Three-Stage Bidirectional Interaction Network for Efficient RGB-D Salient Object Detection

Yang Wang and Yanqing Zhang✉

South China University of Technology, Guangzhou, China
202021045142@mail.scut.edu.cn
zyqcs@scut.edu.cn

Abstract. The addition of depth maps improves the performance of salient object detection (SOD). However, most existing RGB-D SOD methods are inefficient. We observe that existing models take into account the respective advantages of the two modalities but do not fully explore the roles of cross-modality features of various levels. To this end, we remodel the relationship between RGB features and depth features from a new perspective of the feature encoding stage and propose a three-stage bidirectional interaction network (TBINet). Specifically, to obtain robust feature representations, we propose three interaction strategies: bidirectional attention guidance (BAG), bidirectional feature supplement (BFS), and shared network, and use them for the three stages of feature encoder, respectively. In addition, we propose a cross-modality feature aggregation (CFA) module for feature aggregation and refinement. Our model is lightweight (3.7 M parameters) and fast (329 ms on CPU). Experiments on six benchmark datasets show that TBINet outperforms other SOTA methods. Our model achieves the best performance and efficiency trade-off.

1 Introduction

Salient object detection (SOD) aims to locate the object(s) most concerned by human eyes from a given scene. It is the pre-task of many computer vision tasks, such as semantic segmentation [1, 2], tracking [3, 4], image/video compression [5, 6], and image retrieval [7]. Although significant progress has been made in SOD in recent years, it is still challenged to accurately locate objects in complex scenes, such as complex textures, cluttered backgrounds, and low contrast.

With the wide use of depth sensors in smartphones and other devices, RGB-D SOD has attracted the attention of researchers [8–13]. The depth map has illumination invariance and internal consistency, which can provide complementary spatial information for RGB images and improve saliency detection performance. As we all know, RGB and depth are two different modalities. An effective interaction strategy for a two-stream feature encoder can obtain more robust saliency-related features and thereby help the subsequent decoder generate more accurate saliency maps. The existing interaction strategies can be roughly divided into four categories: (i) No interaction mode [9, 13, 14] shown in Fig. 1(a),

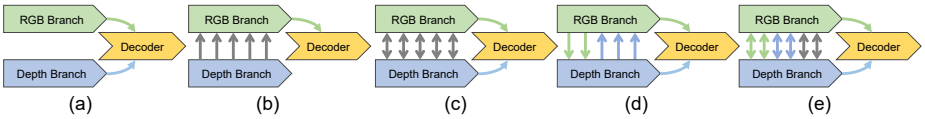

Fig. 1. Comparison of network interaction strategies between existing models and our model. (a) No interaction. (b) Unidirectional interaction. (c) Bidirectional interaction. (d) Cross-modality discrepant interaction. (e) Proposed three-stage bidirectional interaction.

which uses two independent branches to learn features of the two modalities separately, and then feds the features into subsequent feature fusion modules or the decoder. (ii) Unidirectional interaction mode [8, 10, 15] shown in Fig. 1(b), which integrates depth cues into RGB branch, and then feds the integrated features into decoder. (iii) Bidirectional interaction mode [16] shown in Fig. 1(c), which performs the same bidirectional operation on the hierarchical features of the two modalities. (iv) Cross-modality discrepant interaction mode [17] shown in Fig. 1(d), which gives full play to the respective advantages of the two modalities. Most of these interaction strategies are designed based on the modality perspective, while we try to explore the relationship between RGB features and depth features from the perspective of feature encoding stage. The basic observation of hierarchical cross-modality features is that high-level features contain rich global context information, which is conducive to locating salient regions, low-level features contain detailed information that can contribute significantly to refining the boundaries of salient regions [8].

To this end, we propose a novel three-stage bidirectional interaction network (TBINet) for RGB-D SOD. Specifically, the interaction of feature encoding process is divided into three stages (as shown in Fig. 1(e)): the interaction of low-level features in first stage, the interaction of middle-level features in second stage, and the interaction of high-level features in thrid stage. Low-level cross-modality features have specific boundary details, such as RGB image will be difficult to distinguish between salient objects and background in the case of complex texture and low contrast, and depth map will contain misleading information when salient objects and non-salient objects have the same spatial depth. Therefore, in first stage, we propose a bidirectional attention guidance (BAG) module, which can guide the two branches to focus on the important regions of each other while maintaining the modality-specific low-level features. The quality of depth maps tends to be uneven. Decreasing the influence of misleading information from low-quality depth maps is a key and hot issue in RGB-D SOD. We noticed that middle-level features contain approximate location information and rough boundary information. Therefore, in second stage, we propose a bidirectional feature supplement (BFS) module, which extracts cross-modality fusion features and transfers them to two branches separately. The BFS module effectively suppresses the low-quality features of deep branches and helps purify saliency-oriented feature representations. High-level features have the lowest res-

olution and can locate salient objects. After the abstraction of cross-modality information by the previous layers, the high-level cross-modality features have similar global context information, and the features of the two modalities have strong commonality. Inspired by JL-DCF [18], in thrid stage, we use shared network based on shared CNN layers, which can extract high-level cross-modality features with fewer parameters. The three-stage bidirectional interaction strategy effectively utilizes the characteristics of the three encoding stages. It helps the encoder finally generate multi-level cross-modality feature representations with specificity, purity, and commonality.

In addition, to integrate multi-level cross-modality features, we implement a three-stage refinement decoder. The three stages correspond one-to-one with the three stages of the encoder. Each decoder stage contains a cross-modality feature aggregation (CFA) module. The CFA module performs alternate feature fusion and refinement through two steps to effectively fuse and refine cross-modality features. The decoder generates final accurate saliency maps through feature fusion and refinement of the three CFA modules. Inspired by the channel split and channel shuffle operations in ShuffleNet-v2 [19], we redesign an efficient receptive field block (ERFB) module for the CFA module to expand the receptive field and extract multi-scale features.

Our network adopts the lightweight MobileNet-v3 [20] as the backbone network, and all modules adopt a lightweight design. Our model is lightweight (15.1 MB model size and 3.7 M parameters) and fast (329 ms inference time on CPU and 93 FPS inference speed on GPU). Our main contributions are as follows:

1. We propose a novel three-stage bidirectional interaction network (TBINet) for RGB-D SOD. TBINet adopts different interaction strategies in different stages of the feature encoding process so that the cross-modality features of various levels can give full play to their advantages.
2. We propose a three-stage refinement decoder and a cross-modality feature aggregation (CFA) module. Each decoding stage utilizes a CFA module for feature aggregation. The decoder continuously refines the saliency-oriented feature representation through three-stage feature aggregation and finally generates accurate saliency maps.
3. Our model is based on a lightweight design with fewer parameters and faster speed than cumbersome models. Experiments on six public datasets show that our model outperforms 15 state-of-the-art models and achieves a good balance between efficiency and performance.

2 Related Work

2.1 RGB-D SOD

In some complex scenes, salient objects in RGB images are indistinguishable from the background. Adding depth information may help overcome this challenge. Traditional RGB-D SOD models extract handcrafted features from RGB images and depth maps and fuse them for saliency detection [21–24]. However, due to the

limited expressive power of handcrafted features, the performance of traditional methods is not satisfactory.

With the rapid development of deep convolutional neural networks (CNNs), researchers have begun to focus on CNNs-based RGB-D SOD work and push the performance to new peaks [8, 12, 13, 18, 25–28]. Two key challenges facing current RGB-D SOD research are dealing with low-quality depth maps and effectively aggregating cross-modality multi-level features. For low-quality depth maps, for example, Fan *et al.* [29] proposed a depth depurator unit to filter low-quality depth maps. Jin *et al.* [9] proposed a complementary depth network, which estimates a depth map from the RGB image, and fuses the estimated depth map with the original depth map. Ji *et al.* [26] proposed a depth calibration and fusion framework capable of calibrating the depth image and correcting the latent bias in the original depth maps. Zhang *et al.* [15] proposed a depth feature manipulation network that can control depth features and avoid feeding misleading depth features. For cross-modality multi-level feature aggregation, for example, Fu *et al.* [18] developed a densely cooperative fusion strategy that uses dense connections to facilitate the fusion of depth and RGB features at different scales. Li *et al.* [30] proposed an adaptive feature selection module that emphasizes the importance of channel features in self-modality and cross-modality while fusing multi-modality spatial features. For more inspiring related works, refer to the recent survey [31, 32].

2.2 Efficient RGB-D SOD

Efficiency is also important for models besides performance. Recently, researchers have started to propose some efficient models for RGB-D SOD with lighter size and faster speed. Zhao *et al.* [33] proposed an early fusion single-stream network to make the network lighter. Chen *et al.* [34] constructed a lightweight deep stream to make the network more compact and efficient. More and more computer vision applications are adapting to mobile devices. To this end, many lightweight networks for image classification have been proposed, such as MobileNets [20, 35, 36] and ShuffleNets [19, 37]. Unlike classic cumbersome networks, such as VGG [38] and ResNet [39], lightweight networks can be well adapted to mobile devices due to their extremely high efficiency. Some RGB-D SOD models attempt to use a lightweight network as the backbone network. Wu *et al.* [40] proposed a network named MobileSal, which uses MobileNet-v2 [36] as the backbone network and fuses RGB features with depth features only on the coarsest layers. Zhang *et al.* [15] proposed an efficient model DFMNet based on MobileNet-v2 [36] and a tailored depth backbone. The current efficient RGB-D SOD models still lacks performance compared with cumbersome models. In this paper, we propose an efficient model that uses MobileNet-v3 [20] as the model backbone network and achieves a good balance between accuracy and efficiency.

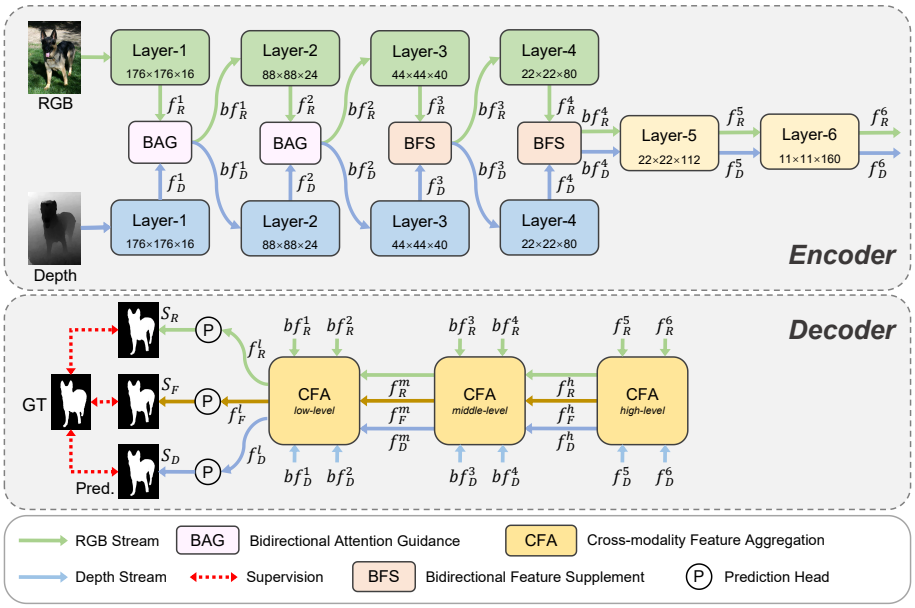

Fig. 2. Overview of our network architecture. Three-stage bidirectional interaction is shown in the upper part of this illustration, and the three stages use the BAG module, BFS module and shared network as the interaction strategy respectively. Three-stage refinement decoder is shown in the lower part of this illustration, and it consists of three CFA modules.

3 Proposed Method

3.1 Overiew

Fig. 2 shows the framework of the proposed three-stage bidirectional interaction network for RGB-D SOD. Our network consists of encoder and decoder. The encoder generates saliency-related features through a three-stage bidirectional interaction strategy, and the decoder aggregates these features and generates the final saliency map. MobileNet-v3 large [20] is used to build the feature encoder. we divide the encoder into six layers, the output stride is 2 for each layer except 1 for the 5^{th} layer, this means that the feature resolution does not change in the 5^{th} layer, so the 5^{th} layer has the same output resolution as the fourth layer. We denote the features output by the i-th layer of the RGB branch and the depth branch as $f_M^i (M \in \{R, D\}, i = 1, ..., 6)$. We take the 1^{st} and 2^{nd} layers as first stage, the 3^{rd} and 4^{th} layers as second stage, and the 5^{th} and 6^{th} layers as thrid stage. We use bidirectional attention guidance (BAG) strategy in first stage and bidirectional feature supplement (BFS) strategy in second stage. The features output by the BAG module or BFS module are denoted as $bf_M^i (M \in \{R, D\}, i = 1, ..., 4)$. After encoding, $bf_M^i (M \in \{R, D\}, i = 1, ..., 4)$ and $f_M^i (M \in \{R, D\}, i = 5, 6)$ are fed into the three-stage refinement decoder. As

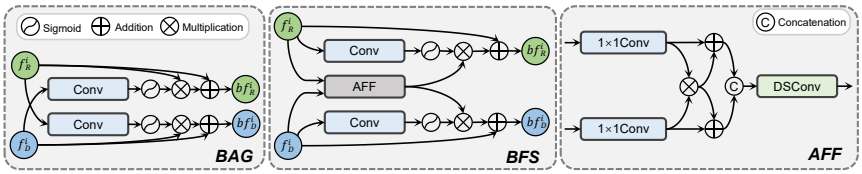

Fig. 3. Illustration of the proposed bidirectional attention guidance (BAG) module, bidirectional feature supplement (BFS) module, and adaptive feature fusion (AFF) module.

shown in Fig. 2, the decoder consists of three cross-modality feature aggregation (CFA) modules, denoted as CFA-high, CFA-middle and CFA-low, respectively.

3.2 Three-Stage Bidirectional Interation (TBI)

The encoder part of Fig. 2 shows the TBI strategy. For the processing of features output by two encoders at different levels, previous works such as SPNet [13], CMWNet [41] and DCFNet [26] tend to fuse the cross-modality features and feed them directly to the decoder. Unlike these works, we process the cross-modality features at each layer and then fed them to the next layer, which enables the use of cross-modality information to improve the network in the feature encoding stage.

First Stage: Bidirectional Attention Guidance (BAG). The detailed structure of the BAG module is shown in Fig. 3(a). The BAG module is based on the spatial attention mechanism. Given the features $f_R^i (i = 1, 2)$ or $f_D^i (i = 1, 2)$, we use a 3×3 convolutional (output channel number is 1) with *Sigmoid* activation function to generate the spatial attention map. To guide one modality to focus on important areas of the other modality. The attention map from one modality is used to enhance another modality. Then, a residual connection is used to combine the enhanced features with their original features. Take the case that depth information enhances RGB information as an example. The process can be described as:

$$bf_R^i = f_R^i + \sigma(Conv_{3\times3}(f_D^i)) \otimes f_R^i, \tag{1}$$

where $Conv_{3\times3}(\cdot)$ denotes a 3×3 convolution, $\sigma(\cdot)$ is the *Sigmoid* actication function, and $\otimes$ represents element-wise multiplication. The features $bf_R^i (i = 1, 2)$ and $bf_D^i (i = 1, 2)$ will be fed into the decoder and the next layer of encoder.

The low-level features of the two modalities have complementary boundary details, so modality specificity should be maintained. The BAG module ensures that the modality specificity is not destroyed while mining more modality correlations.

Second Stage: Bidirectional Feature Supplement (BFS). The detailed structure of the BFS module is shown in Fig. 3(b), an adaptive feature fusion (AFF) module is included in a BFS module. As shown in Fig. 3(c), AFF module is simple and effective, it can adaptively fuse cross-modality features. Given the features $f_R^i(i=3,4)$ and $f_D^i(i=3,4)$, they are first fed into a 1×1 convolution layer with BatchNorm and ReLU activation to adjust their channel number and obtain their smooth feature representations (*i.e.*, $\hat{f_R^i} = Conv_{1\times1}(f_R^i)$ and $\hat{f_D^i} = Conv_{1\times1}(f_D^i)$, where $Conv_{1\times1}(\cdot)$ denotes a 1×1 convolution with BatchNorm and ReLU activation). Then, we use element-wise multiplication to emphasize the shared feature representation, which can be described as $\hat{f_F^i} = \hat{f_R^i} \otimes \hat{f_D^i}$, where $\otimes$ represents element-wise multiplication. We add $\hat{f_F^i}$ with $\hat{f_R^i}$ and $\hat{f_D^i}$ respectively to get the enhanced features. Finally, the enhanced features are concatenated and fed into a depth-wise separable convolution layer to obtain the final fused features, the process can be described as:

$$f_F^i = DSConv_{3\times3}([\hat{f_F^i} + \hat{f_R^i}, \hat{f_F^i} + \hat{f_D^i}]), \tag{2}$$

where $DSConv_{3\times3}(\cdot)$ denotes a 3×3 depth-wise separable convolution with BatchNorm and ReLU activation, and $[\cdot]$ donates feature concatenation. After these operations, the AFF module adaptively fuses cross-modality features. After obtaining the fused features f_F^i, we use the spatial attention mechanism to enhance f_F^i and then combine the enhanced features with the original features of the two modalities. The entire process can be described as:

$$\begin{cases} bf_R^i = f_R^i + \sigma(Conv_{3\times3}(f_R^i)) \otimes f_F^i, \\ bf_D^i = f_D^i + \sigma(Conv_{3\times3}(f_D^i)) \otimes f_F^i, \end{cases} \tag{3}$$

the features $bf_R^i(i=3,4)$ and $bf_D^i(i=3,4)$ will be fed into the decoder and the next layer of encoder.

In a word, we first adaptively fuse cross-modality features to obtain pure fused features, then use spatial attention mechanism to enhance the fused features, and finally transfer them to the two modality branches as supplements. The BFS module can effectively suppress low-quality depth information and transfer high-quality cross-modality shared information between branches.

Third Stage: Shared Network. High-level features have rich global contextual information, which is beneficial for localizing salient objects. The saliency-related high-level features of the two modalities have strong commonality. Inspired by JL-DCF [18], we adopt shared network in the third stage as shown in Fig. 2. Unlike JL-DCF, which uses the strategy of the shared network on the entire feature encoding network, we only share parameters in the most appropriate third stage. Following [18], we concatenate RGB features and depth features in the 4^{th} dimension. The features generated by the 5^{th} and 6^{th} layers of the encoder will be split in the 4^{th} dimension for the decoder. By employing shared network, the two branches share the same parameters in the final stage of

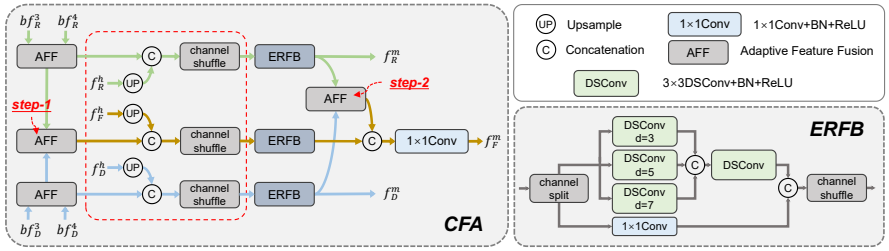

Fig. 4. Illustration of the proposed cross-modality feature aggregation (CFA) module and efficient receptive field block (ERFB) module. The CFA-high module does not contain the part circled by the red dotted line in (a). The two AFF modules pointed to by the red arrows are the two-step cross-modality fusion in the CFA module, AFF modules labelled "step-1" and "step-2" denote "fusion before refinement" and "fusion after refinement", respectively.

the encoder, so the parameters are greatly reduced. Shared network can exploit cross-modality commonality and complementarity, which match the properties of high-level cross-modality features.

3.3 Three-Stage Refinement Decoder

Fig. 2 shows the three-stage refinement decoder, whose three stages correspond one-to-one with the three stages of the encoder. The CFA-high module aggregates high-level cross-modality features, CFA-middle and CFA-low are the same.

Cross-Modality Feature Aggregation (CFA). The detailed structure of the CFA module is shown in Fig. 4(a), which consists of AFF modules and efficient receptive field block (ERFB) modules. There are three branches in the CFA module, namely RGB branch, depth branch, and fusion branch. Take CFA-middle as an example, we first fuse the features bf_M^3 with $bf_M^4 (M \in R, D)$, and obtain the features:

$$\begin{cases} bf_R^m = AFF(bf_R^3, bf_R^4), \\ bf_D^m = AFF(bf_D^3, bf_D^4), \end{cases} \tag{4}$$

where $AFF(\cdot)$ is the AFF module. Note that when fusing features of different layers of the same modality, we added squeeze-and-excitation (SE) modules [42] after the 1×1 convolution layer of the AFF module.

After fusing the features from two levels, we conduct the first-step cross-modality feature fusion ("fusion before refinement"):

$$bf_F^m = AFF(bf_R^m, bf_D^m), \tag{5}$$

we concatenate $bf_M^m (M \in R, D, F)$ with the features $f_M^h (M \in R, D, F)$ generated by the CFA-high of the previous stage. Then, we do channel shuffle [37]

operations and finally feed them into the ERFB modules of the three branches. The outputs of the three ERFB modules in the CFA-middle are defined by:

$$\begin{cases} f_R^m = ERFB([bf_R^m, up(f_R^h)]), \\ f_D^m = ERFB([bf_D^m, up(f_D^h)]), \\ f_F^{m1} = ERFB([bf_F^m, up(f_F^h)]), \end{cases} \tag{6}$$

where $ERFB(\cdot)$ is the ERFB module, and $up(\cdot)$ represents upsample operation. f_R^m and f_D^m are the final refined features of the RGB branch and the depth branch, respectively.

ERFB is a variant of receptive field block (RFB) module [43] as shown in Fig.4(b). It has the basic function of the RFB module and has a lower computational cost. Inspired by ShuffleNet-v2 [19], we use channel split and channel shuffle operations on the ERFB module. Features are divided into two parts in the channel dimension. One half is fed into 1×1 convolution as residuals, and the other half is fed into a dilated convolution block with multiple branches to extract multi-scale features. Finally, we concatenate these two parts and use the channel shuffle operation to ensure information communication between different groups of channels.

We conduct the second-step cross-modality feature fusion ("fusion after refinement"), fusing features f_R^m and f_D^m, and obtain the fused features:

$$f_F^{m2} = AFF(f_R^m, f_D^m), \tag{7}$$

then, the features f_F^{m1} and f_F^{m2} are concatenated, and we use a 1×1 convolution to generate the final fused features of the fusion branch:

$$f_F^m = Conv_{1\times1}([f_F^{m1}, f_F^{m2}]), \tag{8}$$

where $Conv_{1\times1}(\cdot)$ denotes a 1×1 convolution with BatchNorm and ReLU activation. Finally, take the fusion branch as an example, the model generates the final saliency maps:

$$S_F = up(\sigma(Conv_{1\times1}(f_F^l))), \tag{9}$$

where $Conv_{1\times1}(\cdot)$ denotes a 1×1 convolution layer.

Two cross-modality feature fusion steps are included in the CFA module, these two steps serve different purposes. The first step fuses the original cross-modality features and uses them to refine the coarse features generated by the previous stage's fusion branch ("fusion before refinement"), the second step fuses the refined features of the current stage's RGB branch and depth branch ("fusion after refinement"). Finally, cross-modality features are effectively fused and refined.

3.4 Loss Function

We employ the pixel position aware loss $\mathcal{L}_{ppa}$ [44] to implement supervision on the three prediction maps S_F, S_R and S_D.

$$\mathcal{L}_{total} = \mathcal{L}_{ppa}(S_F, G) + \mathcal{L}_{ppa}(S_R, G) + \mathcal{L}_{ppa}(S_D, G), \tag{10}$$

where $\mathcal{L}_{total}$ is the overall loss and G is the ground truth.

Table 1. Quantitative results on seven widely-used datasets. **Red**, **blue** and **bold** indicate the best, second best, and third best performances respectively. ↑/↓ for a metric denotes that a larger/smaller value is better

Model		Non-efficient model											Efficient model				
		D3Net 2020	UCNet 2020	S2MA 2020	BBSNet 2020	JL-DCF 2021	HAINet 2021	CDINet 2021	DCFNet 2021	DSA2F 2021	RD3D 2021	SPNet 2021	DANet 2020	PGAR 2020	MobileSal 2021	DFMNet 2021	TBINet Ours
Params (M) ↓		45.2	31.3	86.6	49.8	70.7	59.8	54.4	108.5	36.5	46.9	175.3	26.7	16.2	**6.5**	**2.2**	**3.7**
CPU (ms) ↓		862	471	2097	633	3136	3019	1585	1069	2288	701	1217	1139	709	**115**	**87**	**329**
GPU (FPS) ↑		52	**99**	25	54	6	9	37	36	21	28	31	46	69	**227**	**299**	93
STERE	$F_\beta^{\max}$ ↑	.891	.899	.882	.903	**.904**	**.906**	.901	.901	.900	**.906**	**.906**	.881	.898	.892	.892	**.910**
	$E_\xi^{\max}$ ↑	.938	.944	.932	.942	**.947**	.944	.942	.945	.942	**.947**	**.949**	.930	.939	.939	.941	**.952**
	S_α ↑	.899	.903	.890	**.908**	.903	.907	.905	.902	.898	**.911**	.907	.892	.907	.901	.898	**.911**
	MAE ↓	.046	.039	.051	.041	.040	.040	.040	**.039**	**.039**	**.037**	**.037**	.048	.041	.042	.045	**.034**
NJU2K	$F_\beta^{\max}$ ↑	.900	.895	.889	.920	.904	.915	**.921**	.915	.907	.914	**.928**	.893	.907	.895	.910	**.928**
	$E_\xi^{\max}$ ↑	.938	.936	.929	.949	.943	.944	**.951**	**.951**	.939	.947	**.957**	.936	.940	.937	.947	**.958**
	S_α ↑	.900	.897	.894	**.921**	.902	.912	**.919**	.912	.904	.916	**.925**	.897	.909	.896	.906	**.924**
	MAE ↓	.047	.043	.054	**.035**	.041	.038	**.035**	.036	.039	.036	**.028**	.047	.042	.045	.042	**.029**
NLPR	$F_\beta^{\max}$ ↑	.897	.903	.902	**.918**	**.918**	.915	.916	.912	.906	**.919**	**.919**	.893	.916	.907	.908	**.932**
	$E_\xi^{\max}$ ↑	.953	.956	.953	.961	**.963**	.960	.960	**.963**	.952	**.965**	.962	.949	.961	.957	.957	**.970**
	S_α ↑	.912	.920	.916	.930	.925	.924	**.927**	.924	.919	**.930**	**.927**	.909	.930	.919	.923	**.937**
	MAE ↓	.030	.025	.030	.023	**.022**	.024	.024	**.022**	.024	**.022**	**.021**	.031	.024	.025	.026	**.018**
SIP	$F_\beta^{\max}$ ↑	.861	.879	.877	.883	.889	**.892**	.884	.884	.875	.889	**.904**	.884	.876	.872	.887	**.905**
	$E_\xi^{\max}$ ↑	.909	.919	.918	.922	.924	.922	.915	.922	.912	.924	**.933**	.920	.915	.911	**.926**	**.937**
	S_α ↑	.860	.875	.872	.879	.880	.880	.875	.876	.862	**.885**	**.894**	.878	.876	.866	.883	**.894**
	MAE ↓	.063	.051	.057	.055	.049	.053	.054	.052	.057	**.048**	**.043**	.054	.055	.058	.051	**.041**
SSD	$F_\beta^{\max}$ ↑	.834	.854	.847	**.859**	.832	.838	.846	.851	**.863**	.772	**.863**	.849	.798	.835	.851	**.872**
	$E_\xi^{\max}$ ↑	.911	.907	.909	**.919**	.902	.903	.899	.909	.913	.859	**.920**	.905	.872	.905	.918	**.921**
	S_α ↑	.857	.865	.868	**.882**	.860	.857	.853	.864	**.877**	.803	.871	.868	.832	.861	.865	**.874**
	MAE ↓	.059	.049	.053	**.044**	.053	.052	.056	.050	**.048**	.082	**.044**	.050	.068	.053	.051	**.042**
DES	$F_\beta^{\max}$ ↑	.884	.930	**.934**	.927	.923	**.936**	**.934**	.893	.915	.929	**.946**	.894	.902	.899	.922	**.934**
	$E_\xi^{\max}$ ↑	.945	**.976**	.973	.966	.968	.973	.970	.951	.954	.972	**.983**	.957	.945	.951	.972	**.974**
	S_α ↑	.897	.934	**.940**	.934	.931	.935	**.937**	.905	.917	.935	**.945**	.904	.913	.909	.931	.935
	MAE ↓	.031	**.018**	.021	.021	.020	**.018**	.020	.024	.023	**.019**	**.014**	.029	.026	.025	.021	**.018**

4 Experiments

4.1 Experimental Setup

Datasets and Evaluation Metrics. We evaluate the proposed model on six widely-used datasets to validate its effectiveness, inculding STERE [45], NJU2K [46], NLPR [47], SIP [29], SSD [48] and DES [49]. Following previous works [8, 13, 29], we use 1,485 samples of NJU2K [46] and 700 samples of NLPR [47] for training, and the remaining samples of NJU2K (500) and NLPR (300) for testing. The datasets of STERE (1,000), SIP (929), SSD (80), and DES (135) are used for testing.

We employ four metrics to evaluate various methods, including maximum F-measure ($F_\beta^{\max}$) [50], maximum E-measure ($E_\xi^{\max}$) [51], S-measure (S_α) [52], and mean absolute error (MAE) [53]. Model parameters, CPU inference time (ms, millisecond) and GPU inference FPS (frame-per-second) are used to evaluate the efficiency of the model.

Implementation Details. We implement our model in PyTorch [54]. Parameters of the backbone network (MobileNet-v3 large [20]) are initialized from the model pre-trained on ImageNet [55]. RGB and depth images are both resized to 352×352 for input. We use a single Nvidia Tesla P100-16GB for training and testing and Intel Xeon (4) @2.199GHz for CPU inference speed test. The training images are augmented using various strategies, including random flipping, rotating, colour enhancement, and border clipping. The initial learning rate is

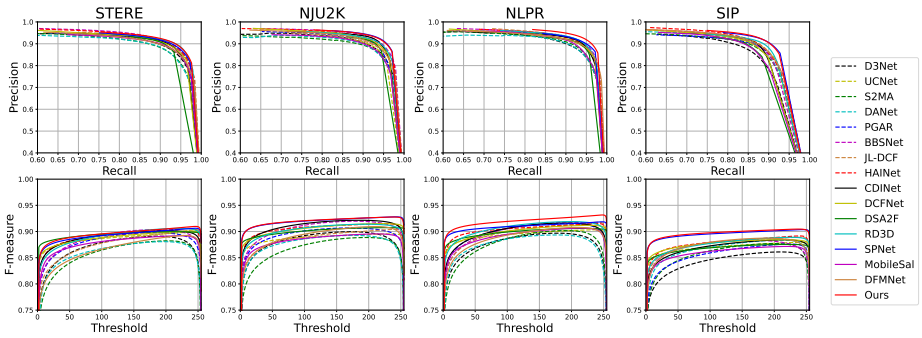

Fig. 5. PR curves [56] and F-measure curves on STERE [45], NJU2K [46], NLPR [47], and SIP [29].

set to 1e-4 and is divided by 5 every 60 epochs. The Adam optimizer is used, and the batch size is 10. It takes about 5 hours to train our model for 160 epochs.

4.2 Comparisons with SOTA Methods

Quantitative Evaluation. We compare the proposed method with 15 RGB-D SOD methods, including 11 non-efficient models (*i.e.*, D3Net [29], UCNet [57], S2MA [58], BBSNet [8], JL-DCF [18], HAINet [14], CDINet [17], DCFNet [26], DSA2F [27], RD3D [25], and SPNet [13]), and 4 efficient models (*i.e.*, DANet [33], PGAR [34], MobileSal [40], and DFMNet [15]). As shown in Table 1, our method outperforms all of the comparison state-of-the-art methods. On the STERE, NLPR, and SIP datasets, our method achieves the best performance on all four evaluation metrics. Our model outperforms most compared RGB-D SOD methods on the NJU2K and SSD datasets except SPNet and BBSNet. As shown in Fig. 5, we plot the PR curves [56] and F-measure curves. For readability, We chose the larger four datasets of the six datasets. The comparison method is still the 15 methods mentioned earlier. In terms of efficiency, among all the compared methods, our method ranks 2^{nd}, 3^{rd} and 4^{th} in model parameters, CPU inference speed, and GPU inference speed, respectively, and is more efficient than most of the compared methods. Overall, our RGB-D SOD method (TBINet) achieves promising performance and efficiency.

Qualitative Evaluation. Fig. 6 shows the saliency maps predicted by the proposed method and several state-of-the-art methods on six representative examples. The first row shows a simple example with a single salient object but some misleading information in the depth map. The salient objects predicted by our method, MobileSal, RD3D, and CDINet, have complete boundary details, while results of other methods appear smeared and incomplete to varying degrees. The 2^{nd} and 3^{rd} rows show multiple salient objects, and it is not easy to detect all salient objects accurately. Only our method, DFMNet, and S2MA

Fig. 6. Visual comparisons of our method (TBINet) with SOTA methods including DFMNet [15],MobileSal [40], SPNet [13], RD3D [25], DSA2F [27], DCFNet [26], CDINet [17], HAINet [14], JL-DCF [18], PGAR [34], S2MA [58], D3Net [29].

can completely detect three salient objects in the second row. At the same time, other compared models have missing objects or incomplete segmentation, and the third row is similar. The 4^{th} row shows a salient object with a complex structure. Thanks to the clear depth map, most methods achieve good results. However, some compared methods make poor use of depth information and confuse the background as a salient object. The 5^{th} row shows the low-contrast scene, and it can be observed that our model segments salient objects sharply. The 6^{th} row shows a scene with complex textures. In this example, the depth map is ambiguous. Our model is not misled by low-quality depth information, and accurately locates salient objects.

4.3 Ablation Studies

To verify the effectiveness of the modules and strategies we use in the model, we conduct ablation studies by removing or replacing relevant modules from the full model and reformulating the strategies. We conduct experiments on NJU2K dataset and NLPR dataset.

Table 2. The effectiveness analyses of TBI strategy.

Strategy		**Ours**	A1	A2	A3	A4	B1	B2	B3	B4	C1	C2	D1	D2	E1	E2	E3	E4
First Stage		BAG		UAG-r	UAG-d	BFS	BAG	BAG	BAG	BAG	BAG	BAG	UAG-r	UAG-d		BAG	BFS	shared
Second Stage		BFS	BFS	BFS	BFS	BFS		UFS-r	UFS-d	BAG	BFS	BFS	UFS-r	UFS-d		BAG	BFS	shared
Thrid Stage		shared	shared	shared	shared	shared	shared	shared	shared	shared		BFS	shared	shared		BAG	BFS	shared
Param (M)		3.7	3.7	3.7	3.7	3.8	3.7	3.7	3.7	3.7	6.4	6.5	3.7	3.7	6.3	6.3	6.5	3.5
NJU2K	$F_\beta^{\max}\uparrow$	**.928**	.926	.925	.925	.927	.920	.921	.926	.925	.926	.924	.921	.926	.919	.923	.924	.914
	MAE $\downarrow$	**.029**	**.029**	.031	.030	.030	.032	.033	.031	.032	**.029**	.031	.033	.031	.033	.032	.030	.034
NLPR	$F_\beta^{\max}\uparrow$	**.932**	.929	.931	.929	.926	.923	.923	.929	.927	**.932**	.928	.923	.929	.929	**.932**	.928	.924
	MAE $\downarrow$	**.018**	.020	.019	.019	.021	.022	.021	.019	.021	.019	.019	.021	.019	.020	.019	.019	.021

Effectiveness of TBI Strategy. Our three-stage interaction strategy uses BAG, BFS, and shared network strategies in the first, second and third stages of the encoding process, respectively. For first stage, we first remove the BAG module, this evaluation is denoted as 'A1' in Table 2. Then, we replace the BAG modules with unidirectional attention guidance (UAG) modules. The RGB-enhanced UAG module is abbreviated as UAG-r, and the depth-enhanced UAG module is abbreviated as UAG-d. We denote the RGB-enhanced and depth-enhanced strategies as 'A2' and 'A3', respectively. Finally, the replacement of the BAG module with the BFS module is denoted as 'A4'. Table 2 shows that the BAG module is effective in guiding the network to learn cross-modality correlations. second stage is similar to first stage, we compare the BFS module with four baselines: removing the BFS module (denoted as 'B1'), replacing the BFS module with a unidirectional feature supplement (UFS) module (denoted as 'B2' and 'B3'), and replacing the BFS module with a BAG module (denoted as 'B4'). Table 2 shows the effectiveness of the BFS module. It is worth noting that the performance of 'B3' is significantly better than that of 'B2' and 'B1', which shows that feature supplement to the deep branch can improve the performance very well. The deep reason may be that the BFS module reduces the interference of low-quality depth information. For thrid stage, we do not use shared network strategy (denoted as 'C1') or change to use the BFS module (denoted as 'C2'), the performance of 'C1' is not much different from our strategy, but the parameters are much more. We also changed the BAG module and BFS module to unidirectional interaction at the same time (denoted as 'D1' and 'D2'). These strategies have gaps compared with our strategy. Finally, we evaluate the cases of using the same interaction strategy in all stages (denoted as 'E1', 'E2', 'E3', and 'E4'). As shown in the Table 2, our three-stage bidirectional interaction strategy outperforms the ordinary bidirectional interaction strategy.

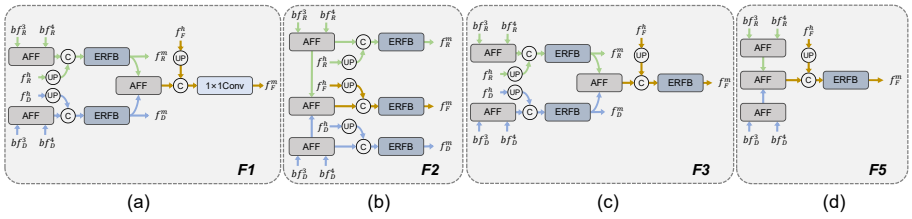

Fig. 7. Illustration of other feature aggregation strategies compared with the CFA module. Channel shuffle operations are hidden for a clear view.

Effectiveness of CFA Module. The CFA module is proposed to aggregate and refine cross-modality features, which adopts a two-step feature fusion and refinement. To verify the effectiveness of the CFA module, we evaluate some different cross-modality feature aggregation strategies, as shown in Fig. 7. We

Table 3. The effectiveness analyses of CFA module.

Strategy		**Ours**	F1	F2	F3	F4	F5	F6
Param(M)		3.7	3.5	3.7	3.7	3.7	3.7	3.6
NJU2K	$F_\beta^{\max}$ ↑	**.928**	.927	.925	.924	.917	.919	.921
	MAE ↓	**.029**	.030	.031	.031	.052	.032	.032
NLPR	$F_\beta^{\max}$ ↑	**.932**	.928	.931	.930	.914	.923	.923
	MAE ↓	**.018**	.020	.020	.019	.037	.021	.021

first remove the step of "fusion before refinement" (denoted as 'F1') as shown in Fig. 7(a) or remove the step of "fusion after refinement" (denoted as 'F2') as shown in Fig. 7(b). Table 3 shows that the performance of 'F1' and 'F2' is reduced to varying degrees. Our proposed two-step cross-modality fusion strategy can better fuse and refine cross-modality features. We formulate a strategy denoted as 'F3' as shown in Fig. 7(c): we fuse the refined features of the RGB branch with the refined features of the depth branch and then use the fused features to refine the features generated by the previous stage's fusion branch. The result shows that our strategy outperforms this "refinement-by-refinement" strategy. We remove the supervision of the saliency maps generated by the RGB branch and the depth branch (denoted as 'F4'), the result shows that the supervision of the RGB branch and the depth branch is effective. The direct removal of the RGB branch and the depth branch in the CFA module is denoted as 'F5' as shown in Fig. 7(d), and Table 3 shows the effectiveness of the three branches in the CFA module. The above evaluation of different cross-modality feature aggregation strategies can conclude that our CFA module can effectively aggregate and refine cross-modality features and generate more accurate saliency maps.

5 Conclusion

We propose a three-stage bidirectional interaction network for RGB-D SOD. Existing works have not explored the relationship between cross-modality features of various levels. Our model employs appropriate interaction strategies at different stages of the encoding process to generate more robust feature representations. In addition, our proposed cross-modality feature aggregation module can effectively aggregate and refine saliency-oriented features to generate accurate saliency maps. Evaluations on six benchmark datasets show promising performance of our TBINet. Our model is lightweight and efficient, which may help the application of RGB-D SOD on mobile devices.

References

1. Shimoda, W., Yanai, K.: Distinct class-specific saliency maps for weakly supervised semantic segmentation. In: European conference on computer vision. pp. 218–234. Springer (2016)
2. Zeng, Y., Zhuge, Y., Lu, H., Zhang, L.: Joint learning of saliency detection and weakly supervised semantic segmentation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 7223–7233 (2019)

3. Hong, S., You, T., Kwak, S., Han, B.: Online tracking by learning discriminative saliency map with convolutional neural network. In: International conference on machine learning. pp. 597–606. PMLR (2015)
4. Mahadevan, V., Vasconcelos, N.: Saliency-based discriminant tracking. In: 2009 IEEE conference on computer vision and pattern recognition. pp. 1007–1013. IEEE (2009)
5. Guo, C., Zhang, L.: A novel multiresolution spatiotemporal saliency detection model and its applications in image and video compression. IEEE transactions on image processing **19**(1), 185–198 (2009)
6. Ji, Q.G., Fang, Z.D., Xie, Z.H., Lu, Z.M.: Video abstraction based on the visual attention model and online clustering. Signal Processing: Image Communication **28**(3), 241–253 (2013)
7. Cheng, M.M., Hou, Q.B., Zhang, S.H., Rosin, P.L.: Intelligent visual media processing: When graphics meets vision. Journal of Computer Science and Technology **32**(1), 110–121 (2017)
8. Fan, D.P., Zhai, Y., Borji, A., Yang, J., Shao, L.: Bbs-net: Rgb-d salient object detection with a bifurcated backbone strategy network. In: European Conference on Computer Vision. pp. 275–292. Springer (2020)
9. Jin, W.D., Xu, J., Han, Q., Zhang, Y., Cheng, M.M.: Cdnet: Complementary depth network for rgb-d salient object detection. IEEE Transactions on Image Processing **30**, 3376–3390 (2021)
10. Liu, Z., Wang, Y., Tu, Z., Xiao, Y., Tang, B.: Tritransnet: Rgb-d salient object detection with a triplet transformer embedding network. In: Proceedings of the 29th ACM International Conference on Multimedia. pp. 4481–4490 (2021)
11. Luo, A., Li, X., Yang, F., Jiao, Z., Cheng, H., Lyu, S.: Cascade graph neural networks for rgb-d salient object detection. In: European Conference on Computer Vision. pp. 346–364. Springer (2020)
12. Piao, Y., Rong, Z., Zhang, M., Ren, W., Lu, H.: 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. pp. 9060–9069 (2020)
13. Zhou, T., Fu, H., Chen, G., Zhou, Y., Fan, D.P., Shao, L.: Specificity-preserving rgb-d saliency detection. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 4681–4691 (2021)
14. Li, G., Liu, Z., Chen, M., Bai, Z., Lin, W., Ling, H.: Hierarchical alternate interaction network for rgb-d salient object detection. IEEE Transactions on Image Processing **30**, 3528–3542 (2021)
15. Zhang, W., Ji, G.P., Wang, Z., Fu, K., Zhao, Q.: Depth quality-inspired feature manipulation for efficient rgb-d salient object detection. In: Proceedings of the 29th ACM International Conference on Multimedia. pp. 731–740 (2021)
16. Zhang, W., Jiang, Y., Fu, K., Zhao, Q.: Bts-net: Bi-directional transfer-and-selection network for rgb-d salient object detection. In: 2021 IEEE International Conference on Multimedia and Expo (ICME). pp. 1–6. IEEE (2021)
17. Zhang, C., Cong, R., Lin, Q., Ma, L., Li, F., Zhao, Y., Kwong, S.: Cross-modality discrepant interaction network for rgb-d salient object detection. In: Proceedings of the 29th ACM International Conference on Multimedia. pp. 2094–2102 (2021)
18. Fu, K., Fan, D.P., Ji, G.P., Zhao, Q., Shen, J., Zhu, C.: Siamese network for rgb-d salient object detection and beyond. IEEE Transactions on Pattern Analysis and Machine Intelligence (2021)

19. Ma, N., Zhang, X., Zheng, H.T., Sun, J.: Shufflenet v2: Practical guidelines for efficient cnn architecture design. In: Proceedings of the European conference on computer vision (ECCV). pp. 116–131 (2018)
20. Howard, A., Sandler, M., Chu, G., Chen, L.C., Chen, B., Tan, M., Wang, W., Zhu, Y., Pang, R., Vasudevan, V., et al.: Searching for mobilenetv3. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 1314–1324 (2019)
21. Desingh, K., Krishna, K.M., Rajan, D., Jawahar, C.: Depth really matters: Improving visual salient region detection with depth. In: BMVC. pp. 1–11 (2013)
22. Feng, D., Barnes, N., You, S., McCarthy, C.: Local background enclosure for rgb-d salient object detection. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2343–2350 (2016)
23. Lang, C., Nguyen, T.V., Katti, H., Yadati, K., Kankanhalli, M., Yan, S.: Depth matters: Influence of depth cues on visual saliency. In: European conference on computer vision. pp. 101–115. Springer (2012)
24. Ren, J., Gong, X., Yu, L., Zhou, W., Ying Yang, M.: Exploiting global priors for rgb-d saliency detection. In: Proceedings of the IEEE conference on computer vision and pattern recognition workshops. pp. 25–32 (2015)
25. Chen, Q., Liu, Z., Zhang, Y., Fu, K., Zhao, Q., Du, H.: Rgb-d salient object detection via 3d convolutional neural networks. In: Proceedings of the AAAI Conference on Artificial Intelligence. pp. 1063–1071 (2021)
26. Ji, W., Li, J., Yu, S., Zhang, M., Piao, Y., Yao, S., Bi, Q., Ma, K., Zheng, Y., Lu, H., et al.: Calibrated rgb-d salient object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9471–9481 (2021)
27. Sun, P., Zhang, W., Wang, H., Li, S., Li, X.: Deep rgb-d saliency detection with depth-sensitive attention and automatic multi-modal fusion. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1407–1417 (2021)
28. Zhang, M., Fei, S.X., Liu, J., Xu, S., Piao, Y., Lu, H.: Asymmetric two-stream architecture for accurate rgb-d saliency detection. In: European Conference on Computer Vision. pp. 374–390. Springer (2020)
29. Fan, D.P., Lin, Z., Zhang, Z., Zhu, M., Cheng, M.M.: Rethinking rgb-d salient object detection: Models, data sets, and large-scale benchmarks. IEEE Transactions on neural networks and learning systems **32**(5), 2075–2089 (2020)
30. Li, C., Cong, R., Piao, Y., Xu, Q., Loy, C.C.: Rgb-d salient object detection with cross-modality modulation and selection. In: European Conference on Computer Vision. pp. 225–241. Springer (2020)
31. Zhou, T., Fan, D.P., Cheng, M.M., Shen, J., Shao, L.: Rgb-d salient object detection: A survey. Computational Visual Media **7**(1), 37–69 (2021)
32. Wang, W., Lai, Q., Fu, H., Shen, J., Ling, H., Yang, R.: Salient object detection in the deep learning era: An in-depth survey. IEEE Transactions on Pattern Analysis and Machine Intelligence (2021)
33. Zhao, X., Zhang, L., Pang, Y., Lu, H., Zhang, L.: A single stream network for robust and real-time rgb-d salient object detection. In: European Conference on Computer Vision. pp. 646–662. Springer (2020)
34. Chen, S., Fu, Y.: Progressively guided alternate refinement network for rgb-d salient object detection. In: European Conference on Computer Vision. pp. 520–538. Springer (2020)
35. Howard, A.G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., Andreetto, M., Adam, H.: Mobilenets: Efficient convolutional neural networks for mobile vision applications. arXiv preprint arXiv:1704.04861 (2017)

36. Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., Chen, L.C.: Mobilenetv2: Inverted residuals and linear bottlenecks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 4510–4520 (2018)
37. Zhang, X., Zhou, X., Lin, M., Sun, J.: Shufflenet: An extremely efficient convolutional neural network for mobile devices. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 6848–6856 (2018)
38. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014)
39. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
40. Wu, Y.H., Liu, Y., Xu, J., Bian, J.W., Gu, Y.C., Cheng, M.M.: Mobilesal: Extremely efficient rgb-d salient object detection. IEEE Transactions on Pattern Analysis and Machine Intelligence (2021). https://doi.org/10.1109/TPAMI.2021.3134684
41. Li, G., Liu, Z., Ye, L., Wang, Y., Ling, H.: Cross-modal weighting network for rgb-d salient object detection. In: European Conference on Computer Vision. pp. 665–681. Springer (2020)
42. Hu, J., Shen, L., Sun, G.: Squeeze-and-excitation networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 7132–7141 (2018)
43. Liu, S., Huang, D., et al.: Receptive field block net for accurate and fast object detection. In: Proceedings of the European conference on computer vision (ECCV). pp. 385–400 (2018)
44. Wei, J., Wang, S., Huang, Q.: F^3net: fusion, feedback and focus for salient object detection. In: Proceedings of the AAAI Conference on Artificial Intelligence. pp. 12321–12328 (2020)
45. Niu, Y., Geng, Y., Li, X., Liu, F.: Leveraging stereopsis for saliency analysis. In: 2012 IEEE Conference on Computer Vision and Pattern Recognition. pp. 454–461. IEEE (2012)
46. Ju, R., Ge, L., Geng, W., Ren, T., Wu, G.: Depth saliency based on anisotropic center-surround difference. In: 2014 IEEE international conference on image processing (ICIP). pp. 1115–1119. IEEE (2014)
47. Peng, H., Li, B., Xiong, W., Hu, W., Ji, R.: Rgbd salient object detection: A benchmark and algorithms. In: European conference on computer vision. pp. 92–109. Springer (2014)
48. Zhu, C., Li, G.: A three-pathway psychobiological framework of salient object detection using stereoscopic technology. In: Proceedings of the IEEE International Conference on Computer Vision Workshops. pp. 3008–3014 (2017)
49. Cheng, Y., Fu, H., Wei, X., Xiao, J., Cao, X.: Depth enhanced saliency detection method. In: Proceedings of international conference on internet multimedia computing and service. pp. 23–27 (2014)
50. Achanta, R., Hemami, S., Estrada, F., Susstrunk, S.: Frequency-tuned salient region detection. In: 2009 IEEE conference on computer vision and pattern recognition. pp. 1597–1604. IEEE (2009)
51. Fan, D.P., Gong, C., Cao, Y., Ren, B., Cheng, M.M., Borji, A.: Enhanced-alignment measure for binary foreground map evaluation. In: Proceedings of the 27th International Joint Conference on Artificial Intelligence. pp. 698–704 (2018)
52. Fan, D.P., Cheng, M.M., Liu, Y., Li, T., Borji, A.: Structure-measure: A new way to evaluate foreground maps. In: Proceedings of the IEEE international conference on computer vision. pp. 4548–4557 (2017)

53. Perazzi, F., Krähenbühl, P., Pritch, Y., Hornung, A.: Saliency filters: Contrast based filtering for salient region detection. In: 2012 IEEE conference on computer vision and pattern recognition. pp. 733–740. IEEE (2012)
54. Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., et al.: Pytorch: An imperative style, high-performance deep learning library. Advances in neural information processing systems **32** (2019)
55. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. Advances in neural information processing systems **25** (2012)
56. Borji, A., Cheng, M.M., Jiang, H., Li, J.: Salient object detection: A benchmark. IEEE transactions on image processing **24**(12), 5706–5722 (2015)
57. Zhang, J., Fan, D.P., Dai, Y., Anwar, S., Saleh, F.S., Zhang, T., Barnes, N.: Uc-net: Uncertainty inspired rgb-d saliency detection via conditional variational autoencoders. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 8582–8591 (2020)
58. Liu, N., Zhang, N., Han, J.: Learning selective self-mutual attention for rgb-d saliency detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 13756–13765 (2020)