



Guided residual network for RGB-D salient object detection with efficient depth feature learning

Jian Wang¹ · Shuhan Chen¹ · Xiao Lv² · Xiuqi Xu¹ · Xuelong Hu¹

Accepted: 4 March 2021 / Published online: 17 April 2021
© The Author(s) 2021

Abstract

RGB-D salient object detection aims at identifying the most attractive parts from a RGB image and its corresponding depth image, which has been widely applied in many computer vision tasks. However, there are still two challenges: (1) how to quickly and effectively integrate the cross-modal features from the RGB-D data; and (2) how to mitigate the negative impact from the low-quality depth map. The previous methods mostly employ a two-stream architecture which adopts two backbone network to process RGB-D data and ignore the quality of depth map. In this paper, we propose a guided residual network to address these two issues. On the one hand, we design a simpler and efficient depth branch only using one convolutional layer and three residual modules to extract depth features instead of employing a pre-trained backbone to handle the depth data, and fuse RGB features and depth features in a multi-scale manner for refinement with top-down guidance. On the other hand, we add adaptive weight to depth maps to control the fusion between them, which mitigates the negative influence of unreliable depth map. Experimental results compared with 13 state-of-the-art methods on 7 datasets demonstrate the validity of the proposed approach both quantitatively and qualitatively, especially in efficiency (**102 FPS**) and compactness (**64.2 MB**).

Keywords RGB-D salient object detection · Guided residual network · Efficient depth feature learning · Adaptive depth weight

1 Introduction

Salient object detection (SOD) aims at locating most attractive regions in an image. As a pre-processing technique, SOD benefits a variety of computer vision tasks including

re-identification [1], image understanding [2], object tracking [3] and video object segmentation [4], to name a few.

In the past years, CNN-based methods (Convolutional Neural Networks, CNN) have achieved excellent performance in the SOD tasks due to its powerful ability to extract and represent features. Most of them [5–9] pay attention to extract features from RGB images to detect the salient objects. However, it is difficult to accurately locate the salient objects in some challenging and complex scenarios only using single modal data, such as similar appearance between the foreground and background, the cluttered background.

Recently, with the development of depth camera, depth information can be easily obtained and various RGB-D SOD methods [10–16,48] have risen. Depth maps provide SOD with location and spatial structural information and help the network accurately locate the salient object from the complex background.

As shown in Fig. 1a, the existing RGB-D SOD methods [10–14] mostly explore RGB-D data relying on a traditional two-stream architecture, in which an extra backbone network is required to process depth information. This architecture

✉ Shuhan Chen
c.shuhan@gmail.com

Jian Wang
haixiaoqu@163.com

Xiao Lv
lvxiao87@126.com

Xiuqi Xu
frequency.xu@foxmail.com

Xuelong Hu
xlhu@yzu.edu.cn

¹ School of Information Engineering, Yangzhou University, Yangzhou, China

² Chongqing Special Equipment Inspection and Research Institute, Chongqing, China

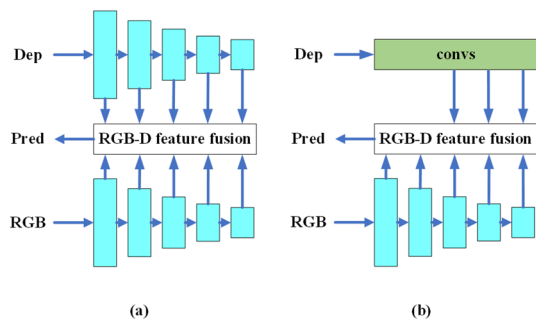


Fig. 1 **a** Existing methods mostly exploiting cross-modal complementarity by a two-stream architecture using two backbone. **b** Our proposed method which designs a lower-complexity depth branch instead of a backbone network and details is illustrated in Fig. 3

brings additional calculation and memory consumption. And the model parameters of the network are huge, which hindered its practical application. To solve this issue, we design a simpler and efficient depth branch to deal with depth information as shown in Fig. 1b instead of using an extra backbone network. Because the depth map itself contains rich location and spatial structure information, there is no need to extract depth information through a complex network. Therefore, we choose to design a simpler depth branch instead of using a complex backbone network and treat depth features extracted from our proposed depth branch as depth prior maps to fuse them with RGB features. In our network, depth data are divided into three different-scales pieces by the improved depth branch to be fused with multi-scales RGB features extracted from the backbone network. The depth map only participates in the high-level stages of the network to reduce the model parameters and calculation cost. We also propose a guided residual module (GRM) to integrate features from the cross-modal RGB-D data seamlessly in a multi-scales and channel-wise manner.

As mentioned above, depth plays an important role in RGB-D SOD and provides the network with location and spatial structural information. However, due to the immaturity of the technology for obtaining depth maps, depth maps sometimes are inaccurate and would contaminate the results of SOD. Previous work generally integrates the RGB and depth information in an indiscriminate manner, which may induce negative results when encountering the inaccurate or blurred depth maps. Hence, we design a depth correction module (DCM) to introduce the adaptive weight to depth map to filter out unreliable depth map. When DCM judges a depth map to be of low quality, this depth map is unreliable and DCM adds a low weight to this depth map to mitigate its negative impact.

In summary, our main contributions are listed as follows:

- (1) Instead of employing a backbone network, we design a simpler and efficient depth branch to extract comple-

mentary features which are fused with RGB features for guided refinement.

- (2) We design DCM to judge the quality of depth maps and add adaptive weight to each depth map to mitigate the negative influence of unreliable depth maps.
- (3) Experiments show competitive performance against 13 state-of-the-art methods on 7 datasets, especially in advantages of efficiency (**102 FPS**) and compactness (**64.2 MB**).

2 Related work

2.1 RGB salient object detection

In recent years, we have witnessed the rapid development of SOD for the RGB image. Numerous models have been presented to explore SOD in terms of boundaries, feature fusion, multi-supervision, pooling, *etc.* Su et al. [5] proposed a boundary-aware network with successive dilation to enhance the feature selectivity at boundaries and guarantee the feature at interiors; Zhao et al. [6] employed pyramid feature attention network to focus on effective high-level context features and low-level spatial structural features; Zheng et al. [7] used multi-source weak supervision for saliency detection; Liu et al. [8] expanded the role of pooling in CNNs to detect saliency; Zhao et al. [9] focused on the complementarity between salient edge information and salient object information. However, in the face of SOD in the challenge and complex scenes, such as low contrast and multi-salient objects, the single-modal SOD models did not perform well.

2.2 RGB-D salient object detection

The pioneering work for RGB-D SOD was produced by Niu et al. [10], who introduced disparity contrast and domain knowledge into stereoscopic photography to measure stereo saliency. After Niu's work, various handcrafted features are originally applied for RGB SOD were extended to RGB-D, such as center-surround difference, contrast and background enclosure.

In the past five years, deep learning-based RGB-D methods have achieved outstanding performance due to the powerful ability of CNNs in extracting salient object representations. Many methods begin to adopt a two-stream architecture which uses two backbone networks (*e.g.*, VGG [17], ResNet [18]) to explore the mining and fusion of the cross-modal RGB-D information. Zhu et al. [11] employed an independent encoder network to take advantage of location and spatial structural information in depth maps and assist the RGB-stream network. Chen et al. [42] exploited the cross-modal complement across all the levels by a complementarity-aware fusion module based on a two-

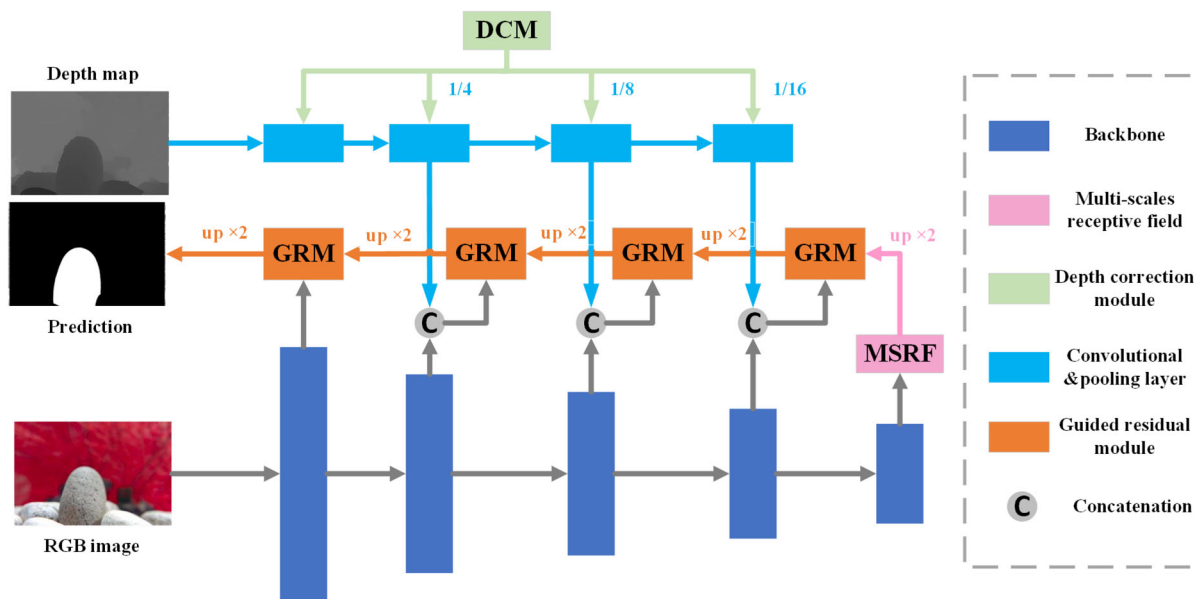


Fig. 2 The schematic illustration of our proposed network. We design an efficient depth branch to improve depth feature learning, MSRF to capture more multi-scale context, DCM to reduce the negative impact

of low-quality depth map and GRM to fuse RGB-D features, which will be introduced in Sects.3.1 to 3.4, respectively

stream structure. Chen et al. [34] proposed a multi-scale multi-path fusion network with cross-modal interactions to enable sufficient and efficient fusion.

Recently, Piao et al. [12] proposed an adaptive and attentive depth distiller to transfer depth information from depth-stream to RGB-stream. Hence, their network needed no more depth maps when testing, which promotes the practical application of RGB-D SOD approaches. Li et al. [14] proposed an attention steered interweave fusion network to detect salient object, which progressively integrates cross-modal and cross-level complementarity from RGB-D images via steering of an attention mechanism. Fu et al. [15] consumed that RGB data and depth information are common and propose a single backbone network to learn from both RGB and depth inputs. Chen et al. [13] introduced depth potentiality-aware mechanism to explicitly model the potentiality of the depth map and effectively integrate the cross-modal complementarity of RGB-D data. Zhang et al. [16] were inspired by the saliency data labeling process and propose probabilistic RGB-D saliency detection network via conditional variational autoencoders to model human annotation uncertainty and generate multiple saliency maps for each input image by sampling in the latent space.

However, most of the above RGB-D approaches [10–14, 16,34,42] focus on the way of integrating RGB-D cross-modal information and employ two pre-trained backbone networks to detect salient object, which require an additional backbone network to process depth data. Different from them, we improve two-stream network by designing

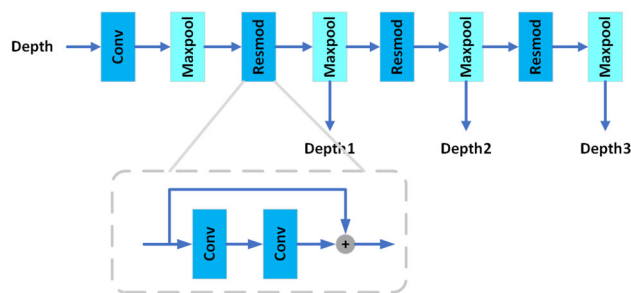


Fig. 3 The structure of the proposed depth branch. “Conv” means a 3×3 convolutional layer. Residual module is denoted as “Resmod”

a simpler and efficient depth branch, which achieved with less computation cost and memory consumption than a traditional two-stream network.

3 The proposed network

3.1 Architecture overview

As shown in Fig. 2, our proposed network is an asymmetric two-stream end-to-end architecture and employs ResNet-50 [18] as the RGB branch. We design a simple and efficient depth branch to extract depth features, which avoids employing an additional backbone network to process depth data and reduces extra calculation cost and memory consumption.

The structure of proposed depth branch is illustrated in Fig. 3. It uses one 3×3 convolutional layer and three residual modules to deal with depth data. The residual modules are composed of two 3×3 convolutional layers. Compared to a backbone network like VGG-16 [17], our proposed depth branch has less convolutional layers and lower complexity. But this does not mean that the performance would be decreased. Because of the rich information of the depth map, a simple depth branch is sufficient to extract the complementary features, which will be verified in the ablation experiment. By improving the way of depth feature learning, we reduce the complexity of the depth branch and use less convolutional layers to achieve better performance.

Since the qualities of depth maps may vary due to the limitations of the depth sensors, we design DCM to add adaptive weight to depth maps by judging the quality of depth maps before concatenating depth and RGB features. After that, we divide depth maps into multi-scales maps to be fused with multi-level RGB features extracted from the backbone network. To gain more high-level semantic information, we adopt a multi-scales receptive field (MSRF) module on the top of the backbone network to enlarge receptive field of the network. On the top-down pathway, we put in multi-GRMs to help integrate depth maps and RGB features seamlessly to acquire predictions in a multi-scales and channel-wise manner. In what follows, we describe the structures of the above-mentioned components and explain their functions in detail.

3.2 Multi-scales receptive field module

The last convolutional layer of the backbone has strong ability to capture semantic information. Therefore, we usually adopt it for global saliency perception to obtain a coarse prediction. Since the scale of salient objects vary from large to small, which implies that the model needs to capture information at different contexts in order to detect objects reliably. However, [43,44] shows that the empirical receptive fields of CNNs are much smaller than the ones in theory especially for deeper layers. The receptive fields of the whole network are not large enough to capture the multi-scale context of the input images. We notice that there are some outstanding structures proposed to solve this issue, such PPM (Pyramid Pooling Module [43]), ASSP (Atrous Spatial Pyramid Pooling [45]), RFB (Receptive Field Block[46]). Different from these parallel concatenation-based methods, we build a hierarchical multi-scales receptive field module to sequentially aggregate the multi-scale contexts. By adding a skip connection into each of the two parallel streams, we not only make pixel sampling denser but also provide larger receptive field.

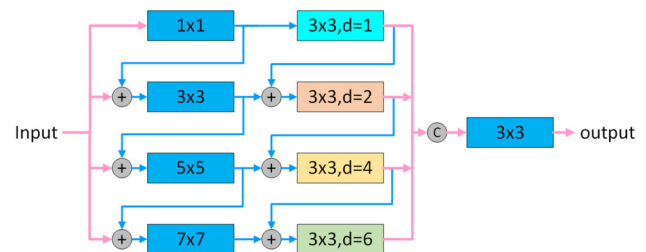


Fig. 4 The structure of the MSRF module. “C” denotes concatenation operation. $k \times k, d$ represents convolutional layer with kernel size k and dilation rate d

3.2.1 Architecture details

In specific, we first reduce the channel into 64 for saving memory. Then, we add four separate branches to capture multi-scale context cues after that, which is inspired from RFB [46]. Each branch consist of two convolutional layers, the first layer is standard convolution with $1 \times 1, 3 \times 3, 5 \times 5, 7 \times 7$ kernel size for dense sampling, the second layer is dilated 3×3 convolution with $\{1, 2, 4, 6\}$ dilation rate for sparse sampling. Different from the previous works, each input is added with the output of its previous branch in each layer except for the first branch. Finally, we concatenate them together and feed it into a 3×3 convolutional layer to generate the single-channel coarse prediction. The whole architecture is illustrated in Fig. 4.

3.3 Depth correction module

The existing approaches [10,12,14,15,34,42] generally integrate RGB and depth features undifferentiated. However, due to the different ways to obtain the depth maps, the quality of them is uneven. The low-quality depth maps cannot guide the network to learn salient regions but have a negative effect on the predictions. To solve this issue, we design a DCM to add adaptive weight to depth maps by judging the quality of them, which reduces the negative impact of low-quality depth maps on prediction results. If the quality of depth map is high, DCM would add a high weight to it and vice versa.

Usually, a high-quality depth map should have many similarities with the ground truth. Based on that, we employ Structural Similarity (SSIM [19]) function to calculate the similarity between depth map and ground truth. SSIM function is designed to calculate the luminance, contrast and structure similarity between two images, which is defined as:

$$SSIM(x, y) = \frac{(2\mu_x\mu_y + C_1)(2\sigma_{xy} + C_2)}{(\mu_x^2 + \mu_y^2 + C_1)(\sigma_x^2 + \sigma_y^2 + C_2)} \quad (1)$$

where C_1 and C_2 mean two constants that prevents the denominator from 0, μ_x to μ_y are denoted as mean of pixel

values, σ_x and σ_y , represents standard deviation of pixel values.

Then, we use the value calculated by SSIM function as adaptive weight and add it to the depth map to get an improved depth map. In addition to screening the quality of the depth maps, DCM can also filter out redundant information. When a depth map is low-quality, the weight added by DCM is low, which means less information of it participating in network computing.

There are also several existing methods proposed to handle the issue of uneven-quality depth maps. Chen and Huang [13] concatenate high-level RGB-D features and compute depth weight value between RGB-D features and ground truth. Zhang et al. [16] design a sub-network to refine depth maps. Although these methods alleviate the negative effects of low-quality depth maps in some ways, they increase the complexity of the network and also the running time of the model. Different from [13,16], our proposed DCM can be treated as pre-processing operations because of using original RGB-D data to calculate adaptive depth weight, which brings little extra computing cost to the network.

3.4 Guided residual module

After getting multi-level RGB-D features, we also need to fuse them to acquire saliency prediction. Low-level features contain rich detail information, such as boundary, texture and spatial structure information, and high-level features capture rich semantic information. We employ U-shape architecture to combine multi-level RGB-D features. However, one of the problems to this type of U-shape architectures is that the high-level features will be gradually diluted when they are transmitted to shallow layers. To solve this issue and help the network integrate RGB-D features seamlessly, we design GRM to mitigate the dilution of confident semantic information with the help of top-down guidance.

As can be seen in Fig. 5, the N -channel feature maps are firstly split into N non-overlapped groups, each of which consists of 1-channel feature map. Then, the side output prediction is used as a guidance feature map to be concatenated with the 1-channel feature map, thus, we can get $2N$ -channel feature maps in total. Several 3×3 convolution layers are applied for guided learning to obtain a 1-channel feature map, which are added with the input guidance map as new side-output prediction. By concatenating high-level prediction with each channel of the side-output features, our approach can well relieve the dilution of the high-level semantic information in the feature fusion process of the U-Net architecture.

3.5 Loss function

We apply the binary cross-entropy (BCE) loss [20] and intersection over union (IoU) loss [21] to optimize the network,

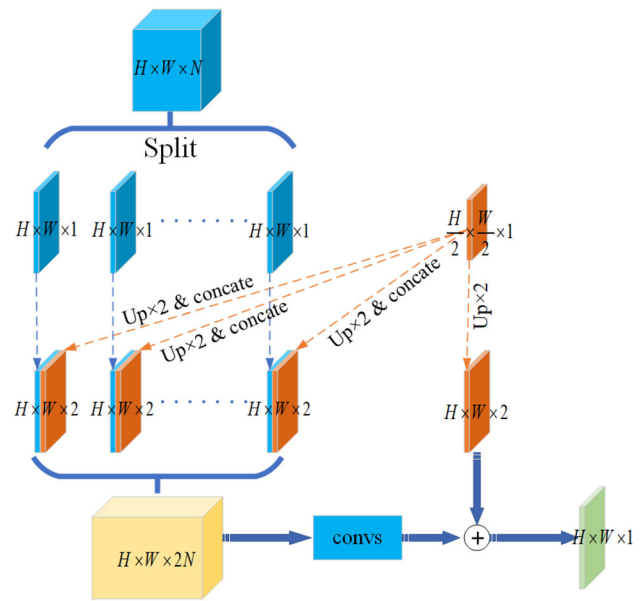


Fig. 5 The construction of the GRM. $up \times 2 \& concatenate$ represents upsampling operation with a factor of 2 and concatenation operation. $H \times W \times N$ means images with length H and width W with N channels

where the BCE and IoU are used to constrain the saliency prediction in pixel-level and image-level, respectively.

In SOD, BCE loss is commonly employed to measure the relation between predicted saliency map and the ground truth, which is defined as:

$$l_{bce} = - \sum_{(r,c)} [G(r,c) \log(S(r,c)) + (1 - G(r,c)) \log(1 - S(r,c))] \quad (2)$$

where $G(r,c) \in \{0, 1\}$ is the ground truth label of the pixel (r,c) and $S(r,c)$ is the predicted probability of being salient object.

IoU loss is originally proposed for measuring the similarity of two sets and then used as a standard evaluation measure for object detection and segmentation. Recently, it has been used in SOD, which is defined as:

$$l_{iou} = 1 - \frac{\sum_{r=1}^H \sum_{c=1}^W S(r,c)G(r,c)}{\sum_{r=1}^H \sum_{c=1}^W [S(r,c) + G(r,c) - S(r,c)G(r,c)]} \quad (3)$$

where $G(r,c) \in \{0, 1\}$ is the ground truth label of the pixel (r,c) and $S(r,c)$ is the predicted probability of being salient object.

In our training process, we combine BCE loss with IoU loss to obtain BCE-IoU loss for optimization, which is defined as:

$$L_{bce+iou} = l_{bce} + l_{iou}. \quad (4)$$

4 Experiments

4.1 Experimental setup

4.1.1 Datasets

We evaluate the proposed approach on 7 public RGB-D SOD datasets. NJUD [22] consists of 1985 RGB images and corresponding depth images with various objects and complex scenarios. The depth images are estimated from the stereo images. NLPR [23] contains 1000 RGB-D images with pixel-wise ground truth, where the depth maps are captured by Microsoft Kinect under different illumination conditions and acquisition scenes. STERE [10] collects 1000 paired RGB-D images, where the depth maps are also estimated from the stereo images. LFSO [24] is constructed for light field saliency detection, which contains 100 all-focus RGB images, the corresponding depth maps and the pixel-wise ground truth. The depth map is captured by the Lytro light field camera. SSD [25] contains 80 images picked up from three stereo movies, where the depth map is generated by depth estimation approach. DUT [26] consists of 1200 paired images containing more complex scenarios, such as multiple or transparent objects. SIP [27] is a new released dataset which contains 929 high-resolution person RGB-D images captured by Huawei Meta10.

4.1.2 Implementation details

Training dataset is the same as [26], which selects 1487 images from NJUD [22], 700 images from NLPR [23] and 800 images from DUT [26]. To prevent the overfitting, we augment the training set by flipping, cropping, rotating and light changing. In this work, we train two versions which employ VGG-16 [17] and ResNet-50 [18] as backbone network, respectively. The RGB branch is initialized by VGG-16 and ResNet-50, and the others are using the default setting of the PyTorch. We implement the proposed network with PyTorch on a PC with an Intel i9 9900K CPU, 32GB RAM, and an NVIDIA GeForce 2080Ti GPU. All the experiments are performed using the Adam [47] optimizer with an initial learning rate of $5e-5$ which is divided by 10 after 20 epochs. Our network is trained for 30 epochs in total.

4.1.3 Evaluation metrics

We employ the precision–recall (P–R) curve, F-measure [28], S-measure [29,30], E-measure [27,31] and mean absolute error (MAE) for quantitative evaluations. Thresholding the saliency map at a series of values, pairs of precision–recall

value can be computed by comparing the binary saliency map with the ground truth. The F-measure is a comprehensive measurement, which is defined as the weighted harmonic mean of precision and recall:

$$F_{\beta} = \frac{(1+\beta^2) \cdot \text{Precision} \cdot \text{Recall}}{\beta^2 \cdot \text{Precision} + \text{Recall}} \quad (5)$$

where we set β^2 to be 0.3 as set in [28]. In this paper, we only report the maximum F-measure score.

S-measure evaluates the structure similarity between the saliency map and ground truth, which is defined as:

$$S_{\alpha} = \alpha \times S_o + (1 - \alpha) \times S_r \quad (6)$$

where α is set to 0.5 for assigning equal contribution to both region (S_r) and object (S_o) similarity.

E-measure considers the local pixel-wise values and the image-level mean value together, which is consistent with cognitive vision studies. It is defined as:

$$E_{\xi} = \frac{1}{W \times H} \sum_{x=1}^W \sum_{y=1}^H f\left(\frac{2\varphi_{GT} \circ \varphi_{FM}}{\varphi_{GT} \circ \varphi_{GT} + \varphi_{FM} \circ \varphi_{FM}}\right) \quad (7)$$

where φ is the bias matrix as the distance between each pixel-wise value of ground truth and its image-level mean, *i.e.*, φ_{GT} and φ_{FM} are for ground truth and binary foreground map, respectively, and $f(\cdot)$ is a quadratic function. “ \circ ” denotes dot product.

MAE is adopted to evaluate the non-salient region average per-pixel difference, which is defined as

$$MAE = \frac{1}{H \cdot W} \sum_{x=1}^H \sum_{y=1}^W |S(x, y) - G(x, y)| \quad (8)$$

where H and W are the height and width of the saliency map S and G denotes its ground truth. It needs to be pointed out that higher F-measure and lower MAE score denote better performance.

4.2 Compared with the state-of-the-arts

We compare our network with other 13 state-of-the-art methods, including DF [32], PCAN [36], CTMF [33], MMCI [34], TANet [35], CPFP [37], DMRA [26], D3Net [27], ICNet [38], S2MA [40], SSF [39], UCNet [16], JLDCF [15]. For fair comparisons, we use their released saliency maps or adopt the released code and their default parameters to reproduce their results.

Table 1 Quantitative measures: S-measure (S_α), F-measure (F_β), E-measure (E_ξ), MAE (M) of SOTA methods and our proposed approach on seven RGB-D datasets

Metric	DF [32]	PCAN [36]	CTMF [33]	MMCI [34]	TANet [35]	CPPP [37]	DMRA [26]	D3Net [27]	ICNet [38]	S2MA [40]	SSF [39]	UCNet [16]	JLDCF [16]	Ours (vgg)	Ours (res)
DUT	$S_\alpha \uparrow$	0.733	0.801	0.831	0.791	0.808	0.749	0.889	0.852	0.903	0.916	-	0.905	0.923	0.929
	$F_\beta \uparrow$	0.764	0.760	0.792	0.753	0.779	0.736	0.884	0.830	0.866	0.914	-	0.883	0.924	0.932
	$E_\xi \uparrow$	0.848	0.858	0.882	0.855	0.866	0.815	0.927	-	0.897	0.921	0.946	-	0.931	0.946
LFSD	$M \downarrow$	0.144	0.100	0.097	0.113	0.093	0.100	0.048	0.072	0.044	0.034	-	0.043	0.031	0.029
	$S_\alpha \uparrow$	0.791	0.794	0.796	0.787	0.801	0.828	0.847	0.868	0.837	0.859	0.864	0.861	0.858	0.879
	$F_\beta \uparrow$	0.806	0.792	0.782	0.779	0.794	0.813	0.849	0.861	0.820	0.867	0.859	0.854	0.863	0.878
NJUD	$E_\xi \uparrow$	0.844	0.842	0.851	0.840	0.845	0.867	0.899	0.891	0.863	0.895	0.897	0.882	0.887	0.912
	$M \downarrow$	0.138	0.112	0.119	0.132	0.111	0.088	0.075	0.071	0.094	0.066	0.066	0.070	0.073	0.064
	$S_\alpha \uparrow$	0.768	0.877	0.849	0.859	0.878	-	0.886	0.894	0.894	0.894	0.898	0.897	0.902	0.912
NLPR	$F_\beta \uparrow$	0.783	0.844	0.788	0.813	0.844	-	0.872	0.868	0.865	0.885	0.889	0.885	0.904	0.905
	$E_\xi \uparrow$	0.839	0.896	0.864	0.878	0.893	-	0.908	0.905	0.896	0.912	0.903	0.913	0.924	0.926
	$M \downarrow$	0.136	0.059	0.085	0.079	0.060	-	0.051	0.052	0.053	0.043	0.043	0.041	0.037	0.036
SIP	$S_\alpha \uparrow$	0.806	0.874	0.860	0.856	0.886	0.888	0.899	0.906	0.915	0.914	0.919	0.925	0.927	0.932
	$F_\beta \uparrow$	0.759	0.795	0.723	0.730	0.796	0.822	0.854	0.870	0.852	0.873	0.890	0.878	0.901	0.906
	$E_\xi \uparrow$	0.884	0.916	0.869	0.872	0.916	0.924	0.941	0.932	0.944	0.937	0.949	0.953	0.958	0.961
SSD	$M \downarrow$	0.079	0.044	0.056	0.059	0.041	0.036	0.031	0.034	0.030	0.026	0.025	0.022	0.022	0.020
	$S_\alpha \uparrow$	0.653	0.842	0.716	0.833	0.835	0.850	0.806	0.864	0.690	0.874	0.875	0.880	0.881	0.884
	$F_\beta \uparrow$	0.673	0.825	0.684	0.795	0.809	0.819	0.819	0.831	0.836	0.615	0.868	0.873	0.856	0.859
STERE	$E_\xi \uparrow$	0.794	0.899	0.824	0.886	0.893	0.899	0.863	0.902	0.796	0.916	0.913	0.921	0.912	0.913
	$M \downarrow$	0.185	0.071	0.139	0.086	0.075	0.064	0.085	0.063	0.159	0.053	0.051	0.049	0.050	0.049
	$S_\alpha \uparrow$	0.747	0.841	0.776	0.813	0.839	0.807	0.857	-	0.868	0.844	-	-	0.863	0.875
SSD	$F_\beta \uparrow$	0.724	0.791	0.710	0.748	0.767	0.726	0.821	-	0.818	0.812	-	-	0.826	0.854
	$E_\xi \uparrow$	0.812	0.886	0.838	0.860	0.879	0.832	0.892	-	0.891	0.889	-	-	0.898	0.906
	$M \downarrow$	0.142	0.062	0.099	0.082	0.063	0.082	0.058	-	0.052	0.058	-	-	0.056	0.045
STERE	$S_\alpha \uparrow$	0.757	0.875	0.848	0.873	0.871	0.879	0.834	0.891	0.890	0.893	0.903	0.903	0.911	0.911
	$F_\beta \uparrow$	0.742	0.826	0.771	0.829	0.835	0.830	0.844	0.865	0.855	0.880	0.885	0.869	0.896	0.897
	$E_\xi \uparrow$	0.838	0.897	0.864	0.901	0.906	0.903	0.899	0.904	0.915	0.907	0.922	0.919	0.927	0.928
STERE	$M \downarrow$	0.141	0.064	0.086	0.068	0.060	0.051	0.066	0.054	0.051	0.044	0.039	0.040	0.035	0.034

The best three scores are highlighted in **bold italic**, **bold** and **italic**, respectively

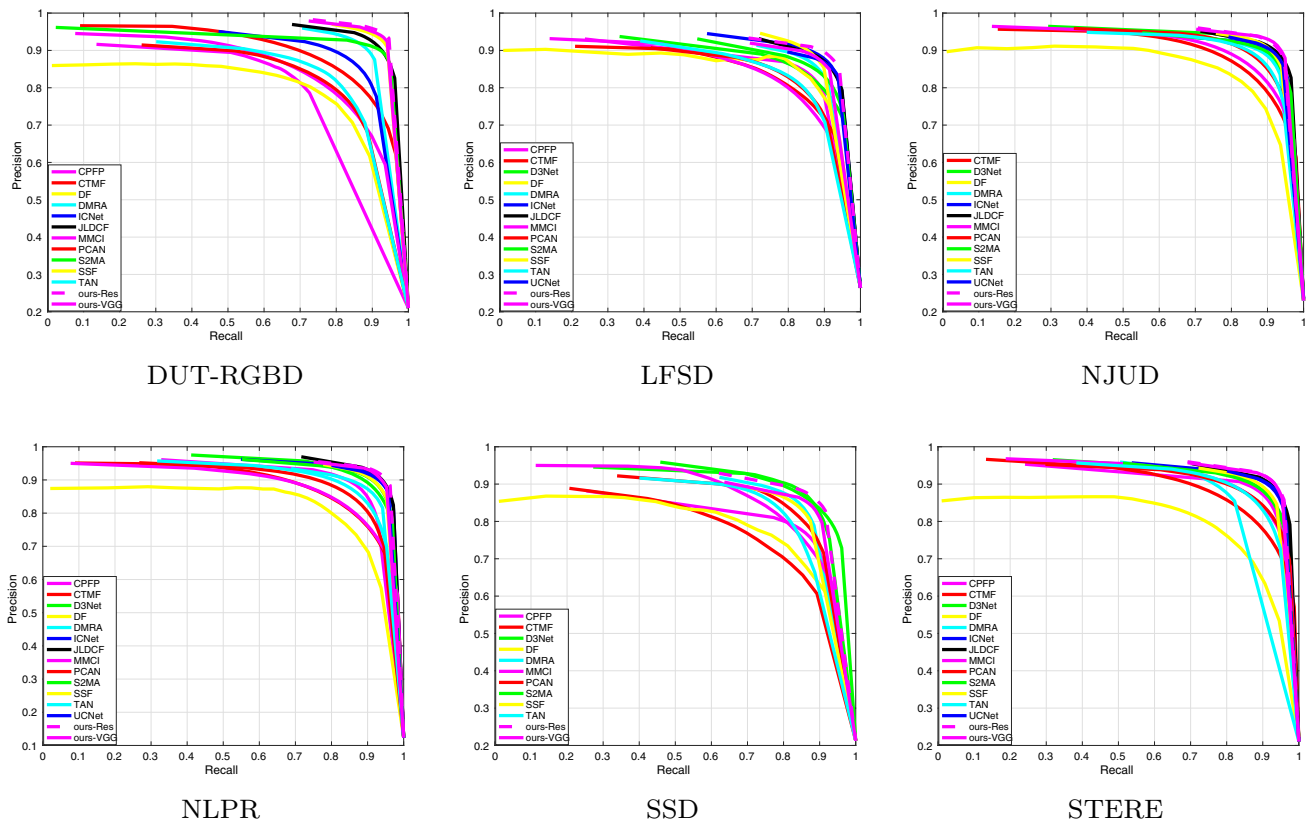


Fig. 6 Comparison of PR curves on six datasets. Best viewed on the screen

4.2.1 Quantitative evaluation

We present the quantitative comparison results in Table 1 and Fig. 6. We can clearly find that our method consistently outperforms the other methods on all the seven datasets with respect to *S*-measure, *F*-measure, *E*-measure and MAE scores, except JLDCF on SIP dataset. It needs to point that JLDCF adopted a backbone network initialized by the pre-trained parameters of DSS [41], which is trained on a RGB SOD dataset with 2500 images. Nevertheless, we still perform better than JLDCF on the other datasets. The PR curves also indicate the superior performance of our method, which are shown in Fig. 5. Thus, the above quantitative evaluation demonstrates the effectiveness and superiority of our proposed method on detecting salient objects.

4.2.2 Qualitative evaluation

In order to show our results more intuitively, we provide some visual representative saliency maps of different methods to demonstrate the superiority of our proposed network. As can be seen in Fig. 7, the salient regions are highlighted more accurately by our method even in some challenging cases. And our results have more clear boundaries compared to other methods.

4.2.3 Complexity evaluation

Moreover, we further compare the FPS (Frames Per Second) and model size with other models for complexity evaluation as shown in Table 2. It can be observed that both running speed and model size of our proposed model performs better than other models. Our best performing model ResNet-50 [18] version runs 33% faster than S2MA and minimizes the model size 30% than SSF. The training process of our proposed work takes about 4h and it runs at a speed of 90 FPS for 352×352 images without any other post-processing.

4.3 Ablation study

In this section, we carry out the ablation studies to demonstrate the effectiveness of the proposed module components. We use VGG-16 as backbone for RGB branch in the following experiments.

4.3.1 Module components

To verify the effectiveness of the modules proposed in this paper, we conduct some experiments to evaluate their performance with different combinations. We select the network which removes DCM, MSRF and GRM as baseline. In the

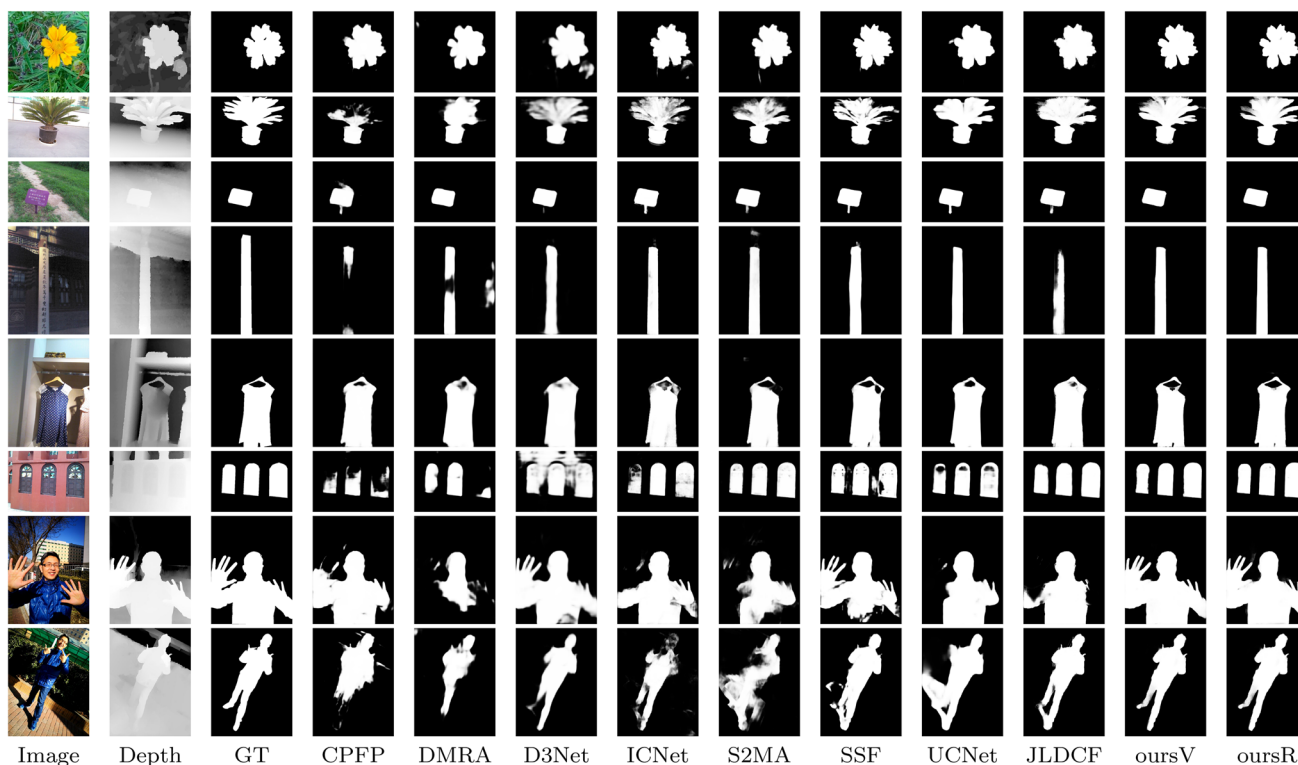


Fig. 7 Visual comparisons of GRN with SOTA RGB-D saliency models. “oursV” means GRN with VGG-16 version. “oursR” represents GRN with ResNet-50 version

baseline, GRMs are replaced with element-wise summation and 3×3 convolution operation. The results are shown in Table 2. We find that when using a single module, the best network performance is GRM, indicating that GRM plays a great role in fully integrating RGB-D features. When using two modules, GRM+MSRF achieved the best performance, indicating MSRF has expanded the receptive field of the network thus captured more advanced semantic information. When all three modules are combined, it can be seen that DCM alleviated the negative impact of low-quality depth maps on the network thus further improved the performance.

4.3.2 Fusion of RGB and depth

As shown in Table 4, summation and concatenation operation in the RGB-D fusion process almost achieved the same performance. It is worthy to mention that even such simple fusion can obtain good performance contributed by the proposed effective depth branch. Considering the test speed of the network, the concatenation operation is better. So we employ a simpler and faster concatenation operation as our fusion strategy.

Table 2 Running speed and model size comparisons with recent models

Method	Platform	FPS \uparrow	MS(MB) \downarrow
CPFP	Caffe	20	291.9
DMRA	PyTorch	16	238.8
S2MA	PyTorch	61	346.8
SSF	PyTorch	26	131.8
Ours(Vgg)	PyTorch	102	64.2
Ours(Res)	PyTorch	90	108.4

4.3.3 Depth branch

We select three designs of the depth branch for comparisons, which are VGG-16 [17], the proposed depth branch and without using depth branch. It can be seen from Table 5 that the depth branch we designed has a better performance compared with VGG-16 and our speed is nearly twice of it. Further compared with the model without using depth branch, we perform much better, and the speed is only reduced by nearly 10 FPS.

Table 3 Quantitative evaluation for ablation studies about the effectiveness of DCM, MSRF and GRM

	DCM	MSRF	GRM	DUT				NJUD			
				$S_\alpha \uparrow$	$F_\beta \uparrow$	$E_\xi \uparrow$	$M \downarrow$	$S_\alpha \uparrow$	$F_\beta \uparrow$	$E_\xi \uparrow$	$M \downarrow$
✓				0.863	0.876	0.908	0.055	0.868	0.873	0.907	0.055
		✓		0.895	0.904	0.935	0.042	0.884	0.889	0.911	0.048
			✓	0.905	0.912	0.937	0.039	0.904	0.900	0.923	0.041
✓	✓			0.897	0.906	0.935	0.042	0.885	0.886	0.917	0.048
✓			✓	0.909	0.917	0.940	0.036	0.906	0.903	0.924	0.039
		✓	✓	0.918	0.918	0.941	0.033	0.909	0.897	0.917	0.039
✓	✓	✓	✓	0.923	0.924	0.946	0.031	0.912	0.904	0.924	0.037

Table 4 Quantitative evaluation for ablation studies about different fusion methods of RGB features and depth maps

Metric	DUT		NLPR	
	Sum	Cat	Sum	Cat
$S_\alpha \uparrow$	0.922	0.923	0.928	0.927
$F_\beta \uparrow$	0.924	0.924	0.901	0.901
$E_\xi \uparrow$	0.945	0.946	0.957	0.958
$M \downarrow$	0.031	0.031	0.021	0.022
FPS \uparrow	83	102	87	102

Table 5 Quantitative evaluation for ablation studies about different designs of the depth branch

Metric	Ours	VGG-16	w/o
SIP			
$S_\alpha \uparrow$	0.881	0.876	0.866
$F_\beta \uparrow$	0.856	0.848	0.830
$E_\xi \uparrow$	0.912	0.907	0.887
$M \downarrow$	0.050	0.053	0.060
FPS \uparrow	102	57	110

5 Conclusion

In this paper, we proposed guided residual network for RGB-D SOD. Instead of employing a traditional two-stream architecture which uses two backbone networks, we propose a simpler and efficient depth branch for extracting depth features. Based on this depth branch, we can quickly and efficiently integrate depth and RGB features. We also proposed DCM by introducing adaptive weight to each depth map to mitigate the negative influence of unreliable depth map. In terms of feature fusion, multi-scale RGB-D features are fused in the top-down pathway by the proposed GRM. Contributed by them, the proposed network located the salient regions and object boundaries accurately and efficiently. Experiments with 13 state-of-the-art methods on seven datasets demonstrated the superior performance of the proposed network. Our proposed network is simple and effective with fast running speed and compact mode size. Thus, we believe our

designed depth branch can be applied into other two-stream RGB-D SOD methods to further improve their performance. In our future works, we will continue to explore more effective fusion strategy to further improve performance.

Acknowledgements This work was supported by the Natural Science Foundation of China (No. 61802336), Jiangsu Province 7th Projects for Summit Talents in Six Main Industries, Electronic Information Industry (DZXX-149, No.110).

Declarations

Conflict of interest We declare that we have no financial and personal relationships with other people or organizations that can inappropriately influence our work.

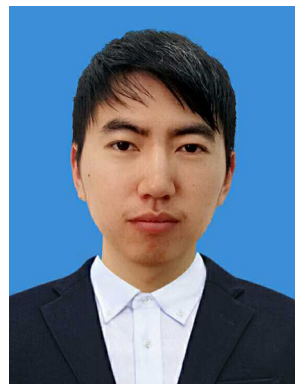
Open Access This article is distributed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made.

References

- Zhao, R., Ouyang, W., Wang, X.: Unsupervised saliency learning for person re-identification. In: CVPR, pp. 3586–3593 (2013)
- Zhang, F., Bo, D., Zhang, L.: Saliency-guided unsupervised feature learning for scene classification. IEEE TGRS **53**(4), 2175–2184 (2014)
- Hong, S., You, T., Kwak, S., Han, B.: Online tracking by learning discriminative saliency map with convolutional neural network. In: ICML, pp. 597–606 (2015)
- Wang, W., Shen, J., Yang, R., Porikli, F.: Saliency-aware video object segmentation. IEEE TPAMI **40**(1), 20–33 (2017)
- Su, J., Li, J., Zhang, Y., Xia, C., Tian, Y.: Selectivity or Invariance: Boundary-aware Salient Object Detection. In: ICCV, pp. 3799–3808 (2019)
- Zhao, T., Wu, X.: Pyramid Feature Attention Network for Saliency detection. In CVPR, pp. 3085–3094 (2019)
- Zeng, Y., Zhuge, Y., Lu, H., Zhang, L., Qian, M., Yu, Y.: Multi-source weak supervision for saliency detection. In: CVPR, pp. 6074–6083 (2009)
- Liu, J.-Ji., Hou, Q., Cheng, M.-M., Feng, J., Jiang, J.: A Simple Pooling-Based Design for Real-Time Salient Object Detection. In: CVPR, pp. 3917–3926 (2019)

9. Zhao, J.-X., Liu, J.-J., Fan, D.-P., Cao, Y., Yang, J.-F., Cheng, M.-M.: EGNNet: Edge Guidance Network for Salient Object Detection. In: ICCV, pp. 8779–8788 (2019)
10. Niu, Y., Geng, Y., Li, X., Liu, F.: Leveraging stereopsis for saliency analysis. In: CVPR, pp. 454–461 (2012)
11. Zhu, C., Cai, X., Huang, K., Li, T.H., Li, G.: Pdnet: Prior-model guided depth-enhanced network for salient object detection. In: ICME, pp. 199–204. IEEE (2019)
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: CVPR, pp. 9060–9069 (2020)
13. Chen, Z., Huang, Q.: Depth Potentiality-Aware Gated Attention Network for RGB-D Salient Object Detection. In: CVPR (2020)
14. Li, C., Cong, R., Kwong, S., Hou, J., Fu, H., Zhu, G., Zhang, D., Huang, Q.: ASIF-Net: attention steered interweave fusion network for RGB-D salient object detection. IEEE Trans. Cybern **51**(1), 88–100 (2021)
15. Fu, K., Fan, D.-P., Ji, G.-P., Zhao, Q.: JL-DCF: Joint learning and densely-cooperative fusion framework for RGB-D salient object detection. In: CVPR, pp. 3052–3062 (2020)
16. Zhang, J., Fan, D.-P., Dai, Y., Anwar, S.: UC-Net: Uncertainty inspired RGB-D saliency detection via conditional variational autoencoders. In: CVPR, pp. 8582–8591 (2020)
17. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. In: ICLR (2015)
18. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR, pp. 770–778 (2016)
19. Wang, Z., Simoncelli, E.P., Bovik, A.C.: Multi-scale structural similarity for image quality assessment. In: the Thirty-Seventh Asilomar Conference on Signals, Systems and Computers, 2003, vol. 2, pp. 1398–1402. Ieee (2003)
20. de Boer, P.-T., Kroese, D.P., Mannor, S., Rubinstein, R.Y.: A tutorial on the cross-entropy method. Ann. OR **134**(1), 19–67 (2005)
21. Mattyus, G., Luo, W., Urtasun, R.: Deep-roadmapper: Extracting road topology from aerial images
22. Ju, R., Ge, L., Geng, W., Ren, T., Wu, G.: Depth saliency based on anisotropic center-surround difference. In: ICIP, pp. 1115–1119 (2014)
23. Peng, H., Li, B., Xiong, W., Hu, W., Ji, R.: RGBD salient object detection: A benchmark and algorithms. In: ECCV, pp. 92–109 (2014)
24. Li, N., Ye, J., Ji, Y., Ling, H., Yu, J.: Saliency detection on light field. In: CVPR, pp. 2806–2813 (2014)
25. Zhu, C., Li, G.: A three-pathway psychobiological framework of salient object detection using stereoscopic technology. In: ICCV, pp. 3008–3014 (2017)
26. Piao, Y., Ji, W., Li, J., Zhang, M., Lu, H.: Depth-induced multi-scale recurrent attention network for saliency detection. In: ICCV (2019)
27. Fan, D.-P., Lin, Z., Zhang, Z., Zhu, M., Cheng, M.-M.: Rethinking RGB-D salient object detection: Models, datasets, and large-scale benchmarks. IEEE TNNLS (2020)
28. Borji, A., Cheng, M.-M., Jiang, H., Li, J.: Salient object detection: A benchmark. IEEE TIP **24**(12), 5706–5722 (2015)
29. Fan, D.-P., Cheng, M.-M., Liu, Y., Li, T., Borji, A.: Structure-measure: A New Way to Evaluate Foreground Maps. In: ICCV, pp. 4548–4557 (2017)
30. Zhao, J.-X., Cao, Y., Fan, D.-P., Cheng, M.-M., Li, X.-Y., Zhang, L.: Contrast prior and fluid pyramid integration for rgb-d salient object detection. In: CVPR, pp. 3927–3936 (2019)
31. Fan, D.-P., Gong, C., Cao, Y., Ren, B., Cheng, M., Borji, A.: Enhanced-alignment measure for binary foreground map evaluation. In: IJCAI, pp. 698–704 (2018)
32. Liangqiong, Q., He, S., Zhang, J., Tian, J., Tang, Y., Yang, Q.: Rgb-d salient object detection via deep fusion. IEEE TIP **26**(5), 2274–2285 (2017)
33. Han, J., Chen, H., Liu, N., Chenggang Y, Xuelong L.: CNNs-based RGB-D saliency detection via cross-view transfer and multiview fusion. IEEE TCYB, pp. 3171–3183 (2018)
34. Chen, H., Li, Y., Dan, S.: Multi-modal fusion network with multi-scale multi-path and cross-modal interactions for RGB-D salient object detection. PR **86**, 376–385 (2019)
35. Chen, H., Li, Y.: Three-stream Attention-aware Network for RGB-D Salient Object Detection. IEEE TIP, pp. 2825–2835 (2019)
36. Chen, H., Li, Y.: Progressively complementarity-aware fusion network for RGB-D Salient Object Detection. In: IEEE CVPR, pp. 3051–3060 (2018)
37. Zhao, J.-X., Cao, Y., Fan, D.-P., Cheng, M.-M., Li, X.-Y., Zhang, L.: Contrast Prior and Fluid Pyramid Integration for RGBD Salient Object Detection. In: IEEE CVPR (2019)
38. Li, G., Liu, Z., Ling, H.: ICNet: information conversion Nnetwork for RGB-D based salient object detection. IEEE Trans. Image Process. **29**, 4873–4884 (2020)
39. Zhang, M., Ren, W., Piao, Y., Rong, Z., Lu, H.: Select, Supplement and Focus for RGB-D Saliency Detection. In: CVPR, pp. 3472–3481 (2020)
40. Liu, N., Zhang, N., Han, J.: Learning Selective Self-Mutual Attention for RGB-D Saliency Detection. In: CVPR, pp. 13756–13765 (2020)
41. Hou, Q., Cheng, M.-M., Hu, X., Borji, A., Tu, Z., Torr, P.H.S.: Deeply supervised salient object detection with short connections. IEEE TPAMI **41**(4), 815–828 (2019)
42. Chen, H., Li, Y.: Progressively complementarity-aware fusion network for rgb-d salient object detection. In: CVPR, pp. 3051–3060 (2018)
43. Zhao, H., Shi, J., Qi, X., Wang, X., Jia, J.: Pyramid scene parsing network. In: CVPR (2017)
44. Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., Torralba, A.: Object detectors emerge in deep scene cnns. In: ICLR (2015)
45. Chen, L.-C., Zhu, Y., Papandreou, G., Schroff, F., Adam, H.: Encoder-Decoder with Atrous Separable Convolution for Semantic Image Segmentation. In: ECCV (2018)
46. Liu, S., Huang, D., Wang, Yu.: Receptive Field Block Net for Accurate and Fast Object Detection. In: ECCV, pp. 385–400 (2018)
47. Diederik, Kingma, P., Ba, J.: Adam: A method for stochastic optimization. In: ICLR (2015)
48. Liu, Z., Duan, Q., Shi, S., et al.: Multi-level progressive parallel attention guided salient object detection for RGB-D images. Vis Comput (2020). <https://doi.org/10.1007/s00371-020-01821-9>

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Jian Wang received the B.E. degree from Yangzhou University, Yangzhou, China, in 2018, where he is currently pursuing the master's degree. His current research interests include salient object detection and shadow detection.



Shuhan Chen is currently working as a lecturer at College of Information Engineering in Yangzhou University. Before that, he received his B.E. and Ph.D. degrees from Chongqing University, China, in 2009 and 2013, respectively. His research interests include deep learning-based saliency detection and binary segmentation.



Xiuqi Xu received the B.E. degree from Yangzhou University, Yangzhou, China, in 2017, where he is currently pursuing the master's degree. His current research interests include camouflage object detection.



Xiao Lv is currently working as a senior engineer at Chongqing Special Equipment Inspection and Research Institute. Before that, he received his B.E. and Ph.D. degrees from Chongqing University, China, in 2009 and 2013, respectively. His research interests include digital image processing.



Xuelong Hu is a professor and discipline leader of Information and Communication Engineering in Yangzhou University, China. He is also a council member of China Society of Image Graphics (CSIG). He has undertaken one of Jiangsu Province Seventh Batch of Projects, for "Summit Talents in Six Main Industries." His current research interests include computer vision, image processing and Internet of things.