IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 23 February 2024, accepted 18 March 2024, date of publication 25 March 2024, date of current version 29 March 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3381524

RESEARCH ARTICLE

RGB-D Salient Object Detection Based on Cross-Modal and Cross-Level Feature Fusion

YANBIN PENG, (Member, IEEE), ZHINIAN ZHAI, AND MINGKUN FENG
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 National Natural Science Foundation of China under Grant 61972357, and in part by the Basic Public Welfare Research Program of Zhejiang Province under Grant LGF22F020017.

ABSTRACT Existing RGB-D saliency detection models have not fully considered the differences between features at various levels, and lack an effective mechanism for cross-level feature fusion. This article proposes a novel cross-modality cross-level fusion learning framework. The framework mainly contains three modules: Attention Enhancement Module (AEM), Modality Feature Fusion Module (MFM), and Graph Reasoning Module (GRM). AEM is used to enhance the features of the two modalities. MFM is used to integrate the features of the two modalities to achieve cross-modality feature fusion. Subsequently, the modality fusion features are divided into high-level features and low-level features. The high-level features contain the semantic localization information of salient objects, and the low-level features contain the detailed information of salient objects. GRM extends the semantic localization information of salient objects in the high-level features from pixel features to the entire salient object area, thereby achieving cross-level feature fusion. This framework can effectively eliminate background noise and enhance the model's expressiveness. Extensive experiments were conducted on seven widely used datasets, and the results show that the new method outperforms nine current state-of-the-art RGB-D SOD methods.

INDEX TERMS Salient object detection, RGB-D, attention mechanism.

I. INTRODUCTION

In the field of computer vision, RGB-D Salient Object Detection (SOD) has progressively evolved into a significant research direction, playing a crucial role in numerous application domains. For instance, in robot navigation [1], [2], salient object detection aids robots in gaining a more profound understanding of their environment, thereby informing their decision-making. In the realm of object tracking [3], [4], salient object detection can effectively assist the system in accurately locating and tracing objects of interest. In terms of scene understanding [5], [6], salient object detection helps the system to highlight and comprehend the key objects and events within a scene. These applications collectively propel the research and innovative development of RGB-D salient object detection technology.

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

The primary task of RGB-D salient object detection [7], [8] is to identify and highlight the primary objects or events within an image. To accomplish this goal, it is usually necessary to integrate the information from RGB images and depth maps. The combination of RGB images and depth maps offers a comprehensive view of both color and spatial attributes, which is crucial for accurately detecting salient objects in complex environments. In Figure 1, we showcase an intricate image with a cluttered background, where traditional RGB-based methods struggle to produce precise saliency maps. By incorporating depth information, RGB-D salient object detection can more effectively discern and highlight the salient regions amidst the challenging backdrop. Despite the important progress that RGB-D salient object detection has made [9], [10], [11], [12], there are still some challenges in this field that need to be overcome.

Firstly, existing RGB-D salient object detection models often fail to fully consider the differences between features at different levels. More specifically, these models

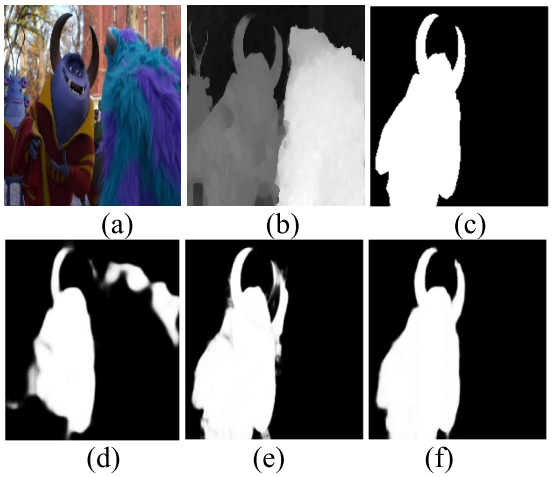


FIGURE 1. Visual comparison of RGB and RGB-D SOD results. (a) RGB image, (b) Depth image, (c) Ground truth saliency map, (d) Predicted saliency map from RGB-based method (MINet [21]), (e) Predicted saliency map from RGB-D-based method (BTSNet [59]), (f) Predicted saliency map from our proposed method.

frequently overlook the complementarity between high-level features (which contain semantic location information of salient objects) and low-level features (which contain detailed information of salient objects). As a result, these models are unable to fully exploit and utilize these two types of features, which adversely affects their detection performance. Secondly, existing RGB-D salient object detection models lack an effective cross-level fusion mechanism, and are unable to fully extract and integrate information from features at different levels. This issue further constrains these models' ability to accurately detect and locate salient objects.

To address the aforementioned issues, we propose a novel Cross-Modal and Cross-Level Fusion Learning Framework (CMCL). This framework comprises three main modules, namely the Attention Enhancement Module (AEM), the Modal Feature Fusion Module (MFM), and the Graph Reasoning Module (GRM). AEM primarily focuses on enhancing the features of the two modalities. By incorporating an attention mechanism, AEM can more precisely focus on the objects of interest, thereby improving the model's detection performance. MFM is responsible for fusing features from the two modalities. With the aid of upsampling and fusion techniques, MFM can effectively integrate information from RGB images and depth maps, achieving cross-modal feature fusion. GRM is chiefly tasked with implementing cross-level feature fusion. By extending the semantic localization information of salient objects from high-level features to the entire salient object area, GRM can effectively achieve cross-level feature integration. The design of this framework takes into account various factors in the salient object detection process, effectively eliminates background noise, and simultaneously enhances the model's representational capacity and detection accuracy. By comprehensively utilizing advanced techniques like attention mechanisms, deep learning, and graph convolutions, our framework can achieve more precise and robust salient object detection.

To validate the effectiveness of our proposed method, we conducted extensive experiments on seven widely used datasets. These datasets encompass a variety of scenes and targets, thoroughly evaluating the performance of our method in diverse situations. The experimental results reveal that our method surpasses nine of the most advanced current methods across four evaluation metrics. These results robustly demonstrate the superiority of our method in the field of RGB-D salient object detection.

II. RELATED WORKS

In recent years, there has been extensive research in the field of RGB-D salient object detection. In this section, we will primarily discuss three significant topics: salient object detection, cross-modal feature fusion, and graph convolutional networks.

A. SALIENT OBJECT DETECTION

Salient object detection is a significant research direction in the field of computer vision, aiming to identify and highlight the most important objects or events in an image. Traditional methods primarily rely on visual features such as color, texture, and contrast for target detection [13], [14], [15], [16], [17]. For instance, Feng et al. [15] proposed a novel saliency feature, namely, Local Background Enclosure (LBE), for salient object detection. The LBE feature effectively addresses the contrast issue in depth scenes by capturing the angle distribution range of candidate regions and their belonging objects relative to the background. Cheng et al [17] introduced a salient region detection algorithm based on global contrast, which extracts saliency by simultaneously evaluating global contrast differences and spatial coherence, thereby enhancing the performance of salient object detection. However, the performance of these methods is limited as they struggle to handle complex scenes and changing environments. When dealing with complex backgrounds or situations where the target color closely resembles the background, the performance of these methods declines.

To address these issues, deep learning techniques have been introduced to salient object detection [18], [19], [20], [21]. These methods, mainly based on Convolutional Neural Networks (CNNs), can automatically extract more abundant and distinctive features by learning from a large number of training samples. For example, Qin et al. [20] proposed the BASNet (Boundary-Aware Salient Object Detection) model, which uses a densely supervised encoder-decoder network for saliency prediction in the prediction phase, and a residual refinement module to refine the predicted saliency map during the refinement phase. Pang et al. [21] proposed a model named MINet (Multi-scale Interactive Network), which consists of Aggregate Interaction Modules (AIM) and Self-Interaction Modules (SIM). The AIM incorporates features from adjacent levels, while the SIM extracts more

effective multi-scale features from the aggregated features. Moreover, MINet also introduced a loss function named Consistency-enhanced Loss (CEL), which is used to highlight the foreground/background differences and maintain intra-class consistency. However, these methods still primarily rely on RGB image information, ignoring the important role that depth information plays in salient object detection.

B. CROSS-MODAL FEATURE FUSION

With the proliferation of depth cameras, RGB-D images (comprising RGB images and depth images) have begun to be widely used in computer vision tasks [22], [23]. As depth images can provide additional spatial and shape information, this opens up new possibilities for salient object detection.

Effectively fusing RGB features and depth features has become a crucial issue in RGB-D salient object detection. Some early methods adopted simple stacking or concatenation approaches for feature fusion [24], [25], [26], but these methods could not fully account for the complementarity and differences between RGB and depth features. To better fuse cross-modal features, some researchers proposed depth-learning-based fusion methods [27], [28], [29], [30], which use multi-task learning or attention mechanisms to simultaneously learn and fuse RGB and depth features. For instance, Gao et al. [27] introduced an innovative fusion network known as MMNet, which is comprised of a dual-module design featuring a Cross-Modal Multi-Stage Fusion Module (CMFM) and a Bi-directional Multi-Scale Decoder (BMD). The CMFM is designed to enhance crucial features during the response phase and merge them with cross-modal characteristics at the stage of adversarial fusion. In contrast, the BMD is structured to assimilate fused features from multiple levels, capturing both micro and macro details of salient objects and thereby elevating the efficacy of multi-modal saliency detection. Chen et al. [28] proposed a multi-scale multi-path fusion network with cross-modal interactions (MMCI), which diversifies the fusion path into a global inference path and another local capturing path, while introducing cross-modal interactions at multiple levels. This improves the traditional dual-stream fusion architecture of a single fusion path. Compared to traditional dual-stream architectures, the MMCI network can provide a more adaptive, flexible fusion process, thereby simplifying the optimization process and achieving more effective fusion.

Sun et al. [60] introduces CATNet, a novel cascaded and aggregated Transformer network for RGB-D salient object detection, which excels in integrating multi-scale features and enhancing feature representation through key modules like AFEM, CMFM, and CCD. This design significantly improves detection performance across various benchmarks. Wang et al. [61] introduces an Attention-guided Multi-modality Interaction Network (AMINet) for RGB-D Salient Object Detection, focusing on addressing challenges such as low-quality depth maps and ineffective salient map predictions with clear boundaries. It proposes novel components like the Depth Enhancement Module (DEM), Cross-Modality Attention Module (CMAM), and Boundary-Aware Module (BAM) to enhance depth quality, locate salient regions effectively, and preserve boundary details, respectively. Wang et al. [62] presents DCMNet, a Discriminant and Cross-Modality Network for RGB-D Salient Object Detection, which includes a novel Depth Decomposition and Recomposition Module (DDRM) for enhancing low-quality depth maps and a Multi-Cross Attention Module (MCAM) for effectively leveraging spatial and channel attention. By integrating these modules with the Res2Net model as an Image Pretraining Model (IPM), DCMNet significantly improves detection accuracy. Tang et al. [63] introduces HRTransNet, a model for two-modality salient object detection leveraging the HRFormer architecture to maintain high-resolution representations and achieve superior detection performance. It incorporates innovative modules for effective fusion of primary and supplementary modalities, enhancing detail and accuracy in salient object detection across various conditions like RGB-D, RGB-T, and light field scenarios.

C. GRAPH CONVOLUTIONAL NETWORKS

Graph Convolutional Networks (GCNs) are deep learning models designed specifically for graph-structured data [31], [32], [33]. Unlike traditional Convolutional Neural Networks that operate on regular grids, GCNs perform convolution operations on graph structures, enabling them to directly handle graph-structured data, including social networks, protein networks, citation networks, etc. The main idea behind GCNs is to capture local patterns in the graph through convolution operations while preserving the graph's global structure. In GCNs, each node updates its features by aggregating information from its neighbor nodes, a process that can be viewed as a convolution operation on the graph.

Given their ability to directly handle graph-structured data, Graph Convolutional Networks have been widely applied in many fields, such as social network analysis, bioinformatics, recommendation systems, and more [34], [35], [36]. For example, in social network analysis, GCNs can be used to predict links within the social network or infer user attributes. In bioinformatics, GCNs can be employed to predict interactions between proteins. In recommendation systems, GCNs can be used to establish complex interaction relationships between users and items. Furthermore, Graph Convolutional Networks (GCNs) have been introduced into image segmentation. Li and Gupta [37] proposed a method for learning graph representations from two-dimensional feature maps, which transforms the 2D image into a graph structure. The vertices of the graph define clusters of pixels (i.e., ''regions''), while the edges measure the similarity between these clusters in the feature space. This method further learns to propagate information across all vertices of the graph and can project the learned graph representation back to the 2D grid. Lu et al. [38] proposed a graph model initialized by a fully

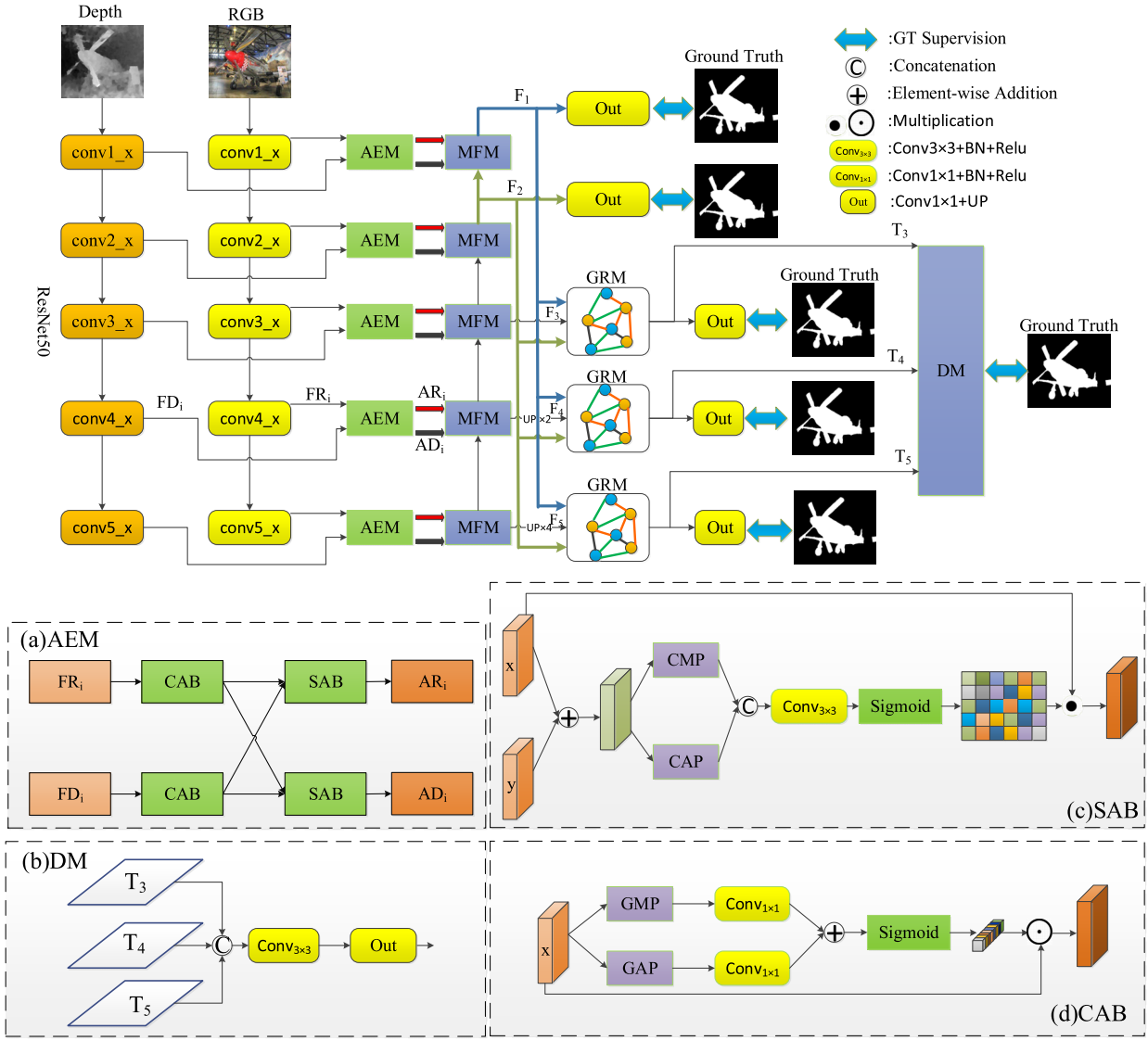


FIGURE 2. The framework of the proposed method.

convolutional network (FCN), named Graph-FCN, for image semantic segmentation. Te et al. [39] introduced Hypergraph Convolutional Neural Networks (HGCNN) to address 3D facial anti-spoofing issues. Zhang et al. [40] proposed the Dual Graph Convolutional Network (DGCNet) to leverage long-range contextual information for semantic segmentation, by modeling two orthogonal graphs within a single framework to represent the global context of input features.

However, GCNs do have limitations, such as certain requirements for the scale and complexity of the graph, and the need for specific techniques to preprocess the graph's structure and node features. Moreover, how to design more effective graph convolution operations to capture complex patterns in the graph remains an open research question.

III. THE PROPOSED METHOD

This section first introduces the overall architecture of the CMCL model, followed by a detailed description of the key components within the model, including the Attention Enhanced Module (AEM), the Modal Feature Fusion Module (MFM), the Graph Reasoning Module (GRM), and the Decoding Module (DM). Finally, we will elaborate on the overall loss function of the model.

A. OVERVIEW

As depicted in Figure 2, the CMCL model proposed in this paper primarily consists of five key components: the ResNet50 feature encoder, the Attention Enhancement Module (AEM), the Modality Feature Fusion Module (MFM), the Graph Reasoning Module (GRM), and the Decoding Module (DM). Specifically, the ResNet50 encoder is responsible for extracting features from the input RGB and depth images. The AEM module is designed to eliminate noise and enhance salient object information. The MFM module is utilized for the fusion of modality features. The GRM module accomplishes cross-level feature fusion between high-level

and low-level features. This module, based on the semantic location information of high-level features, expands salient object information from the pixel level to the image level. Lastly, the DM module is entrusted with the task of generating the final predicted saliency map.

Given an RGB image $\mathbf{R}_i \in \mathbb{R}^{H\times W\times 3}$ and a depth image $\mathbf{D}_i \in \mathbb{R}^{H\times W\times 1}$, this paper employs ResNet50 [41] as the backbone network to extract pyramid features across five scales. The extracted features from the RGB image are denoted as $\mathbf{FR}_i \in \mathbb{R}^{H/2^i\times W/2^i\times C_i}$ while those from the depth image are denoted as $\mathbf{FD}_i \in \mathbb{R}^{H/2^i\times W/2^i\times C_i}$. Here, $i \in \{1, 2, 3, 4, 5\}$, $C_i \in \{64, 256, 512, 1024, 2048\}$, while H and W stand for the height and width of the image, respectively. After the attention enhancement module and the modality feature fusion module process these features, we obtain a fused feature map $\mathbf{F}_i \in \mathbb{R}^{H/2^i\times W/2^i\times C_i}$. Next, using the GRM module, we achieve cross-level fusion between low-level features $\{\mathbf{F}_1, \mathbf{F}_2\}$ and high-level features $\mathbf{F}_3$, $\mathbf{F}_4$, and $\mathbf{F}_5$, further generating three feature maps $\mathbf{T}_3$, $\mathbf{T}_4$, and $\mathbf{T}_5$. Finally, these three feature maps are input into the decoding module to produce the ultimate predicted saliency map.

B. ATTENTION ENHANCEMENT MODULE

Following the extraction of features from RGB and depth images, we employ attention mechanisms to augment the two distinct types of features. As each feature set encapsulates disparate information, the RGB image features convey the color, appearance, and texture details of objects within the scene, whereas the depth image features encapsulate the distance, shape, and spatial relationships of the objects. Consequently, the significance of channels varies between these two modalities of features. For this reason, we apply two channel attention modules to enhance the channels of the RGB image and depth image features, respectively. Subsequently, considering that the spatial location of the salient objects in the two modal images is consistent, we combine the two modal features to generate a shared location map. Using this location map, we enhance the spatial attention of each modal feature, thereby obtaining attention-enhanced modal features. Specifically, given the RGB image features $\mathbf{FR}_i$ and the depth image features $\mathbf{FD}_i$, we perform channel attention operations on these two modal features separately. The process is as follows:

$$\mathbf{CR}_i = \mathrm{CAB}(\mathbf{FR}_i) \tag{1}$$
$$\mathbf{CD}_i = \mathrm{CAB}(\mathbf{FD}_i) \tag{2}$$

where $\mathbf{CR}_i$ and $\mathbf{CD}_i$ represent the channel-enhanced RGB image features and depth image features, respectively. The symbol $\mathrm{CAB}(\cdot)$ denotes the channel attention module, defined as follows:

$$\mathrm{CAB}(x) = x \odot \mathrm{Sigmoid}(\mathrm{M}(x) \oplus \mathrm{A}(x)) \tag{3}$$
$$\mathrm{M}(x) = \mathrm{Conv}_{1\times 1}(\mathrm{GMP}(x)) \tag{4}$$
$$\mathrm{A}(x) = \mathrm{Conv}_{1\times 1}(\mathrm{GAP}(x)) \tag{5}$$

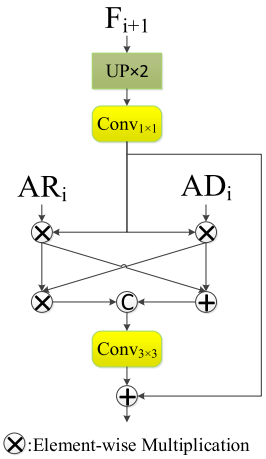


FIGURE 3. Modal feature fusion module.

where $\mathrm{GMP}(\cdot)$ denotes the Global Max Pooling operation, $\mathrm{GAP}(\cdot)$ represents the Global Average Pooling operation, and $\mathrm{Conv}_{1\times 1}$ refers to the convolution operation with a 1×1 kernel. $\oplus$ symbolizes the element-wise addition operation, $\odot$ signifies the element-wise multiplication operation with spatial expansion, and $\mathrm{Sigmoid}(\cdot)$ indicates the Sigmoid activation function.

Next, for the channel-enhanced RGB image features $\mathbf{CR}_i$ and depth image features $\mathbf{CD}_i$, we first employ an element-wise addition operation to combine these two modal features, resulting in a shared location map. We then use this shared location map to adjust the spatial location weights of the RGB image features and depth image features. This leads to the attention-enhanced RGB image features $\mathbf{AR}_i$ and depth image features $\mathbf{AD}_i$. The process is articulated as follows:

$$\mathbf{AR}_i = \mathrm{SAB}(\mathbf{CR}_i, \mathbf{CD}_i) \tag{6}$$
$$\mathbf{AD}_i = \mathrm{SAB}(\mathbf{CD}_i, \mathbf{CR}_i) \tag{7}$$
$$\mathrm{SAB}(x, y) = x \bullet \mathrm{Sigmoid}(\mathrm{Conv}_{3\times 3}(C(x, y))) \tag{8}$$
$$\mathrm{C}(x, y) = \mathrm{Concat}(\mathrm{CMP}(x \oplus y), \mathrm{CAP}(x \oplus y)) \tag{9}$$

where $\mathrm{CMP}(\cdot)$ denotes the Max Pooling operation along the channel, $\mathrm{CAP}(\cdot)$ represents the Average Pooling operation along the channel, and $\mathrm{Concat}(\cdot)$ indicates the concatenation operation. $\mathrm{Conv}_{3\times 3}$ refers to the convolution operation with a 3×3 kernel, $\bullet$ symbolizes the element-wise multiplication operation with channel expansion, and $\mathrm{SAB}(\cdot)$ stands for the spatial attention module.

C. MODAL FEATURE FUSION MODULE

After obtaining the attention-enhanced modal features, we employ the modal feature fusion module to further generate modal fusion features. More specifically, we use $\mathbf{F}_{i+1}$ to represent the modal fusion feature of the $(i+1)$th layer, while $\mathbf{AR}_i$ and $\mathbf{AD}_i$ denote the attention-enhanced RGB image features and depth image features of the ith layer, respectively. As shown in Figure 3, we multiply these two features, $\mathbf{AR}_i$

and $\mathbf{AD}_i$, with $\mathbf{F}_{i+1}$, as expressed below:

$$\mathbf{F}_{i+1}^{\mathrm{U}} = \mathrm{Conv}_{1\times1}(\mathrm{UP}(\mathbf{F}_{i+1})) \tag{10}$$

$$\mathbf{m}_i = \mathbf{AR}_i \otimes \mathbf{F}_{i+1}^{\mathrm{U}} \tag{11}$$

$$\mathbf{n}_i = \mathbf{AD}_i \otimes \mathbf{F}_{i+1}^{\mathrm{U}} \tag{12}$$

where $\mathrm{UP}(\cdot)$ denotes the operation of upsampling by a factor of two, and $\mathrm{Conv}_{1\times1}$ convolution is used to adjust the number of channels to match those of $\mathbf{AR}_i$ and $\mathbf{AD}_i$. $\otimes$ symbolizes the element-wise multiplication operation. Subsequently, the feature maps $\mathbf{m}_i$ and $\mathbf{n}_i$ are fused through addition, multiplication, and concatenation operations, as articulated below:

$$\mathbf{k}_i = \mathrm{Conv}_{3\times3}(\mathrm{Concat}((\mathbf{m}_i \oplus \mathbf{n}_i),(\mathbf{m}_i \otimes \mathbf{n}_i))) \tag{13}$$

Finally, drawing inspiration from the widely adopted UNet architecture [42], we generate multi-scale fusion features. In accordance with the decoding strategy of UNet, we iteratively merge the high-level fusion features down to the lower levels. The process is detailed as follows:

$$\mathbf{F}_i = \begin{cases} \mathbf{k}_i \oplus \mathbf{F}_{i+1}^{\mathrm{U}}, & i = 1, 2, 3, 4 \\ \mathrm{Conv}_{3\times3}(\mathrm{Concat}((\mathbf{AR}_5 \oplus \mathbf{AD}_5), (\mathbf{AR}_5 \otimes \mathbf{AD}_5))), & i = 5 \end{cases} \tag{14}$$

It is worth noting that for the fusion feature of the fifth layer, since there are no higher-level fusion features, we only perform fusion on $\mathbf{AR}_5$ and $\mathbf{AD}_5$.

D. GRAPH REASONING MODULE

We divide the modal fusion features into two groups, namely high-level features Q2=$\{\mathbf{F}_3, \mathbf{F}_4, \mathbf{F}_5\}$ and low-level features Q1=$\{\mathbf{F}_1, \mathbf{F}_2\}$. The high-level features contain rich semantic locational information, whereas the low-level features encompass a wealth of detailed information, such as texture, color, and boundary information. In many salient object scenes, there are often situations where the foreground is similar to the background, or the background is complex. To enhance the representation of salient object detail information in the low-level features and to eliminate noise, we use actual saliency maps to supervise the low-level features.

To make comprehensive use of the advantages of high-level features and low-level features, we designed a graph reasoning module. This module introduces non-local operations [43] into graph convolutional networks [37], [38] for inference, and simultaneously performs cross-level fusion. This design extends the salient object localization information in the high-level features to the low-level features, thereby obtaining a complete and clear salient object. Specifically, we perform cross-level fusion of the low-level features $\{\mathbf{F}_1, \mathbf{F}_2\}$ with the high-level features $\mathbf{F}_3$, $\mathbf{F}_4$, and $\mathbf{F}_5$ respectively, generating three optimized feature maps $\mathbf{T}_3$, $\mathbf{T}_4$, and $\mathbf{T}_5$. We will elaborate on this with $\mathbf{F}_3$ as an example.

As shown in Figure 4, given high-level fusion feature map $\mathbf{F}_3$, which contains semantic localization information, and two low-level fusion feature maps $\mathbf{F}_1$ and $\mathbf{F}_2$, which have rich detail information. H and W represent the height and width

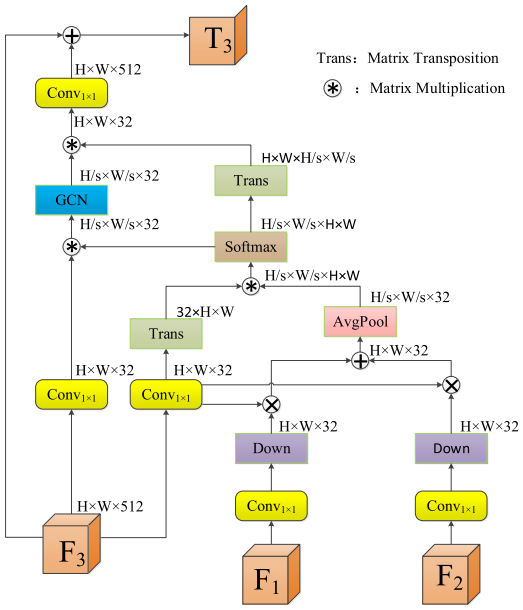


FIGURE 4. Graph reasoning module.

of the feature map $\mathbf{F}_3$, respectively. Our aim is to construct a projection matrix P that can combine the information of $\mathbf{F}_1$ and $\mathbf{F}_2$, and map the feature map $\mathbf{F}_3$ into a graph structure, in which each vertex represents a pixel region, and the edges between vertices represent the relationships between these regions. Then, we use a single-layer graph convolutional network (GCN) [37] to perform inference on the graph structure, and project the learned graph structure back to the pixel grid to obtain an optimized feature map $\mathbf{T}_3$ of the same size. Specifically, we first reduce the number of channels of $\mathbf{F}_3$ to 32 through two convolution operations with 1×1 convolution kernels. This process is expressed as:

$$\mathbf{U} = \mathrm{Conv}_{1\times1}(\mathbf{F}_3) \tag{15}$$

$$\mathbf{V} = \mathrm{Conv}_{1\times1}(\mathbf{F}_3) \tag{16}$$

where $\mathbf{U}$ and $\mathbf{V}$ represent the results of the two convolution operations, respectively. For $\mathbf{F}_1$ and $\mathbf{F}_2$, we also use 1×1 convolution kernels to reduce their number of channels to 32, and adjust their size to the same as $\mathbf{F}_3$ through downsampling operations. This process can be expressed as:

$$\mathbf{W}_1 = \mathrm{Down}(\mathrm{Conv}_{1\times1}(\mathbf{F}_1)) \tag{17}$$

$$\mathbf{W}_2 = \mathrm{Down}(\mathrm{Conv}_{1\times1}(\mathbf{F}_2)) \tag{18}$$

where $\mathrm{Down}(\cdot)$ denotes the downsampling operation. Next, in order to incorporate the detailed information of the low-level features into the projection, we perform element-wise multiplication operations on $\mathbf{U}$ with $\mathbf{W}_1$ and $\mathbf{W}_2$, and then perform element-wise addition operations on the multiplication results. This multiplication and addition operation allocates different weights to different pixels, emphasizing the detailed information of the salient objects. Subsequently, we use an average pooling operation with stride s to obtain the vertices of the graph. Each vertex represents a pixel region in the feature map, and we calculate the similarity between

the vertex and each pixel by taking the product of $\mathbf{U}$ and the vertex set, thereby obtaining the similarity matrix. Then, we apply the softmax function to normalize the similarity matrix, obtaining the projection matrix. The calculation process can be expressed as follows:

$$\mathbf{Z} = \mathrm{AP}((\mathbf{U} \otimes \mathbf{W}_1) \oplus (\mathbf{U} \otimes \mathbf{W}_2)) \tag{19}$$

$$\mathbf{P} = \mathrm{softmax}(\mathbf{Z} \times \mathbf{U}^{\mathrm{T}}) \tag{20}$$

where $\mathrm{AP}(\cdot)$ denotes the average pooling operation. After obtaining the projection matrix $\mathbf{P}$, we multiply it with the feature map $\mathbf{V}$, projecting the feature map onto the graph structure. We then pass the graph structure to a single-layer graph convolutional network $\mathrm{GCN}(\cdot)$ [37], inferring the relationships between regions by propagating information between vertices, thereby obtaining higher-level semantic information. This can be expressed as:

$$\mathbf{G}=\mathrm{GCN}(\mathbf{P} \times \mathbf{V}) \tag{21}$$

Next, we use $\mathbf{P}^{\mathrm{T}}$ to transform the graph convolutional features $\mathbf{G}$ back into the original feature map structure $\mathbf{F}_{\mathrm{c}}$. Through a 1×1 convolution operation, we adjust the number of channels of the reconstructed feature map $\mathbf{F}_{\mathrm{c}}$ to the same size as $\mathbf{F}_3$, and then perform element-wise addition operations with $\mathbf{F}_3$, thereby obtaining the final optimized feature map $\mathbf{T}_3$. This process can be expressed as:

$$\mathbf{F}_{\mathrm{c}} = \mathbf{P}^{\mathrm{T}}\mathbf{G} \tag{22}$$

$$\mathbf{T}_3 = \mathbf{F}_3 \oplus \mathrm{Conv}_{1\times1}(\mathbf{F}_{\mathrm{c}}) \tag{23}$$

E. DECODING MODULE

The graph reasoning module outputs three optimized feature maps $\mathbf{T}_3$, $\mathbf{T}_4$, and $\mathbf{T}_5$. We first supervise these three feature maps with the real saliency map $\mathbf{GT}$. Subsequently, we fuse these three optimized feature maps to generate the final predicted saliency map, which is also supervised with $\mathbf{GT}$. As shown in Figure 2(b), the fusion operation involves concatenating $\mathbf{T}_3$, $\mathbf{T}_4$, and $\mathbf{T}_5$ along the channel direction, followed by a fusion through a convolution layer with a 3×3 kernel. Finally, we adjust the channel number to 1 through a convolution layer with a 1×1 kernel, and upscale to the same size as $\mathbf{GT}$ to generate the final predicted saliency map $\mathbf{S}_{\mathrm{pre}}$.

F. LOSS FUNCTION

In this study, we employ three loss functions: the predicted saliency map loss function, the low-level feature loss function, and the optimized feature loss function. Specifically, the predicted saliency map loss function $\mathrm{L}_{\mathrm{pre}}$ is used to compute the loss value between the final predicted saliency map $\mathbf{S}_{\mathrm{pre}}$ and the real saliency map $\mathbf{GT}$. The low-level feature loss function $\mathrm{L}_{\mathrm{low}}$ supervises the two low-level fused feature maps $\{\mathbf{F}_1, \mathbf{F}_2\}$, and the optimized feature loss function $\mathrm{L}_{\mathrm{opt}}$ supervises the three optimized feature maps $\{\mathbf{T_3}, \mathbf{T}_4, \mathbf{T}_5\}$ generated by the GRM module. As illustrated in Figure 2, we use the output layer "Out" to convert the two low-level fused feature maps and the three optimized feature maps into a single-channel predicted saliency map. The output layer adjusts the channel number of the feature maps to 1 through a convolution operation with a 1×1 kernel, and upscales the feature maps to the same size as $\mathbf{GT}$. The overall loss function is computed as follows:

$$\mathrm{L} = \mathrm{L}_{\mathrm{pre}} + \mathrm{L}_{\mathrm{low}} + \mathrm{L}_{\mathrm{opt}} \tag{24}$$

$$\mathrm{L}_{\mathrm{pre}} = \mathrm{L}_{\mathrm{BCE}}(\mathbf{S}_{\mathrm{pre}}, \mathbf{GT}) + \mathrm{L}_{\mathrm{Dice}}(\mathbf{S}_{\mathrm{pre}}, \mathbf{GT}) \tag{25}$$

$$\mathrm{L}_{\mathrm{low}} = \sum_{i=1}^{2} \mathrm{L}_{BCE}(\mathrm{Out}(\mathbf{F}_i), \mathbf{GT}) + \sum_{i=1}^{2} \mathrm{L}_{\mathrm{Dice}}(\mathrm{Out}(\mathbf{F}_i), \mathbf{GT}) \tag{26}$$

$$\mathrm{L}_{\mathrm{opt}} = \sum_{i=3}^{5} \mathrm{L}_{\mathrm{BCE}}(\mathrm{Out}(\mathbf{T}_i), \mathbf{GT}) + \sum_{i=3}^{5} \mathrm{L}_{\mathrm{Dice}}(\mathrm{Out}(\mathbf{T}_i), \mathbf{GT}) \tag{27}$$

where $\mathrm{L}_{\mathrm{BCE}}$ denotes the Binary Cross Entropy loss function, and $\mathrm{L}_{\mathrm{Dice}}$ denotes the Dice loss function [44].

IV. EXPERIMENT

A. DATASETS

In this study, we employ seven challenging RGB-D SOD public datasets for experimental validation, to demonstrate the effectiveness of the proposed framework. These seven datasets include: NJU2K [13], NLPR [14], DES [45], LFSD [46], SSD [47], STERE [48], and SIP [22].

The NJU2K dataset was collected and organized using the FujiW3 camera and advanced optical flow technology, offering a total of 1985 samples. The SSD dataset was collected using Sun's optical flow technology, and it contains 80 samples. The NLPR and DES datasets were collected using the Microsoft Kinect input device, with respective sample sizes of 1000 and 135. The LFSD dataset was collected using the Lytro camera, and it includes 100 samples. The SIP dataset was collected using Huawei Meta10, and it contains 929 high-resolution RGB-D images of individuals. Lastly, the STERE dataset was collected using a stereo camera and sift flow technology [7], and it offers a total of 1000 samples. The collection methods and sample sizes of these datasets provide a rich data resource for our research.

B. EXPERIMENTAL DETAILS

In our experiment, we employed ResNet50 [41] as the backbone network to extract features from RGB images and depth images. To meet the input requirements of ResNet50, we replicated depth images three times along the channel dimension. Our CMCL model was implemented using PyTorch, and all experiments were conducted on a single NVIDIA RTX A6000 GPU. During the model training phase, we utilized the Adam optimizer, setting the learning rate to 0.0001, batch size to 16, and weight decay to 0.0001. For data augmentation, we performed random flipping, random

TABLE 1. Comparison of evaluation results on four evaluation 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 three results are shown in red, green and blue fonts, respectively. '-' indicates the code or result is not available.

Datasets	Metrics	Comparison Methods													
		DANet	JL-DCF	UCNet	BBS-Net	DMRA	ICNet	CFIDNet	AFNet	BTS	CATNet	AMINet	DCMNet	HRTransNet	Ours
DES	MAE↓	0.029	0.021	0.019	0.021	0.031	0.027	0.023	0.022	0.018	0.016	0.017	-	0.014	0.018
	maxF↑	0.894	0.918	0.930	0.928	0.889	0.913	0.911	0.923	0.940	0.914	0.915	-	0.938	0.938
	maxE↑	0.957	0.957	0.976	0.966	0.941	0.960	0.940	0.953	0.979	0.979	0.973	-	0.983	0.982
	S↑	0.904	0.928	0.934	0.934	0.903	0.920	0.917	0.925	0.943	0.945	0.931	-	0.947	0.948
LFSD	MAE↓	0.083	0.081	0.067	0.072	0.074	0.071	0.071	0.056	0.071	0.051	0.056	0.064	-	0.053
	maxF↑	0.846	0.854	0.863	0.858	0.858	0.870	0.865	0.888	0.873	0.894	0.883	0.867	-	0.899
	maxE↑	0.886	0.887	0.905	0.900	0.905	0.903	0.903	0.923	0.906	0.908	0.906	0.906	-	0.938
	S↑	0.845	0.849	0.864	0.864	0.845	0.868	0.869	0.890	0.867	0.894	0.871	-	-	0.903
NJU2K	MAE↓	0.047	0.039	0.035	0.035	0.049	0.052	0.038	0.032	0.037	0.025	0.035	0.036	0.026	0.030
	maxF↑	0.893	0.915	0.910	0.920	0.892	0.891	0.915	0.928	0.902	0.929	0.912	0.899	0.928	0.933
	maxE↑	0.936	0.951	0.949	0.949	0.937	0.926	0.946	0.958	0.942	0.933	0.928	0.920	0.931	0.964
	S↑	0.897	0.913	0.911	0.921	0.889	0.894	0.914	0.926	0.910	0.937	0.904	-	0.933	0.931
NLPR	MAE↓	0.029	0.022	0.025	0.023	0.030	0.028	0.026	0.020	0.023	0.018	0.019	0.024	0.016	0.021
	maxF↑	0.901	0.918	0.903	0.918	0.875	0.908	0.905	0.925	0.923	0.916	0.916	0.883	0.919	0.931
	maxE↑	0.953	0.965	0.956	0.961	0.942	0.952	0.955	0.968	0.965	0.968	0.963	0.954	0.969	0.973
	S↑	0.915	0.931	0.920	0.931	0.898	0.923	0.922	0.936	0.934	0.939	0.922	-	0.942	0.940
SIP	MAE↓	0.054	0.049	0.051	0.055	0.082	0.070	0.060	0.043	0.044	0.034	-	0.047	0.035	0.041
	maxF↑	0.884	0.894	0.879	0.884	0.835	0.857	0.870	0.909	0.901	0.918	-	0.883	0.916	0.914
	maxE↑	0.920	0.931	0.919	0.922	0.883	0.903	0.909	0.939	0.933	0.944	-	0.926	0.943	0.944
	S↑	0.878	0.885	0.875	0.879	0.816	0.854	0.864	0.896	0.896	0.913	-	-	0.909	0.903
SSD	MAE↓	0.051	0.052	0.049	0.044	0.057	0.064	0.050	0.038	0.077	-	-	-	-	0.037
	maxF↑	0.849	0.839	0.849	0.860	0.849	0.841	0.871	0.885	0.758	-	-	-	-	0.887
	maxE↑	0.905	0.909	0.921	0.920	0.911	0.903	0.926	0.943	0.867	-	-	-	-	0.946
	S↑	0.868	0.864	0.869	0.882	0.855	0.848	0.879	0.897	0.796	-	-	-	-	0.899
STERE	MAE↓	0.048	0.044	0.039	0.041	0.064	0.045	0.043	0.034	0.038	0.030	0.036	-	0.030	0.032
	maxF↑	0.881	0.895	0.899	0.903	0.852	0.898	0.897	0.918	0.911	0.902	0.895	-	0.904	0.921
	maxE↑	0.930	0.942	0.944	0.942	0.917	0.942	0.942	0.957	0.949	0.935	0.928	-	0.930	0.960
	S↑	0.892	0.900	0.903	0.908	0.838	0.903	0.901	0.918	0.915	0.925	0.902	-	0.921	0.922

cropping, random rotation, and color enhancement operations on the training images.

C. EVALUATION METRICS

To validate the effectiveness of our proposed CMCL model, we employ four widely used evaluation metrics: F-measure [49], E-measure [50], S-measure [51], and Mean Absolute Error (MAE).

The F-measure serves as a metric blending precision and recall into a weighted harmonic mean, and it can be formulated in the following manner:

$$\mathrm{F}_\beta = \frac{(1+\beta^2)\mathrm{Pre}\times\mathrm{Rec}}{\beta^2\times\mathrm{Pre}+\mathrm{Rec}} \tag{28}$$

where β^2 is empirically set to 0.3.

The E-measure comprehensively considers the global mean of the image and the values of local pixels, and it is specifically defined as follows:

$$\mathrm{E_m} = \frac{1}{\mathrm{W}\times\mathrm{H}}\sum_{i=1}^{\mathrm{W}}\sum_{j=1}^{\mathrm{H}}\xi(i,j) \tag{29}$$

where ξ represents the enhanced alignment matrix, which describes the correlation between the predicted saliency map and the ground truth saliency map.

The S-measure is used to assess the structural similarity between the predicted saliency map and the ground truth saliency map, defined as follows:

$$\mathrm{S} = \alpha\times\mathrm{S_{object}} + (1-\alpha)\mathrm{S_{region}} \tag{30}$$

where $\mathrm{S_{object}}$ and $\mathrm{S_{region}}$ respectively represent the object-aware similarity and region-aware similarity, and α is empirically set to 0.5.

The computation of the pixel-level average absolute error between the saliency map generated by our approach and the ground truth saliency map is facilitated by the Mean Absolute Error (MAE), which is determined via the given expression:

$$\mathrm{MAE} = \frac{1}{\mathrm{W}\times\mathrm{H}}\sum_{x=1}^{\mathrm{W}}\sum_{y=1}^{\mathrm{H}}|\mathbf{S}(x,y)-\mathbf{GT}(x,y)| \tag{31}$$

In this context, **S** corresponds to the predicted saliency map produced by our model, while **GT** denotes the ground truth saliency map for comparison. W refers to the map's width,

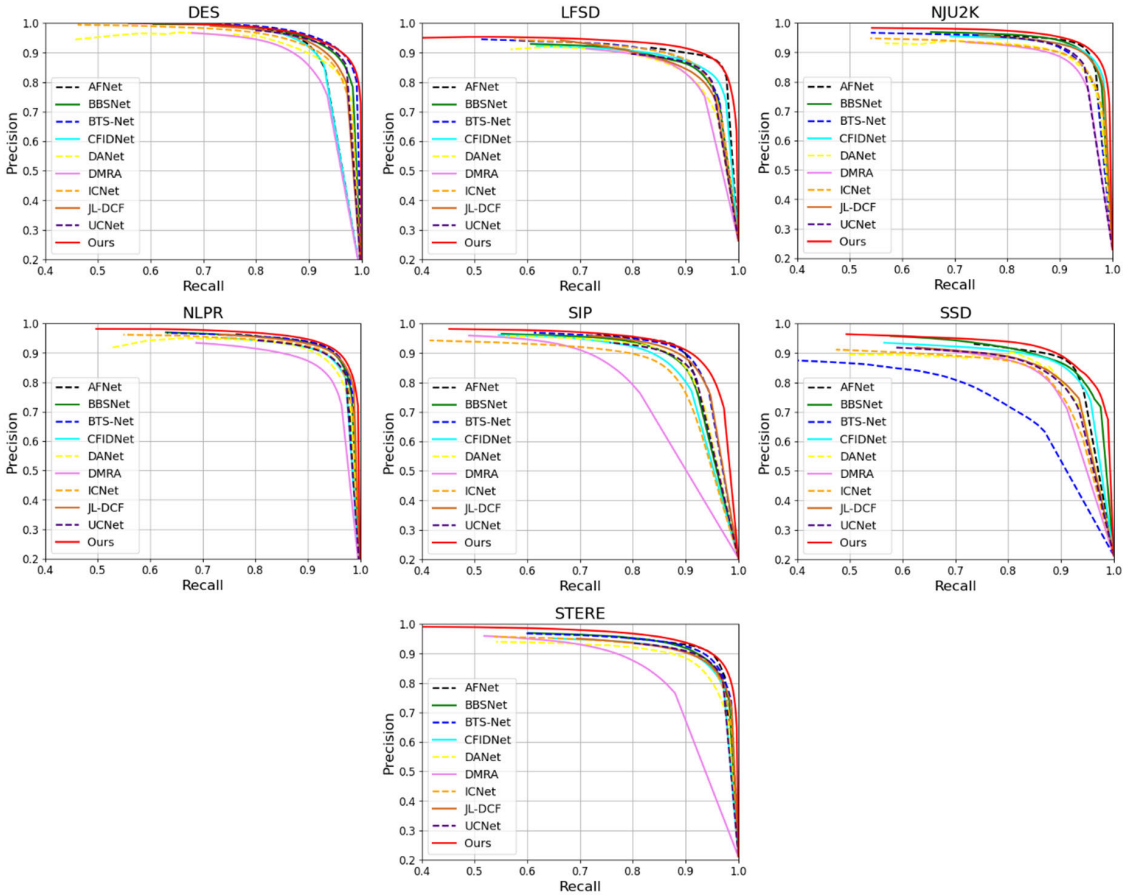


FIGURE 5. Comparison of P-R curves for different methods on seven RGB-D datasets. Our CMCL method is represented by the red solid line.

while H indicates its height, both dimensions defining the spatial resolution of the saliency maps in question.

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

To fully verify the effectiveness of our proposed CMCL model, we compared it with thirteen existing deep-learning-based RGB-D salient object detection methods, including AFNet [52], CFIDNet [53], ICNet [23], DMRA [54], UCNet [55], JLDCF [56], DANet [57], BBS-Net [58], and BTS-Net [59],CATNet [60],AMINet [61], DCMNet [62] and HRTransNet [63]. To ensure the fairness of the comparison, we used the saliency maps provided by the authors. If these were not available, we generated the saliency maps using the source code and model files provided by the authors.

1) QUANTITATIVE COMPARISON

Table 1 presents the quantitative evaluation results for four evaluation metrics, clearly showing the exceptional performance of our proposed CMCL method. On the SSD dataset, CMCL achieved the best results across all four metrics. For the metrics maxE and maxF, our method outperformed others on all datasets except for DES and SIP. Overall, our method achieved the best results in half of the metrics and ranked in the top three for the rest.

Figure 5 shows the comparison results of the PR curves of different methods. The PR curves also reflect the fact that our method performed excellently on seven datasets, proving that our method performed the best among all the compared methods.

2) QUALITATIVE COMPARISON

In Figure 6, we offer a visual comparison that highlights the efficacy of our CMCL model against a range of RGB-D SOD methods. The selected outcomes demonstrate the model's capabilities. Particularly noticeable in the first two rows are scenarios where low-quality depth cues are present; our model distinctly outperforms others by successfully delineating the salient objects despite the poor depth information, which proves challenging for other methods like BTS-Net, CFIDNet, DANet, DMRA, and ICNet, leading to their impaired object segmentation.

Further on, instances involving multiple objects are displayed in rows three and four. Namely, the salient object boundaries in JL-DCF's results appear diffused in the third row. AFNet and ICNet recognize only a single object of interest in the fourth row. Our approach, however, shows a precise delineation of all objects of interest across these cases.

In rows five and six, where complex backgrounds are prevalent, other techniques struggle, often producing

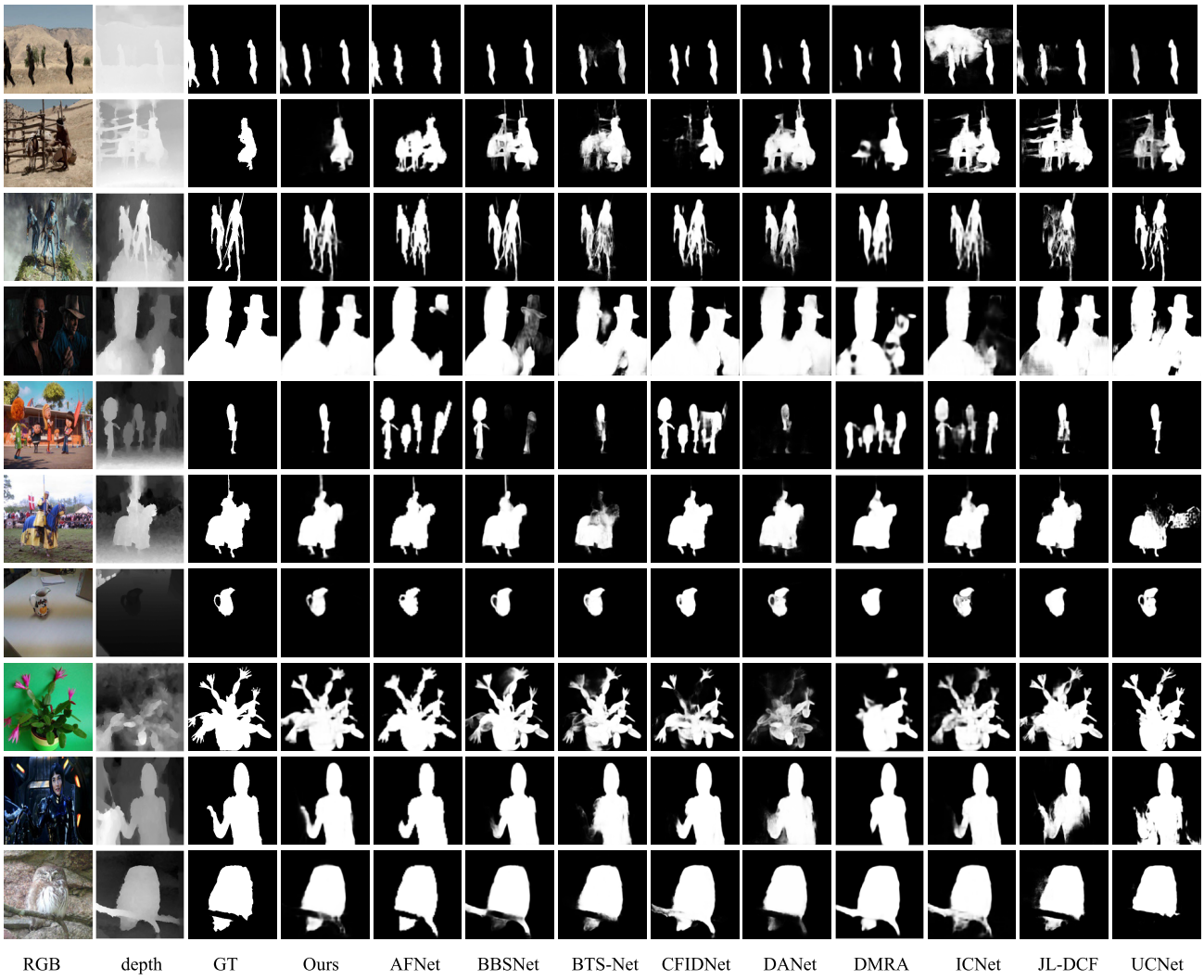


FIGURE 6. Visual comparison between CMCL and the state-of-the-art RGB-D models.

imprecise saliency maps. In contrast, our approach renders sharp, clear saliency maps that retain the integrity of object structures against intricate backdrops. The seventh row illustrates a scenario with a small object, while the eighth row features an example with fine-grained objects. Even in these complex conditions, our method maintains robust performance.

Lastly, the examples in the ninth and tenth rows demonstrate low-contrast scenarios where foreground and background bear close resemblance. Nearly all competing methods fail to extract the entire object of interest, whereas ours succeeds, thanks to the effective fusion of cross-modal and cross-level data that helps in suppressing extraneous information from the background.

E. ABLATION STUDY

As shown in Table 2, we conducted an in-depth ablation analysis to verify the effectiveness of each module. AEM represents the Attention Enhancement Module, MFM stands for the Modal Feature Fusion Module, and GRM is the Graph Reasoning Module. ''Without AEM'', ''without MFM'', and ''without GRM'' respectively represent the models obtained after removing the AEM, MFM, and GRM modules from the CMCL model. By comparing the data in the third and sixth columns, we can clearly see that the introduction of the AEM module significantly improves the model's performance. Similarly, comparing the data in the fourth and sixth columns also shows that the introduction of the MFM module can significantly enhance the model's performance. By comparing the data in the fifth and sixth columns, we can see that the addition of the GRM module enhances the model's performance. These results validate the importance of the three modules: the AEM module enhances attention to the color image and depth map features, the MFM module realizes cross-modal feature fusion, and the GRM module implements cross-level feature fusion. All three functional modules have significantly improved the model's performance. In the last column, we could see that the CMCL model that combines these three modules achieved the best results.

TABLE 2. Comparison of ablation study results. The best performance in each row is highlighted in bold.

datasets	evaluation metrics	Without AEM	Without MFM	Without GRM	CMCL
DES	MAE ↓	0.029	0.027	0.026	**0.018**
	maxF ↑	0.873	0.904	0.902	**0.938**
	maxE ↑	0.928	0.941	0.941	**0.982**
	S ↑	0.899	0.917	0.916	**0.948**
LFSD	MAE ↓	0.094	0.079	0.090	**0.053**
	maxF ↑	0.818	0.853	0.833	**0.899**
	maxE ↑	0.866	0.890	0.878	**0.938**
	S ↑	0.829	0.858	0.837	**0.903**
NJU2K	MAE ↓	0.041	0.038	0.037	**0.030**
	maxF ↑	0.909	0.915	0.916	**0.933**
	maxE ↑	0.944	0.949	0.951	**0.964**
	S ↑	0.913	0.917	0.919	**0.931**
NLPR	MAE ↓	0.029	0.025	0.024	**0.021**
	maxF ↑	0.901	0.913	0.917	**0.931**
	maxE ↑	0.950	0.962	0.964	**0.973**
	S ↑	0.919	0.927	0.930	**0.940**
SIP	MAE ↓	0.062	0.055	0.054	**0.041**
	maxF ↑	0.867	0.886	0.887	**0.914**
	maxE ↑	0.911	0.923	0.925	**0.944**
	S ↑	0.868	0.880	0.881	**0.903**
SSD	MAE ↓	0.053	0.053	0.054	**0.037**
	maxF ↑	0.838	0.864	0.847	**0.887**
	maxE ↑	0.899	0.912	0.908	**0.946**
	S ↑	0.863	0.880	0.868	**0.899**
STERE	MAE ↓	0.046	0.045	0.047	**0.032**
	maxF ↑	0.888	0.892	0.888	**0.921**
	maxE ↑	0.935	0.939	0.936	**0.960**
	S ↑	0.897	0.899	0.896	**0.922**

TABLE 3. Quantitative comparisons of computational analysis.

Model	Backbone	Input Size	Param(M)↓	FLOPs(G)↓	FPS↑
DMRA	VGG-19	256 × 256	59.7	120.95	20
AMINet	Swin Transformer	384 × 384	199.1	124.7	30
CATNet	Swin Transformer	384 × 384	262.6	172.1	24
JL-DCF	VGG-16	320 × 320	143.5	861.2	18
HRTransNet	HRFormer	224 × 224	58.9	17.1	12
BTSNet	ResNet-50	352 × 352	99	250.8	30
UCNet	VGG-16	352 × 352	27	16.2	36
Ours	ResNet-50	352 × 352	150.2	62.5	33

F. COMPUTATIONAL ANALYSIS

In Table 3, we present a comparison of the efficiency of our proposed model. The data for the AMINet and DMRA models are directly cited from their original papers, while the data for the other models were computed on an NVIDIA 3090 GPU. Compared to the BTSNet model, which uses a similar backbone network, our model has slightly more parameters, but significantly lower FLOPs, and its processing speed is also 10% faster than that of BTSNet. In comparison to the JL-DCF model, the number of parameters between the two models is similar, but our model has a distinct advantage in processing speed. Overall, our model ranks second in terms of running speed, just behind the UCNet model. Therefore, in our future research, we plan to introduce lightweight mech-

anisms aimed at further increasing processing speed while ensuring detection performance.

V. CONCLUSION

In this paper, we proposed a novel cross-modal cross-level fusion learning framework to solve the problem of RGB-D salient object detection. Our framework is composed of three parts: the Attention Enhancement Module (AEM), the Modal Feature Fusion Module (MFM), and the Graph Reasoning Module (GRM). These three modules work together, effectively integrating features of various modalities and levels, achieving efficient feature fusion. We conducted extensive experiments on our method across seven widely used datasets. The experimental results proved that our method surpassed nine of the current state-of-the-art methods on four evaluation metrics. These experimental results fully demonstrate the superiority of our method in handling salient object detection tasks and its robustness in dealing with complex backgrounds and cases where the target color is similar to the background.

In summary, the cross-modal cross-level fusion learning framework we proposed provided a new and effective solution in the field of RGB-D salient object detection. We anticipate that this framework will further promote the research of RGB-D salient object detection and bring more possibilities to related application fields. However, there is room for further improvement in our method. For example, we will further explore how to more effectively fuse cross-modal and cross-level features, as well as how to better handle different types of images. We also look forward to applying our method to other computer vision tasks.

REFERENCES

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

[2] Z. Wang, Y. Zhang, Y. Liu, Z. Wang, S. Coleman, and D. Kerr, "TF-SOD: A novel transformer framework for salient object detection," *Neural Comput. Appl.*, vol. 34, no. 14, pp. 11789–11806, Jul. 2022.

[3] Y. Wang, F. Wang, C. Wang, F. Sun, and J. He, "Learning saliency-aware correlation filters for visual tracking," *Comput. J.*, vol. 65, no. 7, pp. 1846–1859, Jul. 2022.

[4] C. Fu, J. Xu, F. Lin, F. Guo, T. Liu, and Z. Zhang, "Object saliency-aware dual regularized correlation filter for real-time aerial tracking," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 12, pp. 8940–8951, Dec. 2020.

[5] J. Han, D. Zhang, G. Cheng, N. Liu, and D. Xu, "Advanced deep-learning techniques for salient and category-specific object detection: A survey," *IEEE Signal Process. Mag.*, vol. 35, no. 1, pp. 84–100, Jan. 2018.

[6] Y. Liu, J. Han, Q. Zhang, and C. Shan, "Deep salient object detection with contextual information guidance," *IEEE Trans. Image Process.*, vol. 29, pp. 360–374, 2020.

[7] 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.

[8] N. Zhang, J. Han, and N. Liu, "Learning implicit class knowledge for RGB-D co-salient object detection with transformers," *IEEE Trans. Image Process.*, vol. 31, pp. 4556–4570, 2022.

[9] Y. Yang, Q. Qin, Y. Luo, Y. Liu, Q. Zhang, and J. Han, "Bi-directional progressive guidance network for RGB-D salient object detection," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 8, pp. 5346–5360, Aug. 2022.

[10] G. Feng, J. Meng, L. Zhang, and H. Lu, "Encoder deep interleaved network with multi-scale aggregation for RGB-D salient object detection," *Pattern Recognit.*, vol. 128, Aug. 2022, Art. no. 108666.

[11] 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.

[12] X. Jin, C. Guo, Z. He, J. Xu, Y. Wang, and Y. Su, "FCMNet: Frequency-aware cross-modality attention networks for RGB-D salient object detection," *Neurocomputing*, vol. 491, pp. 414–425, Jun. 2022.

[13] 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)*, Oct. 2014, pp. 1115–1119.

[14] 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.*, 2014, pp. 92–109.

[15] D. Feng, N. Barnes, S. You, and C. McCarthy, "Local background enclosure for RGB-D salient object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 2343–2350.

[16] J. Guo, T. Ren, and J. Bei, "Salient object detection for RGB-D image via saliency evolution," in *Proc. IEEE Int. Conf. Multimedia Expo. (ICME)*, Jul. 2016, pp. 1–6.

[17] M.-M. Cheng, N. J. Mitra, X. Huang, P. H. S. Torr, and S.-M. Hu, "Global contrast based salient region detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 3, pp. 569–582, Mar. 2015.

[18] A. Borji, M.-M. Cheng, Q. Hou, H. Jiang, and J. Li, "Salient object detection: A survey," *Comput. Vis. Media*, vol. 5, no. 2, pp. 117–150, Jun. 2019.

[19] J. Zhao, J.-J. Liu, D.-P. Fan, Y. Cao, J. Yang, and M.-M. Cheng, "EGNet: Edge guidance network for salient object detection," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 8778–8787.

[20] X. Qin, Z. Zhang, C. Huang, C. Gao, M. Dehghan, and M. Jagersand, "BASNet: Boundary-aware salient object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 7471–7481.

[21] Y. Pang, X. Zhao, L. Zhang, and H. Lu, "Multi-scale interactive network for salient object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 9410–9419.

[22] 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.

[23] G. Li, Z. Liu, and H. Ling, "ICNet: Information conversion network for RGB-D based salient object detection," *IEEE Trans. Image Process.*, vol. 29, pp. 4873–4884, 2020.

[24] H. Chen and Y. Li, "Progressively complementarity-aware fusion network for RGB-D salient object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 3051–3060.

[25] H. Chen and Y. Li, "Three-stream attention-aware network for RGB-D salient object detection," *IEEE Trans. Image Process.*, vol. 28, no. 6, pp. 2825–2835, Jun. 2019.

[26] C. Zhu, X. Cai, K. Huang, T. H. Li, and G. Li, "PDNet: Prior-model guided depth-enhanced network for salient object detection," in *Proc. IEEE Int. Conf. Multimedia Expo. (ICME)*, Jul. 2019, pp. 199–204.

[27] W. Gao, G. Liao, S. Ma, G. Li, Y. Liang, and W. Lin, "Unified information fusion network for multi-modal RGB-D and RGB-T salient object detection," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 4, pp. 2091–2106, Apr. 2022.

[28] 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 Recognit.*, vol. 86, pp. 376–385, Feb. 2019.

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

[30] Z. Liu, K. Wang, H. Dong, and Y. Wang, "A cross-modal edge-guided salient object detection for RGB-D image," *Neurocomputing*, vol. 454, pp. 168–177, Sep. 2021.

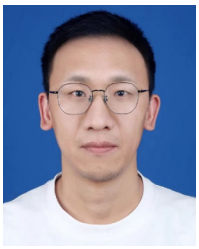
[31] S. Zhang, H. Tong, J. Xu, and R. Maciejewski, "Graph convolutional networks: A comprehensive review," *Comput. Social Netw.*, vol. 6, no. 1, pp. 1–23, Dec. 2019.

[32] F. Wu, A. H. Souza, T. Zhang, C. Fifty, T. Yu, and K. Q. Weinberger, "Simplifying graph convolutional networks," in *Proc. Int. Conf. Mach. Learn. (ICML)*, 2019, pp. 6861–6871.

[33] H. Gao, Z. Wang, and S. Ji, "Large-scale learnable graph convolutional networks," in *Proc. 24th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, Jul. 2018, pp. 1416–1424.
[34] M. Schlichtkrull, T. N. Kipf, P. Bloem, R. van den Berg, I. Titov, and M. Welling, "Modeling relational data with graph convolutional networks," in *Proc. Eur. Semantic Web Conf.*, 2018, pp. 593–607.
[35] H. Cai, S. Lv, G. Lu, and T. Li, "Graph convolutional networks for fast text classification," in *Proc. 4th Int. Conf. Natural Language Process. (ICNLP)*, Mar. 2022, pp. 420–425.
[36] S. Zhang, H. Tong, J. Xu, and R. Maciejewski, "Graph convolutional networks: Algorithms, applications and open challenges," in *Proc. Comput. Data Social Netw., 7th Int. Conf.*, 2018, pp. 79–91.
[37] Y. Li and A. Gupta, "Beyond grids: Learning graph representations for visual recognition," in *Proc. Adv. Neural Inf. Process. Syst.*, 2018, pp. 1–11.
[38] Y. Lu, Y. Chen, D. Zhao, and J. Chen, "Graph-FCN for image semantic segmentation," in *Proc. Int. Symp. Neural Netw.*, 2019, pp. 97–105.
[39] G. Te, W. Hu, and Z. Guo, "Exploring hypergraph representation on face anti-spoofing beyond 2D attacks," in *Proc. IEEE Int. Conf. Multimedia Expo. (ICME)*, Jul. 2020, pp. 1–6.
[40] L. Zhang, X. Li, A. Arnab, K. Yang, Y. Tong, and P. H. S. Torr, "Dual graph convolutional network for semantic segmentation," 2019, *arXiv:1909.06121*.
[41] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.
[42] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Medical Image Computing and Computer-Assisted Intervention*. Cham, Switzerland: Springer, 2015, pp. 234–241.
[43] A. Buades, B. Coll, and J.-M. Morel, "A non-local algorithm for image denoising," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, Jun. 2005, pp. 60–65.
[44] F. Milletari, N. Navab, and S.-A. Ahmadi, "V-Net: Fully convolutional neural networks for volumetric medical image segmentation," in *Proc. 4th Int. Conf. 3D Vis. (3DV)*, Oct. 2016, pp. 565–571.
[45] 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.
[46] N. Li, J. Ye, Y. Ji, H. Ling, and J. Yu, "Saliency detection on light field," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 2806–2813.
[47] 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)*, Oct. 2017, pp. 3008–3014.
[48] Y. Niu, Y. Geng, X. Li, and F. Liu, "Leveraging stereopsis for saliency analysis," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2012, pp. 454–461.
[49] A. Borji, M.-M. Cheng, H. Jiang, and J. Li, "Salient object detection: A benchmark," *IEEE Trans. Image Process.*, vol. 24, no. 12, pp. 5706–5722, Dec. 2015.
[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, *arXiv:1805.10421*.
[51] 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)*, Oct. 2017, pp. 4548–4557.
[52] T. Chen, J. Xiao, X. Hu, G. Zhang, and S. Wang, "Adaptive fusion network for RGB-D salient object detection," *Neurocomputing*, vol. 522, pp. 152–164, Feb. 2023.
[53] T. Chen, X. Hu, J. Xiao, G. Zhang, and S. Wang, "CFIDNet: Cascaded feature interaction decoder for RGB-D salient object detection," *Neural Comput. Appl.*, vol. 34, no. 10, pp. 7547–7563, May 2022.
[54] 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.
[55] 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.
[56] 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.
[57] 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 *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 646–662.
[58] D.-P. Fan, Y. Zhai, A. Borji, J. Yang, and L. Shao, "BBS-Net: RGB-D salient object detection with a bifurcated backbone strategy network," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 275–292.
[59] W. Zhang, Y. Jiang, K. Fu, and Q. Zhao, "BTS-Net: Bi-directional transfer-and-selection network for RGB-D salient object detection," in *Proc. IEEE Int. Conf. Multimedia Expo. (ICME)*, Jul. 2021, pp. 1–6.
[60] F. Sun, P. Ren, B. Yin, F. Wang, and H. Li, "CATNet: A cascaded and aggregated transformer network for RGB-D salient object detection," *IEEE Trans. Multimedia*, vol. 26, pp. 1–14, 2023.
[61] R. Wang, F. Wang, Y. Su, J. Sun, F. Sun, and H. Li, "Attention-guided multi-modality interaction network for RGB-D salient object detection," *ACM Trans. Multimedia Comput., Commun., Appl.*, vol. 20, no. 3, pp. 1–22, Oct. 2023.
[62] F. Wang, R. Wang, and F. Sun, "DCMNet: Discriminant and cross-modality network for RGB-D salient object detection," *Expert Syst. Appl.*, vol. 214, Mar. 2023, Art. no. 119047.
[63] B. Tang, Z. Liu, Y. Tan, and Q. He, "HRTransNet: HRFormer-driven two-modality salient object detection," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 33, no. 2, pp. 728–742, Feb. 2023.

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

ZHINIAN ZHAI received the Ph.D. degree from South China University of Technology. He was a C++ Developer with Guangdong Beidian Communication Equipment Company Ltd. Currently, he is a Lecturer with the School of Information and Electronic Engineering, Zhejiang University of Science and Technology. His research interests include information security, machine learning, and deep learning.



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