Multimedia Tools and Applications
https://doi.org/10.1007/s11042-022-12799-y

BGRDNet: RGB-D salient object detection with a bidirectional gated recurrent decoding network

Zhengyi Liu[1] · Yuan Wang[1] · Zhili Zhang[1] · Yacheng Tan[1]

Received: 26 July 2021 / Revised: 18 October 2021 / Accepted: 25 January 2022

Abstract

Traditional U-Net framework generates multi-level features by the successive convolution and pooling operations, and then decodes the saliency cue by progressive upsampling and skip connection. The multi-level features are generated from the same input source, but quite different with each other. In this paper, we explore the complementarity among multi-level features, and decode them by Bi-GRU. Since multi-level features are different in the size, we first propose scale adjustment module to organize multi-level features into sequential data with the same channel and resolution. The core unit SAGRU of Bi-GRU is then devised based on self-attention, which can effectively fuse the history and current input. Based on the designed SAGRU, we further present the bidirectional decoding fusion module, which decoding the multi-level features in both down-top and top-down manners. The proposed bidirectional gated recurrent decoding network is applied in the RGB-D salient object detection, which leverages the depth map as a complementary information. Concretely, we put forward depth guided residual module to enhance the color feature. Experimental results demonstrate our method outperforms the state-of-the-art methods in the six popular benchmarks. Ablation studies also verify each module plays an important role.

Keywords Salient object detection · RGB-D · Bi-GRU · Depth guided · Self-attention

✉ Zhengyi Liu
liuzywen@ahu.edu.cn

Yuan Wang
wangyuan.ahu@qq.com

Zhili Zhang
528419003@qq.com

Yacheng Tan
tan.yacheng@qq.com

[1] Key Laboratory of Intelligent Computing and Signal Processing of Ministry of Education, School of Computer Science and Technology, Anhui University, Hefei, China

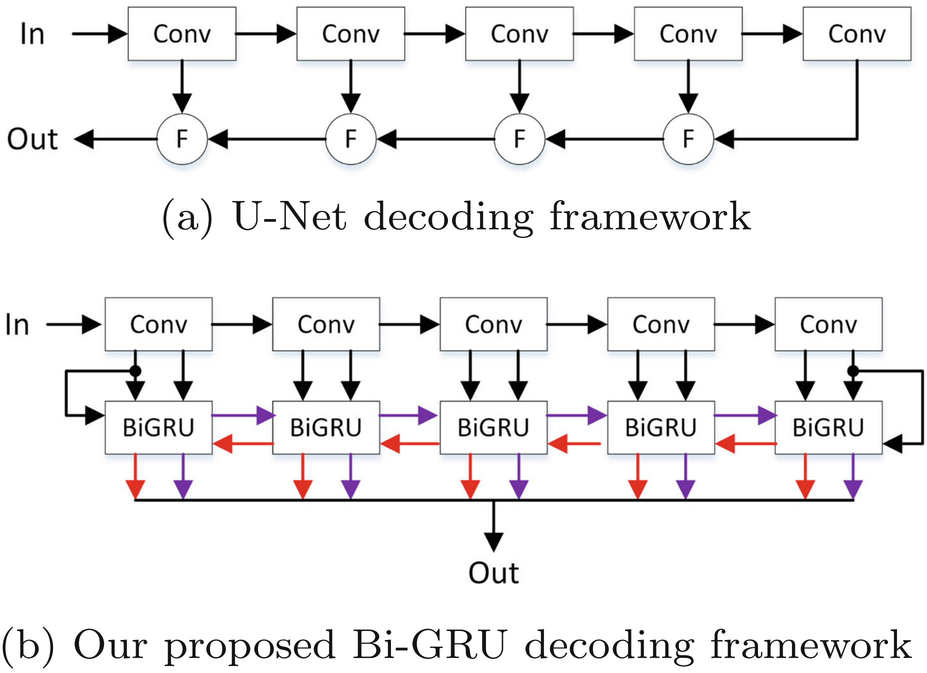
Published online: 23 March 2022

1 Introduction

Salient object detection simulates visual attention mechanism to capture prominent object in a scene. It has been widely applied in computer vision tasks, such as image segmentation [28, 77], tracking [84], retargeting [2], cropping [76], quality assessment [32], video object segmentation [68] and activity prediction [72].

As a pixel-level dense prediction task, salient object detection usually uses U-Net framework [61] shown in Fig. 1(a) which recover the image size by progressive 2×upsampling and communication with the encoder by skip connections. Multi-level features extracted from the encoder show the different attributions. High-layer feature reveals more semantic information but lack of accurate detailed information, while shallow-layer feature exhibits more spatial details but full of noises. Both are different with each other but complementary. The combination among multi-level features can comprehensively consider their respective advantages and improve the accuracy of detection. But the traditional U-Net framework [61] uses element-addition or concatenation fusion without discrimination, and inevitably brings some noises to the result. Therefore, in order to better fuse multi-level features, we adopt the bidirectional gated recurrent unit (Bi-GRU) to decode the multi-level features, which is shown in Fig. 1(b). It can adaptively memorize the useful information and discard the irrelevant part to find the common information which is hidden in the multi-level features, achieving the better fusion by learning long-range dependencies across levels.

There are two important issues which should be solved. Bi-GRU is a particular type of recurrent neural networks (RNN) [18]. It is good at processing the sequential data with the same scales. Multi-level features are quite different from each other in the scales due to



(a) U-Net decoding framework

(b) Our proposed Bi-GRU decoding framework

Fig. 1 Comparison between U-Net framework and our proposed framework

the repeated stride and pooling operations of backbone network. How to adjust the multi-level features adaptive to the input of Bi-GRU is the first question. Simple and straight upsampling with large ratio is certainly not the best way. We design scale adjustment module to solve the problem. Another question is the structure design of gated recurrent unit. Since our idea is to take the multi-level features as the sequential data, and then decode them by Bi-GRU, the design of each gated recurrent unit is so important as to better combine the advantages of multi-level features to achieve the perfect fusion. We introduce self-attention into gated recurrent unit to solve the problem.

In addition, with the widespread use of depth camera, salient object detection for RGB-D image attracts a lot of attention. Depth image is the supplement to the color image for accurate detection, especially in some challenging and complex scenarios, e.g. the low color contrasts between salient objects and background, the cluttered background. There is two main supplementary ways [30, 58]. One adopts two-stream architecture that processes RGB and depth images separately with various cross-modal fusion strategies [8, 10, 15, 16, 29, 31, 38–41, 44, 65, 81–83, 86, 91, 95]. The other utilizes subnetworks tailored for depth image to compensate for RGB representations [13, 14, 22, 52, 58, 63, 78, 89, 96]. Following the latter approach, we design depth guided color encoder, which uses depth information to enhance the color features which will be viewed as sequential data and fed into our proposed decoder.

Our main contributions can be summarized as follows:

- A bidirectional gated recurrent decoding network is proposed for salient object detection in RGB-D image. The multi-level features with different size, which are extracted from backbone network, are viewed as sequential data and fed into bidirectional gated recurrent unit to achieve parallel and recurrent decoding and finally generate accurate saliency map.
- In bidirectional gated recurrent decoding process, scale adjustment module is proposed to adjust the channel and resolution of features, and avoid excessive upsampling. Meanwhile, self-attention gated recurrent unit is designed to better fuse multi-level features by gate structure and memory unit.
- Depth image is viewed as the supplement to color feature, and attached to color feature to enhance the feature representation by depth guided residual learning which introduces spatial attention and channel attention.
- The proposed network is validated on six public datasets, including NLPR [54], NJU2K [33], STERE [48], DES [17], SIP [21] and DUT [57]. The experimental results show that our network exhibits the excellent performance on these RGB-D saliency datasets.

2 Related work

2.1 RGB-D saliency detection

Salient object detection captures the prominent object in the image. The cue of color image is generally not enough to accurately find the salient objects, especially in the challenging scenes [4]. Depth image as a complementary modality is widely used to enhance the color image because it can provide a preponderance of discriminative power in location and spatial structure.

Early RGB-D saliency detection methods use original single dimensional depth value [50] or design hand-crafted features (e.g., depth contrast [17, 54, 60, 66], ACSD [34], LBE [24], HOSO [23], etc). With the wide application of convolution neural network (CNN), CNN based methods [93] have achieved the impressive progress.

The important issue of RGB-D saliency detection is the fusion of color feature and depth feature. Early fusion [12, 45, 88, 92], middle fusion and late fusion [15, 46, 65] are three classic frameworks. There are two main middle fusion ways. One is two-stream model, and the other is depth guided model. PCFN [10], PGAR [13], ICNet [40], CMMS [39] and ASIF-Net [38] are classic two-stream U-Net framework. PDNet [96], CFPF [89], A2dele [58], MobileSal [74] and HDFNet [53] are classic depth guided network which uses depth information to enhance RGB stream.

Since the depth images are not always with high-quality, many pioneering works tries to solve the problem. For example, D3Net [21] uses gate mechanism to filter the poor depth map, EF-Net [11] enhances the depth maps by color hint map, DQSD [8] integrates a depth quality aware subnet into the classic bi-stream structure, assigning the weight of depth feature before conducting the selective RGB-D fusion. In addition, CoNet [30], DASNet [90], SSDP [70] and MobileSal [74] introduce depth estimation, learning to detect the salient object simultaneously.

In the paper, we adopt depth guided manner. Depth information is viewed as the supplement to the color feature. It enhances the color feature by attention mechanism and residual connection.

2.2 Decoding strategy

In the traditional decoding process, the features of high layers are progressively transmitted and fused with the features of shallow layers by element-wise addition [27, 53, 85], concatenation [38, 40, 41, 59], fluid pyramid integration [29, 89], dense aggregation [25, 26, 49], gated fusion [55, 83, 94], bi-directional fusion [42, 87], residual fusion [13, 88], cascaded partial fusion [22, 75] or other fusion operations [39]. The semantic information of deep layer may be gradually diluted in the progressive fusion process, and low-quality depth image also may make the fused result deviate from the right direction.

Gated recurrent unit (GRU) is a particular type of recurrent neural networks (RNN). It is originally used for machine translation [18], and is later developed into a convolutional version for video representation [3]. Its gate mechanism and memory unit ensure that current output depends on history memory and current input, which can get the comprehensive judgement by long-range dependence.

In the salient object detection, GRU has been widely used [5, 6, 36, 37, 51, 57, 62, 67, 79]. But the gated recurrent structure is applied in the one feature or two features with the same layer. Different from them, our method will decode the multi-level features with different size from backbone network.

3 Proposed method

3.1 Overview

The overall framework of the proposed bidirectional gated recurrent decoding Network (BGRDNet) is depicted in Fig. 2, which consists of depth guided color encoder and bidirectional gated recurrent decoder. The details can be seen in the following sections.

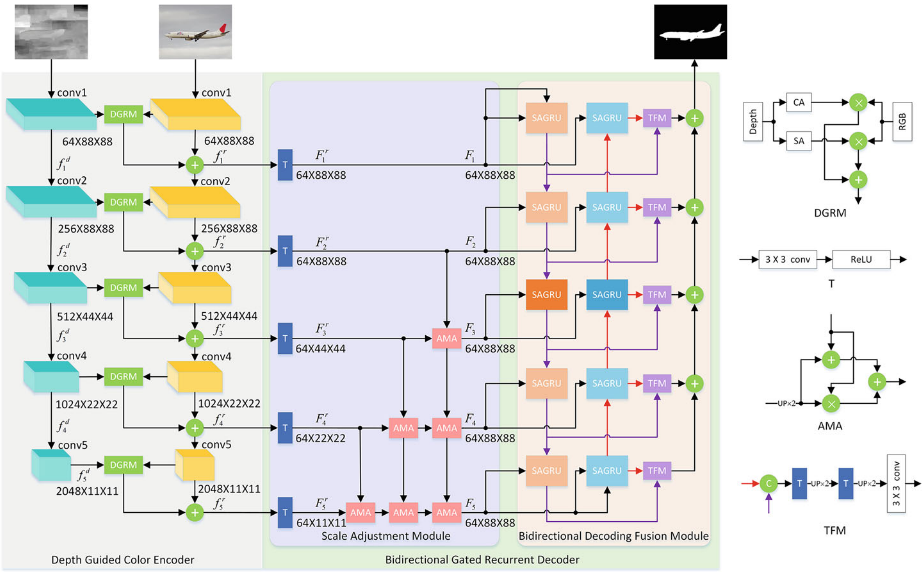


Fig. 2 Main framework

3.2 Depth guided color encoder

Depth information is proved to supply the useful cues and boost the performance for saliency detection [50]. But depth image with poor quality, which likes a noise, brings some negative influences [21]. Therefore, we propose depth guided color encoder which enhances the color feature by dual attention maps of depth feature. To be specific, two ResNet-50 backbone networks are used to extract multi-level features of RGB image and depth image, which are denoted as $f_i'^r$ and $f_i^d (i = 1, 2, \cdots, 5)$, respectively. In order to attach the useful depth information to color feature, we introduce residual connection [27] to enhance the hierarchical color features. Residual part is designed as depth guided residual module (DGRM), and shortcut connection part is used to preserve more original color information. In DGRM, depth feature is fed into two parallel attention modules, then a channel attention map and a spatial attention map are obtained to show what and where are the informative part of depth feature. Next, two attention maps are served as the weight, and multiplied with color feature to achieve the optimization of color feature from channel and spatial aspect. At last two optimized color features are further fused by element-wise addition operation to retain two kinds of information. Thus the color feature guided by the depth feature is updated by the following formula:

$$f_i^r = (CA(f_i^d) \times f_i'^r + SA(f_i^d) \times f_i'^r) + f_i'^r \tag{1}$$

where $CA(\cdot)$ and $SA(\cdot)$ denote the channel and spatial attention module in the CBAM [73], “×” is element-wise multiplication operation, “+” is element-wise addition operation.

3.3 Bidirectional gated recurrent decoder

In the traditional U-Net decoding process, the multi-level features are overlaid one by one from high layer to low layer by the element-wise addition [27, 53, 85] or concatenation operation [38, 40, 41, 43, 59]. The semantic information in the high layer is progressively diluted in the data transmission flow, and inevitable mixed the noise in the shallow layer. To better fuse the multi-level features, we propose the bidirectional gated recurrent decoder, which can parallel and recurrently decode multi-level features with the different scales in the down-top and top-down manners. The process of down-top manner can boost the spatial information in the low layer by the semantic cue in the high layer, and the process of top-down manner can enhance the semantic feature representation in the high layer by the spatial information in the low layer. Both recurrent decoding processes promote each other. The proposed decoder consists of scale adjustment module and bidirectional decoding fusion module. The detail is described as follows.

3.3.1 Scale adjustment module (SAM)

Gated recurrent unit (GRU) and bidirectional gated recurrent unit (BiGRU) are widely used to process the sequential data with the same size. But the sizes of the multi-level features f_i^r from depth guided color encoder are the different. Therefore, the first important task is to adjust the sizes of multi-level features.

At first, a transition layer which contains a 3×3 convolution and a ReLU activation function is applied on f_i^r. It can adjust the number of channels of multi-level features to the same size. It can be described as:

$$F_i^r = T(f_i^r) \tag{2}$$

where $T(\cdot)$ is defined as:

$$T(X) = \sigma(Conv(X)) \tag{3}$$

where $Conv(\cdot)$ is 3×3 convolution operation, and $\sigma(\cdot)$ is ReLU activation function.

Then, we design a lower-triangle upsampling module (LTUM) which is used to adjust the resolution of multi-level features. Since salient object detection is a pixel-level prediction task, simple and straight idea is to upsample the features in the high layer to the same resolution with the feature in the shallow layer. But the upsampling with large ratio, for example, 2×, 4× or 8×upsampling on F_3^r, F_4^r and F_5^r respectively, will bring some noises. Therefore, the features in the high layer are progressively upsampled and remedied by the features in the adjacent shallow layer. To be specific, as shown in Fig. 2, feature in the third layer F_3^r is upsampled and remedied with feature in the second layer F_2^r, to generate F_3; feature in the fourth layer F_4^r is upsampled and remedied with feature in the third layer F_3^r, and then be upsampled and remedied with the upsampling result F_3 of the third layer, to generate F_4; feature in the fifth layer F_5^r is upsampled and remedied with feature F_4^r in the fourth layer, and then be upsampled and remedied with the upsampling result $AMA(F_3^r, F_4^r)$, and continue to be upsampled and remedied with the upsampling result F_4 of the fourth layer, to generate F_5. The upsampling process can be described as:

$$\begin{aligned} F_1 &= F_1^r \\ F_2 &= F_2^r \\ F_3 &= AMA(F_2^r, F_3^r) \\ F_4 &= AMA(F_3, AMA(F_3^r, F_4^r)) \\ F_5 &= AMA(F_4, AMA(AMA(F_3^r, F_4^r), AMA(F_4^r, F_5^r))) \end{aligned} \tag{4}$$

where $AMA(\cdot)$ denotes add-multiply-add feature fusion block [47], which has been verified effective to achieve the fusion. The detail can be described as:

$$AMA(F_{low}, F_{high}) = Up(F_{high}) + F_{low} + Up(F_{high}) \times F_{low} \tag{5}$$

where F_{low} and F_{high} denote the feature from the lower level with high resolution and the feature from the higher level with low resolution, respectively, and $Up(\cdot)$ denotes 2×upsampling operation.

The resolution adjustment process need 3 AMA modules in the fifth layer; 2 AMA modules in the fourth layer; 1 AMA module in the third layer. They constitute a lower-triangle shape. Therefore, we name it lower-triangle upsampling module.

Compared with direct 2×, 4× or 8×upsampling on F_3^r, F_4^r and F_5^r, low-triangle upsampling module can not only adjust the features to the same resolution but also increase the spatial detail of features in the high layer by progressive remedy process.

Thus, the feature sequence $\{F_i|i = 1, \cdots, 5\}$ with the same scales will be viewed as the input sequence and fed into next Bi-GRU.

To better clarify the scale adjustment process, we list the specific size of three kinds of features f_i^r, F_i^r, F_i in Table 1. From the table, we can see that the channel numbers of the features $f_i^r(i = 1, \cdots, 5)$ are adjusted to the same size by the transition layer, to form the features $F_i^r(i = 1, \cdots, 5)$, and then the resolutions of the features $F_i^r(i = 1, \cdots, 5)$ are further adjusted to the same size by the lower-triangle upsampling module, to form the features $F_i(i = 1, \cdots, 5)$. The transition layer and the lower-triangle upsampling module compose scale adjustment module together.

3.3.2 Bidirectional decoding fusion module (BDFM)

The proposed bidirectional decoding fusion module recurrently and parallel decodes multi-level features in down-top and top-down manners. The bidirectional manner can comprehensively combine the multi-level features and reduce the deviation from the order of recurrent fusion. Each directional fusion achieves the interaction among inter-layer by the proposed self-attention gated recurrent unit (SAGRU). The parameters of SAGRU in each manner are shared. Figure 2 uses orange blocks to represent the down-top manner, uses blue blocks to represent the top-down manner, and uses dark and light background color to represent the sharing strategy.

In the down-top dataflow, the initial hidden state is the feature F_1 in the lowest layer. The output of each time step is represented as $P_j(j = 1, \cdots, 5)$. In the top-down dataflow, the initial hidden state is the feature F_5 in the highest layer which has been fused with some features in the shallow layer by lower-triangle upsampling module. The output of each time step is represented as $H_j(j = 5, \cdots, 1)$. The bidirectional output features with the same level are fused to generate the saliency map $S_j(j = 1, \cdots, 5)$, which is described as:

$$S_j = sig(TFM(P_j, H_{6-j})) \tag{6}$$

where P_j and H_{6-j} represent the hidden state output with the same level, $sig(\cdot)$ is sigmoid function, and $TFM(\cdot, \cdot)$ is defined as:

$$TFM(X, Y) = Conv(Up(T(Up(T(Cat(X, Y)))))) \tag{7}$$

Table 1 The size of features f_i^r, F_i^r, F_i which is denoted as channel×height×width

	i=1	i=2	i=3	i=4	i=5
f_i^r	64×88×88	256×88×88	512×44×44	1024×22×22	2048×11×11
F_i^r	64×88×88	64×88×88	64×44×44	64×22×22	64×11×11
F_i	64×88×88	64×88×88	64×88×88	64×88×88	64×88×88

In addition, all the output features are fused to generate the final saliency map S_{final}, which is described as:

$$S_{final} = sig(\sum_{j=1}^{5} TFM(P_j, H_{6-j})) \tag{8}$$

Six saliency maps are supervised by the ground truth maps using pixel position aware loss L_{ppa}^{s} [71] for end-to-end training.

Different from the traditional Bi-GRU, we design self-attention gated recurrent unit to adaptively memorize the useful information and discard the irrelevant part across different levels by self-attention mechanism. Its detail is shown in Fig. 3.

Each SAGRU includes a reset gate, an update gate and a memory unit. Initial hidden state and output hidden state are denoted as h_{j-1} and h_j respectively, and they represent fusion result of multi-level features. The input F is features from scale adjustment module, that is to say, $F \in \{F_i | i = 1, \cdots, 5\}$. The reset gate r_t controls the degree of ignoring the status information at the previous moment. To be specific, the previous fusion result h_{j-1} and current input feature F are concatenated together, and then performed a 3×3 convolution operation and a self-attention operation, and last served as the weight to reset the previous fusion result by the following process:

$$\begin{aligned} r_t &= Self(Conv(Cat(h_{j-1}, F))) \\ h'_j &= r_t \times h_{j-1} \end{aligned} \tag{9}$$

where $Cat(\cdot)$ denotes concatenation operation, $Self(\cdot)$ is self-attention operation, which is designed in Fig. 4. The self-attention operation models interdependencies along the channel dimensions to enhance the feature representation and increase the difference between salient and non-salient region. That is to say, self-attention operation can make the salient region more prominent. The detail can be described as:

$$Self(X) = (X \cdot X^{\mathrm{T}}) \cdot X \tag{10}$$

where "·" is matrix multiplication operation.

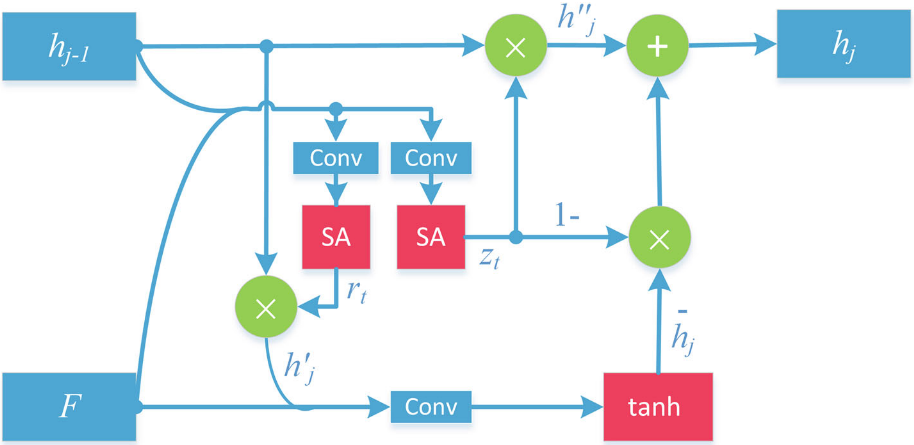


Fig. 3 The structure of Self-Attention Gated Recurrent Unit (SAGRU)

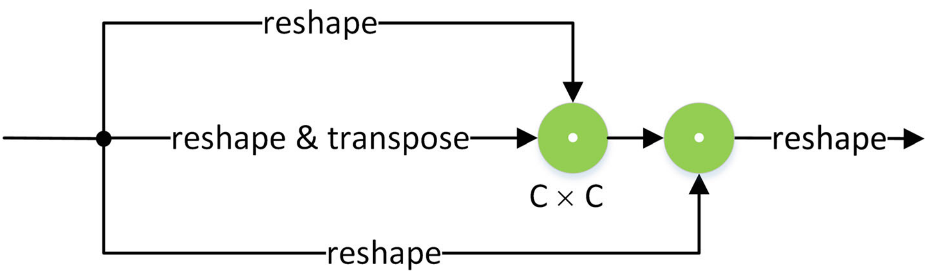


Fig. 4 The structure of Self-Attention

The reset hidden state h'_j select the more salient information in the previous hidden state h_{j-1} and discard some non-common salient part on the comprehensive consideration of previous fusion result h_{j-1} and current input F.

Then the current input F and reset hidden state h'_j are concatenated to form memory unit $\bar{h}_j$, which indicates the current memorized fusion result. The process can be described as:

$$\bar{h}_j = tanh(Conv(Cat(F, h'_j))) \tag{11}$$

where $tanh(\cdot)$ is $tanh$ function. The memory unit $\bar{h}_j$ helps SAGRU to memorize long-term context information in the fusion process.

The update gate z_t determines the extent to which the status information at the previous time is brought into the current status. To be specific, the previous fusion result h_{j-1} and current input feature F are concatenated together, and then performed a convolution operation and a self-attention operation, and last served as the weight to update the previous fusion result by the following process:

$$\begin{aligned} z_t &= Self(Conv(Cat(h_{j-1}, F))) \\ h''_j &= z_t \times h_{j-1} \end{aligned} \tag{12}$$

The updated fusion result h''_j discards the non-common salient regions in the previous fusion result considering the current input. Meanwhile, in order to supplement the salient part in the long-term context information throughout the memory, the output hidden state h_j need to append the non-salient part deleted by update gate but prominent in the memory unit $\bar{h}_j$. The process can be described as:

$$h_j = h''_j + (1 - z_t) \times \bar{h}_j \tag{13}$$

In a word, SAGRU comprehensively optimizes the history hidden state based on current input. The use of self-attention makes the reset and update of history information more accuracy.

4 Experiments

4.1 Datasets and evaluation metrics

4.1.1 Datasets

We evaluate the proposed method on six challenging RGB-D SOD datasets. NLPR [54] includes 1000 images with single or multiple salient objects. NJU2K [33] consists of 2003 stereo image pairs and ground-truth maps with different objects, complex and challenging scenes. STERE [48] incorporates 1000 pairs of binocular images downloaded from the Internet. DES [17] has 135 indoor images collected by Microsoft Kinect. SIP [21] contains 1000 high-resolution images of multiple salient persons. DUT [57] contains 1200 images captured by Lytro camera in real life scenes.

Training/Testing For the sake of fair comparison, we use the same training dataset as in [13, 21], which consists of 1,485 images from the NJU2K dataset and 700 images from the NLPR dataset. The remaining images in the NJU2K and NLPR datasets and the whole datasets of STERE, DES and SIP are used for testing. In addition, on the DUT dataset, we follow the same protocols as in [30, 39, 57, 58, 92] to add additional 800 pairs from DUT for training and test on the remaining 400 pairs. In summary, our training set contains 2,185 paired RGB and depth images, but when testing is conducted on DUT, our training set contains 2,985 paired ones.

4.1.2 Evaluation metrics

We adopt saliency evaluation toolbox[1] to evaluate the performance of our model and state-of-the-art RGB-D SOD models, including the precision-recal(PR) curve [7], S-measure [19], F-measure [1], E-measure [20] and mean absolute error (MAE) [56]. Specifically, the PR curve plots precision and recall values by setting a series of thresholds on the saliency maps to get the binary masks and further comparing them with the ground truth maps. The S-measure can evaluate both region-aware and object-aware structural similarity between saliency map and ground truth. The F-measure is the weighted harmonic mean of precision and recall, which can evaluate the overall performance. The E-measure simultaneously captures global statistics and local pixel matching information. The MAE measures the average of the per-pixel absolute difference between the saliency maps and the ground truth maps. In our experiment, F-measure and E-measure adopts adaptive ones.

4.2 Implementation details

During the training and testing phase, the input RGB and depth images are resized to 352×352. Multiple enhancement strategies are used for all training images, i.e., random flipping, rotating and border clipping. Parameters of the backbone network are initialized with the pretrained parameters of ResNet-50 network [27]. The rest of parameters are initialized to PyTorch default settings. We employ the Adam optimizer [35] to train our network with a batch size of 5 and an initial learning rate 1e-4, and the learning rate will be divided by 10 every 60 epochs. Our model is implemented with PyTorch toolbox and trained on a

[1] https://github.com/jiwei0921/Saliency-Evaluation-Toolbox

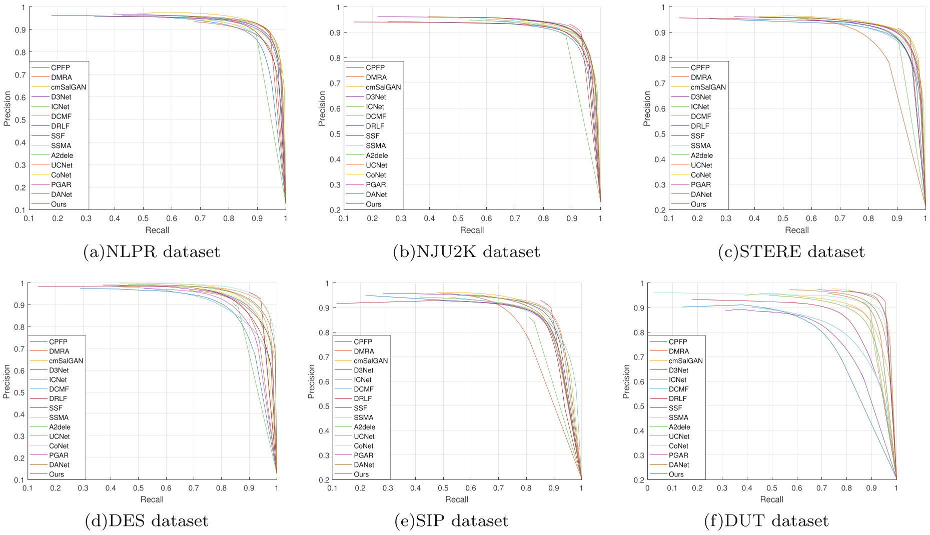
machine with a single NVIDIA GTX 1080Ti GPU. The model converges within 200 epochs, which takes nearly 8 hours.

4.3 Comparisons with state-of-the-art

Our model is compared with 14 state-of-the-art RGB-D salient object detection algorithms, including CPFP [89], DMRA [57], cmSalGAN [31], D3Net [21], ICNet [40], DCMF [9], DRLF [69], SSF [82], SSMA [43], A2dele [58], UC-Net [80], CoNet [30], PGAR [13] and DANet [92]. To ensure the fairness of the comparison results, the saliency maps of the evaluation are provided by the authors or generated by running source codes.

Quantitative evaluation Figure 5 shows the comparison results on PR curve. Table 2 shows the quantitative comparison results of four evaluation metrics. As can be clearly observed from figure that our curves are very short, which means that our recall is very high. Furthermore, from the table, we can see that our F-measure values are also the best on six datasets, so as to verify the effectiveness and advantages of our proposed method. In addition, our MAE and E-measure values are both superior to other state-of-the-art methods. S-measure value is a little bit worse in three datasets, but the majority are the best.

Qualitative evaluation To make the qualitative comparisons, we show some visual examples in Fig. 6. It can be observed that our method has better detection results than other methods in some challenging cases: similar foreground and background(1^{st}-2^{nd} rows), complex scene(3^{rd}-4^{th} rows), low quality depth map(5^{th}-6^{th} rows), small object(7^{th}-8^{th} rows) and multiple objects(9^{th}-10^{th} rows). In addition, our approach can produce more fine-grained details as highlighted in the salient region(11^{th}-12^{th} rows). These indicate that our approach can better locate salient objects and produce more accurate saliency maps.



(a)NLPR dataset (b)NJU2K dataset (c)STERE dataset

(d)DES dataset (e)SIP dataset (f)DUT dataset

Fig. 5 P-R curves comparison of different models on six datasets

Table 2 S-measure, adaptive F-measure, adaptive E-measure, MAE comparisons with different models

Datasets	Metric	CPFP CVPR19 [89]	DMRA ICCV19 [57]	cmSalGAN TMM20 [31]	D3Net TNNLS20 [21]	ICNet TIP20 [40]	DCMF TIP20 [9]	DRLF TIP20 [69]	SSF CVPR20 [82]	SSMA CVPR20 [43]	A2dele CVPR20 [58]	UC-Net CVPR20 [80]	CoNet ECCV20 [30]	PGAR ECCV20 [13]	DANet ECCV20 [92]	BGRDNet Ours
NLPR	S↑	.888	.899	.922	.912	**.923**	.900	.903	.914	.915	.896	.920	.908	.918	.920	.917
	F_β ↑	.823	.854	.863	.861	.870	.839	.843	.875	.853	.878	.890	.846	.871	.875	**.902**
	E_ξ ↑	.924	.941	.947	.944	.944	.933	.936	.949	.938	.945	.953	.934	.948	.951	**.957**
	MAE↓	.036	.031	.027	.030	.028	.035	.032	.026	.030	.028	.025	.031	.028	.027	**.023**
NJU2K	S↑	.878	.886	.903	.901	.894	.889	.886	.899	.894	.869	.897	.895	.906	.899	**.907**
	F_β ↑	.837	.872	.874	.865	.868	.859	.849	.886	.865	.874	.889	.872	.883	.871	**.908**
	E_ξ ↑	.895	.908	.907	.914	.905	.897	.901	.913	.896	.897	.903	.912	.914	.908	**.922**
	MAE↓	.053	.051	.046	.046	.052	.052	.055	.043	.053	.051	.043	.046	.045	.045	**.034**
STERE	S↑	.879	.835	.896	.899	.903	.883	.888	.887	.890	.878	.903	**.905**	.903	.901	.902
	F_β ↑	.830	.845	.863	.859	.865	.841	.845	.867	.855	.874	.885	.884	.872	.868	**.895**
	E_ξ ↑	.903	.900	.914	.920	.915	.904	.915	.921	.907	.915	.922	.927	.917	.921	**.928**
	MAE↓	.051	.066	.050	.046	.045	.054	.050	.046	.051	.044	.039	.037	.044	.043	**.036**
DES	S↑	.872	.901	.913	.898	.920	.877	.895	.905	**.941**	.885	.933	.911	.894	.924	.933
	F_β ↑	.829	.857	.869	.870	.889	.820	.868	.876	.906	.865	.917	.861	.870	.899	**.931**
	E_ξ ↑	.927	.945	.949	.951	.959	.923	.954	.948	.974	.922	.974	.945	.935	.968	**.976**
	MAE↓	.037	.029	.028	.031	.027	.040	.030	.025	.021	.028	.018	.027	.032	.023	**.016**
SIP	S↑	.850	.806	.865	.860	.854	.859	.850	.868	.872	.826	.875	.858	.875	.875	**.882**
	F_β ↑	.819	.819	.849	.835	.836	.819	.813	.851	.854	.825	.868	.842	.848	.855	**.889**
	E_ξ ↑	.899	.863	.902	.902	.899	.898	.891	.911	.911	.892	.913	.909	.908	.914	**.923**
	MAE↓	.064	.085	.064	.063	.069	.068	.071	.056	.057	.070	.051	.063	.059	.054	**.045**
DUT	S↑	.749	.889	.867	.775	.852	.798	.826	.916	.903	.886	.864	.919	.920	.899	**.927**
	F_β ↑	.736	.884	.844	.756	.830	.750	.803	.914	.866	.890	.856	.909	.914	.888	**.935**
	E_ξ ↑	.815	.927	.897	.847	.897	.848	.870	.946	.921	.924	.903	.948	.944	.934	**.951**
	MAE↓	.100	.048	.067	.097	.072	.104	.080	.034	.044	.043	.056	.033	.035	.043	**.029**

The best result is in bold

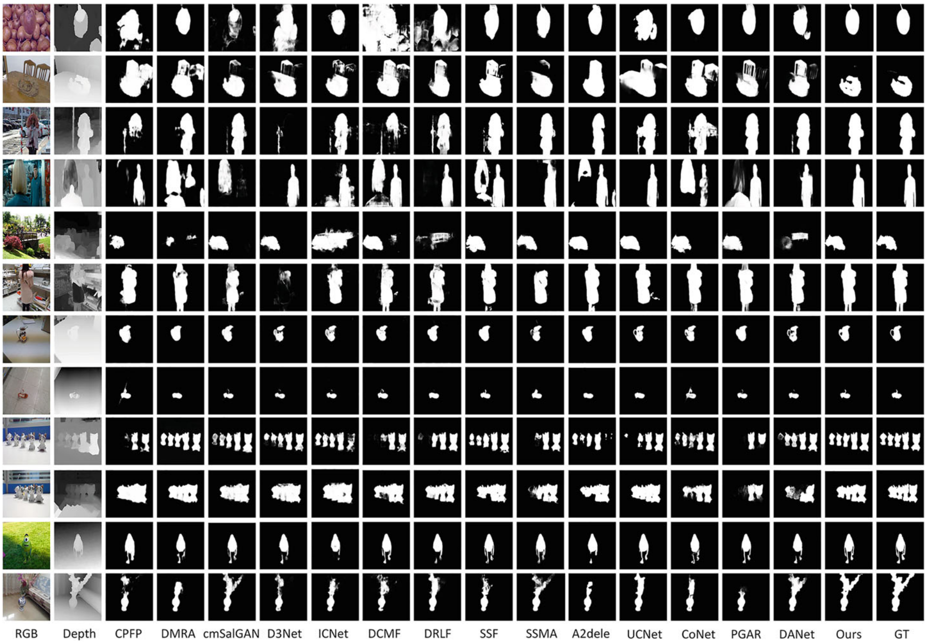


Fig. 6 Visual comparison results with other state-of-the-art models

Model complexity Table 3 shows the complexity comparison of some RGB-D models. It involves model size and computation cost. From the result, we can conclude that the complexity of our model is acceptable.

4.4 Ablation studies

We conduct ablation studies on NLPR, NJU2K, SIP and STERE datasets to investigate the contributions of different modules in the proposed method. Meanwhile, we also test and verify the effectiveness of designed gated recurrent unit.

4.4.1 The Effectiveness of BDFM, SAM and DGRM

The baseline model used here removes depth guided residual module (DGRM), scale adjustment module (SAM) and bidirectional decoding fusion module (BDFM). It attaches the depth feature to color feature by element-wise addition operation in the encoder, and adopts traditional U-Net decoding process by element-wise addition operation. Its performance is illustrated in Table 4 No.1. In order to verify the effectiveness of BDFM, ablation experiments are conducted. The variant No.2 in Table 4 denotes the model which adopts BDFM instead of traditional U-Net additional decoding process based on the baseline. The multi-level features are adjusted to the same size by the progressive 2× upsamling operation. Compared with the variant No.1, the performance of the variant No.2 is significantly improved in NLPR, NJU2K and STERE dataset, and slight enhanced in SIP dataset. Furthermore, we also discuss the role of the scale adjustment module, which adjusts the size of multi-level features. Compared with No.2, the variant No.3 in Table 4 uses low-triangle upsampling module to effectively enhance the detection performance, and avoid the error

Table 3 The complexity comparisons of different RGB-D models

Method	DMRA [57] ICCV19	D3Net [21] TNNLS20	SSF [82] CVPR20	SSMA [43] CVPR20	A2dele [58] CVPR20	UC-Net [80] CVPR20	CoNet [30] ECCV20	PGAR [13] ECCV20	DANet [92] ECCV20	BGRDNet Ours
Params(M)	59.7	45.2	32.9	86.6	30.1	28.9	43.7	16.2	26.7	51.5
FLOPs(G)	120.9	55.1	46.5	141.1	42.8	32.3	27.4	44.6	66.2	61.1

Table 4 Ablation experiments of different modules

Variant	Candidate				NLPR				NJU2K				SIP				STERE			
	Baseline	BDFM	SAM	DGRM	S↑	F_β ↑	E_ξ ↑	MAE↓	S↑	F_β ↑	E_ξ ↑	MAE↓	S↑	F_β ↑	E_ξ ↑	MAE↓	S↑	F_β ↑	E_ξ ↑	MAE↓
No.1	✓				.896	.864	.944	.030	.875	.860	.898	.051	.859	.845	.907	.057	.859	.843	.905	.054
No.2	✓	✓			.910	.897	.952	.026	.897	.905	.921	.039	.865	.878	.908	.055	.883	.886	.918	.042
No.3	✓	✓	✓		.912	.899	.953	.024	.903	.906	.921	.036	.878	.887	.915	.049	.893	.893	.920	.040
No.4	✓	✓	✓	✓	**.917**	**.902**	**.957**	**.023**	**.907**	**.908**	**.922**	**.034**	**.882**	**.889**	**.923**	**.045**	**.902**	**.895**	**.928**	**.036**

The best result is in bold

Table 5 Effectiveness analysis of the designed SAGRU

Variant	NLPR				NJU2K				SIP				STERE			
	S↑	F_β ↑	E_ξ ↑	MAE↓	S↑	F_β ↑	E_ξ ↑	MAE↓	S↑	F_β ↑	E_ξ ↑	MAE↓	S↑	F_β↑	E_ξ ↑	MAE↓
ConvGRU	.914	.901	.955	.024	.904	.905	.920	.036	.877	.881	.919	.047	.898	.887	.924	.039
SAGRU	**.917**	**.902**	**.957**	**.023**	**.907**	**.908**	**.922**	**.034**	**.882**	**.889**	**.923**	**.045**	**.902**	**.895**	**.928**	**.036**

The best result is in bold

from the simple and straight upsampling in the variant No.2. At last, we replace traditional addition operation with DGRM on the fusion of the depth and color feature. The result of the variant No.4 in Table 4 achieves the better performance by the optimization of attention mechanism on the depth feature.

4.4.2 The Effectiveness of SAGRU

We replace the SAGRU with ConGRU [3] to check the effectiveness of the designed SAGRU. From Table 5, we can see that the use of SAGRU improves the detection performance to some extent. It benefits from the feature representation enhancement ability by the self-attention operation.

5 Conclusions

Salient object detection is a pixel-wise dense prediction task. U-Net framework is widely adopted to solve the problem, in which multi-level features from encoder are progressively combined from high layer to shallow layer, so as to further recover image size. Our method decodes the saliency cues from multi-level features by Bi-GRU. We devise scale adjustment module (SAM) to make the multi-level features adaptive to the input of Bi-GRU, and meanwhile, we propose a self-attention gated recurrent unit (SAGRU) which achieves the better fusion of multi-level features by adaptive selection and selective memory. Based on SAGRU, we construct bidirectional decoding fusion module (BDFM) to achieve bidirectional and parallel decoding process. Furthermore, we use depth information to enhance RGB features by depth guided residual module (DGRM). Finally, we achieve the salient object detection in RGB-D images. Experimental results show our method outperforms the state-of-the-art method, and ablation studies verify the role of each module. In the future, we will explore the decoding method based on the Transformer [64] which allows for significantly more parallelization of attention mechanism instead of recurrent structure of GRU to better depict long-range dependencies among multi-level features.

Acknowledgment This work is supported by Natural Science Foundation of Anhui Province (1908085MF182) and University Natural Science Research Project of Anhui Province(KJ2019A0034).

Declarations

Conflict of Interests We declare that we have no conflict of interest.

References

1. Achanta R, Hemami S, Estrada F, Susstrunk S (2009) Frequency-tuned salient region detection. In: 2009 IEEE conference on computer vision and pattern recognition, IEEE, pp 1597–1604
2. Ahmadi M, Karimi N, Samavi S (2021) Context-aware saliency detection for image retargeting using convolutional neural networks. Multi Tools and Appl 80(8):11917–11941
3. Ballas N, Yao L, Pal C, Courville A (2015) Delving deeper into convolutional networks for learning video representations. arXiv:1511.06432
4. Bani NT, Fekri-Ershad S (2019) Content-based image retrieval based on combination of texture and colour information extracted in spatial and frequency domains. The electronic library
5. Bardhan S (2020) Salient object detection by contextual refinement. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops, pp 356–357

6. Bardhan S, Das S, Jacob S (2019) Visual Saliency Detection via Convolutional Gated Recurrent Units. In: International conference on neural information processing, Springer, pp 162–174
7. Borji A, Cheng M-M, Jiang H, Li J (2015) Salient object detection: A benchmark. IEEE Trans Image Processing 24(12):5706–5722
8. Chen C, Wei J, Peng C, Qin H (2021) Depth-quality-aware salient object detection. IEEE Trans Image Process 30:2350–2363
9. Chen H, Deng Y, Li Y, Hung T-Y, Lin G (2020) RGBD salient object detection via disentangled cross-modal fusion. IEEE Trans Image Process 29:8407–8416
10. Chen H, Li Y (2018) Progressively complementarity-aware fusion network for RGB-D salient object detection. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 3051–3060
11. Chen Q, Fu K, Liu Z, Chen G, Du H, Qiu B, Shao L (2020) EF-Net: A novel enhancement and fusion network for RGB-D saliency detection. Pattern Recogn, p 107740
12. Chen Q, Liu Z, Zhang Y, Fu K, Zhao Q, Du H (2021) Rgb-d salient object detection via 3d convolutional neural. AAAI
13. Chen S, Fu Y (2020) Progressively guided alternate refinement network for RGB-D salient object detection. In: European conference on computer vision, Springer, pp 520–538
14. Chen S, Zhu X, Liu W, He X, Liu J (2021) Global-Local Propagation Network for RGB-D Semantic Segmentation. arXiv:2101.10801
15. Chen Z, Cong R, Xu Q, Huang Q (2020) DPANet: Depth Potentiality-Aware Gated Attention Network for RGB-D Salient Object Detection. IEEE Trans Image Process
16. Cheng Y, Cai R, Li Z, Zhao X, Huang K (2017) Locality-sensitive deconvolution networks with gated fusion for rgb-d indoor semantic segmentation. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 3029–3037
17. Cheng Y, Fu H, Wei X, Xiao J, Cao X (2014) Depth enhanced saliency detection method. In: Proceedings of international conference on internet multimedia computing and service, pp 23–27
18. Cho K, Van Merriënboer B, Gulcehre C, Bahdanau D, Bougares F, Schwenk H, Bengio Y (2014) Learning phrase representations using RNN encoder-decoder for statistical machine translation. arXiv:1406.1078
19. Fan D-P, Cheng M-M, Liu Y, Li T, Borji A (2017) Structure-measure: A new way to evaluate foreground maps. In: Proceedings of the IEEE international conference on computer vision, pp 4548–4557
20. Fan D-P, Gong C, Cao Y, Ren B, Cheng M-M, Borji A (2018) Enhanced-alignment measure for binary foreground map evaluation. arXiv:1805.10421
21. Fan D-P, Lin Z, Zhang Z, Zhu M, Cheng M-M (2020) Rethinking RGB-D Salient Object Detection: Models, Data Sets, and Large-Scale Benchmarks. IEEE transactions on neural networks and learning systems
22. Fan D-P, Zhai Y, Borji A, Yang J, Shao L (2020) BBS-Net: RGB-D salient object detection with a bifurcated backbone strategy network. In: European conference on computer vision, Springer, pp 275–292
23. Feng D, Barnes N, You S (2017) HOSO: Histogram Of Surface Orientation for RGB-D Salient Object Detection. In: Digital image computing: techniques and applications (DICTA), IEEE, pp 1–8
24. Feng D, Barnes N, You S, McCarthy C (2016) Local background enclosure for RGB-D salient object detection. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 2343–2350
25. Fu K, Fan D-P, Ji G-P, Zhao Q (2020) JL-DCF: Joint learning and densely-cooperative fusion framework for rgb-d salient object detection. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 3052–3062
26. Fu K, Fan D-P, Ji G-P, Zhao Q, Shen J, Zhu C (2021) Siamese network for RGB-D salient object detection and beyond. IEEE transactions on pattern analysis and machine intelligence
27. He K, Zhang X, Ren S, Sun J (2016) Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 770–778
28. Huang H, Cai M, Lin L, Zheng J, Mao X, Qian X, Peng Z, Zhou J, Iwamoto Y, Han X-H et al (2021) Graph-based Pyramid Global Context Reasoning with a Saliency-aware Projection for COVID-19 Lung Infections Segmentation. In: ICASSP 2021-2021 IEEE international conference on acoustics, speech and signal processing (ICASSP), IEEE, pp 1050–1054
29. Huang Z, Chen H-X, Zhou T, Yang Y-Z, Liu B-Y (2021) Multi-level cross-modal interaction network for RGB-D salient object detection. Neurocomputing 452:200–211
30. Ji W, Li J, Zhang M, Piao Y, Lu H (2020) Accurate rgb-d salient object detection via collaborative learning. In: Computer vision–ECCV 2020: 16th European conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XVIII 16, Springer, pp 52–69

31. Jiang B, Zhou Z, Wang X, Tang J, Luo B (2020) cmSalGAN: RGB-D Salient Object Detection with Cross-View Generative Adversarial Networks. IEEE Trans Multi
32. Jiang Q, Shao F, Lin W, Gu K, Jiang G, Sun H (2017) Optimizing multistage discriminative dictionaries for blind image quality assessment. IEEE Trans Multi 20(8):2035–2048
33. Ju R, Ge L, Geng W, Ren T, Wu G (2014) Depth saliency based on anisotropic center-surround difference. In: 2014 IEEE international conference on image processing (ICIP), IEEE, pp 1115–1119
34. Ju R, Ge L, Geng W, Ren T, Wu G (2014) Depth saliency based on anisotropic center-surround difference. In: Image processing (ICIP), IEEE, pp 1115–1119
35. Kingma DP, Ba J (2014) Adam: A method for stochastic optimization. arXiv:1412.6980
36. Li B, Sun Z, Li Q, Wu Y, Hu A (2019) Group-wise deep object co-segmentation with co-attention recurrent neural network. In: Proceedings of the IEEE international conference on computer vision, pp 8519–8528
37. Li B, Sun Z, Tang L, Sun Y, Shi J (2019) Detecting robust co-saliency with recurrent co-attention neural network. In: IJCAI, pp 818–825
38. Li C, Cong R, Kwong S, Hou J, Fu H, Zhu G, Zhang D, Huang Q (2020) ASIF-Net: Attention steered interweave fusion network for RGB-D salient object detection. IEEE Transactions on Cybernetics
39. Li C, Cong R, Piao Y, Xu Q, Loy CC (2020) RGB-D salient object detection with cross-modality modulation and selection. In: European conference on computer vision, Springer, pp 225–241
40. Li G, Liu Z, Ling H (2020) ICNet: Information conversion network for rgb-d based salient object detection. IEEE Trans Image Process 29:4873–4884
41. Li G, Liu Z, Ye L, Wang Y, Ling H (2020) 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, pp 665–681
42. Liao G, Gao W, Jiang Q, Wang R, Li G (2020) MMNet: Multi-stage and multi-scale fusion network for rgb-d salient object detection. In: Proceedings of the 28th ACM international conference on multimedia, pp 2436–2444
43. Liu N, Zhang N, Han J (2020) 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
44. Liu N, Zhang N, Shao L, Han J (2020) Learning selective mutual attention and contrast for rgb-d saliency detection. arXiv:2010.05537
45. Liu Z, Shi S, Duan Q, Zhang W, Zhao P (2019) Salient object detection for RGB-D image by single stream recurrent convolution neural network. Neurocomputing 363:46–57
46. Liu Z, Zhang W, Zhao P (2020) A cross-modal adaptive gated fusion generative adversarial network for RGB-D salient object detection. Neurocomputing 387:210–220
47. Nie D, Xue J, Ren X (2020) Bidirectional pyramid networks for semantic segmentation. In: Proceedings of the asian conference on computer vision
48. Niu Y, Geng Y, Li X, Liu F (2012) Leveraging stereopsis for saliency analysis. In: 2012 IEEE conference on computer vision and pattern recognition, IEEE, pp 454–461
49. Niu Y, Long G, Liu W, Guo W, He S (2020) Boundary-aware RGBD salient object detection with cross-modal feature sampling. IEEE Trans Image Process 29:9496–9507
50. Ouerhani N, Hugli H (2000) Computing visual attention from scene depth. In: Proceedings 15th international conference on pattern recognition. ICPR-2000, vol 1, IEEE, pp 375–378
51. Pahuja A, Majumder A, Chakraborty A, Babu RV (2019) Enhancing salient object segmentation through attention. In: CVPR workshops, pp 27–36
52. Pan L, Zhou X, Shi R, Zhang J, Yan C (2020) Cross-modal feature extraction and integration based RGBD saliency detection. Image Vis Comput 101:103964
53. Pang Y, Zhang L, Zhao X, Lu H (2020) Hierarchical dynamic filtering network for RGB-D salient object detection. In: Computer Vision–ECCV 2020: 16th European conference, Glasgow, UK, August 23–28, 2020, proceedings, Part XXV 16, Springer, pp 235–252
54. Peng H, Li B, Xiong W, Hu W, Ji R (2014) RGBD salient object detection: a benchmark and algorithms. In: European conference on computer vision, Springer, pp 92–109
55. Peng P, Li Y-J (2020) A unified structure for efficient rgb and rgb-d salient object detection. arXiv:2012.00437
56. Perazzi F, Krähenbühl P, Pritch Y, Hornung A (2012) Saliency filters: Contrast based filtering for salient region detection. In: 2012 IEEE conference on computer vision and pattern recognition, IEEE, pp 733–740
57. Piao Y, Ji W, Li J, Zhang M, Lu H (2019) Depth-induced multi-scale recurrent attention network for saliency detection. In: Proceedings of the IEEE international conference on computer vision, pp 7254–7263

58. Piao Y, Rong Z, Zhang M, Ren W, Lu H (2020) 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
59. Pinheiro PO, Lin T-Y, Collobert R, Dollár P (2016) Learning to refine object segments. In: European conference on computer vision, Springer, pp 75–91
60. Ren J, Gong X, Yu L, Zhou W, Ying Yang M (2015) Exploiting global priors for RGB-D saliency detection. In: Proceedings of the IEEE conference on computer vision and pattern recognition workshops, pp 25–32
61. Ronneberger O, Fischer P, Brox T (2015) U-Net: Convolutional networks for biomedical image segmentation. In: International conference on medical image computing and computer-assisted intervention, Springer, pp 234–241
62. Shi Z, Shen X, Chen H, Lyu Y (2020) Global semantic consistency network for image manipulation detection. IEEE Signal Processing Letters 27:1755–1759
63. Sun L, Yang K, Hu X, Hu W, Wang K (2020) Real-time fusion network for RGB-D semantic segmentation incorporating unexpected obstacle detection for road-driving images. IEEE Robotics and Automation Letters 5(4):5558–5565
64. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser L, Polosukhin I (2017) Attention is all you need. In: Advances in neural information processing systems, pp 5998–6008
65. Wang N, Gong X (2019) Adaptive fusion for RGB-D salient object detection. IEEE Access 7:55277–55284
66. Wang S-T, Zhou Z, Qu H-B, Li B (2016) Visual saliency detection for RGB-D images with generative model. In: Asian conference on computer vision, Springer, pp 20–35
67. Wang W, Shen J, Cheng M-M, Shao L (2019) An iterative and cooperative top-down and bottom-up inference network for salient object detection. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 5968–5977
68. Wang W, Shen J, Yang R, Porikli F (2017) Saliency-aware video object segmentation. IEEE Trans Pattern Analysis and Machine Intelligence 40(1):20–33
69. Wang X, Li S, Chen C, Fang Y, Hao A, Qin H (2020) Data-level recombination and lightweight fusion scheme for RGB-D salient object detection. IEEE Trans Image Process 30:458–471
70. Wang Y, Li Y, Elder JH, Wu R, Lu H, Zhang L (2020) Synergistic saliency and depth prediction for RGB-D saliency detection. In: Proceedings of the asian conference on computer vision, pp 1–17
71. Wei J, Wang S, Huang Q (2020) F^3Net: Fusion, feedback and focus for salient object detection. In: Proceedings of the AAAI conference on artificial intelligence, pp 12321–12328
72. Weng Z, Li W, Jin Z (2021) Human activity prediction using saliency-aware motion enhancement and weighted LSTM network. EURASIP J Image and Video Process 2021(1):1–23
73. Woo S, Park J, Lee J-Y, So Kweon I (2018) CBAM: Convolutional block attention module. In: Proceedings of the European conference on computer vision (ECCV), pp 3–19
74. Wu Y-H, Liu Y, Xu J, Bian J-W, Gu Y, Cheng M-M (2020) MobileSal: Extremely Efficient RGB-D Salient Object Detection. arXiv:2012.13095
75. Wu Z, Su L, Huang Q (2019) Cascaded partial decoder for fast and accurate salient object detection. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 3907–3916
76. Xu Y, Xu W, Wang M, Li L, Sang G, Wei P, Zhu L (2021) Saliency aware image cropping with latent region pair. Expert Syst Appl 171:114596
77. Yarlagadda SK, Montserrat DM, Guerra D, Boushey CJ, Kerr DA, Zhu F (2021) Saliency-aware class-agnostic food image segmentation. arXiv:2102.06882
78. Zeng J, Tong Y, Huang Y, Yan Q, Sun W, Chen J, Wang Y (2019) Deep surface normal estimation with hierarchical rgb-d fusion. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 6153–6162
79. Zhang C, Li G, Lin G, Wu Q, Yao R (2021) Cyclesegnet: Object co-segmentation with cycle refinement and region correspondence. IEEE Trans Image Process
80. Zhang J, Fan D-P, Dai Y, Anwar S, Saleh FS, Zhang T, Barnes N (2020) 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
81. Zhang M, Fei SX, Liu J, Xu S, Piao Y, Lu H (2020) Asymmetric two-stream architecture for accurate RGB-D saliency detection. In: European conference on computer vision, Springer, pp 374–390
82. Zhang M, Ren W, Piao Y, Rong Z, Lu H (2020) Select, supplement and focus for RGB-D saliency detection. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 3472–3481

83. Zhang M, Zhang Y, Piao Y, Hu B, Lu H (2020) Feature reintegration over differential treatment: A top-down and adaptive fusion network for RGB-D salient object detection. In: Proceedings of the 28th ACM international conference on multimedia, pp 4107–4115
84. Zhang P, Liu W, Wang D, Lei Y, Wang H, Lu H (2020) Non-rigid object tracking via deep multi-scale spatial-temporal discriminative saliency maps. Pattern Recogn 100:107130
85. Zhang Q, Cong R, Li C, Cheng M-M, Fang Y, Cao X, Zhao Y, Kwong S (2020) Dense attention fluid network for salient object detection in optical remote sensing images. IEEE Trans Image Process
86. Zhang X, Jin T, Zhou W, Lei J (2021) Attention-based contextual interaction asymmetric network for RGB-D saliency prediction. J Vis Commun Image Represent 74:102997
87. Zhang Y, Zheng J, Li L, Liu N, Jia W, Fan X, Xu C, He X (2021) Rethinking feature aggregation for deep RGB-D salient object detection. Neurocomputing 423:463–473
88. Zhang Z, Lin Z, Xu J, Jin W-D, Lu S-P, Fan D-P (2021) Bilateral attention network for RGB-D salient object detection. IEEE Trans Image Process 30:1949–1961
89. Zhao J-X, Cao Y, Fan D-P, Cheng M-M, Li X-Y, Zhang L (2019) Contrast prior and fluid pyramid integration for RGBD salient object detection. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 3927–3936
90. Zhao J, Zhao Y, Li J, Chen X (2020) Is depth really necessary for salient object detection? In: Proceedings of the 28th ACM international conference on multimedia, pp 1745–1754
91. Zhao X, Pang Y, Zhang L, Lu H, Ruan X (2021) Self-supervised representation learning for rgb-d salient object detection. arXiv:2101.12482
92. Zhao X, Zhang L, Pang Y, Lu H, Zhang L (2020) A single stream network for robust and real-time rgb-d salient object detection. In: European conference on computer vision, Springer, pp 646–662
93. Zhou T, Fan D-P, Cheng M-M, Shen J, Shao L (2021) RGB-D salient object detection: A survey. Computational Visual Media, pp 1–33
94. Zhou W, Chen Y, Liu C, Yu L (2020) GFNet: Gate Fusion Network with Res2Net for Detecting Salient Objects in RGB-D Images. IEEE Signal Process Letters
95. Zhou X, Li G, Gong C, Liu Z, Zhang J (2020) Attention-guided RGBD saliency detection using appearance information. Image Vis Comput 95:103888
96. Zhu C, Cai X, Huang K, Li TH, Li G (2019) PDNet: Prior-model guided depth-enhanced network for salient object detection. In: 2019 IEEE International conference on multimedia and expo (ICME), IEEE, pp 199–204