0.939	0.928	0.925	0.918	0.962
	S↑	0.903	0.884	0.881	0.870	0.925

TABLE 3. Experimental results of different backbone networks.

	DES		LFSD		NJU2K		NLPR		SIP		SSD		STERE	
	S ↑	MAE ↓	S ↑	MAE ↓	S ↑	MAE ↓	S ↑	MAE ↓	S ↑	MAE ↓	S ↑	MAE ↓	S ↑	MAE ↓
Siamese	0.935	0.020	0.872	0.071	0.919	0.034	0.932	0.021	0.891	0.047	0.870	0.051	0.914	0.039
Ours	0.943	0.018	0.890	0.062	0.944	0.026	0.939	0.020	0.904	0.042	0.887	0.043	0.925	0.034

V. CONCLUSION

The RGB-D salient object detection method proposed in this paper aims to effectively integrate multimodal features and, based on this, use contour features to guide the segmentation of salient objects. Our approach ingeniously combines attention mechanisms, interactive cross-modal feature fusion, contour supervision, and contour feature guidance techniques, thereby effectively eliminating background noise and achieving clear boundary saliency maps. Compared with the most advanced methods on various widely-used datasets, our method demonstrated superior performance. Our approach provides a novel perspective for understanding and utilizing the characteristics of RGB-D images, as well as the contour information of salient objects. We believe that this innovative method will offer significant inspiration and impact for both the research and practical application of salient object detection.

REFERENCES

[1] J.-J. Liu, Q. Hou, Z.-A. Liu, and M.-M. Cheng, "PoolNet+: Exploring the potential of pooling for salient object detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 1, pp. 887–904, Jan. 2023.

[2] R. Cong, N. Yang, C. Li, H. Fu, Y. Zhao, Q. Huang, and S. Kwong, "Global-and-Local collaborative learning for co-salient object detection," *IEEE Trans. Cybern.*, vol. 53, no. 3, pp. 1920–1931, Mar. 2023.

[3] X. Zhou, K. Shen, L. Weng, R. Cong, B. Zheng, J. Zhang, and C. Yan, "Edge-guided recurrent positioning network for salient object detection in optical remote sensing images," *IEEE Trans. Cybern.*, vol. 53, no. 1, pp. 539–552, Jan. 2023.

[4] K. Song, J. Wang, and Y. Bao, "A novel visible-depth-thermal image dataset of salient object detection for robotic visual perception," *IEEE/ASME Trans. Mechatronics*, vol. 28, no. 3, pp. 1558–1569, Jun. 2022.

[5] Z. Liu, X. Li, P. Luo, C. C. Loy, and X. Tang, "Deep learning Markov random field for semantic segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 8, pp. 1814–1828, Aug. 2018.

[6] W. Wang, J. Shen, and L. Shao, "Video salient object detection via fully convolutional networks," *IEEE Trans. Image Process.*, vol. 27, no. 1, pp. 38–49, Jan. 2018.

[7] L. Huang, B. Ma, J. Shen, H. He, L. Shao, and F. Porikli, "Visual tracking by sampling in part space," *IEEE Trans. Image Process.*, vol. 26, no. 12, pp. 5800–5810, 2017.

[8] X. Li, L. Zhao, L. Wei, M.-H. Yang, F. Wu, Y. Zhuang, H. Ling, and J. Wang, "DeepSaliency: Multi-task deep neural network model for salient object detection," *IEEE Trans. Image Process.*, vol. 25, no. 8, pp. 3919–3930, Aug. 2016.

[9] P. Zhang, W. Liu, Y. Lei, and H. Lu, "Hyperfusion-Net: Hyper-densely reflective feature fusion for salient object detection," *Pattern Recognit.*, vol. 93, pp. 521–533, Sep. 2019.

[10] L. Qin, Y. Shi, Y. He, J. Zhang, X. Zhang, Y. Li, T. Deng, and H. Yan, "ID-YOLO: Real-time salient object detection based on the driver's fixation region," *IEEE Trans. Intell. Transp. Syst.*, vol. 23, no. 9, pp. 15898–15908, Sep. 2022.

[11] A. R. Pathak, M. Pandey, and S. Rautaray, "Application of deep learning for object detection," *Proc. Comput. Sci.*, vol. 132, no. 1, pp. 1706–1717, Jan. 2018.

[12] G. Dimas, P. Gatoula, and D. K. Iakovidis, "MonoSOD: Monocular salient object detection based on predicted depth," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, Xi'an, China, May 2021, pp. 4377–4383.

[13] C. Craye, D. Filliat, and J.-F. Goudou, "Environment exploration for object-based visual saliency learning," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, Stockholm, Sweden, May 2016, pp. 2303–2309.

[14] Z. Gao, H. Zhang, S. Dong, S. Sun, X. Wang, G. Yang, W. Wu, S. Li, and V. H. C. de Albuquerque, "Salient object detection in the distributed cloud-edge intelligent network," *IEEE Netw.*, vol. 34, no. 2, pp. 216–224, Mar. 2020.

[15] C. Wang, S. Dong, X. Zhao, G. Papanastasiou, H. Zhang, and G. Yang, "SaliencyGAN: Deep learning semisupervised salient object detection in the fog of IoT," *IEEE Trans. Ind. Informat.*, vol. 16, no. 4, pp. 2667–2676, Apr. 2020.

[16] T. Liu, Z. Yuan, J. Sun, J. Wang, N. Zheng, X. Tang, and H.-Y. Shum, "Learning to detect a salient object," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 2, pp. 353–367, Feb. 2011.

[17] S. Y. Kwak, B. C. Ko, and H. Byun, "Automatic salient-object extraction using the contrast map and salient points," in *Proc. PCM*, Tokyo, Japan, 2004, pp. 138–145.

[18] T. Ikeda and M. Ikehara, "RGB-D salient object detection using saliency and edge reverse attention," *IEEE Access*, vol. 11, pp. 68818–68825, 2023.

[19] D. P. Fan, Y. Zhai, and A. Borji, "BBS-Net: RGB-D salient object detection with a bifurcated backbone strategy network," in *Proc. ECCV*, Glasgow, U.K., 2020, pp. 275–292.

[20] Y. Han, L. Wang, A. Du, and S. Jiang, "LIANet: Layer interactive attention network for RGB-D salient object detection," *IEEE Access*, vol. 10, pp. 25435–25447, 2022.

[21] R. Zhao, W. Ouyang, H. Li, and X. Wang, "Saliency detection by multi-context deep learning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Boston, BA, USA, Jun. 2015, pp. 1265–1274.

[22] J. Wu, G. Han, H. Wang, H. Yang, Q. Li, D. Liu, F. Ye, and P. Liu, "Progressive guided fusion network with multi-modal and multi-scale attention for RGB-D salient object detection," *IEEE Access*, vol. 9, pp. 150608–150622, 2021.

[23] 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 Trans. Multimedia*, vol. 24, pp. 1651–1664, 2022.

[24] Y. Endo and C. Premachandra, "Development of a bathing accident monitoring system using a depth sensor," *IEEE Sensors Lett.*, vol. 6, no. 2, pp. 1–4, Feb. 2022.

[25] H. Matsumura and C. Premachandra, "Deep-Learning-Based stair detection using 3D point cloud data for preventing walking accidents of the visually impaired," *IEEE Access*, vol. 10, pp. 56249–56255, 2022.

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

[27] Z. Wu, G. Allibert, F. Meriaudeau, C. Ma, and C. Demonceaux, "HiDAnet: RGB-D salient object detection via hierarchical depth awareness," *IEEE Trans. Image Process.*, vol. 32, pp. 2160–2173, 2023.

[28] J. X. Zhao, J. J. Liu, D. P. Cao, Y. Cao, J. Yang, and M. M. Cheng, "EGNet: Edge guidance network for salient object detection," in *Proc. ICCV*, Seoul, (South) Korea, 2019, pp. 8779–8788.

[29] L. Gao, B. Liu, P. Fu, and M. Xu, "Depth-aware inverted refinement network for RGB-D salient object detection," *Neurocomputing*, vol. 518, pp. 507–522, Jan. 2023.

[30] J. Li, W. Ji, M. Zhang, Y. Piao, H. Lu, and L. Cheng, "Delving into calibrated depth for accurate RGB-D salient object detection," *Int. J. Comput. Vis.*, vol. 131, no. 4, pp. 855–876, Apr. 2023.

[31] D. Theckedath and R. R. Sedamkar, "Detecting affect states using VGG16, ResNet50 and SE-ResNet50 networks," *Social Netw. Comput. Sci.*, vol. 1, no. 2, pp. 1–7, Mar. 2020.

[32] R. Ju, L. Ge, W. Geng, T. Ren, and G. Wu, "Depth saliency based on anisotropic center-surround difference," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Paris, France, Oct. 2014, pp. 1115–1119.

[33] H. Peng, B. Li, W. Xiong, W. Hu, and R. Ji, "RGBD salient object detection: A benchmark and algorithms," in *Proc. Eur. Conf. Comput. Vis.*, Zurich, Switzerland, 2014, pp. 92–109.

[34] Y. Niu, Y. Geng, X. Li, and F. Liu, "Leveraging stereopsis for saliency analysis," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Providence, RI, USA, Jun. 2012, pp. 454–461.

[35] G. Li and C. Zhu, "A three-pathway psychobiological framework of salient object detection using stereoscopic technology," in *Proc. IEEE Int. Conf. Comput. Vis. Workshops (ICCVW)*, Venice, Italy, Oct. 2017, pp. 3008–3014.

[36] 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 Trans. Neural Netw. Learn. Syst.*, vol. 32, no. 5, pp. 2075–2089, May 2021.

[37] Y. Cheng, H. Fu, X. Wei, J. Xiao, and X. Cao, "Depth enhanced saliency detection method," in *Proc. Int. Conf. Internet Multimedia Comput. Service*, Jul. 2014, pp. 23–27.

[38] N. Li, J. Ye, Y. Ji, H. Ling, and J. Yu, "Saliency detection on light field," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Columbus, OH, USA, Jun. 2014, pp. 2806–2813.

[39] R. Cong, J. Lei, H. Fu, M.-M. Cheng, W. Lin, and Q. Huang, "Review of visual saliency detection with comprehensive information," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 29, no. 10, pp. 2941–2959, Oct. 2019.

[40] D.-P. Fan, C. Gong, Y. Cao, B. Ren, M.-M. Cheng, and A. Borji, "Enhanced-alignment measure for binary foreground map evaluation," in *Proc. 27th Int. Joint Conf. Artif. Intell.*, Stockholm, Sweden, Jul. 2018, pp. 698–704.

[41] D.-P. Fan, M.-M. Cheng, Y. Liu, T. Li, and A. Borji, "Structure-measure: A new way to evaluate foreground maps," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Venice, Italy, Oct. 2017, pp. 4548–4557.

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

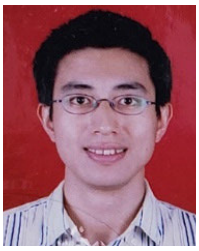
[43] R. Cong, Q. Lin, C. Zhang, C. Li, X. Cao, Q. Huang, and Y. Zhao, "CIR-Net: Cross-modality interaction and refinement for RGB-D salient object detection," *IEEE Trans. Image Process.*, vol. 31, pp. 6800–6815, 2022.

[44] J.-X. Zhao, Y. Cao, D.-P. Fan, M.-M. Cheng, X.-Y. Li, and L. Zhang, "Contrast prior and fluid pyramid integration for RGBD salient object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Los Angeles, CA, USA, Jun. 2019, pp. 3922–3931.

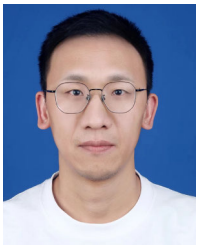
[45] W. Ji, G. Yan, J. Li, Y. Piao, S. Yao, M. Zhang, L. Cheng, and H. Lu, "DMRA: Depth-induced multi-scale recurrent attention network for RGB-D saliency detection," *IEEE Trans. Image Process.*, vol. 31, pp. 2321–2336, 2022.

[46] J. Zhang, D.-P. Fan, Y. Dai, S. Anwar, F. S. Saleh, T. Zhang, and N. Barnes, "UC-Net: Uncertainty inspired RGB-D saliency detection via conditional variational autoencoders," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 8579–8588.

[47] Y. Pang, L. Zhang, X. Zhao, and H. Lu, "Hierarchical dynamic filtering network for RGB-D salient object detection," in *Proc. ECCV*, Glasgow, U.K., 2020, pp. 235–252.
[48] J. Zhang, D.-P. Fan, Y. Dai, X. Yu, Y. Zhong, N. Barnes, and L. Shao, "RGB-D saliency detection via cascaded mutual information minimization," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Montreal, QC, Canada, Oct. 2021, pp. 4318–4327.
[49] K. Fu, D.-P. Fan, G.-P. Ji, Q. Zhao, J. Shen, and C. Zhu, "Siamese network for RGB-D salient object detection and beyond," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 9, pp. 5541–5559, Sep. 2022.
[50] H. Bi, R. Wu, Z. Liu, H. Zhu, C. Zhang, and T.-Z. Xiang, "Cross-modal hierarchical interaction network for RGB-D salient object detection," *Pattern Recognit.*, vol. 136, Apr. 2023, Art. no. 109194.
[51] Z. Chen, H. Zhou, J. Lai, L. Yang, and X. Xie, "Contour-aware loss: Boundary-aware learning for salient object segmentation," *IEEE Trans. Image Process.*, vol. 30, pp. 431–443, 2021.
[52] Y. Wang, X. Zhao, X. Hu, Y. Li, and K. Huang, "Focal boundary guided salient object detection," *IEEE Trans. Image Process.*, vol. 28, no. 6, pp. 2813–2824, Jun. 2019.
[53] W. Guan, T. Wang, J. Qi, L. Zhang, and H. Lu, "Edge-aware convolution neural network based salient object detection," *IEEE Signal Process. Lett.*, vol. 26, no. 1, pp. 114–118, Jan. 2019.
[54] Y. Zhuge, G. Yang, P. Zhang, and H. Lu, "Boundary-guided feature aggregation network for salient object detection," *IEEE Signal Process. Lett.*, vol. 25, no. 12, pp. 1800–1804, Dec. 2018.
[55] Z. Wu, L. Su, and Q. Huang, "Stacked cross refinement network for edge-aware salient object detection," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 7264–7273.
[56] S. Zhou, J. Zhang, J. Wang, F. Wang, and D. Huang, "SE2Net: Siamese edge-enhancement network for salient object detection," 2019, *arXiv:1904.00048*.
[57] J. Su, J. Li, Y. Zhang, C. Xia, and Y. Tian, "Selectivity or invariance: Boundary-aware salient object detection," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 3798–3807.
[58] F. Lin, C. Yang, H. Li, and B. Jiang, "Boundary-aware salient object detection via recurrent two-stream guided refinement network," 2019, *arXiv:1912.05236*.
[59] J. Zhang, Y. Dai, F. Porikli, and M. He, "Deep edge-aware saliency detection," 2017, *arXiv:1708.04366*.
[60] W. Wang, S. Zhao, J. Shen, S. C. H. Hoi, and A. Borji, "Salient object detection with pyramid attention and salient edges," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Los Angeles, CA, USA, Jun. 2019, pp. 1448–1457.
[61] B. Jiang, Z. Zhou, X. Wang, J. Tang, and B. Luo, "CmSalGAN: RGB-D salient object detection with cross-view generative adversarial networks," *IEEE Trans. Multimedia*, vol. 23, pp. 1343–1353, 2021.
[62] M. Zhang, W. Ren, Y. Piao, Z. Rong, and H. Lu, "Select, supplement and focus for RGB-D saliency detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 3469–3478.
[63] Z. Liu, Y. Tan, Q. He, and Y. Xiao, "SwinNet: Swin transformer drives edge-aware RGB-D and RGB-T salient object detection," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 7, pp. 4486–4497, Jul. 2022.
[64] X. Wu, D. Hong, and J. Chanussot, "UIU-Net: U-Net in U-Net for infrared small object detection," *IEEE Trans. Image Process.*, vol. 32, pp. 364–376, 2023.
[65] H. Wang, S. Xie, L. Lin, Y. Iwamoto, X.-H. Han, Y.-W. Chen, and R. Tong, "Mixed transformer U-Net for medical image segmentation," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Shenzhen, China, May 2022, pp. 2390–2394.



YANBIN PENG received the Ph.D. degree from Zhejiang University, China, in 2008. He is currently an Associate Professor with the Zhejiang University of Science and Technology. His current research interests include computer vision, image processing, deep learning, object detection, and their applications.



MINGKUN FENG received the Ph.D. degree in information and communication engineering from the Nanjing University of Posts and Telecommunications, China, in 2016. He is currently an Associate Professor with the Zhejiang University of Science and Technology, China. His research interests include pattern recognition, machine learning, and artificial intelligence.

ZHIJUN ZHENG received the Ph.D. degree from Xi'an Jiaotong University, China. He is currently an Associate Professor with the Zhejiang University of Science and Technology. His research interests include machine learning, multimedia analysis retrieval, image retrieval, and statistical learning.