# Dynamic Selective Network for RGB-D Salient Object Detection

Hongfa Wen, Chenggang Yan, Xiaofei Zhou, Runmin Cong, Yaoqi Sun, Bolun Zheng, Jiyong Zhang, *Member, IEEE*, Yongjun Bao, and Guiguang Ding

***Abstract*— RGB-D saliency detection is receiving more and more attention in recent years. There are many efforts have been devoted to this area, where most of them try to integrate the multi-modal information, *i.e.* RGB images and depth maps, via various fusion strategies. However, some of them ignore the inherent difference between the two modalities, which leads to the performance degradation when handling some challenging scenes. Therefore, in this paper, we propose a novel RGB-D saliency model, namely Dynamic Selective Network (DSNet), to perform salient object detection (SOD) in RGB-D images by taking full advantage of the complementarity between the two modalities. Specifically, we first deploy a cross-modal global context module (CGCM) to acquire the high-level semantic information, which can be used to roughly locate salient objects. Then, we design a dynamic selective module (DSM) to dynamically mine the cross-modal complementary information between RGB images and depth maps, and to further optimize the multi-level and multi-scale information by executing the gated and pooling based selection, respectively. Moreover, we conduct the boundary refinement to obtain high-quality saliency maps with clear boundary details. Extensive experiments on eight public RGB-D datasets show that the proposed DSNet achieves a competitive and excellent performance against the current 17 state-of-the-art RGB-D SOD models.**

***Index Terms*— RGB-D salient object detection, multi-modal, dynamic selection, feature fusion.**

Manuscript received May 28, 2021; revised September 15, 2021; accepted October 11, 2021. Date of publication November 5, 2021; date of current version November 11, 2021. This work was supported in part by the National Key Research and Development Program of China under Grant 2020YFB1406604; in part by the National Natural Science Foundation of China under Grant 61931008, Grant 61671196, Grant 62071415, Grant 62001146, Grant 61701149, Grant 61801157, Grant 61971268, Grant 61901145, Grant 61901150, Grant 61972123, and Grant 62002014; in part by the Zhejiang Province Natural Science Foundation of China under Grant LR17F030006 and Grant Q19F010030; in part by the 111 Project under Grant D17019; and in part by the Beijing Nova Program under Grant Z201100006820016. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Ioan Tabus. *(Corresponding authors: Xiaofei Zhou; Jiyong Zhang.)*

Hongfa Wen, Chenggang Yan, Xiaofei Zhou, Yaoqi Sun, Bolun Zheng, and Jiyong Zhang are with the School of Automation, Hangzhou Dianzi University, Hangzhou 310018, China (e-mail: hf_wen@outlook.com; cgyan@hdu.edu.cn; zxforchid@outlook.com; syq@hdu.edu.cn; blzheng@hdu.edu.cn; jzhang@hdu.edu.cn).

Runmin Cong is with the Institute of Information Science, Beijing Jiaotong University, Beijing 100044, China, and also with the Beijing Key Laboratory of Advanced Information Science and Network Technology, Beijing 100044, China (e-mail: rmcong@bjtu.edu.cn).

Yongjun Bao is with Business Growth BU, JD.com, Beijing 100101, China (e-mail: baoyongjun@jd.com).

Guiguang Ding is with the School of Software, Tsinghua University, Beijing 100084, China (e-mail: dinggg@tsinghua.edu.cn).

Digital Object Identifier 10.1109/TIP.2021.3123548

## I. Introduction

SALIENT object detection (SOD) is a fundamental problem that has received continuous attention in recent years. Its purpose is to pop-out the most attractive regions in images or videos. With the continuous efforts of researchers, SOD has made significant achievements and plays an important role in many applications, such as image segmentation [1], [2], object recognition [3], [4], visual tracking [5], [6], and video analysis [7], [8], to name a few. Therefore, it is of great theoretical value and practical significance to carry out the research on saliency detection.

The traditional SOD model [9], [10] has some certain limitations. Most of them rely on manually designed features, and are in lack of the effective representation of high-level semantic information. Recently, with the rapid development of deep learning technologies, Convolutional Neural Networks (CNNs) [11], [12] has become the protagonist in SOD [13]–[20], which have achieved better performance than traditional methods. However, when handling some complex scenes, such as low contrast and cluttered backgrounds, the performance of existing deep learning based saliency models often degrades to some degree, as shown in Fig. 1. The main reason behind this lies in that RGB images represent appearance information well, but they cannot effectively define spatial information. Meanwhile, depth maps are capable of measuring the distance of objects from the camera, and contain the rich spatial structure information, which is essential for saliency detection. Besides, with the fast development of depth sensors such as Microsoft Kinect and Intel RealSense, the collection of depth information has become easier and the generated depth maps are more accurate. Therefore, researchers try to introduce depth cues into RGB SOD, namely RGB-D SOD, which gives a further performance improvement.

Similar to RGB SOD, most of the early RGB-D SOD methods [22]–[25] focused on using specific prior knowledge to design manual feature descriptors, which ignore the effects of semantic information in SOD. In contrast, deep learning based RGB-D SOD models [26]–[31] try to sufficiently utilize the high-level semantic features and low-level spatial features. However, there is still a large room to elevate the performance of RGB-D SOD models, though the cutting-edge models have achieved stable and reliable results. Generally, there are mainly following challenges to be solved: 1) How to effectively aggregate cross-modal features. Obviously, the

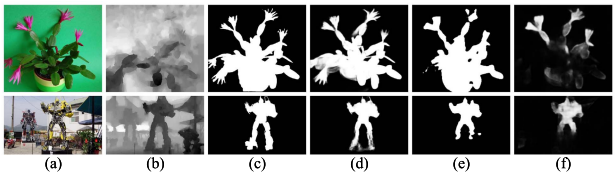


Fig. 1. The results of SOD in different challenging scenes (top: low contrast; bottom: complex background). (a) RGB images, (b) Depth maps, (c) Ground Truth, the saliency maps generated by (d) Ours, (e) A2dele [21] (RGB-D SOD), and (f) SOD100K [17] (RGB SOD).

inherent attributions of RGB images and depth maps are quite different, where the former focuses on texture information and the latter pays more attention to geometric information. Some existing methods [26]–[28] statically mine the complementary information between different modalities, which makes it difficult to comprehensively explore the interaction between cross-modal features. 2) How to efficiently integrate cross-level deep features. The simple fusion methods such as summation and concatenation ignore the particularity of different level features. Some existing methods [29]–[31] indiscriminately fuse cross-level information, which can easily superimpose and amplify inherent noise and lead to counterproductive effects.

To address the above challenges, we propose a Dynamic Selective Network (DSNet) for RGB-D saliency detection, as shown in Fig. 2. It explores the possibility of consistent fusion of cross-modal (RGB images and depth maps), cross-level (obtained from different convolutional blocks), and multi-scale (generated from various pooling rates) cues by means of dynamic selection. The proposed DSNet mainly includes a dynamic selective module (DSM) and a cross-modal global context module (CGCM), where the DSM contains two sub-modules including a cross-modal attention module (CAM) and a bi-directional gated pooling module (BGPM). Specifically, to acquire more comprehensive high-level semantic features, which are helpful to locate salient objects, we first introduce a CGCM to roughly highlight salient objects. Then, inspired by the attention mechanism, we design a CAM to dynamically mine the complementary information between RGB images and depth maps from layer and spatial views, which promotes the cross-modal feature fusion. Next, we deploy a BGPM to pay more concerns on cross-level and multi-scale deep features, where the BGPM bi-directionally optimizes cross-level information with gated-based selection and adaptively strengthens multi-scale information with pooling-based scaling. Furthermore, we deploy the deeply supervision strategy with spatial attention based feedback mechanism. Following this way, we can obtain high-quality saliency maps, which can not only highlight salient objects completely, but also provides clear boundary details.

In summary, the main contributions of the proposed DSNet can be stated as follows:

1) We propose a novel RGB-D saliency model, *i.e.* Dynamic Selective Network (DSNet), which automatically explores the cross-modal complementary information, bi-directionally optimizes the cross-level information, and adaptively strengthens the multi-scale information.
2) To fuse the multi-source information containing object cues, we design an effective dynamic selective module (DSM), where the cross-modal attention module (CAM) and bi-directional gated pooling module (BGPM) aggregate the deep features from the perspective of multi-modal, multi-level and multi-scale.
3) To mine the internal correlation between high-level RGB cues and depth cues, we introduce a cross-modal global context module (CGCM), which can roughly locate salient regions and suppress background regions.
4) We conduct comprehensive experiments on eight public RGB-D datasets, and the experimental results show that the proposed DSNet achieves a comparable performance when compared with 17 state-of-the-art models.

The rest of this paper is arranged as follows. Sec. II first summarizes the related works of RGB-D SOD. Then, Sec. III details the proposed RGB-D SOD model. Next, the experimental results and corresponding analysis are presented in Sec. IV. Eventually, the concise conclusion is drawn in Sec. V.

## II. Related Works

In this section, we briefly introduce some previous saliency detection methods related to our work.

### A. Salient Object Detection for RGB Images

In the past several decades, salient object detection (SOD) has been widely concerned by researchers. The early efforts are carried out on RGB images. Itti *et al.* [32] proposed an IT model that integrates multiple features, which fully considers the color, brightness, direction, and gradient information. Inspired by this, many hand-crafted feature based models [33]–[35] are designed to recognize salient objects. Among them, heuristic prior knowledge such as color contrast [33], background prior [34] and center prior [35] plays a very important role in SOD. However, these methods are heavily depending on specially designed hand-crafted features, namely the low-level appearance information, while they ignore the high-level semantic cues.

Recently, deep learning has played a key role in various computer vision tasks. Obviously, researchers have also introduced deep learning technologies such as convolutional neural networks (CNNs) into SOD, where the performance of SOD has been pushed forward remarkably. These methods [36]–[38] usually adopt an encoder-decoder structure, that is to say, CNNs are used as the encoder to extract multi-level and multi-scale RGB features, and then these features are merged in a specially designed decoder to generate the final saliency map. In particular, the feature selection ability of the attention mechanism [39], [40] is very consistent with the goals of SOD. Therefore, some methods [41]–[43] try to selectively weight key regions to better explore the structure of salient objects.

The RGB images based SOD methods have achieved encouraging performance, but the RGB saliency models encounter bottlenecks when dealing with challenging scenarios such as low contrast and cluttered backgrounds. This is mainly

because RGB images only provide rich appearance cues, which cannot distinguish the foreground from background in some complex scenes. Fortunately, depth maps, which capture abundant spatial cues, are beneficial for locating and segmenting salient objects. Therefore, researchers have tried to introduce depth maps into RGB SOD, namely RGB-D SOD. In this paper, our model focuses on the task of RGB-D SOD.

### *B. Salient Object Detection for RGB-D Images*

Similar to RGB SOD, the traditional RGB-D SOD methods mainly focus on using specific prior knowledge to design hand-crafted features. For example, Feng *et al.* [23] utilized a local background enclosure to capture the angular direction of the background, and calculated the saliency score of each region. Song *et al.* [25] used low-level feature contrasts, mid-level feature weighted factors, and high-level location priors to achieve multi-scale discriminative saliency fusion. The manual feature based models are capable of describing local details accurately, but generally are lack of the representation of high-level semantic information, which limits the performance improvement.

With the widespread application of deep learning technologies in computer vision, the CNNs based RGB-D SOD methods [21], [27], [29], [44]–[51] have effectively pushed forward the progress of RGB-D SOD. At present, many methods have been designed to efficiently learn the complementary information of different modalities (*i.e.*, RGB images and depth maps). For example, Li *et al.* [29] designed a cross-modal depth-weighted combination block to discriminate the cross-modal features from different sources and to enhance RGB features with depth features at each level. Piao *et al.* [21] implemented a RGB stream embedded with a depth distiller, which transfers the depth cues from depth stream to RGB stream. Liu *et al.* [44] proposed a self-mutual attention module to incorporate the long-range global context from different modalities. Li *et al.* [45] introduced a cross-modality feature modulation module, which takes the depth features as prior to enhance the representations of corresponding RGB features. Fan *et al.* [27] employed a depth-enhanced module to capture the informative cues in depth maps and improved the compatibility of RGB and depth features. The above-mentioned methods adopt an asymmetrical structure to enhance RGB information with depth cues. Such a design only regards depth maps as the supplement of RGB images, and ignores the complementarity of them. In contrast, our model treats the two modalities equally and explores their correlations dynamically.

Except the fusion of cross-modal information, the cross-level feature fusion is also crucial for CNNs based RGB-D saliency models, where they attempted to integrate the low-level texture information and high-level semantic information progressively. For example, Chen and Li [46] employed a progressive feature fusion method, which considers the complementarity of cross-modal features and integrates multi-scale information. Fu *et al.* [47] leveraged a densely-cooperative fusion strategy to robustly integrate cross-level features. Both of them adopt progressive and dense integration rules to aggregate features at different levels in a one-directional way. Compared with the bi-directional interaction, an obvious weakness of one-directional interaction is that it is difficult to comprehensively explore the relationship between cross-level features. Besides, to deal with the large discrepancy among different scale salient objects, the weighted fusion of multi-scale features has also become popular [48], [49]. Nevertheless, we use a symmetrical structure with bi-directional interaction to mutually guide adjacent-level features, and adopt pooling operations with different scales to adaptively aggregate multi-scale features.

In addition, there is no doubt that more accurate edge information can better guide the prediction with more clear boundaries. Ji *et al.* [50] introduced an edge collaborator to extract edge cues from low-level RGB features and used it for precise saliency predictions. Zhang *et al.* [51] implemented a boundary supplement unit, which enhanced the edge details of salient objects, to focus on high-level RGB features. Liu *et al.* [52] developed a unified model based on a pure transformer for the RGB and RGB-D SOD, where the saliency and boundary detection are simultaneously performed by introducing task-related tokens and a patch-task-attention mechanism. In this paper, we deploy the deeply supervision strategy with spatial attention based feedback mechanism to sharp the boundaries of salient regions at each level.

## III. Proposed Method

We first briefly describe the overall architecture of the proposed dynamic selective network (DSNet) in Sec. III-A. Second, we give a detailed description for the cross-modal global context module (CGCM) in detail in Sec. III-B, and present the dynamic selective module (DSM) in Sec. III-C. Eventually, we provide the implementation details of DSNet in Sec. III-D.

### *A. Architecture Overview*

The overall architecture of DSNet is shown in Fig. 2, which can be regarded as a typical encoder-decoder architecture. Briefly, the encoder contains a symmetric dual-stream backbone network, which is constructed based on ResNet-50 [12], and is used to extract the multi-level appearance features of RGB images and spatial features of depth maps. Noted, for each branch of encoder part, we discard the last pooling layer and fully connected layer, and only retain five convolutional blocks, which are downsampled by 2,4,8,16, and 16 times respectively, and converted the number of channels at each level from 64, 256, 512, 1024, 2048 to 64, 128, 256, 512, 512. For the decoder part, we design a dynamic selective module (DSM) and a cross-modal global context module (CGCM) for high-quality saliency predictions.

Specifically, we first encode the single-channel depth map into three-channel HHA [53], which represent the horizontal disparity, height above ground, and the angle of local surface normal with the inferred gravity direction, respectively. Therefore, we take the image pair composed of RGB image $I$ and HHA image $D$ as the input of our model. Then, we use a dual-stream encoder, namely the feature extraction network, to extract the multi-level texture features $\{\mathbf{F}_i^I\}_{i=1}^5$ and geometric features $\{\mathbf{F}_i^D\}_{i=1}^5$. After that, we design an effective and

efficient decoder. It mainly contains a cross-modal attention module (CAM) that dynamically integrates the complementary information of different modalities, a cross-modal global context module (CGCM) that emphasizes high-level semantic features, and a bi-directional gated pooling module (BGPM) that adaptively explores cross-level and multi-scale cues. Following this way, the proposed DSNet can generate accurate and high-quality saliency predictions.

### *B. Cross-Modal Global Context Module (CGCM)*

High-level features contain abundant semantic cues, which can effectively characterize global information. To fully integrate the high-level semantic features between different modalities, *i.e.*, RGB images and depth maps, we try to model the cross-modal long-range dependency. As a simplified non-local block, Global Context (GC) [54] block has two advantages including effective global context modeling and efficient lightweight computation. Inspired by [54], we propose an efficient cross-modal global context module (CGCM), which can roughly locate the salient objects. Concretely, according to Fig. 2, for the high-level feature $\mathbf{F}_5^I$ and $\mathbf{F}_5^D$ from the RGB and depth branches, a parameter sharing convolution operation is first used to convert the number of channels to 1, where the kernel size is $1 \times 1$ and stride is 1. Next, for RGB branch, we deploy the Sigmoid function to scale the value of high-level semantic feature $\mathbf{F}_5^{D'}$ to [0, 1], and then multiply it with $\mathbf{F}_5^I$ to obtain the interactive feature $\mathbf{F}_{inter}^I$. We call it the "inter-modal attention mechanism". Subsequently, we adopt the *Conv-ReLU-Conv* structure to further enhance deep RGB features, and multiply the feature maps normalized by Sigmoid function with $\mathbf{F}_5^I$ to obtain the enhanced RGB feature $\mathbf{F}_{intra}^I$. We call it the "intra-modal attention mechanism". Different from the GC block, which adds transformed features to original features, our CGCM further highlights the salient regions through multiplication, where the CGCM is built based on the spatial attention mechanism [40]. In addition, the GC block only models the single-modal long-range dependency, while the CGCM effectively achieves the interaction of global information between different modalities. The above calculation processes are formulated as,

$$\begin{cases} \mathbf{F}_{inter}^I = \mathbf{F}_5^I \otimes \delta\left(C_{1\times1}\left(\mathbf{F}_5^D\right)\right) \\ \mathbf{F}_{intra}^I = \mathbf{F}_5^I \odot \delta\left(CRC\left(\mathbf{F}_{inter}^I\right)\right), \end{cases} \tag{1}$$

where $\delta(\cdot)$ is the Sigmoid function, $C_{1\times1}(\cdot)$ denotes the convolutional layer with $1 \times 1$ kernel size, $CRC(\cdot)$ means the *Conv-ReLU-Conv* structure, and $\otimes$ and $\odot$ represent matrix multiplication and element-wise multiplication, respectively. Similarly, the enhanced depth feature $\mathbf{F}_{intra}^D$ is defined as,

$$\begin{cases} \mathbf{F}_{inter}^D = \mathbf{F}_5^D \otimes \delta\left(C_{1\times1}\left(\mathbf{F}_5^I\right)\right), \\ \mathbf{F}_{intra}^D = \mathbf{F}_5^D \odot \delta\left(CRC\left(\mathbf{F}_{inter}^D\right)\right). \end{cases} \tag{2}$$

Consequently, we obtain the depth-guided high-level RGB feature $\mathbf{F}_{intra}^I$ and the RGB-guided high-level depth feature $\mathbf{F}_{intra}^D$, which contain sufficient spatial texture cues and geometric structure information.

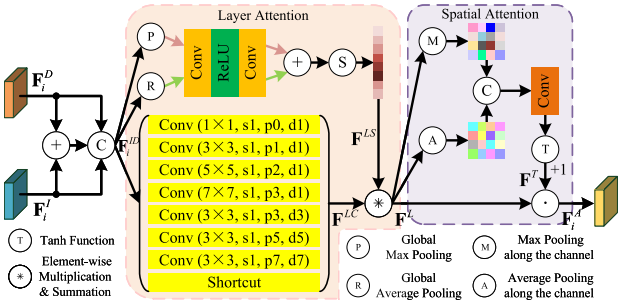


Fig. 3. Illustration of the cross-modal attention module (CAM). **Best viewed by zooming in**.

To further integrate RGB and depth high-level semantic features, we concatenate $\mathbf{F}_{intra}^I$ and $\mathbf{F}_{intra}^D$ to obtain the hybrid feature $\mathbf{F}^{ID}$, which is denoted as,

$$\mathbf{F}^{ID} = \left[\mathbf{F}_{intra}^I; \mathbf{F}_{intra}^D\right], \tag{3}$$

where $[\cdot;\cdot]$ represents the concatenation operation. Meanwhile, as shown in Fig. 2, we introduce the spatial attention mechanism [40] to expand the GC block-like structure, where the spatial attention mechanism applies average pooling and max pooling operations along the channel axis and concatenates the pooling results to locate salient objects. Finally, CGCM generates the cross-modal hybrid feature $\mathbf{F}^C$ that can sufficiently characterize global context information, which is calculated as,

$$\mathbf{F}^C = C_{1\times1}\left(\mathbf{F}^{ID}\right) \odot \delta\left(C_{7\times7}\left(\left[M\left(\mathbf{F}^{ID}\right); A\left(\mathbf{F}^{ID}\right)\right]\right)\right), \tag{4}$$

where $C_{n\times n}(\cdot)$ denotes the convolution operation with $n \times n$ kernel size, $M(\cdot)$ and $A(\cdot)$ represent the max pooling operation and average pooling operation along the channel axis. In general, CGCM ensures that our model can make complete saliency prediction since it integrates the high-level semantic features of different modalities, which gives the salient regions a coarse prediction.

### *C. Dynamic Selective Module (DSM)*

To improve the robustness of our model, we propose a dynamic selective module (DSM). It can not only automatically select and merge cross-modal features, namely RGB images and depth maps, but also autonomously optimize and strengthen cross-level and multi-scale deep features. Formally, DSM consists of two sub-modules: cross-modal attention module (CAM) and bi-directional gated pooling module (BGPM).

*1) Cross-Modal Attention Module (CAM):* How to effectively mine the complementarity between the cross-modal information is a crucial problem in multi-modal learning tasks, including RGB-D SOD. Therefore, we design a cross-modal attention module (CAM) shown in Fig. 3 to fully explore the information correlation between RGB images and depth maps, where the CAM efficiently highlights spatial features and fuses cross-modal information.

As shown in Fig. 2, let $\mathbf{F}_i^I$ and $\mathbf{F}_i^D$ denote the output feature maps of the $i^{th}(i = 1, \cdots, 5)$ convolutional block of RGB branch and depth branch, respectively, and each group

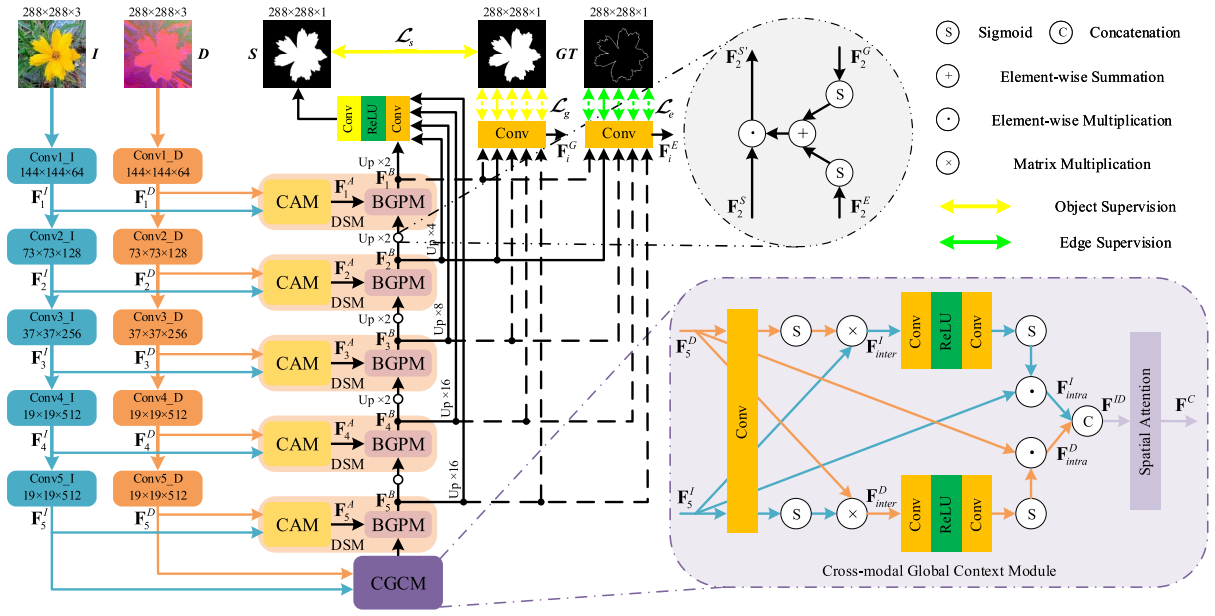


Fig. 2. The overall architecture of the proposed DSNet. We first adopt a variant of ResNet-50 for feature extraction. 'Conv1_I' ∼ 'Conv5_I' represent different convolutional blocks of the RGB branch, and 'Conv1_D' ∼ 'Conv5_D' represent different convolutional blocks of the depth branch. Then, the generated multi-level features $\mathbf{F}_i^I (i = 1, \cdots, 5)$ and $\mathbf{F}_i^D (i = 1, \cdots, 5)$ are selected and merged by the CAM to obtain the hybrid feature $\mathbf{F}_i^A (i = 1, \cdots, 5)$. Next, the CGCM further mines and integrates the high-level semantic information in a global view. Lastly, starting from $\mathbf{F}^C$, the hybrid feature $\mathbf{F}_i^A (i = 1, \cdots, 5)$ is gradually fed into the BGPM to obtain the final saliency map $S$. Details of our model are introduced in Sec. III.

of cross-modal features including $\mathbf{F}_i^I$ and $\mathbf{F}_i^D$ are sent to CAM. Concretely, for each CAM, according to Fig. 3, we first initially merge RGB feature $\mathbf{F}_i^I$ and depth feature $\mathbf{F}_i^D$, which not only achieves feature interaction but also retains their own unique information. The fusion process of the two modalities can be depicted as,

$$\mathbf{F}_i^{ID} = \left[\mathbf{F}_i^I; \mathbf{F}_i^D; \mathbf{F}_i^I \oplus \mathbf{F}_i^D\right], \tag{5}$$

where $\mathbf{F}_i^{ID}$ denotes the preliminary cross-modal feature of the $i^{th}$ block, and $\oplus$ represents element-wise summation.

Furthermore, to further mine cross-modal features, we explore the relationship between the output features of parallel convolutional layers with different settings. To be specific, we first set up seven different convolutional layers, as shown in Fig. 3, where '$1 \times 1$' means that the convolution kernel size is $1 \times 1$, '$s1$' means that the stride is 1, '$p0$' means that the padding is 0, and '$d1$' means that the dilation coefficient is 1. Particularly, we employ a shortcut branch to preserve the original information. Meanwhile, referring to the architecture of [40], we use both global max pooling operation and global average pooling operation simultaneously to calculate spatial statistics. The difference is that we generate a vector $\mathbf{F}^{LS}$ with eight elements to match the number of the parallel convolutional layers (including the shortcut branch). Then, we optimize the cross-layer deep features by adaptive selection, which is formulated as,

$$\mathbf{F}^L = \sum_{j=1}^{8} \mathbf{F}_j^{LS} \odot \mathbf{F}_j^{LC}, \tag{6}$$

where $\mathbf{F}_j^{LS}$ is the $j^{th}$ element in $\mathbf{F}^{LS}$, and $\mathbf{F}_j^{LC}$ represents the output feature maps of the $j^{th}$ convolutional layer (or the shortcut branch). Here, we treat the above computational procedure as "layer attention mechanism", and $\mathbf{F}^L$ denotes the output feature maps of layer attention mechanism. Noted, the different convolution settings ensure the feature diversity, which is beneficial for mining the complementarity of cross-modal information.

Besides, to strengthen the spatial structure features, we adopt spatial attention mechanism again, where we replace Sigmoid function with Tanh function, to expand the gap between foreground features and background features, suppress background regions, and highlight salient regions. The whole calculation process is implemented as,

$$\begin{cases} \mathbf{F}^T = Tanh\left(C_{7\times7}\left(\left[M\left(\mathbf{F}^L\right); A\left(\mathbf{F}^L\right)\right]\right)\right) + 1 \\ \mathbf{F}_i^A = \mathbf{F}^L \odot \mathbf{F}^T, \end{cases} \tag{7}$$

where $Tanh(\cdot)$ denotes the Tanh function, and $\mathbf{F}^T$ represents the spatial feature. $\mathbf{F}_i^A$ is the output of spatial attention block, that is, the overall output of the $i^{th}$ CAM. Noted, the CAM holds on the spatial resolution and the number of channel of feature maps. Generally, CAM adequately exploits the complementarity between RGB modality and depth modality, and achieves adaptive fusion of cross-modal features.

*2) Bi-Directional Gated Pooling Module (BGPM):* To explore the correlation between cross-level and multi-scale features, we leverage a bi-directional gated pooling module (BGPM) to autonomously merges different features. As shown in Fig. 4, our BGPM is divided into two steps. Firstly, the differences between cross-level features lie in that the features from shallow layers and deeper layers focus on representing spatial texture cues and semantic context information, respectively. Meanwhile, we have also noticed that the recurrent neural networks (RNNs) have achieved excellent performance when processing sequence signals, such as long short-term memory units (LSTM) [55] and gated

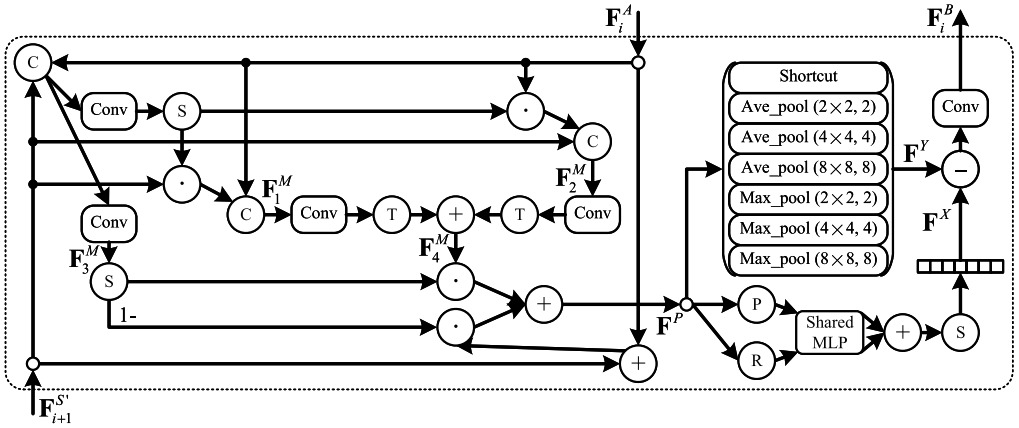


Fig. 4. The overall architecture of the bi-directional gated pooling module (BGPM), where ⊖ represents the element-wise multiplication and concatenation operation.

recurrent units (GRU) [56]. To associate features at different levels, we treat them as adjacent sequence signals and try to learn their correlation using RNNs-like structures. Concretely, we regard the features from the previous layer BGPM and the corresponding layer CAM as bi-directional data. Different from [56], which processes sequence signals sequentially in a one-directional way, the previous layer BGPM is used to strengthen semantic information of the corresponding layer CAM which is in turn applied to enhance the spatial cues of BGPM. Thus, BGPM explores the internal relationship between cross-level features in a bi-directional way, as shown in Fig. 4, where the computational process can be formulated as,

$$\begin{cases} \mathbf{F}_1^M = \left[\delta\left(C_{1\times1}\left(\left[\mathbf{F}_{i+1}^{S'};\mathbf{F}_i^A\right]\right)\right)\odot\mathbf{F}_{i+1}^{S'};\mathbf{F}_i^A\right] \\ \mathbf{F}_2^M = \left[\delta\left(C_{1\times1}\left(\left[\mathbf{F}_{i+1}^{S'};\mathbf{F}_i^A\right]\right)\right)\odot\mathbf{F}_i^A;\mathbf{F}_{i+1}^{S'}\right] \\ \mathbf{F}_3^M = C_{1\times1}\left(\left[\mathbf{F}_{i+1}^{S'};\mathbf{F}_i^A\right]\right) \\ \mathbf{F}_4^M = Tanh\left(C_{1\times1}\left(\mathbf{F}_1^M\right)\right)\oplus Tanh\left(C_{1\times1}\left(\mathbf{F}_2^M\right)\right), \end{cases} \tag{8}$$

where $\mathbf{F}_i^A (i = 1, 2, 3, 4, 5)$ denotes the output feature maps of the corresponding CAM, and $\mathbf{F}_{i+1}^{S'}$ denotes the enhanced output of the previous BGPM (we will introduce the $\mathbf{F}_{i+1}^{S'}$ in detail in Sec. III-D). Especially, when $i = 5$, $\mathbf{F}_6^{S'} = \mathbf{F}^C$, which is the output of CGCM. And $\mathbf{F}_1^M$, $\mathbf{F}_2^M$, $\mathbf{F}_3^M$ and $\mathbf{F}_4^M$ denote the hybrid features. In this way, we can obtain the deep feature maps,

$$\mathbf{F}^P = \delta\left(\mathbf{F}_3^M\right)\odot\mathbf{F}_4^M\oplus\left(1-\delta\left(\mathbf{F}_3^M\right)\right)\odot\left(\mathbf{F}_{i+1}^{S'}\oplus\mathbf{F}_i^A\right), \tag{9}$$

which combines the cross-level features.

Secondly, in order to be able to accurately detect the salient objects with different scales, we improve the layer attention mechanism proposed in CAM, and replace convolutional layers with different pooling layers, as shown in the right part of Fig. 4. It contains three average pooling layers, three max pooling layers and one shortcut connection. The kernel sizes of different pooling layers are set to 2, 4 and 8, respectively. Therefore, we not only retain the original information, but also extract effective features with different scale. Then, we integrate the deep features with different scale (including the shortcut branch) using following formulation,

$$\mathbf{F}_i^B = C_{3\times3}\left(\left[\mathbf{F}_j^X\odot\mathbf{F}_j^Y\right]\right), \tag{10}$$

where $\mathbf{F}_j^X$ represents the $j^{th}$ element in $\mathbf{F}^X$, and $\mathbf{F}_j^Y$ represents the output features of the $j^{th}$ pooling layer (including the shortcut branch). After that, we concatenate the product of each group of corresponding features (element) and pass them to a $3 \times 3$ convolutional layer, yielding the feature maps $\mathbf{F}_i^B$, where $i$ denotes the $i^{th}$ BGPM. Eventually, the output of each BGPM are first resized to the same resolution and same number of channel, and then they will be concatenated and feed into a convolutional block. Following this way, we can obtain the final saliency maps $\mathbf{S}$.

### D. Implementation Details

*1) Training Loss:* In this work, we adopt the cross entropy loss to train the proposed DSNet. To be specific, given binary object ground truth map $\mathbf{G}^s \in \{0, 1\}$ and final saliency map $\mathbf{S}^s \in [0, 1]$, the loss $\mathcal{L}_s$ can be defined as,

$$\mathcal{L}_s\left(\mathbf{S}^s,\mathbf{G}^s\right) = -\sum_{i=1}^{\mathcal{H}\times\mathcal{W}}\left[\mathbf{G}_i^s\log\left(\mathbf{S}_i^s\right)+\left(1-\mathbf{G}_i^s\right)\log\left(1-\mathbf{S}_i^s\right)\right], \tag{11}$$

where $\mathcal{H}$ and $\mathcal{W}$ respectively denote the height and width of input image, and $i$ denotes each pixel index. Furthermore, to ensure that each side-outputs in decoder part can accurately locate the salient objects, we adopt deeply supervision strategy [29], [30]. Concretely, we upsample each side-output to the same resolution as input image, yielding coarse saliency map $\mathbf{S}_j^g (j = 1, \cdots, 5)$. Similar as Eq. 11, the loss function $\mathcal{L}_g$ can be defined as follows:

$$\mathcal{L}_g\left(\mathbf{S}^g,\mathbf{G}^s\right) = \sum_{j=1}^{5} -\left[\mathbf{G}^s\log\left(\mathbf{S}_j^g\right)+\left(1-\mathbf{G}^s\right)\log\left(1-\mathbf{S}_j^g\right)\right], \tag{12}$$

where $\mathbf{S}_j^g$ denotes the $j^{th}$ coarse object saliency map.

Besides, to give a sharp boundary detail for salient objects, we deploy the supervision for edge computation. First of

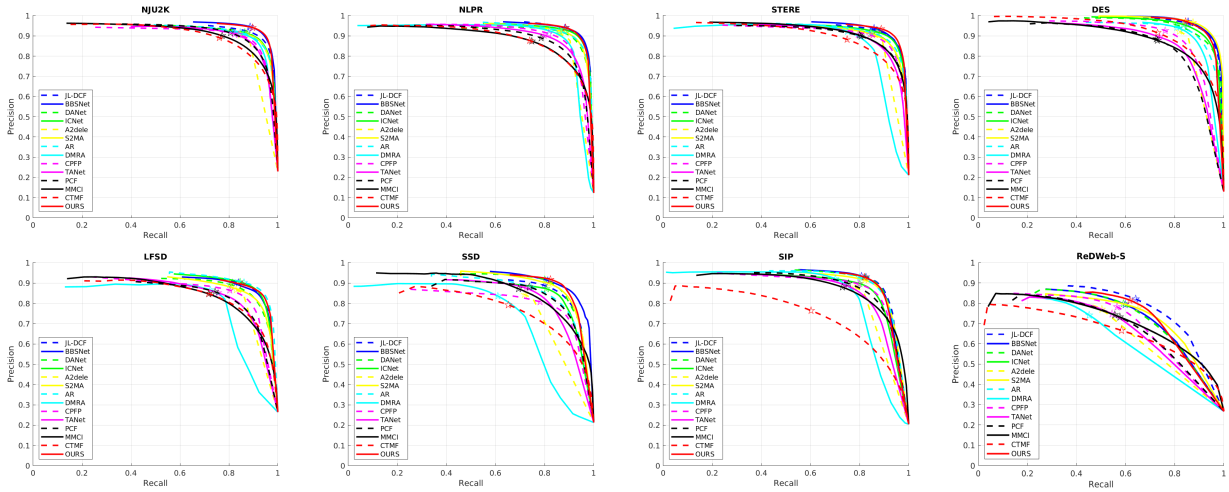


Fig. 5. The PR curves of our model with 13 state-of-the-art models on eight challenging RGB-D datasets. The solid red lines represent our model and stars on the curves represent the corresponding value of maximum F-measure.

all, referring to [61], we can obtain edge ground truth maps $\mathbf{G}^e \in \{0, 1\}$. Meanwhile, to remit the imbalance between positive and negative samples, we adopt the balanced cross-entropy loss for each side-output of edge maps $\mathbf{S}^e \in [0, 1]$,

$$\mathcal{L}_e\left(\mathbf{S}^e, \mathbf{G}^e\right) = \sum_{j=1}^{5} -\left[\theta \mathbf{G}^e \log\left(\mathbf{S}_j^e\right) + \eta\left(1-\mathbf{G}^e\right)\log\left(1-\mathbf{S}_j^e\right)\right], \quad (13)$$

where $\mathbf{S}_j^e$ denotes the $j^{th}$ coarse edge map, $\eta$ represents the ratio of negative pixels, and $\theta = \lambda(1-\eta)$. In our experiment, we set the hyperparameter $\lambda$ to 1.1, which will pay more attention on edge pixels than background. According to Eq. 11, Eq. 12 and Eq. 13, the total loss $\mathcal{L}_{total}$ is formulated as,

$$\mathcal{L}_{total} = \alpha\mathcal{L}_s\left(\mathbf{S}^s, \mathbf{G}^s\right) + \beta\mathcal{L}_g\left(\mathbf{S}^g, \mathbf{G}^s\right) + \gamma\,\mathcal{L}_e\left(\mathbf{S}^e, \mathbf{G}^e\right), \quad (14)$$

where we set $\alpha$, $\beta$ and $\gamma$ to 1, 0.5 and 0.5 for balancing the importance of each part in total loss, respectively.

In addition, to further promote all side-outputs, we implement a multi-level feedback mechanism shown in Fig. 2, where the feature maps are enhanced before feeding it to BGPM,

$$\mathbf{F}_i^{S'} = \mathbf{F}_i^S \odot \left(\delta\left(\mathbf{F}_i^G\right) \oplus \delta\left(\mathbf{F}_i^E\right)\right), \quad (15)$$

where $i \in \{2, 3, 4, 5\}$, $\mathbf{F}^S$ and $\mathbf{F}^{S'}$ denote the feature maps before and after enhancement, respectively. And $\mathbf{F}^G$ is object mask and $\mathbf{F}^E$ is edge mask, which are generated by $\mathbf{F}^B$ with different convolutional blocks at each level. As shown in Eq. 16, both of them have the relationship as,

$$\begin{cases} \mathbf{S}_i^g = Up\left(\mathbf{F}_i^G\right) \\ \mathbf{S}_i^e = Up\left(\mathbf{F}_i^E\right), \end{cases} \quad (16)$$

where $Up(\cdot)$ denotes the bilinear interpolation operation.

*2) Training Protocol:* We implement the proposed DSNet using PyTorch toolbox [62] and trained it on a high-performance server with a E5-2678 CPU (64 GB memory) and a NVIDIA GeForce RTX 2080Ti GPU (11 GB memory). During training process, the backbone parts of DSNet is initialized with pre-trained ResNet-50 [12], and the remaining parameters are initialized with He initialization [63]. It should also be noted that the dual-stream encoder doesn't share parameters. Furthermore, we adopt Adam algorithm [64] to optimize the model, where the momentum, weight decay, and initial learning rate are set to [0.9, 0.999], $1e$-4, and $1e$-5, respectively. In addition, we train the model for 50 epochs with a mini-batch size of 4, and divide learning rate by 10 after 35 epochs. Meanwhile, to prevent the model from overfitting, we also leverage multiple augmentation strategies, such as rotation and flipping. Specifically, we adopt a fixed augmentation strategy, namely first rotating the original images with angles 90°, 180° and 270°, and then flipping all of them respectively. In the end, the augmented data is 8 times the original data in total, *i.e.*, 17,480 image pairs. Notably, we resize the input images (*i.e.*, RGB images and depth maps) to $288 \times 288$ in both training and testing phases.

## IV. Experiments

In this section, we first introduce RGB-D datasets and evaluation metrics in Sec. IV-A. Then, we provide the quantitative and qualitative comparison between the proposed method and state-of-the-art RGB-D SOD methods in Sec. IV-B. Next, in Sec. IV-C, we conduct comprehensive ablation studies to demonstrate the rationality of the design of our model. Finally, we give a detailed analysis of the failure cases in Sec. IV-D.

### A. Experimental Settings

*1) Datasets:* We conduct our experiments on eight public RGB-D datasets: NJU2K [65], NLPR [66], STERE [67], DES [68], LFSD [69], SSD [70], SIP [71] and ReDWeb-S [72].

For a fair comparison, we follow the same dataset settings as [27]. Concretely, the training set before augmentation consists of 2,185 samples in total, including 1,485 samples from NJU2K and 700 samples from NLPR. The test set contains the remaining samples in the NJU2K, NLPR, ReDWeb-S test set and the whole of STERE, DES, LFSD, SSD and SIP.

*2) Evaluation Metrics:* We leverage six widely used evaluation metrics, including mean absolute error (MAE), mean

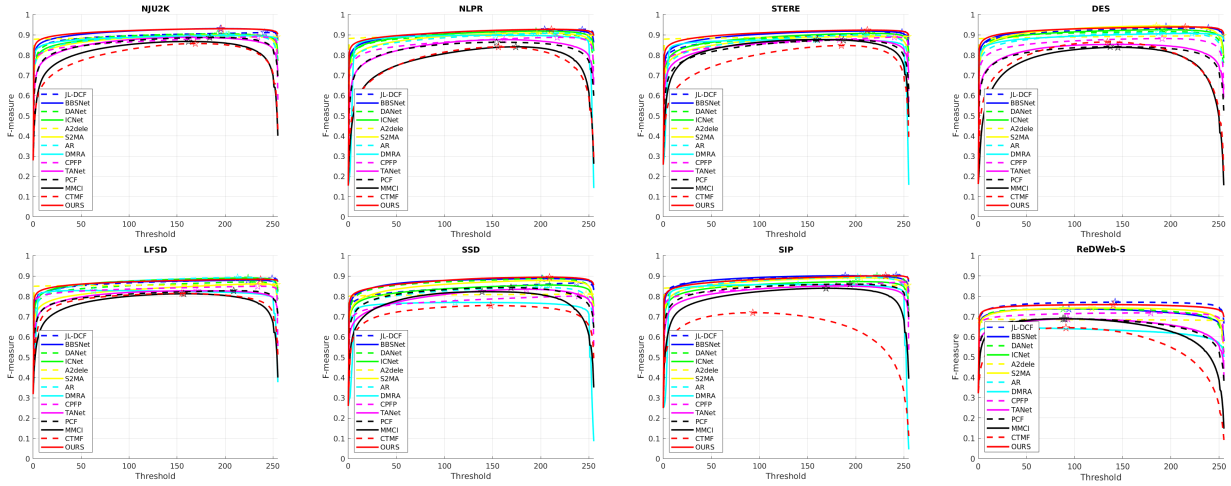


Fig. 6. The F-measure curves of our method and 13 CNNs-based methods over eight benchmark datasets. The solid red lines represent our method and stars on the curves represent the corresponding value of maximum F-measure.

F-measure ($F_\beta$) [73], mean E-measure ($E_\xi$) [74], S-measure ($S_\alpha$) [75], precision-recall (PR) curves and the recently proposed weighted F-measure ($F_\beta^w$) [76], to evaluate the performance of different methods.

### B. Comparison With the State-of-the-Arts

We compare our DSNet with 17 state-of-the-art RGB-D SOD models, including four manual features based models: CDCP [22], LBE [23], SE [24] and MDSF [25]; and 13 deep learning based models: CTMF [57], MMCI [58], PCF [46],TANet [59], CPFP [60], DMRA [48], AR [30], ICNet [29], A2dele [21], S2MA [44], DANet [26] and BBSNet [27] and JL-DCF [47]. Besides, we adopt the officially released source code and parameters to generate saliency maps or directly use the results provided by the author for comparisons. In particular, we resize the predicted saliency maps of different methods to the same resolution, and use the same evaluation toolbox for fair evaluation.

*1) Quantitative Performance Comparison:* As shown in Tab. I, our method achieves a comparable performance against the state-of-the-art methods on eight challenging datasets, *i.e.*, NJU2K [65], NLPR [66], STERE [67], DES [68], LFSD [69], SSD [70], SIP [71] and ReDWeb-S [72], in terms of five commonly used evaluation metrics. We highlight the best three results in red, blue, and green successively. It can be found that our method obtains competitive results on all benchmark datasets except SIP and ReDWeb-S when compared with the state-of-the-art methods, where the performance of our model is also comparable with JL-DCF [47] on the SIP and ReDWeb-S datasets. In Fig. 5, we display PR curves of various methods on different benchmark datasets. It is clear to see that our method (*i.e.*, solid red line) is superior to other methods, especially on NJU2K and STERE dataset. Besides, we also show the evaluation results of our method and 13 CNNs-based methods in terms of F-measure curves shown in Fig. 6. It can be found that our method achieves great advantages when compared with other models on different datasets. According to the aforementioned experiments, we can clearly demonstrate the effectiveness of our model.

In addition, in Tab. I, we also provide the inference time (seconds per image) of different models. It can be seen that our model takes about 0.046s for an image ($288 \times 288$). This is a remarkable performance when compared with the state-of-the-art methods such as JL-DCF [47], S2MA [44], and ICNet [29]. Therefore, Tab. I demonstrates the efficiency of our model.

*2) Qualitative Performance Comparison:* In this part, we present some saliency maps predicted by our method and several state-of-the-art methods shown in Fig. 7, which contains several representative situations including simple scenes (the $1^{st}$ row), low contrast (the $2^{nd}$ row), small object (the $3^{rd}$ row), multiple objects (the $4^{th}$ row), unclear depth (the $5^{th}$ row), similar background (the $6^{th}$ row), similar depth (the $7^{th}$ row), complex background (the $8^{th}$ row), and complex objects (the $9^{th}$ row).

Specifically, as shown in the $1^{st}$ row of Fig. 7, we present a simple case. It is easy to find that most methods can locate and highlight the airplane in broad sky. When the contrast between salient objects and background regions is low, such as the pink high-heeled shoe and the partition in the $2^{nd}$ row of Fig. 7, some methods mistakenly confuse background regions as salient objects, while our method can accurately detect the high-heeled shoe. In the $3^{rd}$ row of Fig. 7, we find that our method highlights small object more clearly and accurately when compared with other models. In fact, there are often multiple salient objects in a scene, as shown in the $4^{th}$ row of Fig. 7. Some methods only detect a part of salient objects, while our method locates all salient objects, *i.e.*, the two people in conversation. Meanwhile, as an important supplement of RGB images, the quality of depth maps seriously affects the accuracy of salient object detection. In the $5^{th}$ row of Fig. 7, although the depth map contains a lot of noise and has low quality, our method still predicts the salient objects more accurately than other methods. When dealing with the situation that the foreground and background are similarity, some previous methods falsely highlight the confusing background regions. As shown in the $6^{th}$ row of Fig. 7, our method suppresses the similar maple leaves more effectively, while other models falsely pop-out the non-salient leaves. Besides, as an RGB-D saliency detection method, our model utilizes

TABLE I

QUANTITATIVE COMPARISON RESULTS OF MAE ($\mathcal{M}$), MEAN F-MEASURE ($F_\beta$), MEAN E-MEASURE ($E_\xi$), S-MEASURE ($S_\alpha$) AND WEIGHTED F-MEASURE ($F_\beta^w$) ON EIGHT WIDELY USED DATASETS WITH DIFFERENT RGB-D SALIENCY MODELS. NOTED, $*$ REPRESENTS THE TRADITIONAL RGB-D SALIENCY MODEL, 'DSNET' IS THE PROPOSED MODEL, $\downarrow$ & $\uparrow$ DENOTE SMALLER AND LARGER IS BETTER, RESPECTIVELY. THE BEST THREE RESULTS IN EACH ROW ARE HIGHLIGHTED IN RED, BLUE, AND GREEN SUCCESSIVELY

| | Metrics | CDCP* [22] | LBE* [23] | SE* [24] | MDSF* [25] | CTMF [57] | MMCI [58] | PCF [46] | TANet [59] | CPFP [60] | DMRA [48] | AR [30] | ICNet [29] | A2dele [21] | S2MA [44] | DANet [26] | BBSNet [27] | JL-DCF [47] | **DSNet Ours** |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Time(s) | >60.0 | 3.110 | 1.570 | >60.0 | 0.630 | 0.050 | 0.060 | 0.070 | 0.170 | 0.063 | 0.179 | 0.075 | 0.014 | 0.111 | 0.031 | 0.038 | 0.111 | 0.046 |
| *NJU2K* | $\mathcal{M}\downarrow$ | 0.180 | 0.153 | 0.169 | 0.157 | 0.085 | 0.079 | 0.059 | 0.060 | 0.053 | 0.051 | 0.055 | 0.052 | 0.051 | 0.053 | 0.046 | 0.035 | 0.041 | 0.034 |
| | $F_\beta\uparrow$ | 0.595 | 0.606 | 0.583 | 0.628 | 0.779 | 0.793 | 0.840 | 0.841 | 0.850 | 0.873 | 0.866 | 0.869 | 0.869 | 0.865 | 0.874 | 0.902 | 0.885 | 0.909 |
| | $E_\xi\uparrow$ | 0.706 | 0.655 | 0.624 | 0.677 | 0.846 | 0.851 | 0.895 | 0.895 | 0.910 | 0.920 | 0.912 | 0.914 | 0.912 | 0.914 | 0.921 | 0.938 | 0.935 | 0.945 |
| | $S_\alpha\uparrow$ | 0.669 | 0.695 | 0.664 | 0.748 | 0.849 | 0.858 | 0.877 | 0.878 | 0.878 | 0.886 | 0.893 | 0.894 | 0.871 | 0.894 | 0.897 | 0.921 | 0.902 | 0.921 |
| | $F_\beta^w\uparrow$ | 0.511 | 0.552 | 0.507 | 0.560 | 0.720 | 0.738 | 0.803 | 0.804 | 0.828 | 0.846 | 0.840 | 0.843 | 0.843 | 0.842 | 0.853 | 0.884 | 0.869 | 0.895 |
| *NLPR* | $\mathcal{M}\downarrow$ | 0.112 | 0.081 | 0.091 | 0.095 | 0.056 | 0.059 | 0.044 | 0.041 | 0.036 | 0.031 | 0.031 | 0.028 | 0.029 | 0.030 | 0.031 | 0.023 | 0.022 | 0.024 |
| | $F_\beta\uparrow$ | 0.609 | 0.736 | 0.624 | 0.649 | 0.740 | 0.737 | 0.802 | 0.819 | 0.840 | 0.863 | 0.867 | 0.884 | 0.875 | 0.873 | 0.871 | 0.896 | 0.894 | 0.899 |
| | $E_\xi\uparrow$ | 0.781 | 0.719 | 0.742 | 0.745 | 0.840 | 0.841 | 0.887 | 0.902 | 0.918 | 0.939 | 0.934 | 0.941 | 0.941 | 0.937 | 0.933 | 0.950 | 0.955 | 0.951 |
| | $S_\alpha\uparrow$ | 0.727 | 0.762 | 0.756 | 0.805 | 0.860 | 0.856 | 0.874 | 0.886 | 0.888 | 0.899 | 0.914 | 0.923 | 0.898 | 0.915 | 0.909 | 0.930 | 0.925 | 0.926 |
| | $F_\beta^w\uparrow$ | 0.510 | 0.592 | 0.560 | 0.602 | 0.679 | 0.676 | 0.762 | 0.780 | 0.813 | 0.837 | 0.847 | 0.864 | 0.857 | 0.852 | 0.850 | 0.879 | 0.882 | 0.883 |
| *STERE* | $\mathcal{M}\downarrow$ | 0.149 | 0.250 | 0.143 | 0.176 | 0.086 | 0.068 | 0.064 | 0.060 | 0.051 | 0.047 | 0.053 | 0.045 | 0.044 | 0.051 | 0.048 | 0.041 | 0.040 | 0.036 |
| | $F_\beta\uparrow$ | 0.638 | 0.501 | 0.610 | 0.527 | 0.758 | 0.813 | 0.818 | 0.828 | 0.841 | 0.868 | 0.850 | 0.870 | 0.873 | 0.853 | 0.857 | 0.883 | 0.873 | 0.895 |
| | $E_\xi\uparrow$ | 0.751 | 0.601 | 0.665 | 0.614 | 0.841 | 0.873 | 0.887 | 0.893 | 0.912 | 0.930 | 0.908 | 0.926 | 0.924 | 0.914 | 0.915 | 0.928 | 0.935 | 0.940 |
| | $S_\alpha\uparrow$ | 0.713 | 0.660 | 0.708 | 0.728 | 0.848 | 0.873 | 0.875 | 0.871 | 0.879 | 0.886 | 0.893 | 0.903 | 0.878 | 0.890 | 0.892 | 0.908 | 0.903 | 0.915 |
| | $F_\beta^w\uparrow$ | 0.558 | 0.397 | 0.538 | 0.457 | 0.698 | 0.760 | 0.778 | 0.787 | 0.817 | 0.749 | 0.820 | 0.844 | 0.846 | 0.825 | 0.830 | 0.858 | 0.857 | 0.877 |
| *DES* | $\mathcal{M}\downarrow$ | 0.115 | 0.208 | 0.090 | 0.122 | 0.055 | 0.065 | 0.049 | 0.046 | 0.038 | 0.030 | 0.030 | 0.027 | 0.029 | 0.021 | 0.028 | 0.021 | 0.020 | 0.021 |
| | $F_\beta\uparrow$ | 0.585 | 0.576 | 0.617 | 0.523 | 0.756 | 0.735 | 0.765 | 0.790 | 0.824 | 0.872 | 0.867 | 0.893 | 0.865 | 0.908 | 0.877 | 0.910 | 0.907 | 0.913 |
| | $E_\xi\uparrow$ | 0.748 | 0.649 | 0.707 | 0.621 | 0.826 | 0.825 | 0.838 | 0.863 | 0.889 | 0.932 | 0.934 | 0.941 | 0.915 | 0.953 | 0.923 | 0.949 | 0.959 | 0.952 |
| | $S_\alpha\uparrow$ | 0.709 | 0.703 | 0.741 | 0.741 | 0.863 | 0.848 | 0.842 | 0.858 | 0.872 | 0.900 | 0.915 | 0.920 | 0.886 | 0.941 | 0.905 | 0.933 | 0.931 | 0.927 |
| | $F_\beta^w\uparrow$ | 0.486 | 0.341 | 0.541 | 0.434 | 0.686 | 0.650 | 0.714 | 0.739 | 0.787 | 0.842 | 0.847 | 0.867 | 0.836 | 0.892 | 0.848 | 0.890 | 0.894 | 0.894 |
| *LFSD* | $\mathcal{M}\downarrow$ | 0.167 | 0.208 | 0.167 | 0.190 | 0.119 | 0.132 | 0.112 | 0.111 | 0.088 | 0.076 | 0.070 | 0.071 | 0.077 | 0.094 | 0.082 | 0.072 | 0.070 | 0.068 |
| | $F_\beta\uparrow$ | 0.680 | 0.611 | 0.640 | 0.521 | 0.756 | 0.722 | 0.761 | 0.771 | 0.811 | 0.845 | 0.852 | 0.852 | 0.828 | 0.806 | 0.826 | 0.843 | 0.848 | 0.849 |
| | $E_\xi\uparrow$ | 0.754 | 0.670 | 0.653 | 0.588 | 0.810 | 0.775 | 0.818 | 0.821 | 0.863 | 0.893 | 0.894 | 0.892 | 0.871 | 0.855 | 0.872 | 0.883 | 0.894 | 0.890 |
| | $S_\alpha\uparrow$ | 0.717 | 0.736 | 0.698 | 0.700 | 0.796 | 0.787 | 0.794 | 0.801 | 0.828 | 0.847 | 0.876 | 0.868 | 0.833 | 0.837 | 0.845 | 0.864 | 0.861 | 0.867 |
| | $F_\beta^w\uparrow$ | 0.602 | 0.530 | 0.574 | 0.499 | 0.698 | 0.663 | 0.714 | 0.719 | 0.775 | 0.811 | 0.825 | 0.822 | 0.800 | 0.772 | 0.789 | 0.814 | 0.822 | 0.824 |
| *SSD* | $\mathcal{M}\downarrow$ | 0.214 | 0.278 | 0.165 | 0.192 | 0.099 | 0.082 | 0.062 | 0.063 | 0.082 | 0.059 | 0.079 | 0.064 | 0.070 | 0.052 | 0.050 | 0.044 | 0.055 | 0.044 |
| | $F_\beta\uparrow$ | 0.515 | 0.489 | 0.564 | 0.470 | 0.689 | 0.721 | 0.777 | 0.773 | 0.747 | 0.827 | 0.797 | 0.815 | 0.773 | 0.823 | 0.833 | 0.843 | 0.817 | 0.860 |
| | $E_\xi\uparrow$ | 0.676 | 0.574 | 0.631 | 0.576 | 0.796 | 0.796 | 0.856 | 0.861 | 0.839 | 0.897 | 0.862 | 0.887 | 0.856 | 0.890 | 0.894 | 0.904 | 0.898 | 0.907 |
| | $S_\alpha\uparrow$ | 0.603 | 0.621 | 0.675 | 0.673 | 0.776 | 0.813 | 0.841 | 0.839 | 0.807 | 0.857 | 0.842 | 0.848 | 0.802 | 0.868 | 0.869 | 0.882 | 0.862 | 0.885 |
| | $F_\beta^w\uparrow$ | 0.423 | 0.367 | 0.485 | 0.417 | 0.624 | 0.660 | 0.735 | 0.726 | 0.708 | 0.784 | 0.744 | 0.772 | 0.726 | 0.787 | 0.798 | 0.812 | 0.786 | 0.830 |
| *SIP* | $\mathcal{M}\downarrow$ | 0.224 | 0.200 | 0.164 | 0.167 | 0.139 | 0.086 | 0.071 | 0.075 | 0.064 | 0.094 | 0.063 | 0.069 | 0.070 | 0.057 | 0.054 | 0.055 | 0.049 | 0.051 |
| | $F_\beta\uparrow$ | 0.482 | 0.571 | 0.515 | 0.568 | 0.608 | 0.771 | 0.814 | 0.803 | 0.821 | 0.801 | 0.853 | 0.834 | 0.826 | 0.854 | 0.864 | 0.868 | 0.873 | 0.864 |
| | $E_\xi\uparrow$ | 0.683 | 0.651 | 0.592 | 0.645 | 0.705 | 0.845 | 0.878 | 0.870 | 0.893 | 0.835 | 0.899 | 0.890 | 0.886 | 0.905 | 0.910 | 0.906 | 0.918 | 0.911 |
| | $S_\alpha\uparrow$ | 0.595 | 0.727 | 0.628 | 0.717 | 0.716 | 0.833 | 0.842 | 0.835 | 0.850 | 0.816 | 0.872 | 0.854 | 0.828 | 0.872 | 0.878 | 0.879 | 0.880 | 0.876 |
| | $F_\beta^w\uparrow$ | 0.558 | 0.438 | 0.430 | 0.469 | 0.535 | 0.712 | 0.768 | 0.748 | 0.788 | 0.693 | 0.814 | 0.791 | 0.780 | 0.819 | 0.829 | 0.830 | 0.844 | 0.834 |
| *ReDWeb-S* | $\mathcal{M}\downarrow$ | - | - | - | - | 0.204 | 0.176 | 0.166 | 0.165 | 0.142 | 0.188 | - | - | 0.160 | 0.139 | 0.142 | 0.150 | 0.128 | 0.133 |
| | $F_\beta\uparrow$ | - | - | - | - | 0.514 | 0.562 | 0.577 | 0.584 | 0.635 | 0.538 | - | - | 0.596 | 0.677 | 0.647 | 0.648 | 0.712 | 0.676 |
| | $E_\xi\uparrow$ | - | - | - | - | 0.622 | 0.652 | 0.667 | 0.672 | 0.717 | 0.614 | - | - | 0.661 | 0.751 | 0.713 | 0.709 | 0.776 | 0.736 |
| | $S_\alpha\uparrow$ | - | - | - | - | 0.641 | 0.660 | 0.655 | 0.656 | 0.685 | 0.592 | - | - | 0.641 | 0.711 | 0.693 | 0.693 | 0.734 | 0.716 |
| | $F_\beta^w\uparrow$ | - | - | - | - | 0.457 | 0.503 | 0.516 | 0.517 | 0.578 | 0.456 | - | - | 0.528 | 0.621 | 0.587 | 0.585 | 0.656 | 0.624 |

depth information efficiently, but does not heavily rely on depth information. In the $7^{th}$ row of Fig. 7, we show a situation with similar depth, and we can clearly see that our method segments salient objects correctly. In contrast, some other models either falsely highlight the pedestal, or detect salient objects incompletely. As shown in the $8^{th}$ row of Fig. 7, in the scenarios with complex background, our method stably detects the salient car with sharper boundaries. In the $9^{th}$ row of Fig. 7, we also provide an example with complex objects. Many methods are difficult to locate all parts of salient objects accurately and completely, but our method provides reliable predicted results.

Overall, compared with various state-of-the-art methods, our method performs well. When facing various challenging scenarios, our method can not only highlight salient objects accurately and completely, but also provides precise boundary details.

### C. Ablation Studies

*1) Utility of Depth Cues:* To explore whether depth maps can really help RGB images to improve the performance of SOD, we conduct sufficient experiments from different perspectives. On the one hand, we verify the effectiveness of depth cues on the baseline. We remove all CGCM and DSM modules in DSNet, as shown in Tab. II. In the $1^{st}$ row of Tab. II, we only keep the RGB branch, and the variant model achieves unsatisfactory results on NJU2K test set and LFSD dataset. For a comparison, we further add the depth branch (*i.e.*, the $2^{nd}$ row of Tab. II), while holding

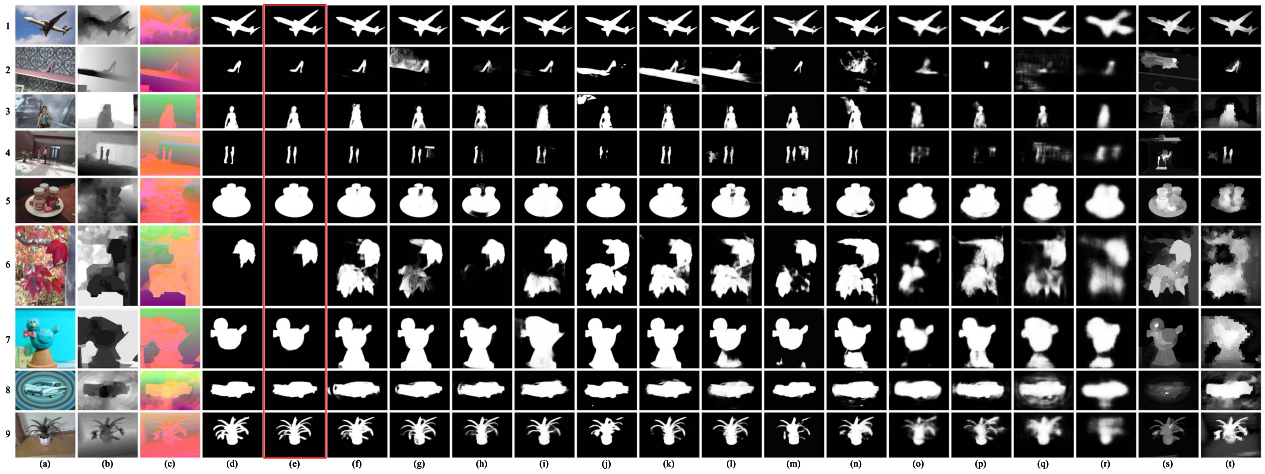

Fig. 7. Qualitative comparison results of different RGB-D saliency models on several challenging scenarios. (a) RGB, (b) Depth, (c) HHA, (d) GT, (e) Ours, (f) JL-DCF [47], (g) BBSNet [27], (h) DANet [26], (i) S2MA [44], (j) A2dele [21], (k) ICNet [29], (l) AR [30], (m) DMRA [48], (n) CPFP [60], (o) TANet [59], (p) PCF [46], (q) MMCI [58], (r) CTMF [57], (s) MDSF [25], (t) LBE [23].

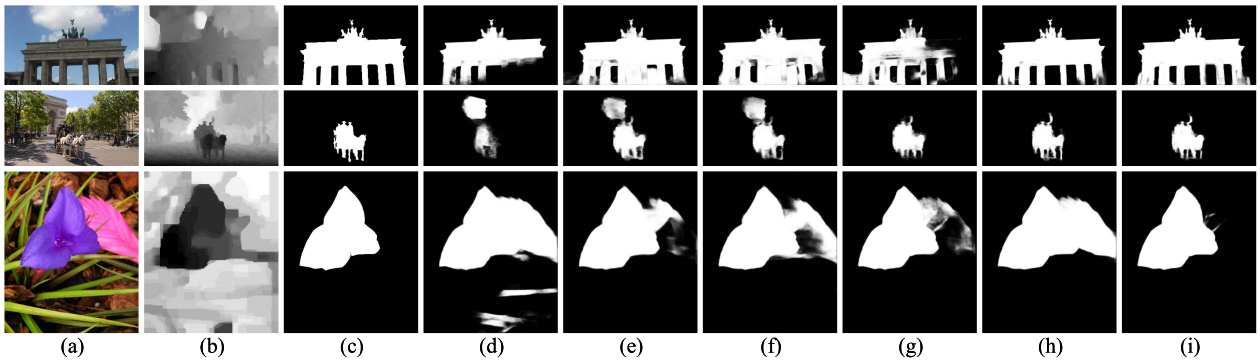

Fig. 8. Visualization results of ablation studies. (a) RGB, (b) Depth, (c) GT, (d) #1, (e) #2, (f) #3, (g) #4, (h) #5, (i) #8. Noted, '#n' corresponds to the $n^{th}$ row in Tab. II.

TABLE II

ABLATION STUDIES FOR OUR DSNET ON WIDELY USED NJU2K TEST SET AND LFSD DATASET. THE BEST RESULT IN EACH COLUMN IS MARKED IN **BOLD**. 'RGB-B': RGB BRANCH, 'DEPTH-B': DEPTH BRANCH, 'CGCM': CROSS-MODAL GLOBAL CONTEXT MODULE, 'CAM': CROSS-MODAL ATTENTION MODULE, 'BGPM': BI-DIRECTIONAL GATED POOLING MODULE, 'OBJECT-S': OBJECT SUPERVISION, AND 'EDGE-S': EDGE SUPERVISION

| # | Settings | | | | | | | NJU2K [65] | | | | | LFSD [69] | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | RGB-B | Depth-B | CGCM | CAM | BGPM | Object-S | Edge-S | $\mathcal{M}\downarrow$ | $F_\beta\uparrow$ | $E_\xi\uparrow$ | $S_\alpha\uparrow$ | $F_\beta^w\uparrow$ | $\mathcal{M}\downarrow$ | $F_\beta\uparrow$ | $E_\xi\uparrow$ | $S_\alpha\uparrow$ | $F_\beta^w\uparrow$ |
| 1 | ✓ | | | | | ✓ | ✓ | 0.049 | 0.868 | 0.916 | 0.889 | 0.842 | 0.112 | 0.753 | 0.806 | 0.785 | 0.716 |
| 2 | ✓ | ✓ | | | | ✓ | ✓ | 0.041 | 0.882 | 0.928 | 0.904 | 0.862 | 0.093 | 0.776 | 0.828 | 0.811 | 0.746 |
| 3 | ✓ | ✓ | ✓ | | | ✓ | ✓ | 0.040 | 0.887 | 0.930 | 0.906 | 0.866 | 0.088 | 0.807 | 0.853 | 0.830 | 0.775 |
| 4 | ✓ | ✓ | ✓ | ✓ | | ✓ | ✓ | 0.038 | 0.894 | 0.934 | 0.911 | 0.875 | 0.087 | 0.809 | 0.862 | 0.833 | 0.779 |
| 5 | ✓ | ✓ | ✓ | | ✓ | ✓ | ✓ | 0.037 | 0.907 | 0.940 | 0.919 | 0.890 | 0.071 | 0.838 | 0.887 | 0.859 | 0.812 |
| 6 | ✓ | | ✓ | ✓ | ✓ | ✓ | ✓ | 0.040 | 0.897 | 0.934 | 0.910 | 0.877 | 0.092 | 0.801 | 0.851 | 0.827 | 0.771 |
| 7 | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | | 0.035 | 0.901 | 0.940 | 0.917 | 0.886 | 0.073 | 0.847 | 0.887 | 0.866 | 0.823 |
| 8 | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | **0.034** | **0.909** | **0.945** | **0.921** | **0.895** | **0.068** | **0.849** | **0.890** | **0.867** | **0.824** |

on the other experimental settings. With the help of depth maps, we clearly see that the model with depth branch achieves a large improvement on five different evaluation metrics (*i.e.*, $\mathcal{M}$, $F_\beta$, $E_\xi$, $S_\alpha$ and $F_\beta^w$). To illustrate the utility of depth cues more intuitively, we visualize the predicted results, as shown in Fig. 8. The saliency maps in Fig. 8 (d) and (e) correspond to the model without/with depth branch, respectively. It is easy to find that the latter is good at highlighting salient regions (*e.g.*, the $1^{st}$ row of Fig. 8) and suppressing background regions (*e.g.*, the $3^{rd}$ row of Fig. 8).

On the other hand, we compare DSNet with several state-of-the-art RGB SOD methods, including PiCANet [41], PAGRN [77], $R^3$Net [78], CPD [79], PoolNet [36], SOD100K [17] and GateNet [18]. As shown in Tab. III, we conduct experiments on three datasets (*i.e.*, NJU2K test set, NLPR test set, and DES dataset) in terms of two evaluation metrics (*i.e.*, $\mathcal{M}$ and $S_\alpha$). Notably, we use official pre-trained model of SOD100K to generate saliency maps, and adopt the public predicted results of GateNet. For other methods, we borrow the evaluation results in [27] and [48]. Meanwhile, we further design another variant of DSNet, namely our model

TABLE III

COMPARISON WITH THE STATE-OF-THE-ART RGB SOD METHODS ON THREE DATASETS. 'W/O D' AND 'W/ D' DENOTE THAT OUR MODEL IS TRAINED AND TESTED WITHOUT/WITH THE DEPTH BRANCH, RESPECTIVELY

| Datasets | Metrics | PiCANet [41] | PAGR [77] | $R^3$Net [78] | CPD [79] | PoolNet [36] | SOD100K [17] | GateNet [18] | DSNet(w/o D) | DSNet(w/ D) |
|---|---|---|---|---|---|---|---|---|---|---|
| NJU2K [65] | $\mathcal{M}\downarrow$ | 0.071 | 0.081 | 0.092 | 0.046 | 0.045 | 0.083 | 0.047 | 0.040 | **0.034** |
| | $S_\alpha\uparrow$ | 0.847 | 0.829 | 0.837 | 0.894 | 0.887 | 0.851 | 0.902 | 0.910 | **0.921** |
| NLPR [66] | $\mathcal{M}\downarrow$ | 0.053 | 0.051 | 0.101 | 0.025 | 0.029 | 0.053 | 0.032 | 0.025 | **0.024** |
| | $S_\alpha\uparrow$ | 0.834 | 0.844 | 0.798 | 0.915 | 0.900 | 0.876 | 0.910 | 0.922 | **0.926** |
| DES [68] | $\mathcal{M}\downarrow$ | 0.042 | 0.044 | 0.066 | 0.028 | 0.034 | 0.048 | 0.030 | 0.028 | **0.021** |
| | $S_\alpha\uparrow$ | 0.854 | 0.858 | 0.847 | 0.897 | 0.873 | 0.882 | 0.905 | 0.899 | **0.927** |

without depth cues, where we replace the original depth branch in DSNet with an additional RGB branch and mark it as 'w/o D'. Comparing the $6^{th}$ row and the $8^{th}$ row of Tab. II, we observe that the model without depth branch has a significant drop. Besides, in Tab. III, it is clear that our method (*i.e.*, DSNet (w/o D)) is obviously superior to the state-of-the-art RGB SOD methods. With the addition of depth cues, our method (*i.e.*, DSNet (w/ D)) achieves further performance improvement.

These experimental results adequately demonstrate the effects of depth cues in SOD. As a complementary modality of RGB images, depth maps provide sufficient spatial structure information, which is beneficial for locating and highlighting salient objects.

*2) Benefits of Different Modules:* To further evaluate the contribution of different modules in the proposed DSNet, we conduct various experiments from quantitative and qualitative perspectives, as shown in Tab. II and Fig. 8. In the $2^{nd}$ row of Tab. II, we test the model without additional modules (*i.e.*, CGCM, CAM, and BGPM), and obtain comparable results to the state-of-the-art methods. To verify the effectiveness of CGCM, we follow the single variable principle, and only add CGCM while leaving other settings as default, as shown in the $3^{rd}$ row of Tab. II. We find that the model with CGCM gains performance improvements on both NJU2K test set and LFSD dataset. The corresponding visualization result is shown in Fig. 8 (f), which pays more attention to global information. Besides, we further add CAM and BGPM, and the results on various evaluation metrics show that they effectively improve the performance shown in the $4^{th}$ and $5^{th}$ rows of Tab. II. Furthermore, it can be found from Fig. 8 (g) and (h) that the predicted saliency maps can effectively suppress backgrounds and detect salient objects. When adding the complete DSM (the $8^{th}$ row of Tab. II), including CAM and BGPM, our model performs best, which not only accurately and completely highlights salient objects, but also depicts sharper edges clearly, as shown in Fig. 8 (i).

Overall, according to Tab. II and Fig. 8, we can make a conclusion that each module in our model is beneficial for improving the performance of the proposed DSNet.

*3) Advantages of Edge Supervision:* To analyze the advantages of edge supervision in training phase, we conduct corresponding experiments. Naturally, we remove all levels of edge supervision (*i.e.*, $\mathcal{L}_e$), which retrains the variant model only under object supervision. As shown in the $7^{th}$ and $8^{th}$ rows of Tab. II, we find that the performance of DSNet on different evaluation metrics is better than the variant model without edge supervision. Meanwhile, we also show the visualization results under different supervision strategies in Fig. 9. Specifically, the $1^{st}$ row is the edge saliency maps of different levels under edge supervision, which clearly describes the boundaries of salient objects. The $2^{nd}$ and $3^{rd}$ rows are the object saliency maps of different levels with/without edge supervision, respectively. Obviously, it can be found that the former shows clearer and sharper edges than the latter under the supervision of edge information. Therefore, we can say that edge supervision is helpful to acquire saliency maps with clear boundaries.

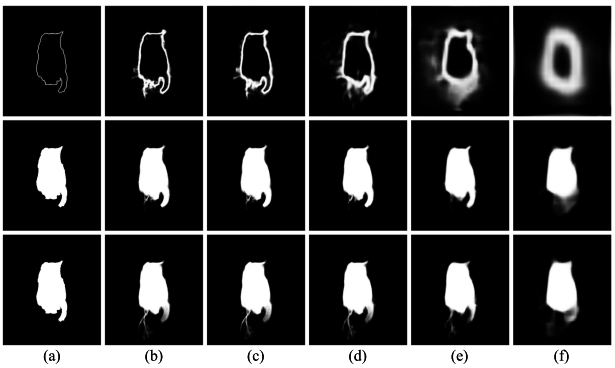


Fig. 9. Multi-level visualization results under object supervision and edge supervision. (a) represents Ground Truth, (b) $\sim$ (f) represent object saliency maps and edge saliency maps at different levels. Note that the first two rows are generated by the model with edge supervision, and the last row is generated by the model without edge supervision.

### *D. Failure Case Analysis*

As aforementioned, we illustrate the effectiveness and advancement of the proposed DSNet through various quantitative and qualitative experiments. However, in some special situations, the proposed DSNet is still difficult to achieve the expected results. As shown in Fig. 10, we present six representative failure cases, which can be divided into three categories. In the first category, the proposed model falsely highlights background regions and cannot correctly segment salient objects, such as the bridge and the traffic sign, as shown in Fig. 10 (a) and (b). This is mainly because our model pays more attention to the global consistency of objects. In the second category, our model either misses the salient object or fails to detect it perfectly when dealing with occlusion scenes. In Fig. 10 (c) and (d), we show two typical examples, one is an owl occluded by a tree branch, and the other is a horse

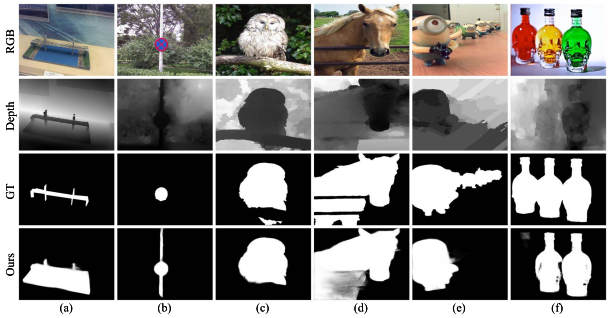


Fig. 10. Some typical failure cases of our model. (a)–(f) Represents three different common scenes, (a) and (b): global consistency, (c) and (d): object occlusion, (e) and (f): depth sensitivity.

occluded by fences. In this case, our model cannot accurately detect salient objects. The reason behind this maybe lies in that the occlusion destroys the integrity of salient objects and greatly interferes with the detection of complete objects. In the third category, it is difficult for the proposed DSNet to focus on the salient object far away from camera, such as the doll in Fig. 10 (e) and the bottle in Fig. 10 (f). We noticed that depth maps tend to concern more on the objects close to camera. This may lead to this phenomenon that the closer object is more likely to be a salient object. Generally speaking, how to further effectively and efficiently mine the complementary information between RGB images and depth maps is still a problem worthy of further exploration.

## V. Conclusion

In this paper, we propose a novel Dynamic Selective Network (DSNet) for RGB-D saliency detection. Our method can not only automatically select and fuse the cross-modal features, *i.e.*, RGB images and depth maps, but also autonomously optimize the cross-level and multi-scale hybrid features. Specifically, to mine the high-level semantic information more effectively, we employ a CGCM to highlight salient objects globally. Besides, in the CAM, we make full use of the attention mechanism and design a layer attention block to dynamically explore the cross-modal complementary information between RGB images and depth maps. Furthermore, we introduce a BGPM to better integrate the cross-level features and optimize the multi-scale features. Finally, the deeply supervision strategy with feedback mechanism ensures that the predicted saliency maps are with clearer objects and sharper edges. Extensive experiments on eight public RGB-D datasets show that the proposed DSNet achieves a comparable performance against 17 state-of-the-art methods in terms of five evaluation metrics.

## References

[1] P. Mukherjee and B. Lall, "Saliency and KAZE features assisted object segmentation," *Image Vis. Comput.*, vol. 61, pp. 82–97, May 2017.

[2] Y. Li, X. Hou, C. Koch, J. Rehg, and A. Yuille, "The secrets of salient object segmentation," in *Proc. CVPR*, Jun. 2014, pp. 4321–4328.

[3] Z. Ren, S. Gao, L.-T. Chia, and I. W.-H. Tsang, "Region-based saliency detection and its application in object recognition," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 24, no. 5, pp. 769–779, May 2014.

[4] A. Shokoufandeh, I. Marsic, and S. J. Dickinson, "View-based object recognition using saliency maps," *Image Vis. Comput.*, vol. 17, nos. 5–6, pp. 445–460, Apr. 1999.

[5] Z. Chi, H. Li, H. Lu, and M.-H. Yang, "Dual deep network for visual tracking," *IEEE Trans. Image Process.*, vol. 26, no. 4, pp. 2005–2015, May 2017.

[6] S. Hong, T. You, S. Kwak, and B. Han, "Online tracking by learning discriminative saliency map with convolutional neural network," in *Proc. Int. Conf. Mach. Learn.*, Jun. 2015, pp. 597–606.

[7] H. Wen, X. Zhou, Y. Sun, J. Zhang, and C. Yan, "Deep fusion based video saliency detection," *J. Vis. Commun. Image Represent.*, vol. 62, pp. 279–285, Jul. 2019.

[8] G.-P. Ji, K. Fu, Z. Wu, D.-P. Fan, J. Shen, and L. Shao, "Full-duplex strategy for video object segmentation," in *Proc. Int. Conf. Comput. Vis.*, 2021, pp. 4922–4933.

[9] C. Yang, L. Zhang, H. Lu, X. Ruan, and M.-H. Yang, "Saliency detection via graph-based manifold ranking," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2013, pp. 3166–3173.

[10] Y. Fang, J. Wang, M. Narwaria, P. Le Callet, and W. Lin, "Saliency detection for stereoscopic images," *IEEE Trans. Image Process.*, vol. 23, no. 6, pp. 2625–2636, Jun. 2014.

[11] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. Int. Conf. Learn. Represent.*, 2015, pp. 1–14.

[12] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 770–778.

[13] D.-P. Fan, W. Wang, M.-M. Cheng, and J. Shen, "Shifting more attention to video salient object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2019, pp. 8554–8564.

[14] W. Wang, J. Shen, J. Xie, M.-M. Cheng, H. Ling, and A. Borji, "Revisiting video saliency prediction in the deep learning era," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 1, pp. 220–237, Jan. 2021.

[15] N. Zhang, J. Han, N. Liu, and L. Shao, "Summarize and search: Learning consensus-aware dynamic convolution for co-saliency detection," in *Proc. Int. Conf. Comput. Vis.*, Oct. 2021, pp. 4167–4176.

[16] N. Liu, W. Zhao, D. Zhang, J. Han, and L. Shao, "Light field saliency detection with dual local graph learning and reciprocative guidance," in *Proc. Int. Conf. Comput. Vis.*, Jun. 2021, pp. 4712–4721.

[17] S.-H. Gao, Y.-Q. Tan, M.-M. Cheng, C. Lu, Y. Chen, and S. Yan, "Highly efficient salient object detection with 100k parameters," in *Proc. ECCV*, Aug. 2020, pp. 702–721.

[18] X. Zhao, Y. Pang, L. Zhang, H. Lu, and L. Zhang, "Suppress and balance: A simple gated network for salient object detection," in *Proc. ECCV*, Aug. 2020, pp. 35–51.

[19] J. Zhang, J. Xie, and N. Barnes, "Learning noise-aware encoder-decoder from noisy labels by alternating back-propagation for saliency detection," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Aug. 2020, pp. 349–366.

[20] Q. Chen, Z. Liu, Y. Zhang, K. Fu, Q. Zhao, and H. Du, "RGB-D salient object detection via 3D con, vol.utional neural networks," in *Proc. Assoc. Adv. Artif. Intell.*, 2021, vol. 35, no. 2, pp. 1063–1071.

[21] Y. Piao, Z. Rong, M. Zhang, W. Ren, and H. Lu, "A2dele: Adaptive and attentive depth distiller for efficient RGB-D salient object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 9060–9069.

[22] C. Zhu, G. Li, W. Wang, and R. Wang, "An innovative salient object detection using center-dark channel prior," in *Proc. ICCV Workshop*, Oct. 2017, pp. 1509–1515.

[23] 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.

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

[25] H. Song, Z. Liu, H. Du, G. Sun, O. Le Meur, and T. Ren, "Depth-aware salient object detection and segmentation via multiscale discriminative saliency fusion and bootstrap learning," *IEEE Trans. Image Process.*, vol. 26, no. 9, pp. 4204–4216, Sep. 2017.

[26] 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.*, Aug. 2020, pp. 646–662.

[27] 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.*, Aug. 2020, pp. 275–292.

[28] 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.*, early access, Apr. 16, 2021, doi: 10.1109/TPAMI.2021.3073689.

[29] 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.

[30] X. Zhou, G. Li, C. Gong, Z. Liu, and J. Zhang, "Attention-guided RGBD saliency detection using appearance information," *Image Vis. Comput.*, vol. 95, Mar. 2020, Art. no. 103888.

[31] W. Zhang, G.-P. Ji, Z. Wang, K. Fu, and Q. Zhao, "Depth quality-inspired feature manipulation for efficient RGB-D salient object detection," in *Proc. ACM MM*, 2021, pp. 731–740.

[32] L. Itti, C. Koch, and E. Niebur, "A model of saliency-based visual attention for rapid scene analysis," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 20, no. 11, pp. 1254–1259, Nov. 1998.

[33] M.-M. Cheng, N. Mitra, X. Huang, P. 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.

[34] G. Li and Y. Yu, "Visual saliency based on multiscale deep features," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 5455–5463.

[35] Z. Jiang and L. S. Davis, "Submodular salient region detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2013, pp. 2043–2050.

[36] J.-J. Liu, Q. Hou, M.-M. Cheng, J. Feng, and J. Jiang, "A simple pooling-based design for real-time salient object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2019, pp. 3917–3926.

[37] 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.

[38] B. Wang, Q. Chen, M. Zhou, Z. Zhang, X. Jin, and K. Gai, "Progressive feature polishing network for salient object detection," in *Proc. Assoc. Adv. Artif. Intell.*, 2020, pp. 12128–12135.

[39] A. Vaswani *et al.*, "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017, pp. 5998–6008.

[40] S. Woo, J. Park, J. Y. Lee, and I. S. Kweon, "CBAM: Convolutional block attention module," in *Proc. Eur. Conf. Comput. Vis.*, Sep. 2018, pp. 3–19.

[41] N. Liu, J. Han, and M. H. Yang, "PiCANet: Learning pixel-wise contextual attention for saliency detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Aug. 2018, pp. 3089–3098.

[42] M. Feng, H. Lu, and E. Ding, "Attentive feedback network for boundary-aware salient object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2019, pp. 1623–1632.

[43] T. Zhao and X. Wu, "Pyramid feature attention network for saliency detection," in *Proc. CVPR*, Jun. 2019, pp. 3085–3094.

[44] N. Liu, N. Zhang, and J. Han, "Learning selective self-mutual attention for RGB-D saliency detection," in *Proc. Comput. Vis. Pattern Recognit.*, Jun. 2020, pp. 13756–13765.

[45] C. Li, R. Cong, Y. Piao, Q. Xu, and C. C. Loy, "RGB-D salient object detection with cross-modality modulation and selection," in *Proc. Eur. Conf. Comput. Vis.*, Aug. 2020, pp. 225–241.

[46] H. Chen and Y. Li, "Progressively complementarity-aware fusion network for RGB-D salient object detection," in *Proc. CVPR*, Jun. 2018, pp. 3051–3060.

[47] K. Fu, D.-P. Fan, G.-P. Ji, and Q. Zhao, "JL-DCF: Joint learning and densely-cooperative fusion framework for RGB-D salient object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 3052–3062.

[48] Y. Piao, W. Ji, J. Li, M. Zhang, and H. Lu, "Depth-induced multi-scale recurrent attention network for saliency detection," in *Proc. IEEE ICCV*, Oct. 2019, pp. 7254–7263.

[49] Y. Pang, L. Zhang, X. Zhao, and H. Lu, "Hierarchical dynamic filtering network for RGB-D salient object detection," in *Proc. Eur. Conf. Comput. Vis.*, Aug. 2020, pp. 235–252.

[50] W. Ji, J. Li, M. Zhang, Y. Piao, and H. Lu, "Accurate RGB-D salient object detection via collaborative learning," in *Proc. Eur. Conf. Comput. Vis.*, Aug. 2020, pp. 52–69.

[51] 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. 3472–3481.

[52] N. Liu, N. Zhang, K. Wan, L. Shao, and J. Han, "Visual saliency transformer," in *Proc. Int. Conf. Comput. Vis.*, Oct. 2021, pp. 4722–4732.

[53] S. Gupta, R. Girshick, P. Arbeláez, and J. Malik, "Learning rich features from RGB-D images for object detection and segmentation," in *Proc. Eur. Conf. Comput. Vis.*, Jul. 2014, pp. 345–360.

[54] Y. Cao, J. Xu, S. Lin, F. Wei, and H. Hu, "GCNet: Non-local networks meet squeeze-excitation networks and beyond," in *Proc. IEEE Int. Conf. Comput. Vis. Workshop*, Oct. 2019, pp. 1971–1980.

[55] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, 1997.

[56] K. Cho *et al.*, "Learning phrase representations using RNN encoder-decoder for statistical machine translation," in *Proc. Conf. Empirical Methods Natural Lang. Process. (EMNLP)*, Oct. 2014, pp. 1724–1734.

[57] J. Han, H. Chen, N. Liu, C. Yan, and X. Li, "CNNs-based RGB-D saliency detection via cross-view transfer and multiview fusion," *IEEE Trans. Cybern.*, vol. 48, no. 11, pp. 3171–3183, Nov. 2017.

[58] 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.

[59] 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.

[60] 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 Conf. Comput. Vis. Pattern Recognit.*, Jun. 2019, pp. 3927–3936.

[61] J.-X. 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. Int. Conf. Comput. Vis.*, Oct. 2019, pp. 8779–8788.

[62] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Commun. ACM*, vol. 60, no. 6, pp. 84–90, May 2017.

[63] K. He, X. Zhang, S. Ren, and J. Sun, "Delving deep into rectifiers: Surpassing human-level performance on ImageNet classification," in *Proc. Int. Conf. Comput. Vis.*, Dec. 2015, pp. 1026–1034.

[64] D. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. Int. Conf. Learn. Represent.*, 2015, pp. 1–15.

[65] 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.

[66] H. Peng, B. Li, W. Xiong, W. Hu, and R. Ji, "RGBD salient object detection: A benchmark and algorithms," in *Proc. ECCV*, Sep. 2014, pp. 92–109.

[67] Y. Niu, Y. Geng, X. Li, and F. Liu, "Leveraging stereopsis for saliency analysis," in *Proc. IEEE Int. Conf. CVPR*, Jun. 2012, pp. 454–461.

[68] 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.

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

[70] 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 (CVPR)*, Oct. 2017, pp. 3008–3014.

[71] 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.

[72] N. Liu, N. Zhang, L. Shao, and J. Han, "Learning selective mutual attention and contrast for RGB-D saliency detection," 2020, *arXiv:2010.05537*.

[73] 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.

[74] 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. IJCAI*, Jul. 2018, pp. 698–704.

[75] D.-P. Fan, M.-M. Cheng, Y. Liu, T. Li, and A. Borji, "Structure-measure: A new way to evaluate foreground maps," in *Proc. ICCV*, Oct. 2017, pp. 4548–4557.

[76] R. Margolin, L. Zelnik-Manor, and A. Tal, "How to evaluate foreground maps," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 248–255.

[77] X. Zhang, T. Wang, J. Qi, H. Lu, and G. Wang, "Progressive attention guided recurrent network for salient object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 714–722.

[78] Z. Deng *et al.*, "$R^3$Net: Recurrent residual refinement network for saliency setection," in *Proc. Int. Joint Conf. Artif. Intell.*, Jul. 2018, pp. 684–690.

[79] Z. Wu, L. Su, and Q. Huang, "Cascaded partial decoder for fast and accurate salient object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2019, pp. 3907–3916.

**Hongfa Wen** received the B.S. and M.S. degrees from Hangzhou Dianzi University, Hangzhou, China, in 2017 and 2020, respectively, where he is currently pursuing the Ph.D. degree with the School of Automation. His research interests include computer vision, pattern recognition, deep learning, and saliency detection.

**Chenggang Yan** received the B.S. degree in computer science from Shandong University in 2008 and the Ph.D. degree in computer science from the Institute of Computing Technology, Chinese Academy of Sciences, in 2013. He is currently a Professor with Hangzhou Dianzi University. Before that, he was an Assistant Research Fellow with Tsinghua University. He has authored or coauthored over 30 referenced journal and conference papers. His research interests include intelligent information processing, machine learning, image processing, computational biology, and computational photography. As the coauthor, he received the Best Paper Candidate from the International Conference on Multimedia and Expo 2011, and the Best Paper Award from the International Conference on Game Theory for Networks 2014 and the SPIE/COS Photonics Asia Conference 2014.

**Xiaofei Zhou** received the Ph.D. degree from Shanghai University, Shanghai, China, in 2018. He is currently a Lecturer with the School of Automation, Hangzhou Dianzi University, Hangzhou, China. His research interests include saliency detection and video segmentation.

**Runmin Cong** received the Ph.D. degree in information and communication engineering from Tianjin University, Tianjin, China, in June 2019. He is currently an Associate Professor with the Institute of Information Science, Beijing Jiaotong University, Beijing, China. He was a Visiting Student/Staff with Nanyang Technological University (NTU), Singapore, and the City University of Hong Kong (CityU), Hong Kong. His research interests include computer vision, multimedia processing and understanding, visual attention perception and saliency computation, remote sensing image interpretation and analysis, and visual content enhancement in an open environment. He was a recipient of the Young Elite Scientist Sponsorship Program by the China Association for Science and Technology, the Beijing Nova Program, the Hong Kong Scholars Program, the IEEE ICME Best Student Paper Runner-Up Award, the ACM SIGWEB China Rising Star Award, the First Prize for Scientific and Technological Progress Award of Tianjin Municipality, the Excellent Doctoral Dissertation Award from the China Society of Image and Graphics (CSIG), and the Excellent Scientific Paper Award for Beijing Youth. He also serves as an Associate Editor for the *Signal, Image and Video Processing*; a Guest Editor for the IEEE JOURNAL OF OCEANIC ENGINEERING, *Signal Processing: Image Communication*, and *Multimedia Tools and Applications*; and the Area Chair/PC Member for NeurIPS, CVPR, ICML, ICCV, ACM MM, AAAI, IJCAI, ACM MM, and ICME.

**Yaoqi Sun** received the B.S. degree from the Zhejiang University of Science and Technology in 2016 and the M.S. degree in electronic and communication engineering from Hangzhou Dianzi University in 2021. He is currently an Assistant Research Fellow with the Department of Automation, Hangzhou Dianzi University. His research interests include intelligent information processing, machine learning, and pattern recognition.

**Bolun Zheng** received the B.S. and Ph.D. degrees in electronic information technology and instrument from Zhejiang University in 2014 and 2019, respectively. He is currently a Lecturer with Hangzhou Dianzi University. His research interests are computer vision, pattern recognition, image processing, and embedded parallel computing.

**Jiyong Zhang** (Member, IEEE) received the B.S. and M.S. degrees in computer science from Tsinghua University in 1999 and 2001, respectively, and the Ph.D. degree in computer sciences from the Swiss Federal Institute of Technology (EPFL), Lausanne, in 2008. He is currently a Distinguished Professor with Hangzhou Dianzi University. His research interests include intelligent information processing, machine learning, data sciences, and recommender systems.

**Yongjun Bao** received the B.S. degree from Southeast University, Nanjing, China, and the M.S. degree from Peking University in 2007. He currently serves as the Vice President for JD Group, the Standing Member for JD Retail Technology Committee, the Chairperson for JD Retail Data Algorithm Channel Committee, and the Head for the Data and Intelligence Department, Technology and Data Center, JD Retail. His areas of expertise include computer vision and natural language processing.

**Guiguang Ding** is currently a Distinguished Researcher with the School of Software, Tsinghua University; a Ph.D. Supervisor; an Associate Dean of the School of Software, Tsinghua University; and the Deputy Director of the National Research Center for Information Science and Technology. His research interests mainly focus on visual perception, theory and method of efficient retrieval and weak supervised learning, neural network compression of vision task under edge computing and power limited scenes, visual computing systems, and platform developing. He was the Winner of the National Science Fund for Distinguished Young Scholars.