This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCSVT.2022.3144852, IEEE Transactions on Circuits and Systems for Video Technology

Bi-directional Progressive Guidance Network for RGB-D Salient Object Detection

Yang Yang, Qi Qin, Yongjiang Luo, Yi Liu*, Qiang Zhang*, and Jungong Han

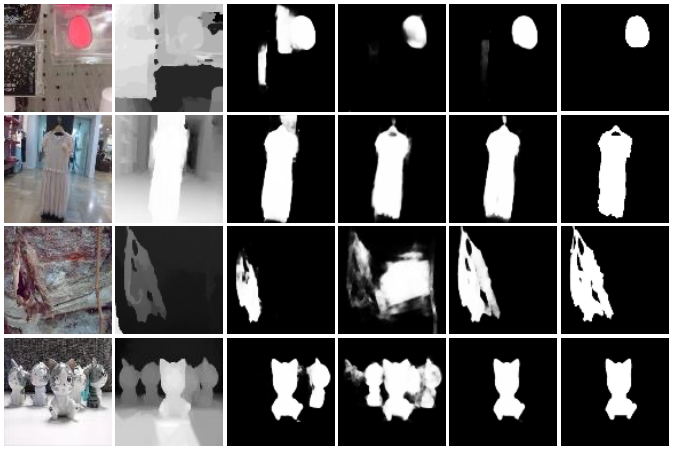
***Abstract*—Most existing RGB-D salient detection models pay more attention to the quality of the depth images, while in some special cases, the quality of RGB images may even have greater impacts on saliency detection, which has long been ignored and underestimated. To address this problem, in this paper, we present a Bi-directional Progressive Guidance Network (BPGNet) for RGB-D salient object detection, where the qualities of both RGB and depth images are invovled. Since it is usually difficult to determine which modality data have low quality in advance, a bi-directional framework based on progressive guidance (PG) strategy is employed to extract and enhance the unimodal features with the aid of another modality data via the alternative interactions between the saliency prediction results and the extracted features from the multi-modality input data. Specifically, the proposed PG strategy is achieved by using the proposed Global Context Awareness (GCA), Auxiliary Feature Extraction (AFE) and Cross-modality Feature Enhancement (CFE) modules. Benefiting from the proposed PG strategy, the disturbing information within the input RGB and depth images can be well suppressed, while the discriminative information within the input images gets enhanced. On top of that, a Fusion Prediction Module (FPM) is further designed to adaptively select those features with higher discriminability as well as enhancing the common information for the final saliency prediction. Experimental results demonstrate that our proposed model is comparable to those of state-of-the-art RGB-D SOD models.**

***Index Terms*—RGB-D images, salient object detection, image quality, bi-directional progressive guidance.**

I. Introduction

THE purpose of salient object detection (SOD) is to automatically detect the most interesting regions of the human eye in an image, and separate them from the background [1], [2]. It has attracted widespread attention and played an important role in many image and video processing tasks, such as quality assessment [3], object recognition [4], semantic segmentation [5], visual tracking [6], video detection [7] and image compression [8].

Many previous works for SOD mainly deal with RGB images and only leverage the appearance cues [9]–[12]. Although these RGB SOD methods have achieved remarkable progress, they still underperform in some complex scenes, e.g., objects and backgrounds sharing similar appearances, disturbing backgrounds, etc. Alternatively, depth images have been proven beneficial for accurate saliency prediction in the above challenging cases. This is owing to rich three-dimensional layout and spatial cues provided by depth images, which can attack light and color changes. Comprehensive consideration of the complementary information from RGB-D images may achieve more desirable SOD results for the same scene.



(a) (b) (c) (d) (e) (f)

Fig. 1. Illustrations of the prediction results on some images with different qualities. (a) RGB images; (b) Depth images; (c) Saliency maps deduced by cross-modality features fused by using simple addition operation; (d) D3Net [13]; (e) BPGNet (OURS); (f) GT.

Unfortunately, depth images are sometimes inaccurate and thus lead to poor fusion results [13]. Similarly, for complex scenarios mentioned above, low discriminative cues from RGB images would also contaminate the results of SOD as noise cues contained in inaccurate depth images do. Most existing works consider inaccurate depth images as low-quality images. To be coherent with them, in this paper, those RGB images with complex scenarios and inaccurate depth images are simultaneously defined as low-quality images.

Earlier RGB-D SOD methods [14]–[18] usually begin with two independent streams to extract the unimodal features from the RGB images and the depth images, respectively. Then, these unimodal features are combined by using some fusion rules for the final saliency detection. However, these methods usually ignore the qualities of the input images, which may lead to some problems. For example, as shown in Fig. 1(c), where two modality data are fused via an addition operation,

Yang Yang, Qi Qin, Qiang Zhang are with Center for Complex Systems, School of Mechano-Electronic Engineering, Xidian University, Xi'an, Shaanxi 710071, China. Email: yang@stu.xidian.edu.cn, qinqi67190304@stu.xidian.edu.cn and qzhang@xidian.edu.cn.

Yongjiang Luo is with the School of Electronic Engineering, Xidian University, Xi'an Shaanxi 710071, China. Email: yjluo@mail.xidian.edu.cn.

Yi Liu is with School of Computer Science and Artificial Intelligence, and Aliyun School of Big Data, Changzhou University, Changzhou, Jiangsu 213164, China. Email: liuyi0089@gmail.com.

Jungong Han is with Computer Science Department, Aberystwyth University, SY23 3FL, UK. Email: jungonghan77@gmail.com.

*Corresponding authors: Qiang Zhang and Yi Liu.

the features extracted from those low-quality input images or local regions may reduce the discriminability of the fused features, thus degrading the final SOD results.

Some recent works have started paying attention to the qualities of input images for the saliency detection [13], [19]–[21]. For example, Fan et al. [13] presented a depth depurator unit to reduce the impacts of low-quality depth images. Due to the attention on the qualities of depth images, these works may work well for those scenes with low-quality depth images, which can be verified in the 1^{st} and 2^{nd} rows of Fig. 1(d). While they may fail for those scenes where the input RGB images have low qualities, e.g., the foregrounds and backgrounds have similar spatial appearances or the backgrounds are complex, which can be shown in the 3^{rd} and 4^{th} rows of Fig. 1(d).

In order to address such issues mentioned above, in this paper, we will present a new Bi-directional Progressive Guidance Network (BPGNet) for RGB-D salient object detection, where the disturbing cues in the RGB images and those in the depth images will be simultaneously taken into consideration via a proposed progressive guidance (PG) strategy.

The key idea behind the PG strategy is as follows. Given the multi-modality input images, one modality data may relatively have better qualities and the other modality data may have lower qualities. High-quality modality data may be more beneficial for saliency than the low-quality modality data. Thus, compared with the saliency map deduced from the low-quality modality data, the saliency map deduced from the high-quality modality data may be more reliable, and will be further conducive for the primitive feature extraction from the low-quality modality data. With the aid of the cross-modal useful information from the low-quality modality data, the features from high-quality modality data may be further refined. This provides us with a feasible way to suppress the disturbing information contained in the low-quality modality data as well as those disturbing information within in the high-quality modality data via the alternative interactions of the saliency maps from one modality data and those features from another modality data.

To cater to the above analyses, a Global Context Awareness (GCA) module is first designed to learn global contextual semantic information for generating reliable coarse saliency prediction results, which will be treated as spatial prior guidance. On top of that, an Auxiliary Feature Extraction (AFE) module and a Cross-modality Feature Enhancement (CFE) module are designed to achieve the alternative interaction between the prediction results from one modality data and the extracted features from another modality data, thus constructing a progressive guidance relationship between the prediction results and the extracted features. More specifically, the coarse saliency results are first obtained from one supposed high-quality modality data (e.g., RGB images). Then the AFE module is employed to guide the extraction of informative unimodal features from the other supposed low-quality modality data (e.g., depth images), which will help to refine the coarse saliency results. The initial unimodal features from the high-quality modality data will be further enhanced with the guidance of the extracted features and their corresponding refined saliency detection results from the low-quality modality data via the proposed CFE module, thus obtaining the enhanced unimodal features that possess more informativeness.

Actually, for real-life scenarios, we could not determine which modality has high or low quality in advance. Considering that, a bi-directional network framework based on the proposed PG strategy is presented for the extraction and enhancement of unimodal features. Specifically, a RGB-Guidance-Depth Network (RGDNet) is proposed to achieve the extraction and enhancement of RGB features with the aid of the input depth images, where the input RGB images are supposed to have better qualities than the input depth images. Similarly, a Depth-Guidance-RGB Network (DGRNet) is designed, where the depth inputs are supposed as high-quality data.

Following that, a Fusion and Prediction Module (FPM) is further designed to combine the enhanced RGB features and the enhanced depth features, which are learned from RGDNet and DGRNet, respectively, for the final saliency prediction. With FPM, those features with higher discriminability from RGDNet or DGRNet will be adaptively selected for the final saliency prediction by using a weighted selection way. As well, those features that are co-captured by RGDNet and DGRNet will also be further enhanced via a multiplication way.

In summary, the main contributions of this paper are summarized as follows:

(1) We introduce a Bi-directional Progressive Guidance Network (BPGNet) for RGB-D SOD, which simultaneously considers the disturbing cues from the RGB images and those from the depth images. Experimental results on six RGB-D saliency detection benchmark datasets demonstrate the effectiveness of our BPGNet.

(2) We design a new PG strategy, which progressively and interactively suppresses the disturbing cues within the multi-modality input images, and is achieved by three proposed modules, i.e., GCA, AFE and CFE.

(3) We propose a FPM module to combine the enhanced RGB and depth features for the final saliency prediction, in which the common information and complementary information within/between multi-modality input images are simultaneously considered.

The rest of this paper is organized as follows. In Sec. II, we briefly review some related work. In Sec. III, we introduce the proposed BPGNet for RGB-D SOD in detail. Extensive experiments are conducted in Sec. IV. Finally, some conclusions are given in Sec. V.

II. Related Work

A. RGB Salient Object Detection

Earlier salient object detection methods mostly predict saliency regions by using some hand-crafted features [22]–[24]. A thorough understanding on such methods can be seen in [25]. Recently, Convolutional Neural Networks (CNNs) have been successfully applied for the task of SOD because of their powerful representation ability, and have achieved substantial improvements [26]–[34]. For typical examples,

Wang *et al.* [30] aimed to infer salient objects from the fixation map within a unified neural network, which offered a deep insight into the confluence between fixation prediction and salient object detection. Zhang *et al.* [31] focused on salient object detection in optical remote sensing images and proposed an end-to-end encoder-decoder network, which showed strong robustness to the changes of scenarios. Wang *et al.* [33] proposed an iterative top-down and bottom-up inference network for salient object detection, which learned top-down, coarse-to-fine saliency inference and bottom-up, shallow-to-deep saliency inference in an iterative and end-to-end manner. Wang *et al.* [34] proposed a deep sub-region network for salient object detection, which simultaneously aggregated multi-scale salient context information of multiple sub-regions and the global context information from the whole image. More recently, some visual attention models have also been extensively explored for SOD [35], [36]. For example, Zhang *et al.* [35] designed an attention-guided network for SOD by jointly employing spatial attention and channel-wise attention. Hu *et al.* [36] presented a saliency detection network based on the spatial attenuation context derived from the intra-modal attention mechanism. These RGB saliency models have achieved appealing results for those real-life scenarios with good spatial appearances.

B. RGB-D Salient Object Detection

For those cases with poor spatial appearances, depth images of the scenes have been involved to solve the problem of SOD [37]–[41]. These methods can be roughly divided into three categories: input fusion, feature fusion and result fusion.

The input fusion based methods usually directly feed the RGB and depth inputs into the networks [38], [42]–[44]. For example, Song *et al.* [42] used the 4-channel data to compute multi-scale saliency maps. Similarly, Liu *et al.* [43] proposed a single stream recurrent CNN with four-channel RGB-D data as inputs to infer salient objects. Recently, Wang *et al.* [38] reconstructed the RGB images and depth images into some new 3-channel input data at the channel-level for saliency detection.

Different from input fusion, the result fusion based RGB-D SOD methods usually obtain the RGB saliency results and depth saliency results separately, which were then integrated by using some fusion strategies [17], [39], [45]. For example, in [45], the final saliency maps were achieved by multiplying the prediction results from the two unimodal saliency detection streams. Differently, Ding *et al.* [39] designed a saliency fusion network, which integrated the color saliency maps with the depth saliency maps into the final saliency maps.

As a better choice, feature fusion based RGB-D SOD models have attracted more attention in recent years [16], [46], [47]. These methods first use a two-stream CNN to extract the unimodal features from RGB images and depth images, respectively, which were then fused to learn more primitive features for further saliency prediction. For example, Han *et al.* [16] learned cross-modality information through some fully-connected layers to infer the final saliency map. Lately, some advanced multi-modality feature fusion modules have been designed for multi-modality SOD tasks [48]–[53]. Li *et al.* [52] proposed a novel Cross-Modal Weighting (CMW) strategy to encourage comprehensive interactions between RGB and depth channels in the process of multi-modal feature fusion. Differently, in [40], more interactions between RGB and depth information were performed on the process of feature extraction for better grabbing the potentially useful information.

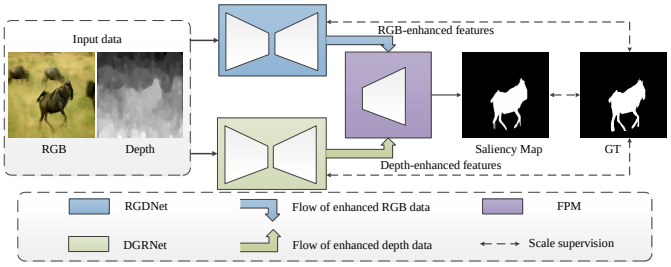


Fig. 2. Overall architecture of the proposed BPGNet, including three parts, i.e., RGDNet, DGRNet and FPM. RGDNet and DGRNet are used to achieve extract and enhance unimodal features, while FPM is used to combine the enhanced unimodal features for final saliency prediction.

Despite some improvements, most feature fusion based RGB-D SOD models mentioned above mainly focus on capturing the complementary information within the multi-modality input images, while ignoring the impacts of image qualities on the representation ability of fused features, thus degrading the subsequent saliency detection performance. Recently, some studies have been carried out on the disturbing problem caused by the low-quality images [13], [19]–[21], [54]–[57]. For example, Zhao *et al.* [19] designed a contrast enhancement module with contrast prior information to enhance the quality of depth images, thus boosting the saliency detection performance. Chen *et al.* [20] introduced a two-phase depth estimation approach to produce more trustworthy depth images. Rather than directly enhancing low-quality depth images, Fan *et al.* [13] proposed a depth depurator unit to reduce the impact of low-quality depth images on the saliency detection performance at the result-level. Bai *et al.* [54] and Li *et al.* [55] employed RGB features to filter distractors in depth features prior to exploiting cross-modal complementarity. Chen *et al.* [56] modeled a task-orientated depth potentiality perception module to weaken the contamination from unreliable depth information. Gao *et al.* [57] used the content-based spatial attention to select the important response of intra-modal information.

Different from the existing methods that mainly focus on the impacts of low-quality depth images on the saliency detection results, we simultaneously consider the disturbing information within the RGB images and the depth images by constructing a bi-directional progressive guidance network.

III. Proposed Method

As shown in Fig. 2, the proposed BPGNet contains three parts: RGB-Guidance-Depth Network (RGDNet), Depth-Guidance-RGB Network (DGRNet) and Fusion Prediction Module (FPM). RGDNet and DGRNet share the same structures but with different parameters. Especially, RGDNet first extracts the unimodal RGB features from the input RGB

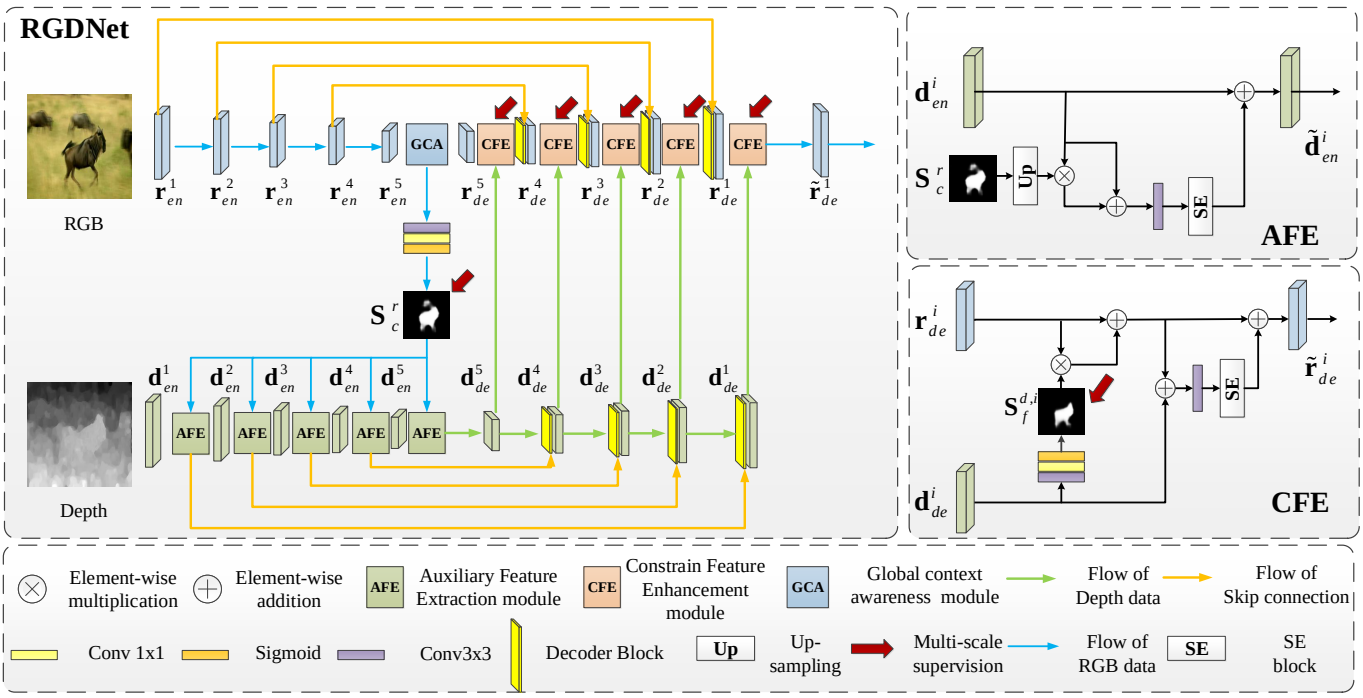


Fig. 3. Architecture of our proposed RGDNet, which contains three key stages: coarse saliency map generation, feature extraction and feature enhancement. First, GCA is used to capture global contextual information from the deepest level of encoded RGB features to obtain a coarse saliency map, which is fed into AFE as the guidance to help extract depth features. Finally, the extracted depth features are transferred to CFE for the enhancement of RGB features. For better visualization, we ignore the convolutional layer in the skip connection for compressing channels.

images and further enhances the extracted RGB features with the aid of the depth images via a progressive guidance strategy. Similarly, DGRNet first extracts the unimodal depth features from the depth images and then enhances the extracted depth features with the aid of RGB input. FPM is used to capture the common information and select more discriminative information between the enhanced RGB features and the enhanced depth features, achieving the fused features for the final saliency prediction. Meanwhile, some auxiliary supervisions are also performed on each unit for better saliency detection results.

A. RGDNet / DGRNet

In this subsection, we will take RGDNet as an example to introduce the unimodal feature extraction and enhancement process in detail. As shown in Fig. 3, RGDNet is a two-stream network, including a primary branch for the feature extraction and enhancement of RGB images and an auxiliary branch for the feature extraction of depth images. In addition, each branch in the two-stream network contains an encoder subnetwork and a decoder subnetwork.

1) ***Backbone network****:* For fair comparisons, we also employ the pretrained VGG-16 as the backbone network [58] of our encoder in RGDNet. Furthermore, we modify the VGG-16 for our saliency detection task by dropping the last pooling layer and three fully-connected layers from the original VGG-16. As well, for convenience, the single-channel depth image is first transformed into three-channel ones with duplication and concatenation as inputs for the encoder of the auxiliary depth branch. On top of that, we adopt a symmetric network structure in the decoder of RGDNet, where the up-sampling operations are implemented by the bilinear interpolation. Outputs of the five convolutional blocks in the RGB encoder are denoted as $\{\mathbf{r}_{en}^i | i = 1, 2, 3, 4, 5\}$, which have $\{1, 1/2, 1/4, 1/8, 1/16\}$ of the input image resolution, respectively. Similarly, the outputs of the five convolutional blocks in the depth encoder are denoted as $\{\mathbf{d}_{en}^i | i = 1, 2, 3, 4, 5\}$, which have the same resolutions as the corresponding levels of RGB features. The outputs from the RGB decoder and depth decoder are denoted as $\{\mathbf{r}_{de}^i | i = 1, 2, 3, 4, 5\}$ and $\{\mathbf{d}_{de}^i | i = 1, 2, 3, 4, 5\}$, respectively.

2) ***Progressive Guidance Strategy****:* We apply a Progressive Guidance (PG) strategy to further enhance of the discrimination of extracted RGB features with the proposed Global Context Awareness (GCA), Auxiliary Feature Extraction (AFE) and Cross-modality Feature Enhancement(CFE) mdoules. The proposed progressive guidance strategy focuses on how to reduce the disturbing information within the input data and enhance the unimodal feature representation ability, which will be illustrated in detail as follows.

The basic idea of the proposed progressive guidance strategy is based on the attention mechanism, which has been applied in many computer vision tasks, including image caption [59], semantic segmentation [60] and translation [61]. Concretely, the coarse saliency detection results from one modality data are employed to guide the feature extraction of the other modality data by providing some spatial position priors. Specifically, in RGDNet, the RGB input images are supposed to have higher qualities than the depth images, and the corresponding saliency results from the RGB images are supposed to be more reliable. These coarse saliency detection results will be used to guide

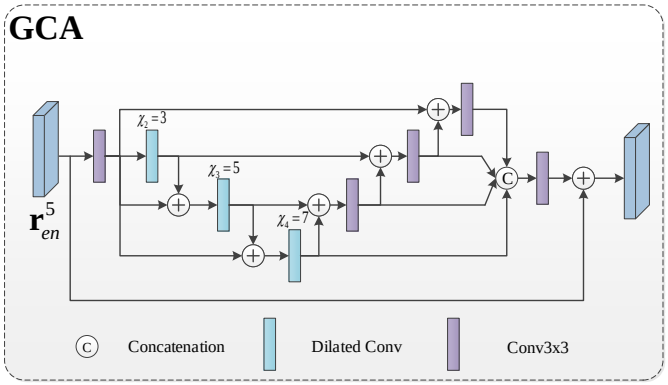


Fig. 4. Architecture of our proposed GCA module. First, some convolutional layers with different dilation rates are employed to extract multi-scale features. Then, some bottom–up and top–down pathways as well as some skip connections are employed to achieve the interactions among multi-scale features. Finally, the original information is added to the aggregated multi-scale features for obtaining the final output features.

the extraction of depth features, which will further refine the coarse saliency detection results. The refined coarse results can reversely guide the enhancement of RGB features. In detail, the process of the proposed progressive guidance strategy is described as follows.

Step 1: **Predict coarse saliency maps from the input RGB images.** Considering that the deepest level of features from the VGG-16 contain rich semantic information, we adopt the 5-th level of RGB features, i.e., $\mathbf{r}_{en}^5$, to produce the coarse saliency map. Besides, we introduce a GCA module to enlarge the receptive field of $\mathbf{r}_{en}^5$ to improve the reliability of coarse saliency map, which can be formulated as:

$$\mathbf{S}_c^r = \mathrm{Sig}\left(\mathrm{Conv}_1\left(\mathrm{Conv}_3\left(\mathrm{GCA}(\mathbf{r}_{en}^5), \theta_r\right), \gamma_r\right)\right), \quad (1)$$

where $\mathrm{Sig}(\cdot)$ denotes the Sigmoid activation function. $\mathrm{Conv}_1(\cdot, \gamma_r)$ denotes a 1×1 convolutional layer with its parameters γ_r. $\mathrm{Conv}_3(\cdot, \theta_\mathrm{r})$ is a 3×3 convolutional layer with parameters θ_r. $\mathrm{GCA}(\cdot)$ refers to the Global Context Awareness module, which will be described later.

Step 2: **Extract useful depth features with the guidance of coarse saliency maps.** As shown in Fig. 3, we append an AFE module on each block in the encoder of the auxiliary branch to guide the depth feature extraction with the aid of the coarse saliency map $\mathbf{S}_c^r$, thus obtaining a new set of enhanced depth features $\tilde{\mathbf{d}}_{en}^i$ $(i = 1, 2, 3, 4, 5)$, i.e.,

$$\tilde{\mathbf{d}}_{en}^i = \mathrm{AFE}(\mathbf{d}_{en}^i, \mathbf{S}_c^r). \quad (2)$$

Here, $\mathrm{AFE}(\cdot)$ denotes the AFE module, which will be described in detail later.

With the enhanced depth features, another five levels of depth features $\mathbf{d}_{de}^i$ $(i = 1, 2, 3, 4, 5)$ are further generated from the decoder of the auxiliary depth branch by using some skip connection, convolution and up-sampling operations, i.e.,

$$\mathbf{d}_{de}^i = \begin{cases} \mathrm{DB}(\mathrm{Conv}_3(\tilde{\mathbf{d}}_{en}^i, \theta_d^i) \oplus \mathrm{Up}(\mathbf{d}_{de}^{i+1}), \eta_d^i), \\ \qquad\qquad\qquad\qquad i = 1, 2, 3, 4, \\ \mathrm{Conv}_3(\mathbf{d}_{en}^i, \theta_d^i), \qquad\qquad i = 5, \end{cases} \quad (3)$$

where $\oplus$ is the element-wise addition operation. $\mathrm{DB}(\cdot, \eta_d^i)$ represents a Decoder Block, which contains three successive 3×3 convolutional layers and η_d^i refers to their parameters.

Step3: **Enhance the RGB features via the extracted depth features.** As discussed earlier, there may still exist some disturbing information within the RGB images, although the input RGB images are supposed to have better qualities than those depth images in RGDNet. Considering that, as shown in Fig. 3, we append a CFE module on each block in the decoder of the primary RGB branch to refine the RGB features with the extracted depth features $\mathbf{d}_{de}^i$ $(i = 1, 2, 3, 4, 5)$ from the auxiliary depth branch, thus obtaining a new set of enhanced RGB features $\tilde{\mathbf{r}}_{de}^i$ $(i = 1, 2, 3, 4, 5)$. Mathematically, the process can be expressed by

$$\tilde{\mathbf{r}}_{de}^i = \mathrm{CFE}(\mathbf{r}_{de}^i, \mathbf{d}_{de}^i). \quad (4)$$

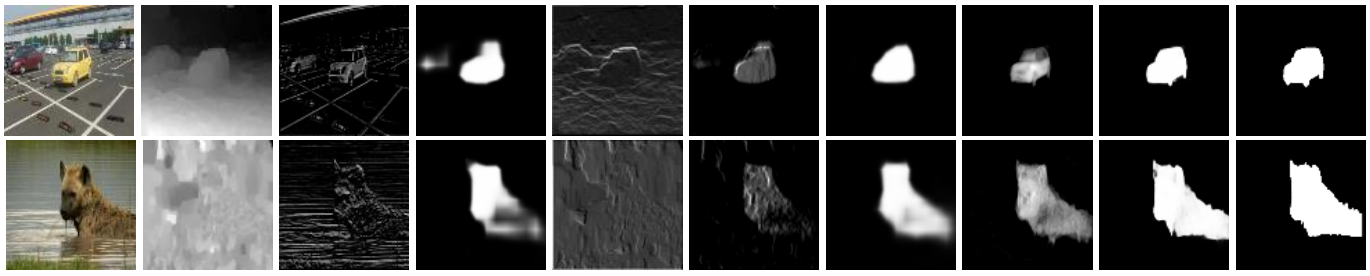
Here, $\mathrm{CFE}(\cdot)$ denotes the proposed CFE module, which will also be discussed later.

3) ***Global Context Awareness (GCA) Module****:* The structure of GCA is shown in Fig. 4. Given the deepest level of RGB features r_{en}^5, a GCA module is specially designed to further capture their multi-scale semantic information by using some dilated convolution operations with the purpose of increasing receptive fields while keeping the feature resolutions unchanged.

Specifically, a 3×3 convolutional layer is first applied to reduce the channel number of the input features r_{en}^5. Then, four parallel branches are used to obtain four scales of features. For the first branch (i.e., k=1), the channel-reduced features are directly seen as one scale of features with dilated rate χ_1 = 1. While, for the rest of branches (k=2, 3, 4), a dilated convolution block with a kernel size of 3 and different dilated rates (3/5/7) are applied. In addition, some bottom–up and top–down pathways as well as some skip connections are employed to achieve the interactions among multi-scale features. This is different from the traditional Atrous Spatial Pyramid Pooling (ASPP) module [62], where the interactions among different scales of features are neglected. On top of that, the outputs from the four branches are concatenated together, which is followed by a 3×3 convolution operation to well combine these multi-scale features. Finally, a residual connection is used to preserve the original information, obtaining the final multi-scale features.

4) ***Auxiliary Feature Extraction (AFE) Module****:* As shown in Fig. 3, the proposed AFE module achieves the enhancement of the extracted features (i.e., the depth features) from the axillary branch of RGDNet based on the spatial and channel attention mechanisms. Specifically, the coarse saliency map $\mathbf{S}_c^r$ deduced from one modality data (i.e., RGB images in RGDNet) is employed as some spatial priors to guide the feature extraction of another modality data (i.e., depth images in RGDNet). Subsequently, a channel-wise attention module via the SE block in [63] is employed to adaptively boost those discriminative channels of features, while suppressing those non-discriminative ones.

More specifically, for the i-th block of the encoder in the auxiliary branch, the coarse saliency map $\mathbf{S}_c^r$ is first interpolated via the bilinear upsampling operation and then



(a) (b) (c) (d) (e) (f) (g) (h) (i) (j)

Fig. 5. Visual examples of some intermediate results from RGDNet. From left to right: (a) RGB images; (b) Depth images; (c) Original RGB features $\mathbf{r}_{en}^{1}$; (d) Coarse saliency maps $\mathbf{S}_{c}^{r}$; (e) Original depth features $\mathbf{d}_{en}^{1}$; (f) Enhanced depth features $\tilde{\mathbf{d}}_{en}^{1}$ obtained by AFE; (g) Refined saliency maps $\mathbf{S}_{f}^{d,5}$; (h) Enhanced RGB features $\tilde{\mathbf{r}}_{de}^{1}$ obtained by CFE; (i) Refined saliency maps $\mathbf{S}_{R}^{1}$ deduced from (h); (j) GT.

is employed as the spatial weights to weigh the features $\mathbf{d}_{en}^{i}$ $(i = 1, 2, 3, 4, 5)$ with the element-wise multiplication, which are further element-wisely added with the original features to obtain the spatially enhanced features. On top of that, the SE block is used to achieve the discriminative features. Finally, to avoid the loss of the details and complementary information in the original depth features, some skip connection operations are employed to preserve original information in the proposed AFE module, thus obtaining enhanced depth features $\tilde{\mathbf{d}}_{en}^{i}$ $(i = 1, 2, 3, 4, 5)$. Mathematically, AFE can be expressed by

$$\tilde{\mathbf{d}}_{en}^{i} = \mathbf{d}_{en}^{i} \oplus \mathrm{SE}(\mathrm{Conv}_3(\mathrm{Up}(S_c^r) \otimes \mathbf{d}_{en}^{i} \oplus \mathbf{d}_{en}^{i}, \theta_{en}^{d,i}), \phi_{en}^{i}), \tag{5}$$

where $i = 1, 2, 3, 4, 5$ denotes the *i*-th depth encoder features. $\mathrm{SE}(*, \phi_{en}^{i})$ refers to the SE block [63] and ϕ_{en}^{i} denotes its parameters. $\otimes$ and $\mathrm{Up}(*)$ refer to the element-wise multiplication and the bilinear interpolation upsampling operations.

Benefiting from the spatial position prior information provided by the coarse saliency maps, more valuable depth information can be learned. As shown in Fig. 5, compared with the original depth features in Fig. 5(e), the depth features guided by the coarse results in Fig. 5(f) contain more valuable information about the salient objects, while suppressing more disturbing information within the background regions.

*5) **Cross-modality Feature Enhancement (CFE) Module**:* CFE is used to enhance the features (i.e., RGB features) from the primary branch of RGDNet, considering that the RGB features may also contain some disturbing information. Similar to that in AFE, spatial and channel attention mechanisms are also jointly employed in CFE to achieve the enhancement of the extracted features.

Specifically, the depth features $\tilde{\mathbf{d}}_{de}^{i}$ $(i = 1, 2, 3, 4, 5)$ from the decoder of the auxiliary branch may contain much more useful information with the aid of the AFE module in the encoder of the auxiliary branch. And these useful depth information may be complementary to those RGB features. Besides, the refined saliency map $\mathbf{S}_{f}^{d,i}$ $(i = 1, 2, 3, 4, 5)$ deduced from the corresponding level of depth decoder features $\mathbf{d}_{de}^{i}$ $(i = 1, 2, 3, 4, 5)$ may be more accurate than the coarse saliency map $\mathbf{S}_{c}^{r}$ deduced from the original RGB features $\mathbf{r}_{en}^{5}$. Considering that, we will simultaneously use the depth decoder features $\mathbf{d}_{de}^{i}$ $(i = 1, 2, 3, 4, 5)$ and their deduced saliency maps $\mathbf{S}_{f}^{d,i}$ $(i = 1, 2, 3, 4, 5)$ to enhance the RGB features in CFE.

Specifically, for the i-th level of decoder in the primary branch, the refined saliency map $\mathbf{S}_{f}^{d,i}$ $(i = 1, 2, 3, 4, 5)$ is first deduced from the corresponding level of depth decoder features $\mathbf{d}_{de}^{i}$ as follows,

$$\mathbf{S}_{f}^{d,i} = \mathrm{Sig}\left(\mathrm{Conv}_1\left(\mathrm{Conv}_3\left(\mathbf{d}_{de}^{i}, \theta_{de}^{d,i}\right), \gamma_{de}^{d,i}\right)\right). \tag{6}$$

Then the refined saliency map $\mathbf{S}_{f}^{d,i}$ is employed as the spatial prior to weigh the RGB features $\mathbf{r}_{de}^{i}$ from the current level of decoder, thus obtaining the initial spatially-enhanced RGB features $\tilde{\mathbf{r}}_{is}^{i}$, i.e.,

$$\tilde{\mathbf{r}}_{is}^{i} = \mathbf{S}_{f}^{d,i} \otimes \mathbf{r}_{de}^{i} \oplus \mathbf{r}_{de}^{i}. \tag{7}$$

On top of that, the depth features with the same level are combined with the initial spatially-enhanced features to further spatially enhance the RGB features, thus obtaining the final spatially-enhanced RGB features $\tilde{\mathbf{r}}_{fs}^{i}$, i.e.,

$$\tilde{\mathbf{r}}_{fs}^{i} = \tilde{\mathbf{r}}_{is}^{i} \oplus \mathbf{d}_{de}^{i}. \tag{8}$$

Finally, $\tilde{\mathbf{r}}_{fs}^{i}$ are further enhanced in a channel-wise way with a SE block [63]. A residual connection is additionally used to achieve the final enhance RGB features $\tilde{\mathbf{r}}_{de}^{i}$, i.e.,

$$\tilde{\mathbf{r}}_{de}^{i} = \mathrm{SE}(\mathrm{Conv}_3(\tilde{\mathbf{r}}_{fs}^{i}, \theta_{de}^{i}), \phi_{de}^{i}) \oplus \tilde{\mathbf{r}}_{is}^{i}. \tag{9}$$

To facilitate the network training, we also use deep supervision for enhanced RGB features, where a 1×1 Conv layer with the Sigmoid activation function is performed on each level of enhanced RGB features to generate corresponding saliency maps $\mathbf{S}_{R}^{i}$ $(i = 1, 2, 3, 4, 5)$, i.e.,

$$\mathbf{S}_{R}^{i} = \mathrm{Sig}\left(\mathrm{Conv}_1\left(\tilde{\mathbf{r}}_{de}^{i}, \gamma_{de}^{r,i}\right)\right). \tag{10}$$

As shown in Fig. 5, the saliency maps (Fig. 5(g)) deduced from the extracted depth features are more accurate than those saliency maps (Fig. 5(d)) deduced from the original RGB features. This is because the coarse saliency maps (Fig. 5(d)) suppress some disturbing information within the original depth features (Fig. 5(e)) to some extents. With the aid of the refined saliency maps and the extracted depth features, the disturbing information within the original RGB features (Fig. 5(c)) can be further suppressed, thus boosting the discriminability of the enhanced RGB features(Fig. 5(h)).

In RGDNet, the input RGB images are supposed to have

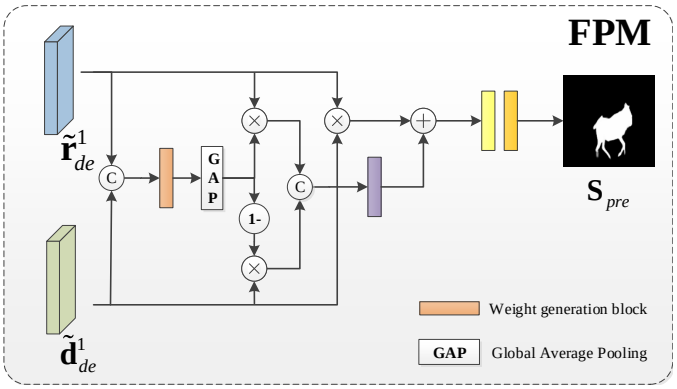


Fig. 6. Architecture of our proposed Fusion Prediction Module (FPM). First, we adopt the concatenation-convolution operation to learn the complementarity weights for different enhanced features. Secondly, we adopt multiplication operation to learn common information and strengthen the response of salient objects. Subsequently, the addition operation is used to obtain the final fused features. Finally, the fused features are mapped to a final saliency map through a 1×1 convolutional layer.

better qualities than the input depth images. The coarse saliency maps are used to guide the feature extraction of depth images, and the refined saliency maps are further to guide the refinement of RGB features. However, such assumptions may fail in some scenes with poor spatial appearances, where RGDNet may not work well. To address such an issue, another network based on the progressive guidance strategy, i.e., DGRNet, which supposes the input depth images have better qualities than the input RGB images, is also employed in our proposed model. DGRNet and RGDNet have the same structures but different parameters. Similarly, a set of enhanced depth features $\tilde{\mathbf{d}}^i_{de}$ $(i = 1, 2, 3, 4, 5)$ will be extracted from DGRNet. Subsequently, the enhanced RGB features and depth features will be fed into the following fusion module to achieve the fused features for the final saliency prediction.

B. Fusion Prediction Module

As discussed above, RGDNet will achieve satisfactory results if the input RGB images have better qualities, and DGRNet will work well when the input depth images have better qualities. Otherwise, the output enhanced features from RGDNet or DGRNet may still contain much disturbing information. For that, FPM is further designed to combine the enhanced features from the outputs of RGDNet and DGRNet for the final salient object detection.

Fig. 6 illustrates the diagram of the proposed FPM. Specifically, the input features $\tilde{\mathbf{r}}^1_{de}$ and $\tilde{\mathbf{d}}^1_{de}$ are concatenated to learn some channel-wise weights via some convolution and global averaging pooling operations, i.e.,

$$\{\mathbf{W}_1, \mathbf{W}_2\} = \mathrm{GAP}(\mathrm{Sig}(\mathrm{WG}(Cat(\tilde{\mathbf{r}}^1_{de}, \tilde{\mathbf{d}}^1_{de}), \xi))), \quad (11)$$

where $\mathrm{GAP}(*)$ denotes the global average pooling. $\mathrm{WG}(*;\xi)$ denotes a Weight generation block with its parameters ξ, which contains a 1×1 convolutional layer for channel compression, a 3×3 convolutional layer for learning the weight values and a Sigmoid activate function. $\mathrm{Cat}(*,*)$ refers to the concatenation operation. $\mathbf{W}_1$and $\mathbf{W}_2$ denote the channel-wise weights for RGB features $\tilde{\mathbf{r}}^1_{de}$ and depth features $\tilde{\mathbf{d}}^1_{de}$, respectively.

Algorithm 1 Bi-directional Progressive Guidance Network

Input: An RGB image $\mathcal{R}$, and a Depth image $\mathcal{D}$.

1: Forward the paired RGB-D images $\{\mathcal{R}, \mathcal{D}\}$ to RGDNet;
2: In RGDNet, the proposed PG strategy follows the RGB-depth-RGB manner to get enhanced RGB features $\tilde{\mathbf{r}}^i_{de}$:
Step1: Predict coarse saliency map from the input RGB image after using GCA;
Step2: Extract useful depth features with the guidance of coarse saliency map by using AFE;
Step3: Enhance the RGB features via the extracted depth features by using CFE;
3: Forward the paired RGB-D images $\{\mathcal{R}, \mathcal{D}\}$ to DGRNet;
4: In DGRNet, the proposed PG strategy follows the depth-RGB-depth manner to get enhanced depth features $\tilde{\mathbf{d}}^i_{de}$:
Step1: Predict coarse saliency map from the input depth image after using GCA;
Step2: Extract useful RGB features with the guidance of coarse saliency map by using AFE;
Step3: Enhance the depth features via the extracted RGB features by using CFE;
5: Perform selective integration on the enhanced RGB features $\tilde{\mathbf{r}}^1_{de}$ and the enhanced depth features $\tilde{\mathbf{d}}^1_{de}$ by using FPM;
6: Predict the saliency prediction result with the selected features $\mathbf{S}_{pre}$.

Output: Saliency prediction result $\mathbf{S}_{pre}$.

With the two sets of weights $\mathbf{W}_1$ and $\mathbf{W}_2$, some features $\mathbf{f}_w$ with higher discriminability are thus selected by

$$\mathbf{f}_w = \mathrm{Conv}_3(\mathrm{Cat}(\mathbf{W}_1 \bullet \tilde{\mathbf{r}}^1_{de}, \mathbf{W}_2 \bullet \tilde{\mathbf{d}}^1_{de}), \theta), \quad (12)$$

where $\bullet$ denotes the channel-wise multiplication.

Moreover, there may exist some common information between the two modality data, which will also benefit the saliency prediction. For that, an element-wise multiplication operation is performed on RGB features $\tilde{\mathbf{r}}^1_{de}$ and depth features $\tilde{\mathbf{d}}^1_{de}$ for capturing their common features $\mathbf{f}_c$, i.e.,

$$\mathbf{f}_c = \tilde{\mathbf{r}}^1_{de} \otimes \tilde{\mathbf{d}}^1_{de}. \quad (13)$$

By using this way, the common foreground regions will be preserved and some irrelevant information will simultaneously be suppressed.

Finally, these common features $\mathbf{f}_c$ and selective feature $\mathbf{f}_w$ are further combined to achieve the final fused features by using an element-wise addition operation. On top of that, the final saliency map $\mathbf{S}_{pre}$ is simply obtained by performing some convolution and Sigmod operations on the fused features, i.e.,

$$\mathbf{S}_{pre} = \mathrm{Sig}(\mathrm{Conv}_1(\mathbf{f}_c \oplus \mathbf{f}_w, \gamma)). \quad (14)$$

Algorithm 1 summarizes our proposed method BPGNet.

C. Network Training and Loss Function

Our network training process contains two phases, i.e., RGDNet/DGRNet pre-training and end-to-end fine-tuning.

RGDNet/DGRNet pre-training: Here, a joint loss function ℓ is used to train our model, which consists of a BCE loss ℓ_{bce} [64] and a IoU loss ℓ_{iou} [65], i.e.,

$$\ell = \ell_{bce} + \ell_{iou}. \tag{15}$$

Specifically, BCE [64] loss is one of the most widely used loss in SOD, which enforces the predicted saliency map as close to the ground truth as possible. IoU is originally used as an evaluation measure for object detection and segmentation. In recent years, it has also been used as the training loss for SOD [26].

When training RGDNet and DGRNet, some intermediate results from the two networks are supervised for accurate location guidance information. Taking RGDNet as an example, the intermediate results to be supervised include the coarse saliency prediction map $\mathbf{S}_c^r$ deduced from the deepest level of RGB features r_{en}^5 after GCA in the encoder of the primary branch, the refined saliency maps $\mathbf{S}_f^{d,i}$ $(i = 1, 2, 3, 4, 5)$ deduced from the corresponding *i*-th level of depth features in the decoder of the auxiliary branch, and the intermediate saliency maps $\mathbf{S}_R^i$ $(i = 1, 2, 3, 4, 5)$ deduced from the i-th level of RGB features in the decoder of the primary branch. Thus, there are 11 loss functions for training RGDNet in total, i.e.,

$$\begin{aligned} L_{RGD} &= L_{Prediction} + \lambda * L_{Giudance} \\ &= \sum\nolimits_{i=1}^{5} \ell(\mathbf{S}_R^i, \mathbf{G}) + \lambda_1 * \ell(\mathbf{S}_c^r, \mathbf{G}) + \lambda_2 * \sum\nolimits_{i=1}^{5} \ell(\mathbf{S}_f^{d,i}, \mathbf{G}). \end{aligned} \tag{16}$$

Here, λ_1 and λ_2 are two hyper-parameters for balancing the losses and are empirically set to 0.5 and 0.2, respectively. Similarly, the total loss function L_{DGR} for DGRNet is computed by

$$\begin{aligned} L_{DGR} &= L_{Prediction} + \lambda * L_{Giudance} \\ &= \sum\nolimits_{i=1}^{5} \ell(\mathbf{S}_D^i, \mathbf{G}) + \lambda_1 * \ell(\mathbf{S}_c^d, \mathbf{G}) + \lambda_2 * \sum\nolimits_{i=1}^{5} \ell(\mathbf{S}_f^{r,i}, \mathbf{G}), \end{aligned} \tag{17}$$

where $\mathbf{S}_D^i$ $(i = 1, 2, 3, 4, 5)$ refers to the saliency map deduced from the *i*-th level of depth features in the decoder of the primary branch. $\mathbf{S}_c^d$ denotes the coarse saliency prediction deduced from the deepest level of depth features $\mathbf{d}_{en}^5$. $\mathbf{S}_f^{r,i}$ $(i = 1, 2, 3, 4, 5)$ is the refined saliency map deduced from the corresponding *i*-th level of RGB features in the decoder of the auxiliary branch.

The stochastic gradient descent (SGD) method [70] is adopted to train RGDNet and DGRNet with batch size 4, momentum 0.9 and weight decay 5e-4. Meanwhile, the initial learning rate is set to 5e-4, which is divided by 10 after 40 epochs. The maximum epochs for RGDNet/DGRNet are 70.

End-to-end fine-tuning: Based on the pre-trained RGDNet and DGRNet, the whole BPGNet is fine-tuned for 20 epochs by employing the same SGD optimizer. The batch size, weight decay and momentum are set to 4, 5e-4 and 0.9, respectively. The learning rate is set to 5e-4, which is divided by 10 after 15 epochs. Here, in addition to the supervision on the final predicted saliency map $\mathbf{S}_{pre}$ from the FPM, some auxiliary supervisions are also performed. As a result, the total loss L_{total} for our model is formulated as:

$$L_{total} = \ell(\mathbf{S}_{pre}, \mathbf{G}) + L_{RGD} + L_{DGR}. \tag{18}$$

IV. Experiments

A. Datasets and Evaluation Metrics

1) Datasets: We conduct our experiments on six widely used RGB-D benchmark datasets, including DUT-RGBD [18], NJU2K [71], NLPR [45], STERE [72], LFSD [73] and RGBD135 [74]. DUT-RGBD [18] contains 1200 images captured by Lytro camera in real-life scenes. This dataset is split into 800 training data and 400 testing data. NJUD [71] includes 1985 RGB-D stereo images (the latest version), which were collected from the Internet, 3D movies and photographs taken by a Fuji W3 stereo camera. NLPR [45] contains 1000 image pairs captured by Kinect under different illumination conditions. STERE [72] contains 1000 stereoscopic images downloaded from the Internet, where the depth images are estimated from the stereo images. LFSD [73] contains 100 images with depth information captured via a Lytro light field camera. RGBD135 [74] contains 135 simple RGB-D images collected by Kinect for testing.

2) Evaluation Metrics: We use some standard metrics for performance evaluation, including Precision-Recall (PR) [75], maximum F-measure (F_β) [75], Mean Absolute Error (MAE) [76], S-measure (S_α) [77] and maximum E-measure (E_ξ) [78]. Larger values of F_β, S_α, E_ξ and smaller values of MAE are more desirable for a SOD method.

B. Implementation Details

We implement our network by using the Pytorch toolbox on an NVIDIA 2080Ti GPU. As the same splitting way in [18], [66], we randomly select 800 samples from DUT-RGBD, 1485 samples from NJU2K and 700 samples from NLPR for training. The remaining images in the three datasets and other three datasets are all for testing to verify the performance of different models. During training and testing, all the input images are resize to 224×224. Random horizontal flipping and random vertical flipping are adopted for data augmentation. The newly added convolution layers are initialized by the normal distribution. The average inference time of our method is 0.034 s for an image based on the above-mentioned configuration. The number of parameters and floating point operations (FLOPs) of our proposed model are 84.31 M and 138.55 G, respectively.

C. Comparison with the State-of-the-Arts

1) Comparison Methods: We compare our model with 13 state-of-the-art RGB-D based SOD methods, including MMCI [14],TANet [15], DMRA [18], ICNet [48], A2dele [66], S2MA [67], DRLF [38], CCAFNet [68], JL-DCF [41], CPFP [19], D3Net [13], DQSD [21] and DFMNet [69]. Specially, the qualities of depth images are also considered in the last three methods, i.e., CPFP [19], D3Net [13], DQSD [21] and DFMNet [69]. For all the methods mentioned here, we use either the implementations with their default parameter

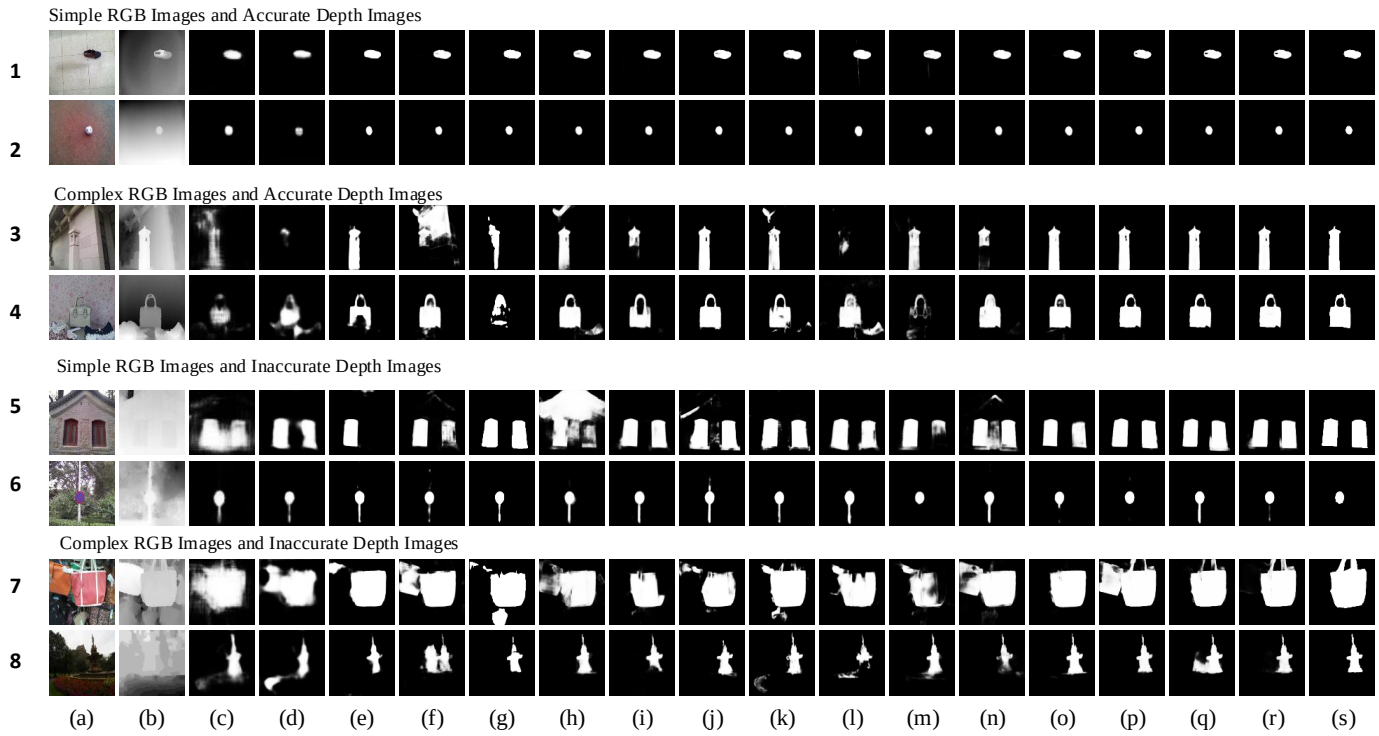


Fig. 7. Some visualization results of different SOD methods. (a) RGB images; (b) Depth images; (c) MMCI [14]; (d) TANet [15]; (e) DMRA [18]; (f) ICNet [48]; (g) A2dele [66]; (h) S2MA [67]; (i) DRLF [38]; (j) CCAFNet [68]; (k) JL-DCF [41];(l) CPFP [19]; (m) D3Net [13]; (n) DQSD [21]; (o) DFMNet [69]; (p) RGDNet(OURS); (q) DGRNet(OURS); (r) BPGNet(OURS); (s) GT.

settings or the saliency maps provided by their authors for fair comparisons.

2) Qualitative Evaluation and Quantitative Evaluation: We denote RGB images with similar foregrounds and backgrounds or cluttered backgrounds as complex RGB images, while denoting those RGB images with high contrast as simple ones. In the similar way, we define those depth images of high visual qualities as accurate depth images, while defining those depth images of low visual qualities as inaccurate ones. Fig. 7 illustrates some saliency maps of different maps under different challenging situations.

As shown in the first two rows of Fig. 7, all of the methods mentioned here perform well for those images with simple scenes. However, as described earlier, when one of the input images contains disturbing cues (e.g., the complex RGB images or the inaccurate depth images), these SOTA methods may not be able to achieve desirable saliency detection results. As shown in the 3^{rd} and 4^{th} rows of Fig. 7, compared with the depth images of high visual qualities, the complex RGB images contain much more disturbing information, which will bring risks for achieving unaccurate segmentation results. Similarly, the inaccurate depth images will also mislead the determination of the foreground objects in the 5^{th} and 6^{th} rows. Intuitively, our method results in a more complete and accurate saliency map structure than other methods in these challenging scenarios. Even in the case of that both the RGB images and the depth images contain some disturbing information, as shown in the 7^{th} and 8^{th} rows, our method can still achieve accurate segmentation of the foreground objects by effectively exploring the useful information within the multi-modality RGB and depth data.

In order to compare the performance of different models more intuitively, we report the PR curves of different models in Fig. 8 and the quantitative metric scores of Maximum F-measure, Maximum E-measure, S-measure and MAE scores in Table I. It can be seen that the proposed model achieves the best performance in most cases on six datasets.

D. Ablation Studies

In this section, we conduct some ablation experiments on the dataset NJU2K [71] to verify the validity of each component in our proposed model for a thorough understanding of the proposed method. Specifically, we will first demonstrate the validities of the proposed AFE, GCA and CFE modules on the extraction and enhancement of unimodal features. Then, we will verify the validities of the proposed bi-directional progressive guidance strategy and the FPM module for cross-modality fusion and saliency prediction. As well, considering that there are two networks, i.e., RGDNet and DGRNet, in our proposed model for the unimodal feature extraction and enhancement, we will accordingly construct two baseline models, i.e., B(R) and B(D), from RGDNet and DGRNet, respectively, before the verification of each component or module. For that, the AFE modules and GCA module are removed from RGDNet and DGRNet, and the CFE modules are replaced with some element-wise addition operations. On top of that, some simple convolution operations and Sigmoid function operations are performed on the features from the decoders in the two baselines to predict the saliency maps. Table II provides the quantitative results by adding different

TABLE I
QUANTITATIVE COMPARISONS OF OUR PROPOSED MODEL WITH 13 STATE-OF-THE-ART RGB-D SALIENCY MODELS ON 6 BENCHMARK DATASETS. THE BEST THREE RESULTS ARE SHOWN IN RED, GREEN AND BLUE COLORS, RESPECTIVELY.

	Metric	MMCI [14]	TANet [15]	DMRA [18]	ICNet [48]	A2dele [66]	S2MA [67]	DRLF [38]	CCAFNet [68]	JL-DCF [41]	CPFP [19]	D3Net [13]	DQSD [21]	DFMNet [69]	RGDNet (OURS)	DGRNet (OURS)	BPGNet (OURS)
NJU2K [71]	$F_\beta \uparrow$	0.852	0.874	0.886	0.891	0.873	0.889	0.883	0.910	0.912	0.877	0.900	0.900	0.913	0.921	0.916	0.926
	$E_\xi \uparrow$	0.915	0.925	0.927	0.926	0.916	0.930	0.926	0.943	0.949	0.923	0.939	0.936	0.949	0.951	0.947	0.953
	$S_\alpha \uparrow$	0.858	0.878	0.886	0.894	0.869	0.894	0.886	0.909	0.910	0.878	0.900	0.899	0.912	0.917	0.915	0.923
	$MAE \downarrow$	0.079	0.060	0.051	0.052	0.051	0.053	0.055	0.937	0.038	0.053	0.046	0.050	0.039	0.036	0.038	0.034
NLPR [45]	$F_\beta \uparrow$	0.815	0.863	0.880	0.908	0.880	0.902	0.880	0.908	0.915	0.867	0.897	0.898	0.912	0.907	0.901	0.914
	$E_\xi \uparrow$	0.913	0.941	0.947	0.952	0.945	0.953	0.939	0.956	0.963	0.932	0.953	0.952	0.961	0.955	0.950	0.959
	$S_\alpha \uparrow$	0.856	0.886	0.899	0.923	0.896	0.915	0.903	0.921	0.926	0.888	0.912	0.916	0.925	0.921	0.918	0.927
	$MAE \downarrow$	0.059	0.041	0.031	0.028	0.028	0.030	0.032	0.026	0.024	0.036	0.030	0.029	0.024	0.026	0.027	0.024
DUT [18]	$F_\beta \uparrow$	0.767	0.790	0.898	0.850	0.892	0.901	0.801	0.913	0.878	0.795	0.793	0.827	-	0.936	0.920	0.938
	$E_\xi \uparrow$	0.859	0.861	0.933	0.899	0.930	0.937	0.856	0.943	0.920	0.859	0.829	0.878	-	0.956	0.946	0.958
	$S_\alpha \uparrow$	0.791	0.808	0.889	0.852	0.885	0.903	0.826	0.903	0.881	0.818	0.773	0.845	-	0.927	0.918	0.930
	$MAE \downarrow$	0.113	0.093	0.048	0.072	0.042	0.043	0.080	0.037	0.055	0.076	0.098	0.072	-	0.031	0.036	0.031
STERE [72]	$F_\beta \uparrow$	0.863	0.861	0.857	0.898	0.879	0.882	0.878	0.887	0.898	0.874	0.891	0.886	0.904	0.897	0.902	0.904
	$E_\xi \uparrow$	0.927	0.923	0.916	0.942	0.928	0.932	0.929	0.934	0.942	0.925	0.938	0.935	0.948	0.939	0.941	0.944
	$S_\alpha \uparrow$	0.873	0.871	0.845	0.903	0.879	0.890	0.888	0.892	0.900	0.879	0.899	0.892	0.908	0.900	0.903	0.907
	$MAE \downarrow$	0.068	0.060	0.063	0.045	0.045	0.051	0.050	0.044	0.042	0.051	0.046	0.051	0.040	0.043	0.042	0.040
LFSD [73]	$F_\beta \uparrow$	0.771	0.796	0.856	0.871	0.835	0.835	0.828	0.832	0.839	0.826	0.810	0.847	0.866	0.861	0.875	0.875
	$E_\xi \uparrow$	0.839	0.847	0.900	0.903	0.879	0.873	0.873	0.876	0.879	0.872	0.862	0.878	0.902	0.903	0.908	0.908
	$S_\alpha \uparrow$	0.787	0.801	0.847	0.868	0.836	0.837	0.834	0.826	0.833	0.828	0.825	0.851	0.870	0.865	0.871	0.874
	$MAE \downarrow$	0.132	0.111	0.075	0.071	0.074	0.094	0.089	0.087	0.084	0.088	0.095	0.085	0.068	0.068	0.065	0.066
RGBD135 [74]	$F_\beta \uparrow$	0.822	0.827	0.888	0.913	0.867	0.935	0.869	0.937	0.917	0.846	0.885	0.927	0.932	0.934	0.918	0.932
	$E_\xi \uparrow$	0.928	0.910	0.945	0.960	0.923	0.973	0.940	0.977	0.960	0.923	0.946	0.973	0.973	0.969	0.960	0.973
	$S_\alpha \uparrow$	0.848	0.858	0.901	0.920	0.885	0.941	0.895	0.938	0.924	0.872	0.898	0.935	0.938	0.935	0.925	0.937
	$MAE \downarrow$	0.065	0.046	0.029	0.027	0.028	0.021	0.030	0.017	0.021	0.038	0.031	0.021	0.019	0.019	0.022	0.020

TABLE II
QUANTITATIVE EVALUATION OF ABLATION STUDIES ON NJU2K DATASET.

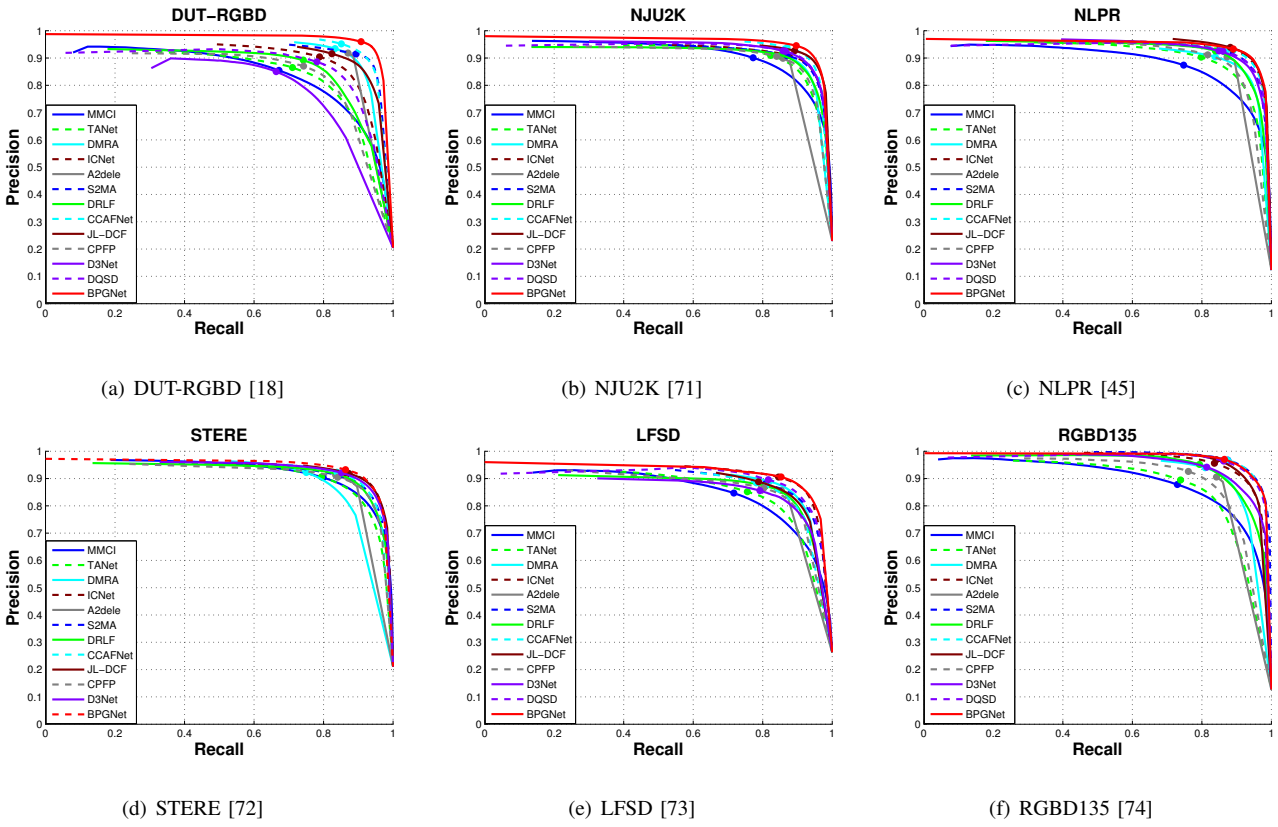
Methods	$F_\beta \uparrow$	$E_\xi \uparrow$	$S_\alpha \uparrow$	$MAE \downarrow$
B (R)	0.902	0.935	0.903	0.043
B+GCA (R)	0.912	0.943	0.910	0.040
B+GCA+AFE (R)	0.917	0.947	0.914	0.038
B+GCA+AFE+CFE (R)	0.921	0.951	0.917	0.036
B (D)	0.904	0.937	0.903	0.042
B+GCA (D)	0.907	0.941	0.905	0.041
B+GCA+AFE (D)	0.914	0.943	0.910	0.040
B+GCA+AFE+CFE (D)	0.916	0.947	0.915	0.038
Bi-A	0.918	0.946	0.916	0.038
Bi-M	0.922	0.949	0.918	0.037
Bi-W	0.922	0.951	0.919	0.036
Bi-FPM (BPGNet)	0.926	0.953	0.923	0.034

components on the two baselines. It should also be noted that, as shown in Table II, the two baselines perform competitively, since they have very similar structures.

(1) Validity of Global Context Awareness (GCA) module. To verify the effectiveness of the proposed GCA module, we construct two networks, i.e., B+GCA (R) and B+GCA (D), by introducing GCA modules into B(R) and B(D), respectively. It can be seen from the quantitative results in Table II that adding GCA can improve the performance of B (R) and B (D) significantly. Intuitively, the visual results in Fig. 9(c) and Fig. 9(d) also illustrate that B+GCA(R) and B+GCA (D) can better locate the salient objects than B (R) and B (D) do, respectively. The improvements benefit from that GCA can effectively extract global contextual semantic information for saliency prediction.

(2) Validity of Auxiliary Feature Extraction (AFE) module. To validate the validity of the coarse saliency detection results deduced from the encoder of the primary branch in RGDNet or DGRNet for guiding the feature extraction process of the auxiliary branch, we compare two model versions, i.e., B+GCA+AFE(R) and B+GCA+AFE (D), which are implemented by introducing the AFE modules into B+GCA(R) and B+GCA(D), respectively. It can be easily observed in Table II that both the performance of B+GCA+AFE (R) and that of B+GCA+AFE (D) can be boosted to different extents by introducing AFE modules. As well, it can also be seen in Fig. 9(d) and Fig. 9(e), B+GCA+AFE (R) and B+GCA+AFE (D) can better suppress such disturbing background regions and achieve more complete foregrounds than B+GCA (R) and B+GCA (D) do. These observations verify that AFE can improve the saliency detection performance by utilizing the coarse saliency map of one modality to guide the feature extraction of another modality.

(3) Validity of Cross-modality Feature Enhancement (CFE) module. To verify the validity of CFE, we compare another two networks, i.e., B+GCA+AFE+CFE (R) and



(a) DUT-RGBD [18] (b) NJU2K [71] (c) NLPR [45]

(d) STERE [72] (e) LFSD [73] (f) RGBD135 [74]

Fig. 8. Quantitative comparisons of our method with other methods on six challenging benchmark datasets.

B+GCA+AFE+CFE (D), by further introducing the CFE modules on top of B+GCA+AFE (R) and B+GCA+AFE (D), respectively. As reported in Table II, the CFE module can further promote the saliency detection performance of B+GCA+AFE+CFE (R) and that of B+GCA+AFE+CFE (D) by enhancing the original unimodal features of one modality data with the aid of the extracted features as well as their refined saliency maps of another modality data. Intuitively, the visual results illustrated in Fig. 9(e) and Fig. 9(f) consistently demonstrate that B+GCA+AFE+CFE (R) and B+GCA+AFE+CFE (D) can better suppress the background regions and highlight the foreground regions than B+GCA+AFE (R) and B+GCA+AFE (D) do, respectively, which indicates the validity of our proposed progressive guidance strategy across multi-modality data for RGB-D salient object detection.

(4) Validity of the Bi-directional progressive guidance framework. In this subsection, we will verify the validity of our proposed bi-directional progressive guidance strategy for multi-modality salient object detection. As reported in Table II, there exist obvious differences between the performance of B+GCA+AFE+CFE (R) and B+GCA+AFE+CFE (D), although the performance of their baselines (i.e., B (R) and B (D)) is competitive. As discussed earlier, in B+GCA+AFE+CFE (R), the RGB images are supposed to have relative better qualities than those depth images and dominate the saliency detection results, while the depth images are supposed to have relative better qualities and dominate the saliency detection results in B+GCA+AFE+CFE (D). As well, in the dataset NJU2K-TEST, the RGB images usually have better visual qualities than those depth images. Therefore, the assumption in B+GCA+AFE+CFE (R) is more reasonable than that in B+GCA+AFE+CFE (D). Accordingly, B+GCA+AFE+CFE (R) outperforms B+GCA+AFE+CFE (D) on NJU2K-TEST.

This indicates that our proposed (single-directional) progressive guidance strategy may be valid only if the assumption on the input multi-modality data is reasonable. Otherwise, some undesirable saliency detection results may also be achieved. This can also be illustrated by the visual saliency detection results in Fig. 10. As shown in the first two rows of Fig. 10, B+GCA+AFE+CFE (R) achieves better saliency maps than B+GCA+AFE+CFE (D) does when the input RGB images have better visual qualities than the input depth images. While, as shown in last two rows of Fig. 10, B+GCA+AFE+CFE (D) outperforms B+GCA+AFE+CFE (R) when the input depth images have better visual qualities. However, for all of the four pairs of input images in Fig. 10, our proposed BPGNet achieves desirable saliency detection results because of the employed bi-directional progressive guidance framework, as illustrated in Fig. 10(e). The quantitative results in Table II also demonstrate that our proposed BPGNet (i.e., Bi-FPM (BPGNet)) significantly outperforms B+GCA+AFE+CFE (R) and B+GCA+AFE+CFE (D). This also indicates the validity of our proposed bi-directional progressive guidance framework.

(5) Validities of Fusion Prediction Module (FPM). In order to further verify the effectiveness of the proposed FPM, we conduct the following several models, i.e., Bi-A, Bi-

Fig. 9. Some visual comparisons on RGDNet and DGRNet. (a) RGB images; (b) Depth images; (c) B; (d) B+GCA; (e) B+GCA+AFE; (f) B+GCA+AFE+CFE; (g) GT. In the first three rows, some high-quality RGB images and relatively low-quality depth images are used to verify the effectiveness of each component on RGDNet. While, in the last three rows, some high-quality depth images and relatively low-quality RGB images are used to verify the effectiveness of each component on DGRNet.

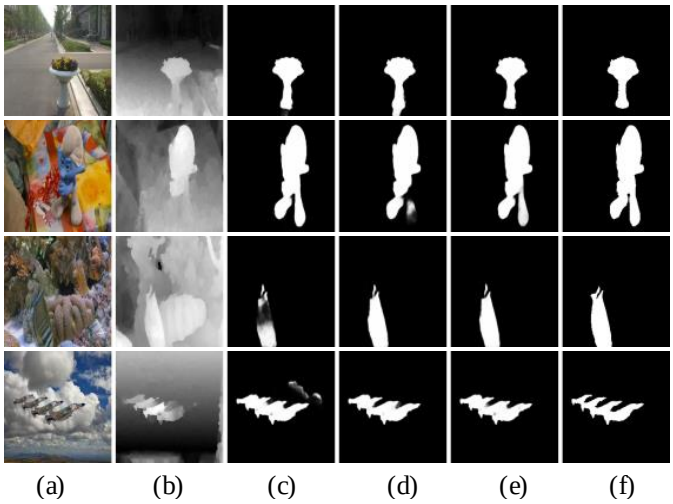


Fig. 10. Visual comparisons of RGDNet and DGRNet under different cases. (a) RGB images; (b) Depth images; (c) Saliency maps S_R^1 deduced from the enhanced RGB features of RGDNet; (d) Saliency maps S_D^1 deduced from the enhanced depth features of DGRNet; (e) Saliency maps S_{pre} deduced from the fused features of BPGNet; (f)GT.

M, Bi-W and Bi-FPM (i.e., BPGNet), which employ the element-wise addition, element-wise multiplication, channel-wise attention and our proposed FPM for the fusion of the enhanced RGB and depth features, respectively.

From Table II, it can be seen that Bi-A and Bi-M are inferior to B+GCA+AFE+CFE (R), although both of them employ the bi-directional progressive guidance strategy during the saliency detection. This may be due to the following fact. As discussed previously, the depth images in the dataset NJU2K-TEST usually have lower visual qualities than the RGB images. Accordingly, the enhanced depth features from DGRNet still contain much disturbing information although the enhanced RGB features from RGDNet contain much more discriminative information. By simply using the element-wise addition based fusion strategy in Bi-A, the disturbing information within the enhanced depth features will degrade the discriminability of the enhanced RGB features during the fusion, thus degrading the final saliency prediction. By simply using the element-wise multiplication based fusion strategy in Bi-M, only the common information between the enhanced RGB features and the enhanced depth features is preserved in the fused features, while some complementary information between the enhanced unimodal features is suppressed during the fusion. And these complementary information may be more important for the multi-modality salient object detection. As a result of that, the saliency detection performance of Bi-M is also degraded to some extents.

Differently, in Bi-W, those unimodal features with higher discriminative ability are adaptively selected for the final saliency prediction with a channel-wise attention based fusion strategy. For that, the bi-directional progressive guidance based model Bi-W performs better than the two single-directional progressive guidance based models, i.e., B+GCA+AFE+CEF (R) and B+GCA+AFE+CFE (D), as shown in Table II. With the proposed FPM, those higher discriminative features are selected and preserved into the fused features. As well, those common information that is simultaneously captured by using RGDNet and DGRNet can also be preserved into the fused features. This further boosts the saliency detection results. As a result, Bi+FPM, i.e., our proposed BPGNet, achieves the best performance among these configurations. This also indicates that our proposed BPGNet can well select higher discriminative features and simultaneously suppress those disturbing information within the input images for saliency detection with the collaboration of the bi-directional progressive guidance strategy and FPM.

V. Conclusion

In this paper, we have presented a bi-directional progressive guidance network, i.e., BPGNet, for RGB-D salient object detection, which consists of a two-stream feature extraction and enhancement network (i.e., RGDNet and DGRNet) and a multi-modality feature Fusion Prediction Module (FPM). Especially, a progressive guidance strategy is employed in RGDNet and DGRNet to explore the alternative interactions between the prediction results from one modality data and the extracted features from another modality data, which can suppress the disturbing information within the two modalities of inputs in a coarse-to-fine manner. In addition, with the proposed FPM, some features with higher discriminative ability are adaptively selected from the outputs of RGDNet and DGRNet. In addition, the common information that are simultaneously captured by RGDNet and DGRNet is further enhanced for the final saliency prediction. With the collaboration of the bi-directional progressive guidance strategy and FPM, our proposed BPGNet achieves competitive results. Experimental results demonstrate the effectiveness and superiority of the proposed framework over some state-of-the-art

methods, especially when one modality data from the multi-modality input images have low visual qualities.

It should be noted that the high performance of our proposed method is at the cost of a complex architecture, which may limit its application in some other vision tasks. For future work, we will pay more attention to achieving a lightweight architecture for SOD task, while maintaining the saliency detection accuracy.

ACKNOWLEDGEMENT

This work is supported by the National Natural Science Foundation of China under Grant No.61773301 and 62001341.

REFERENCES

[1] W. Wang, Q. Lai, H. Fu, J. Shen, H. Ling, and R. Yang, "Salient object detection in the deep learning era: An in-depth survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021.

[2] R. Cong, J. Lei, H. Fu, M.-M. Cheng, W. Lin, and Q. Huang, "Review of Visual Saliency Detection With Comprehensive Information," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 29, no. 10, pp. 2941–2959, 2019.

[3] Q. Jiang, F. Shao, W. Lin, K. Gu, G. Jiang, and H. Sun, "Optimizing multistage discriminative dictionaries for blind image quality assessment," *IEEE Transactions on Multimedia*, vol. 20, no. 8, pp. 2035–2048, 2017.

[4] D. Gao, S. Han, and N. Vasconcelos, "Discriminant saliency, the detection of suspicious coincidences, and applications to visual recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, no. 6, pp. 989–1005, 2009.

[5] Y. Wei, X. Liang, Y. Chen, X. Shen, M.-M. Cheng, J. Feng, Y. Zhao, and S. Yan, "Stc: A simple to complex framework for weakly-supervised semantic segmentation," *IEEE transactions on pattern analysis and machine intelligence*, vol. 39, no. 11, pp. 2314–2320, 2016.

[6] X. Jia, H. Lu, and M.-H. Yang, "Visual tracking via coarse and fine structural local sparse appearance models," *IEEE Transactions on Image processing*, vol. 25, no. 10, pp. 4555–4564, 2016.

[7] W. Wang, J. Shen, and L. Shao, "Video salient object detection via fully convolutional networks," *IEEE Transactions on Image Processing*, vol. 27, no. 1, pp. 38–49, 2017.

[8] C. Guo and L. Zhang, "A novel multiresolution spatiotemporal saliency detection model and its applications in image and video compression," *IEEE transactions on image processing*, vol. 19, no. 1, pp. 185–198, 2009.

[9] S. Wang, M. Wang, S. Yang, and K. Zhang, "Salient region detection via discriminative dictionary learning and joint bayesian inference," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 28, no. 5, pp. 1116–1129, 2017.

[10] Y. Piao, Z. Rong, M. Zhang, X. Li, and H. Lu, "Deep light-field-driven saliency detection from a single view." in *IJCAI*, 2019, pp. 904–911.

[11] Y. Liu, J. Han, Q. Zhang, and L. Wang, "Salient Object Detection via Two-Stage Graphs," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 29, no. 4, pp. 1023–1037, 2019.

[12] P. Zhang, D. Wang, H. Lu, H. Wang, and X. Ruan, "Amulet: Aggregating multi-level convolutional features for salient object detection," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 202–211.

[13] D.-P. Fan, Z. Lin, Z. Zhang, M. Zhu, and M.-M. Cheng, "Rethinking rgb-d salient object detection: Models, data sets, and large-scale benchmarks," *IEEE Transactions on Neural Networks and Learning Systems*, 2020.

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

[15] H. Chen and Y. Li, "Three-stream attention-aware network for rgb-d salient object detection," *IEEE Transactions on Image Processing*, vol. 28, no. 6, pp. 2825–2835, 2019.

[16] J. Han, H. Chen, N. Liu, C. Yan, and X. Li, "Cnns-based rgb-d saliency detection via cross-view transfer and multiview fusion," *IEEE transactions on cybernetics*, vol. 48, no. 11, pp. 3171–3183, 2017.

[17] N. Wang and X. Gong, "Adaptive fusion for rgb-d salient object detection," *IEEE Access*, vol. 7, pp. 55 277–55 284, 2019.

[18] Y. Piao, W. Ji, J. Li, M. Zhang, and H. Lu, "Depth-induced multi-scale recurrent attention network for saliency detection," in *Proceedings of the IEEE International Conference on Computer Vision*, 2019, pp. 7254–7263.

[19] 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 *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 3927–3936.

[20] C. Chen, J. Wei, C. Peng, W. Zhang, and H. Qin, "Improved saliency detection in rgb-d images using two-phase depth estimation and selective deep fusion," *IEEE Transactions on Image Processing*, vol. 29, pp. 4296–4307, 2020.

[21] C. Chen, J. Wei, C. Peng, and H. Qin, "Depth-quality-aware salient object detection," *IEEE Transactions on Image Processing*, vol. 30, pp. 2350–2363, 2021.

[22] H. Peng, B. Li, H. Ling, W. Hu, W. Xiong, and S. J. Maybank, "Salient object detection via structured matrix decomposition," *IEEE transactions on pattern analysis and machine intelligence*, vol. 39, no. 4, pp. 818–832, 2016.

[23] C. Yang, L. Zhang, H. Lu, X. Ruan, and M.-H. Yang, "Saliency detection via graph-based manifold ranking," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2013, pp. 3166–3173.

[24] D. A. Klein and S. Frintrop, "Center-surround divergence of feature statistics for salient object detection," in *2011 International Conference on Computer Vision*. IEEE, 2011, pp. 2214–2219.

[25] A. Borji, M.-M. Cheng, H. Jiang, and J. Li, "Salient object detection: A benchmark," *IEEE transactions on image processing*, vol. 24, no. 12, pp. 5706–5722, 2015.

[26] X. Qin, Z. Zhang, C. Huang, C. Gao, M. Dehghan, and M. Jagersand, "Basnet: Boundary-aware salient object detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 7479–7489.

[27] N. Liu and J. Han, "Dhsnet: Deep hierarchical saliency network for salient object detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 678–686.

[28] Z. Tu, Y. Ma, C. Li, J. Tang, and B. Luo, "Edge-guided non-local fully convolutional network for salient object detection," *IEEE Transactions on Circuits and Systems for Video Technology*, 2020.

[29] M. Huang, Z. Liu, L. Ye, X. Zhou, and Y. Wang, "Saliency detection via multi-level integration and multi-scale fusion neural networks," *Neurocomputing*, vol. 364, pp. 310–321, 2019.

[30] W. Wang, J. Shen, X. Dong, and A. Borji, "Salient object detection driven by fixation prediction," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 1711–1720.

[31] Q. Zhang, R. Cong, C. Li, M.-M. Cheng, Y. Fang, X. Cao, Y. Zhao, and S. Kwong, "Dense Attention Fluid Network for Salient Object Detection in Optical Remote Sensing Images," *IEEE Transactions on Image Processing*, vol. 30, pp. 1305–1317, 2021.

[32] W. Wang, S. Zhao, J. Shen, S. C. Hoi, and A. Borji, "Salient object detection with pyramid attention and salient edges," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 1448–1457.

[33] W. Wang, J. Shen, M.-M. Cheng, and L. Shao, "An iterative and cooperative top-down and bottom-up inference network for salient object detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 5968–5977.

[34] L. Wang, R. Chen, L. Zhu, H. Xie, and X. Li, "Deep sub-region network for salient object detection," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 31, no. 2, pp. 728–741, 2021.

[35] X. Zhang, T. Wang, J. Qi, H. Lu, and G. Wang, "Progressive attention guided recurrent network for salient object detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 714–722.

[36] X. Hu, C.-W. Fu, L. Zhu, T. Wang, and P.-A. Heng, "SAC-Net: Spatial Attenuation Context for Salient Object Detection," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 31, no. 3, pp. 1079–1090, 2021.

[37] R. Cong, J. Lei, H. Fu, J. Hou, Q. Huang, and S. Kwong, "Going From RGB to RGBD Saliency: A Depth-Guided Transformation Model," *IEEE Transactions on Cybernetics*, vol. 50, no. 8, pp. 3627–3639, 2020.

[38] X. Wang, S. Li, C. Chen, Y. Fang, A. Hao, and H. Qin, "Data-Level Recombination and Lightweight Fusion Scheme for RGB-D Salient Object Detection," *IEEE Transactions on Image Processing*, vol. 30, pp. 458–471, 2021.

[39] Y. Ding, Z. Liu, M. Huang, R. Shi, and X. Wang, "Depth-aware saliency detection using convolutional neural networks," *Journal of*

Visual Communication and Image Representation, vol. 61, pp. 1–9, 2019.

[40] C. Li, R. Cong, S. Kwong, J. Hou, H. Fu, G. Zhu, D. Zhang, and Q. Huang, "ASIF-Net: Attention Steered Interweave Fusion Network for RGB-D Salient Object Detection," *IEEE Transactions on Cybernetics*, vol. 51, no. 1, pp. 88–100, 2021.

[41] 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 Transactions on Pattern Analysis and Machine Intelligence*, 2021.

[42] 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 Transactions on Image Processing*, vol. 26, no. 9, pp. 4204–4216, 2017.

[43] Z. Liu, S. Shi, Q. Duan, W. Zhang, and P. Zhao, "Salient object detection for rgb-d image by single stream recurrent convolution neural network," *Neurocomputing*, vol. 363, pp. 46–57, 2019.

[44] L. Qu, S. He, J. Zhang, J. Tian, Y. Tang, and Q. Yang, "Rgbd salient object detection via deep fusion," *IEEE Transactions on Image Processing*, vol. 26, no. 5, pp. 2274–2285, 2017.

[45] H. Peng, B. Li, W. Xiong, W. Hu, and R. Ji, "Rgbd salient object detection: a benchmark and algorithms," in *European conference on computer vision*. Springer, 2014, pp. 92–109.

[46] R. Shigematsu, D. Feng, S. You, and N. Barnes, "Learning rgb-d salient object detection using background enclosure, depth contrast, and top-down features," in *Proceedings of the IEEE International Conference on Computer Vision Workshops*, 2017, pp. 2749–2757.

[47] H. Chen and Y. Li, "Progressively complementarity-aware fusion network for rgb-d salient object detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 3051–3060.

[48] G. Li, Z. Liu, and H. Ling, "Icnet: Information conversion network for rgb-d based salient object detection," *IEEE Transactions on Image Processing*, vol. 29, pp. 4873–4884, 2020.

[49] W. Zhou, Y. Chen, C. Liu, and L. Yu, "Gfnet: Gate fusion network with res2net for detecting salient objects in rgb-d images," *IEEE Signal Processing Letters*, vol. 27, pp. 800–804, 2020.

[50] N. Huang, Y. Yang, D. Zhang, Q. Zhang, and J. Han, "Employing bilinear fusion and saliency prior information for rgb-d salient object detection," *IEEE Transactions on Multimedia*, 2021.

[51] N. Liu, N. Zhang, and J. Han, "Learning selective self-mutual attention for rgb-d saliency detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 13 756–13 765.

[52] G. Li, Z. Liu, L. Ye, Y. Wang, and H. Ling, "Cross-modal weighting network for RGB-D salient object detection," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XVII 16*. Springer, 2020, pp. 665–681.

[53] N. Huang, Y. Liu, Q. Zhang, and J. Han, "Joint cross-modal and unimodal features for rgb-d salient object detection," *IEEE Transactions on Multimedia*, 2020.

[54] Z. Bai, Z. Liu, G. Li, L. Ye, and Y. Wang, "Circular Complement Network for RGB-D Salient Object Detection," *Neurocomputing*, vol. 451, pp. 95–106, 2021.

[55] G. Li, Z. Liu, M. Chen, Z. Bai, W. Lin, and H. Ling, "Hierarchical alternate interaction network for RGB-D salient object detection," *IEEE Transactions on Image Processing*, vol. 30, pp. 3528–3542, 2021.

[56] Z. Chen, R. Cong, Q. Xu, and Q. Huang, "DPANet: Depth Potentiality-Aware Gated Attention Network for RGB-D Salient Object Detection," *IEEE Transactions on Image Processing*, vol. 30, pp. 7012–7024, 2021.

[57] 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 Transactions on Circuits and Systems for Video Technology*, pp. 1–1, 2021.

[58] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[59] L. Chen, H. Zhang, J. Xiao, L. Nie, J. Shao, W. Liu, and T.-S. Chua, "Sca-cnn: Spatial and channel-wise attention in convolutional networks for image captioning," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 5659–5667.

[60] J. Fu, J. Liu, H. Tian, Y. Li, Y. Bao, Z. Fang, and H. Lu, "Dual attention network for scene segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 3146–3154.

[61] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in neural information processing systems*, 2017, pp. 5998–6008.

[62] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs," *IEEE transactions on pattern analysis and machine intelligence*, vol. 40, no. 4, pp. 834–848, 2017.

[63] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 7132–7141.

[64] P.-T. De Boer, D. P. Kroese, S. Mannor, and R. Y. Rubinstein, "A tutorial on the cross-entropy method," *Annals of operations research*, vol. 134, no. 1, pp. 19–67, 2005.

[65] G. Máttyus, W. Luo, and R. Urtasun, "Deeproadmapper: Extracting road topology from aerial images," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 3438–3446.

[66] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 9060–9069.

[67] N. Liu, N. Zhang, and J. Han, "Learning Selective Self-Mutual Attention for RGB-D Saliency Detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 13 756–13 765.

[68] W. Zhou, Y. Zhu, J. Lei, J. Wan, and L. Yu, "CCAFNet: Crossflow and cross-scale adaptive fusion network for detecting salient objects in RGB-D images," *IEEE Transactions on Multimedia*, 2021.

[69] 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 *Proceedings of the 29th ACM International Conference on Multimedia*, 2021, pp. 731–740.

[70] L. Bottou, "Large-scale machine learning with stochastic gradient descent," in *Proceedings of COMPSTAT'2010*. Springer, 2010, pp. 177–186.

[71] R. Ju, L. Ge, W. Geng, T. Ren, and G. Wu, "Depth saliency based on anisotropic center-surround difference," in *2014 IEEE international conference on image processing (ICIP)*. IEEE, 2014, pp. 1115–1119.

[72] Y. Niu, Y. Geng, X. Li, and F. Liu, "Leveraging stereopsis for saliency analysis," in *2012 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2012, pp. 454–461.

[73] N. Li, J. Ye, Y. Ji, H. Ling, and J. Yu, "Saliency detection on light field," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 2806–2813.

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

[75] R. Achanta, S. Hemami, F. Estrada, and S. Susstrunk, "Frequency-tuned salient region detection," in *2009 IEEE conference on computer vision and pattern recognition*. IEEE, 2009, pp. 1597–1604.

[76] F. Perazzi, P. Krähenbühl, Y. Pritch, and A. Hornung, "Saliency filters: Contrast based filtering for salient region detection," in *2012 IEEE conference on computer vision and pattern recognition*. IEEE, 2012, pp. 733–740.

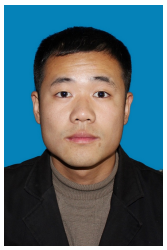
[77] D.-P. Fan, M.-M. Cheng, Y. Liu, T. Li, and A. Borji, "Structure-measure: A new way to evaluate foreground maps," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 4548–4557.

[78] D.-P. Fan, C. Gong, Y. Cao, B. Ren, M.-M. Cheng, and A. Borji, "Enhanced-alignment measure for binary foreground map evaluation," *arXiv preprint arXiv:1805.10421*, 2018.

Yang Yang received his B. S. degree from Chang'an University, Xi'an, China, in 2019. He is currently pursuing the Ph.D. degree in School of Mechano-Electronic Engineering, Xidian University, China. His current research interests include multimodal image processing and deep learning.

Qi Qin received the B.S. degree from Xidian University, Xi'an ,China, in 2018. she is currently pursing M.S. degree in School of Mechano-Electronic Engineering,Xidian University,China. Her current research interests include multimodal image processing and deep learning.

Yongjiang Luo received the B.S. degree in automatic control, the M.S. degree and Ph.D. in circuit and system from Xidian University, China, in 2001, 2004, and 2011, respectively. He was a Visiting Scholar with University of California, Merced, USA. He is currently an associate professor with the School of Electronic Engineering, Xidian University, China. His current research interests include wideband signal processing and intelligent information processing.



Yi Liu received the B. S. degree from Nanjing Institute of Technology, Nanjing, China, in 2012, the M. S. degree from Dalian University, Dalian, China, in 2015, and Ph.D. degree from Xidian University, Xi'an, China, in 2019. He is currently working at Changzhou University. He was a visiting student at Lancaster University from September 2018 to September 2019. His current research interests include computer vision and machine learning.

Qiang Zhang received the B.S. degree in automatic control, the M.S. degree in pattern recognition and intelligent systems, and the Ph.D. degree in circuit and system from Xidian University, China, in 2001,2004, and 2008, respectively. He was a Visiting Scholar with the Center for Intelligent Machines, McGill University, Canada. He is currently a professor with the Automatic Control Department, Xidian University, China. His current research interests include computer vision and image processing..

Jungong Han is currently a Full Professor and Chair in Computer Science at Aberystwyth University, UK. His research interests span the fields of video analysis, computer vision and applied machine learning. He has published over 180 papers, including 40+ IEEE Trans and 40+ A* conference papers.