RGB-D Salient Object Detection with Cross-Modality Modulation and Selection

Chongyi Li[1], Runmin Cong[2(✉)], Yongri Piao[3], Qianqian Xu[4], and Chen Change Loy[1]

[1] Nanyang Technological University, Singapore, Singapore
lichongyi25@gmail.com, ccloy@ntu.edu.sg
[2] Beijing Jiaotong University, Beijing, China
rmcong@bjtu.edu.cn
[3] Dalian University of Technology, Dalian, China
yrpiao@dlut.edu.cn
[4] Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China
xuqianqian@ict.ac.cn
https://li-chongyi.github.io/Proj_ECCV20

Abstract. We present an effective method to progressively integrate and refine the cross-modality complementarities for RGB-D salient object detection (SOD). The proposed network mainly solves two challenging issues: 1) how to effectively integrate the complementary information from RGB image and its corresponding depth map, and 2) how to adaptively select more saliency-related features. *First*, we propose a cross-modality feature modulation (cmFM) module to enhance feature representations by taking the depth features as prior, which models the complementary relations of RGB-D data. *Second*, we propose an adaptive feature selection (AFS) module to select saliency-related features and suppress the inferior ones. The AFS module exploits multi-modality spatial feature fusion with the self-modality and cross-modality interdependencies of channel features are considered. *Third*, we employ a saliency-guided position-edge attention (sg-PEA) module to encourage our network to focus more on saliency-related regions. The above modules as a whole, called cmMS block, facilitates the refinement of saliency features in a coarse-to-fine fashion. Coupled with a bottom-up inference, the refined saliency features enable accurate and edge-preserving SOD. Extensive experiments demonstrate that our network outperforms state-of-the-art saliency detectors on six popular RGB-D SOD benchmarks.

C. Li and R. Cong—equal contribution.

Electronic supplementary material The online version of this chapter (https://doi.org/10.1007/978-3-030-58598-3_14) contains supplementary material, which is available to authorized users.

A. Vedaldi et al. (Eds.): ECCV 2020, LNCS 12353, pp. 225–241, 2020.
https://doi.org/10.1007/978-3-030-58598-3_14

1 Introduction

Depth maps provide useful cues such as depth of field, shape, and boundary to complement RGB images for SOD [3,4,6,32,33,39,46]. However, depth maps are inherently noisy and the cues provided can be inconsistent or misaligned with the RGB modality. The issues make designing an RGB-D algorithm challenging. Contemporary RGB-D SOD detectors, CPFP [46] (Fig. 1(d)) and A2dele [33] (Fig. 1(e)), could still miss salient objects due to cluttered backgrounds or yield incomplete or serrated boundaries of saliency maps.

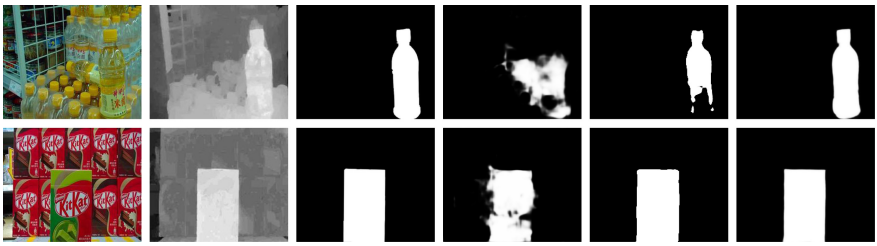

Fig. 1. Two motivating examples of SOD. (a)–(c) represent the input images, the corresponding depth maps, and the ground truth (GT), respectively. (d) and (e) are the results of state-of-the-art RGB-D SOD detectors CPFP (**CVPR'19**) [46] and A2dele (**CVPR'20**) [33], respectively. (f) are our results. *Compared with the latest CPFP and A2dele, our method can yield more complete, sharp, and edge-preserving saliency detection results by effectively integrating cross-modality complementaries and adaptively selecting saliency-related features.*

In this work, we consider addressing the aforementioned problem through more careful investigation on the integration of cross-modality complementaries from RGB image and depth map as well as the selection of saliency-related features. To this end, we present an effective network that achieves complete, sharp, and edge-preserving saliency detection, as shown in Fig. 1(f).

First, we propose a cross-modality feature modulation (cmFM) module that enhances RGB feature representations by taking the corresponding depth features as prior. This is in contrast to popular strategies that perform either input fusion [30], early fusion [19], or late fusion [18], that crudely concatenate or add the multi-modality information. The proposed modulation design enables effective integration of multi-modality information through feature transformation, distinctly models the inseparable cross-modality relations, and reduces the interference caused by the inherent inconsistency of multi-modality data.

Second, we devise an adaptive feature selection (AFS) module that highlights the importance of different channel features in self- and cross-modalities, while fusing multi-modality spatial features in a gated manner. This is different from previous RGB-D SOD algorithms [3–6,22,46] that treat channel features from different modalities equally and independently. Relaxing such assumptions allows

our method to adaptively select more saliency-related features and suppress the inferior ones from both spatial features and channel features. It also mitigates the negative influence of poorly captured depth maps. Hence, our network equips additional flexibility in dealing with different information. We also emphasize the saliency-related positions and edges by introducing a saliency-guided position-edge attention (sg-PEA) module, which collects its attention weights from the predicted saliency maps and saliency edge maps.

Our method is unique in that the feature modulation and attention mechanism are closely coupled in a coarse-to-fine manner. Specifically, fusion is first performed by the cmFM module to provide rich features representations. Coordinated with our AFS module, saliency-related features are emphasized while redundant features are suppressed. The saliency-related features are further refined by the sg-PEA module. A careful design to place the cmFM, AFS, and sg-PEA modules allows the cross-modality complementarities to go through modulation, selection, and refinement in a coarse-to-fine fashion, providing our network with precise saliency features. Coupled with a bottom-up inference, the precise saliency features enable us to perform more accurate and robust SOD.

Contributions. We present an effective approach for RGB-D SOD. Cross-modality complementarities are effectively integrated and saliency-related features are adaptively selected. This is made possible by designing a coarse-to-fine fusion that consists of 1) a cross-modality feature modulation module that enhances RGB feature representations by taking the corresponding depth features as prior, and 2) an adaptive feature selection module that progressively emphasizes the importance of channel features in self- and cross-modalities while fusing the significant multi-modality spatial features. Our method consistently outperforms state-of-the-art SOD methods on six popular RGB-D SOD benchmarks.

2 Related Work

Salient Object Detection. SOD methods range from bottom-up [25,29,42] to top-down models [14,17,19,26,34,47]. In addition to the color appearance, depth maps can provide useful cues such as depth of field, shape, and boundary. The depth map is implicitly used in the unsupervised methods [9,10,21,27,30,37,48]. Whereas for the supervised methods, the discriminative and complementary features are learned from RGB-D images [3–6,12,16,18,22,32,35,43,44,46]. Our work differs from recent works [12,16,32,33,43,44,46], mainly in two aspects: 1) we use depth features as prior to learn optimal affine transformation parameters, which can flexibly modulate multi-level RGB features, and 2) we consider both self-modality and cross-modality channel features as well as multi-modality spatial features, thus effectively capturing relations among different modalities.

Feature Modulation. Inspired by FiLM [31] that first applies linear feature modulation for visual reasoning, feature modulation has been used in few-shot learning [28] and image super-resolution [40]. In our studies, we modulate the

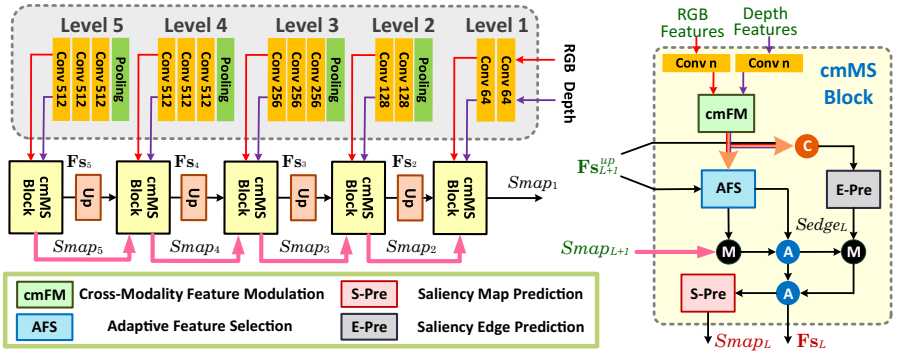

Fig. 2. Overview of our network architecture. The inputs are the RGB image and its depth map. The cmMS block consists of a cmFM module, an AFS module, and an sg-PEA module. Here, the sg-PEA module further contains an S-Pre unit and an E-Pre unit. 'Conv n' represents the convolutional layer that outputs n feature maps, where n is the half number of input feature maps. 'A', 'M', and 'C' represent element-wise addition, element-wise multiplication, and concatenation along with the channel dimension, respectively. 'Up' represents the up-sampling block. Pink line indicates $2\times$ linear interpolation. **Fs** represent the refined features after the cmMS block while $\mathbf{Fs}^{up}$ are the up-sampled **Fs** by the 'Up' block. In this figure, each convolutional layer is followed by the ReLU activation. Our network finally produces five saliency maps ($Smap_L$) and five saliency edge maps ($Sedge_L$) with the resolutions, ranging from 14×14 to 224×224 by a scale of 2. L indicates the level. We treat $Smap_1$ as the final result.

multi-level feature representations conditioned on the corresponding depth features. Besides, we design the cross-modality feature modulation in a pixel-wise manner, which provides elaborate and fine-grained control to the features.

Attention Mechanism. Attention mechanism is increasingly applied in diverse forms such as spatial attention [7], dual-attention [15], self-attention [38], multi-level attention [41], and channel attention [45]. In contrast, we employ the attention mechanism in our adaptive feature selection module, which explores the interdependencies of channel features in the self- and cross-modalities while fusing the significant multi-modality spatial features in a gated manner.

3 Our Method

We first present an overview of our network architecture. Then, we describe the key components including the cross-modality feature modulation module, adaptive feature selection module, and saliency-guided position-edge attention module. At last, we introduce the loss functions.

3.1 Overview of Network Architecture

The overview of our network architecture is illustrated in Fig. 2. After the top-down features extraction from VGG-16 backbone [36], the multi-level RGB features and depth features are fed to a convolutional layer for halving the number

of feature maps, respectively. Then, the dimension reduced RGB-D features are forwarded to the corresponding cmMS block. In each cmMS block, the RGB-D features go through cmFM module, AFS module, and sg-PEA module for feature modulation, selection, and refinement, respectively. Specifically, we introduce modulated features by using our proposed cross-modality feature modulation (cmFM) module. The purpose of cmFM module is to effectively integrate the cross-modality complementarities in a flexible and trainable fashion. After that, RGB features, depth features, modulated features, and up-sampled features from the higher level (if any) are independently forwarded to our proposed adaptive feature selection (AFS) module for selectively emphasizing the informative channel features and fusing the significant spatial features. The AFS module models the relations between different levels and accelerates task-oriented feature integration. Meanwhile, the concatenation of RGB features, depth features, modulated features, and up-sampled features (if any) is applied to predict the saliency edge map via a saliency edge prediction (E-Pre) unit. Then, with the saliency map up-sampled from the higher level (if any) and saliency edge map, we highlight the saliency position and edge regions of the features after the AFS module. After that, we predict the saliency map in the current level via a saliency map prediction (S-Pre) unit by using the refined features. At last, in the bottom-up inference, we progressively integrate and highlight multi-level features to predict the fine-scaled saliency map (*i.e.*, the $Smap_1$ in Fig. 2). We adopt 3×3 kernels for all convolutional layers in our network, except the cmFM module that employs the multi-scale convolutions to enlarge receptive field.

3.2 Cross-Modality Feature Modulation (cmFM)

Inspired by the unsupervised RGB-D SOD algorithms [10,13] which take the depth map as prior information to enrich the saliency cues, we propose a cmFM module conditioned on the depth features. The cmFM module learns pixel-wise affine transformation parameters from the conditioning depth features then modulates the corresponding RGB feature representations in each level of our network. The detailed cmFM module is illustrated in Fig. 3.

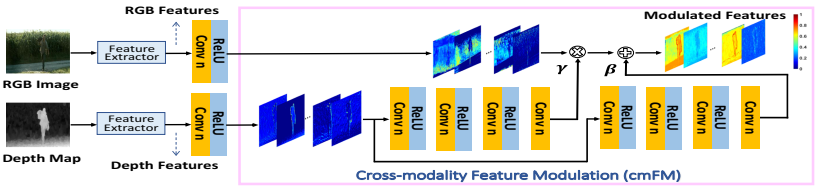

Fig. 3. The proposed cmFM module. For the estimation of both $\boldsymbol{\gamma}$ and $\boldsymbol{\beta}$, the kernels of convolutional layers are 7×7, 5×5, 3×3, and 3×3. The feature extractor represents VGG-16 backbone. The feature maps are illustrated as heatmaps.

Given the dimension halved RGB features $\mathbf{F}_L^{rgb} \in \mathbb{R}^{N\times H\times W}$ and depth features $\mathbf{F}_L^{depth} \in \mathbb{R}^{N\times H\times W}$, the cmFM module learns a mapping function $\mathcal{M}$ conditioned on the depth features to yield a set of affine transformation parameters $(\boldsymbol{\gamma}_L,\boldsymbol{\beta}_L) \in \mathbb{R}^{N\times H\times W}$. Here, N is the number of feature maps; H and W are the height and width of the feature maps, respectively. It can be expressed as:

$$(\boldsymbol{\gamma}_L, \boldsymbol{\beta}_L) = \mathcal{M}(\mathbf{F}_L^{depth}), \tag{1}$$

where the superscript indicates the modality while the subscript represents the level. The mapping function $\mathcal{M}$ is built on two parallel stacked convolutional layers as shown in Fig. 3. With the estimated affine transformation parameters $(\boldsymbol{\gamma}_L, \boldsymbol{\beta}_L)$, we conduct pixel-wise scaling and shifting on the RGB feature representations, which can be expressed as:

$$\mathbf{F}_L^{mod} = \mathbf{F}_L^{rgb} \otimes \boldsymbol{\gamma}_L \oplus \boldsymbol{\beta}_L, \tag{2}$$

where $\mathbf{F}_L^{mod}$ represent the modulated features; $\otimes$ and $\oplus$ indicate the element-wise multiplication and element-wise addition, respectively. As shown in Fig. 3, the cluttered backgrounds of RGB features become clear and the salient object is highlighted with the modulation of depth features.

3.3 Adaptive Feature Selection (AFS)

To make our network focus more on informative features, we propose an AFS module to progressively re-scale channel-wise features. Simultaneously, the AFS module fuses significant spatial features of multi-modalities. To be specific, we first explore the interdependencies of channel features in the self-modality, then further determine the relevance in the cross-modality. After squeezing by a convolutional layer that reduces the redundant features, we achieve the channel attention-on-channel attention features. Such a self-modality and cross-modality channel attention mechanism can model relations of the channel features among different modalities well and adaptively select the informative channel features. The advantages of our channel attention-on-channel attention than the conventional channel attention are verified in the ablation studies.

We simultaneously fuse the multi-modality features to achieve the enhanced feature representations based on a gated spatial fusion mechanism, where the pixel-wise confidence map for each input feature is calculated. In this way, the significant multi-modality spatial features are preserved. As a result, we achieve saliency-related features and filter out irrelevant or misleading features from both spatial and channel aspects. The detail of AFS module is shown in Fig. 4.

Given the features ($\mathbf{F}_L^{rgb}$, $\mathbf{F}_L^{depth}$, $\mathbf{F}_L^{mod}$, $\mathbf{Fs}_{L+1}^{up}$), we first perform global average pooling on each set of features separately, leading to a channel descriptor $\mathbf{z} \in \mathbb{R}^{N\times 1}$ for each one, which is an embedded global distribution of channel-wise feature responses. $\mathbf{Fs}_{L+1}^{up}$ indicate 2× up-sampled features from the $L+1$ level by using the 'Up' block that consists of one 2× linear interpolation followed by

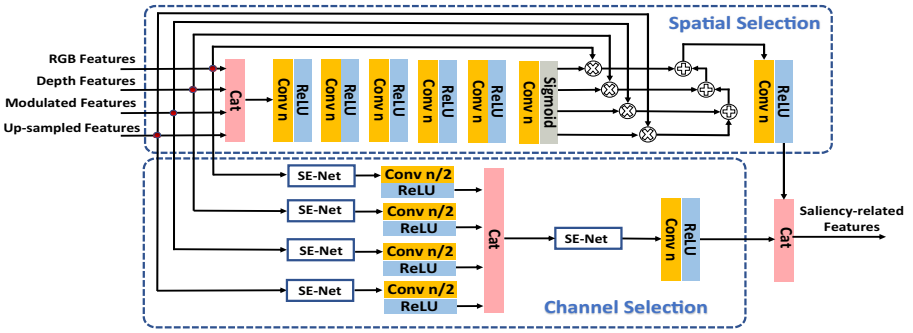

Fig. 4. The detail of AFS module. 'Cat' represents the concatenation operation. 'SE-Net' is the squeeze-and-excitation network.

two convolutional layers, where each convolutional layer is followed by the ReLU activation and outputs n feature maps. The k-th entry of $\mathbf{z}$ is expressed as:

$$z_k = \frac{1}{H \times W} \sum_{i}^{H} \sum_{j}^{W} \mathbf{F}_k(i, j), \tag{3}$$

where $k \in [1, N]$. Then, a self-gating mechanism is used to fully capture channel-wise dependencies $\mathbf{s} \in \mathbb{R}^{N \times 1}$:

$$\mathbf{s} = \sigma(\mathbf{W}_2 * (\delta(\mathbf{W}_1 * \mathbf{z}))), \tag{4}$$

where $\sigma(\cdot)$ represents the Sigmoid activation, $\delta(\cdot)$ represents the ReLU activation, $*$ denotes the convolution operation, and $\mathbf{W}_1$ and $\mathbf{W}_2$ are the weights of two fully-connected layers with their numbers of output channels being $\frac{N}{16}$ and N, respectively. At last, these weights are applied to each set of input features $\mathbf{F}$ to generate re-scaled features $\mathbf{U} \in \mathbb{R}^{N \times H \times W}$: $\mathbf{U} = \mathbf{F} \otimes \mathbf{s}$. This processing is mathematically expressed as an SE mapping function in this paper and can also be implemented by the squeeze-and-excitation network [20]. However, the highlighted channel features may become relatively useless among all channel attention results from multi-modalities.

To emphasize the informative channel features, we first halve the number of feature maps in each channel attention result by a convolutional layer, then concatenate them: $\mathbf{V}_L = Cat\{\mathbf{U}_L^{rgb}, \mathbf{U}_L^{depth}, \mathbf{U}_L^{mod}, \mathbf{U}\mathbf{s}_{L+1}^{up}\}$. After that, we further explore the interdependencies of channel features by $\mathbf{Y}_L = SE(\mathbf{V}_L)$. We finally squeeze the number of channel features by a convolutional layer and achieve the results of channel attention-on-channel attention $\mathbf{Y}_L^{caca}$.

Meanwhile, we fuse the multi-modality input features to achieve enhanced spatial feature representations. First, the input features are concatenated $\mathbf{F}_L^{cat} = Cat\{\mathbf{F}_L^{rgb}, \mathbf{F}_L^{depth}, \mathbf{F}_L^{mod}, \mathbf{F}\mathbf{s}_{L+1}^{up}\}$, and fed to a plain CNN network (indicated as $\mathcal{G}$) to estimate their pixel-wise confidence maps:

$$(\mathbf{C}_L^{rgb}, \mathbf{C}_L^{depth}, \mathbf{C}_L^{mod}, \mathbf{C}_{L+1}^{up}) = \mathcal{G}(\mathbf{F}_L^{cat}), \tag{5}$$

where $\mathbf{C}_L^{rgb}$, $\mathbf{C}_L^{depth}$, $\mathbf{C}_L^{mod}$, and $\mathbf{C}_{L+1}^{up} \in \mathbb{R}^{N\times H\times W}$ represent the confidence maps. The $\mathcal{G}$ is built on six stacked convolutional layers as shown in Fig. 4. The achieved features in the level L can be expressed as:

$$\mathbf{F}_L^{gated} = \mathbf{F}_L^{rgb} \otimes \mathbf{C}_L^{rgb} \oplus \mathbf{F}_L^{depth} \otimes \mathbf{C}_L^{depth} \oplus \mathbf{F}_L^{mod} \otimes \mathbf{C}_L^{mod} \oplus \mathbf{Fs}_{L+1}^{up} \otimes \mathbf{C}_{L+1}^{up} \quad (6)$$

Then, we pass these features to a convolutional layer and achieve the gated fusion features $\mathbf{F}_L^{gated'}$. At last, we combine the enhanced spatial feature representations with the enhanced channel feature representations by:

$$\mathbf{F}_L^{AFS} = Cat\{\mathbf{F}_L^{gated'}, \mathbf{Y}_L^{caca}\}, \quad (7)$$

where the final results $\mathbf{F}_L^{AFS}$ enjoy the most informative features towards saliency detection, called saliency-related features in this paper. The visual examples are presented in Fig. 5. As shown, the saliency-related spatial features and channel features are preserved and highlighted.

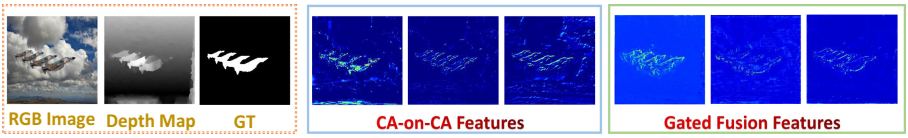

Fig. 5. Visual results of the intermediate features in our AFS module. 'CA-on-CA Features' indicates the features after our channel selection while 'Gated Fusion Features' represents the features after our spatial selection.

3.4 Saliency-Guided Position-Edge Attention (sg-PEA)

After selecting the saliency-related features, we also encourage the network to focus on those positions and edges most essential to the nature of salient objects. The benefits are illustrated as follows: 1) the saliency position attention can better locate the salient objects and accelerate the network convergence, and 2) the saliency edge attention can alleviate the problem of edge blur caused by the repeated pooling operations, which is vital for the pixel-wise saliency prediction.

To the end, we propose a saliency-guided position-edge attention (sg-PEA) module to locate and sharpen salient objects. The sg-PEA module further includes a saliency map prediction (S-Pre) unit and a saliency edge prediction (E-Pre) unit as shown in Fig. 2. The details are provided in Fig. 6, where S-Pre unit and E-Pre unit share the same structure, but different weights.

Position Attention. We employ the up-sampled saliency map from the higher level as the attention weights. Here, the up-sampling is implemented by the simple 2× linear interpolation. In our method, the saliency map is predicted by the S-Pre unit in each level in a supervised learning manner. The benefits of such a side supervision manner lie in four aspects: 1) the convolutional layers in each level have explicit objective towards saliency detection, 2) the side supervision

can accelerate gradient back-propagation, 3) the predicted saliency map works as a guidance and can steer the convolutional layers of lower level to focus more on saliency positions in a low-computational manner, and 4) the multiple side outputs can provide diverse choices based on accuracy and inference speed. We provide more analysis on the side outputs in the supplementary material.

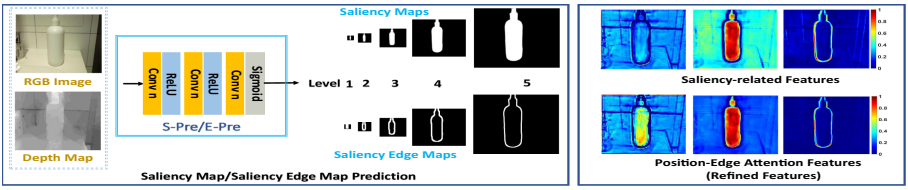

Fig. 6. Visual results of sg-PEA module. Left panel shows the structure of S-Pre/E-Pre unit, and the predicted saliency maps and saliency edge maps in different levels. Right panel shows the intermediate features before and after the sg-PEA module. After the sg-PEA module, the background of features are suppressed, and the edge and position details are assigned more focuses.

To be specific, with the saliency-related features $\mathbf{F}_L^{AFS}$ and the up-sampled saliency map $Smap_{L+1}^{up}$, the position attention results $\mathbf{F}_L^{poa}$ can be expressed as:

$$\mathbf{F}_L^{poa} = \mathbf{F}_L^{AFS} \oplus \mathbf{F}_L^{AFS} \otimes Smap_{L+1}^{up} \tag{8}$$

In contrast to treating all positions of saliency features equally, the position attention can quickly and efficiently employ the saliency property of higher level and enhance the saliency representations of the current level. To avoid gradient diffusion induced by successive attention (the values of feature maps are close to zero), we adopt an identical mapping manner as shown in Eq. (8).

Edge Attention. To obtain the edge attention weights, we first concatenate the RGB-D features, the modulated features, and up-sampled features, then forward them to the E-Pre unit to predict the saliency edge map in each level. The saliency edge maps, also estimated by supervised learning, can be used to emphasize the salient edges of the features by simple element-wise multiplication. For level L, the output features of edge attention can be expressed as:

$$\mathbf{Fs}_L = \mathbf{F}_L^{poa} \oplus \mathbf{F}_L^{poa} \otimes Sedge_L, \tag{9}$$

where $Sedge_L$ is the predicted saliency edge map in the level L. We call $\mathbf{Fs}_L$ as the refined features. At last, with the refined features, the final result (*i.e.*, $Smap_1$) with the same size as the input RGB image can be achieved in a bottom-up manner. In Fig. 6, we present the changes of features before and after sg-PEA module. As shown, the features increasingly focus on the saliency position and edge details, while the cluttered backgrounds are concurrently reduced.

3.5 Loss Function

We employ the standard cross-entropy (SCE) loss [1] to jointly optimize our network for the saliency prediction and saliency edge prediction:

$$Loss = \sum_{i=1}^{L} (\lambda_i SCE_i^{SPre} + \eta_i SCE_i^{EPre}), \tag{10}$$

where L indicates the level, SCE_i^{SPre} and SCE_i^{EPre} represent the losses for predicting the saliency map and saliency edge map in the level i, respectively. λ and η are the corresponding weights.

4 Experiments

4.1 Benchmark Datasets and Evaluation Metrics

We conduct experiments on six popular RGB-D SOD datasets, including **NJUD** [21] (1985 RGB-D images), **NLPR** [30] (1000 RGB-D images), **STEREO** [27] (797 RGB-D images), **LFSD** [24] (100 RGB-D images), **SSD** [23] (80 RGB-D images), and **DUT** [32] (1200 RGB-D images). For quantitative evaluations, Precision-Recall (P-R) curve, F-measure [2], MAE score [8], and S-measure [11] are employed. P-R curve depicts the different combinations of precision and recall scores; the closer the P-R curve is to (1, 1), the better the performance of the method. F-measure is the weighted harmonic mean of precision and recall; it is a comprehensive measurement, with a larger value indicating a better performance. MAE score measures the difference between the continuous saliency map and ground truth; a smaller value indicates a smaller gap hence better. S-measure calculates the structural similarity between the saliency map and ground truth; a larger value indicates a better performance. Additionally, we compare the model sizes of different methods in the supplementary material.

4.2 Implementation Details

We adopt the same training, validation, and testing sets as described in [32,33]. The ground truth of saliency edge map prediction is obtained by using the Canny edge detector on the saliency mask. We implement our network with TensorFlow on a PC with an Nvidia Tesla V100 GPU. During training, the batch size is set to 4, the filter weights of each layer are initialized by Gaussian distribution, and the bias is initialized as a constant. We use ADAM and fix the learning rate to $1e^{-4}$. The weight λ_1 for predicting the final saliency map is set to 1.2 while other weights are set to 1 in Eq. (10). For a pair of RGB-D images of size 224×224, the average runtime of our method is 0.037 s on the aforementioned PC.

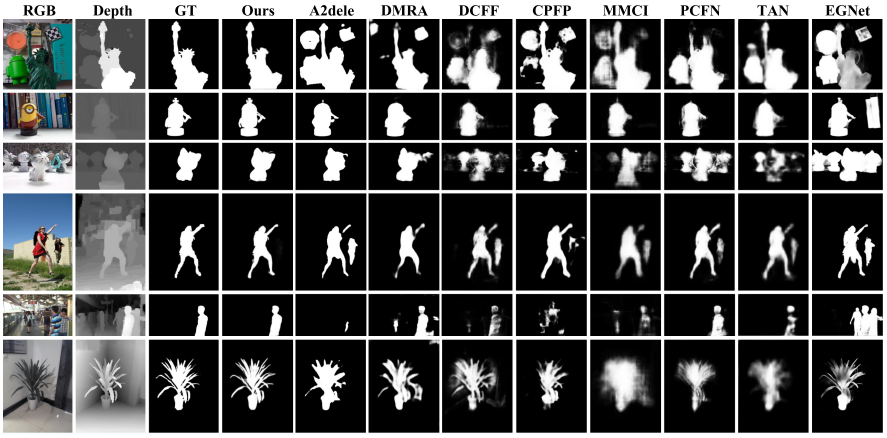

Fig. 7. Visual examples of different methods.

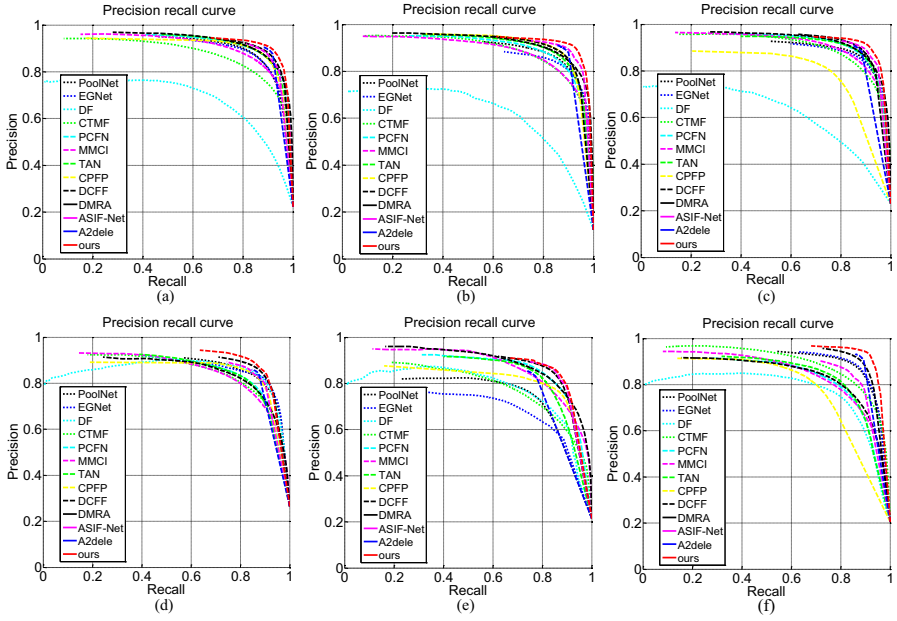

Fig. 8. P-R curves of different methods on the testing datasets. (a)-(f) correspond to STEREO, NLPR-Test, NJUD-Test, LFSD, SSD, and DUT-Test datasets.

4.3 Comparisons with State-of-the-art Methods

We compare our method with 12 state-of-the-art learning-based SOD methods, including two latest RGB-induced SOD methods (*i.e.*, PoolNet [26] and EGNet [47]), and ten RGB-D SOD methods (*i.e.*, DF [35], CTMF [18], MMCI [6], PCFN [3], TAN [4], CPFP [46], DCFF [5], DMRA [32], ASIF-Net [22], and A2dele [33]).

Table 1. Quantitative comparisons on six testing datasets. The bold numbers are performance of our method, also the best across all datasets

	STEREO dataset			NLPR-test dataset			NJUD-test dataset		
	$F_\beta \uparrow$	MAE $\downarrow$	$S_m \uparrow$	$F_\beta \uparrow$	MAE $\downarrow$	$S_m \uparrow$	$F_\beta \uparrow$	MAE $\downarrow$	$S_m \uparrow$
PoolNet [26]	0.8757	0.0655	0.8359	0.8627	0.0448	0.8573	0.8740	0.0676	0.8600
EGNet [47]	0.8717	0.0671	0.8363	0.8452	0.0504	0.8497	0.8667	0.0704	0.8562
DF [35]	0.6961	0.1738	0.6279	0.6480	0.1079	0.6710	0.6355	0.1987	0.5930
CTMF [18]	0.8265	0.1023	0.8230	0.8407	0.0561	0.8549	0.8572	0.0847	0.8493
PCFN [3]	0.8838	0.0606	0.8722	0.8635	0.0437	0.8592	0.8875	0.0592	0.8768
MMCI [6]	0.8610	0.0796	0.8504	0.8412	0.0591	0.8524	0.8684	0.0789	0.8588
TAN [4]	0.8865	0.0591	0.8701	0.8765	0.0410	0.8736	0.8882	0.0605	0.8785
CPFP [46]	0.8856	0.0537	0.8702	0.8878	0.0359	0.8760	0.7994	0.0794	0.7984
DCFF [5]	0.8867	0.0638	0.8706	0.8779	0.0439	0.8695	0.8910	0.0646	0.8774
DMRA [32]	0.8953	0.0474	0.8778	0.8870	0.0339	0.8646	0.9003	0.0529	0.8804
ASIF-Net [22]	0.8939	0.0493	0.8686	0.9002	0.0298	0.8844	0.9007	0.0471	0.8887
A2dele [33]	0.8997	0.0431	0.8713	0.8976	0.0285	0.8770	0.8939	0.0510	0.8704
Ours	**0.9084**	**0.0422**	**0.8895**	**0.9137**	**0.0273**	**0.8999**	**0.9149**	**0.0442**	**0.9040**
	LFSD dataset			SSD dataset			DUT-test dataset		
	$F_\beta \uparrow$	MAE $\downarrow$	$S_m \uparrow$	$F_\beta \uparrow$	MAE $\downarrow$	$S_m \uparrow$	$F_\beta \uparrow$	MAE $\downarrow$	$S_m \uparrow$
PoolNet [26]	0.8474	0.0945	0.8217	0.7644	0.1099	0.7491	0.8828	0.0669	0.8392
EGNet [47]	0.8445	0.0871	0.8300	0.7040	0.1351	0.7072	0.8876	0.0641	0.8439
DF [35]	0.8534	0.1424	0.7791	0.7631	0.1511	0.7422	0.7747	0.1455	0.7051
CTMF [18]	0.8147	0.1202	0.7883	0.7550	0.1003	0.7757	0.8417	0.0971	0.8226
PCFN [3]	0.8290	0.1118	0.7919	0.8447	0.0627	0.8427	0.8094	0.0999	0.7878
MMCI [6]	0.8128	0.1318	0.7793	0.8230	0.0820	0.8133	0.8044	0.1125	0.7818
TAN [4]	0.8275	0.1108	0.7935	0.8350	0.0629	0.8393	0.8236	0.0926	0.7948
CPFP [46]	0.8495	0.0881	0.8200	0.8014	0.0818	0.8067	0.7866	0.0995	0.7335
DCFF [5]	0.8220	0.1191	0.7917	0.8388	0.0769	0.8316	0.8141	0.1014	0.7835
DMRA [32]	0.8723	0.0754	0.8391	0.8579	0.0583	0.8569	0.9082	0.0477	0.8637
ASIF-Net [22]	0.8584	0.0896	0.8144	0.8633	0.0562	0.8566	0.8574	0.0725	0.8141
A2dele [33]	0.8577	0.0740	0.8306	0.8248	0.0691	0.8093	0.9145	0.0426	0.8611
Ours	**0.8882**	**0.0720**	**0.8465**	**0.8650**	**0.0524**	**0.8615**	**0.9328**	**0.0366**	**0.8853**

Visual comparisons are shown in Fig. 7. Our method achieves more competitive performance than the compared methods. **First**, the salient objects in our results are more complete and accurate, and the object boundaries are sharper. In the first image, only our method can accurately and completely detect the salient toy in front, while the competing methods incorrectly reserve the background regions (*e.g.*, Android doll and checkerboard). In the fourth image that comes with an unsatisfactory depth map, our method can still accurately locate salient target with a complete structure and clear boundaries. **Second**, our method preserves more details in the saliency result. In the sixth image, more details of plant leaves are better conserved. **Third**, our method can address some challenging cases, such as a complex background and small object. In the third image, the cat dolls in the back row are successfully suppressed by our method, the

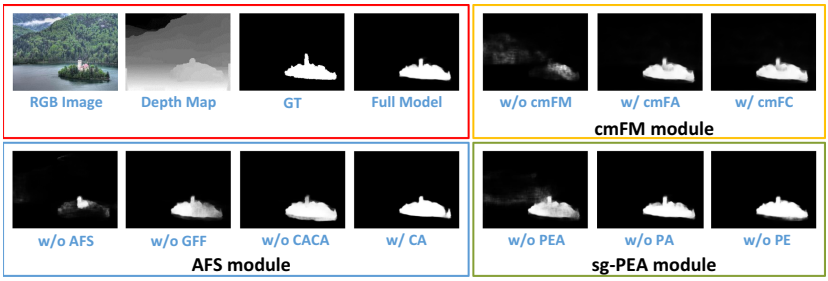

Fig. 9. Visual comparison with different baselines. **(1)** The baseline *w/o cmFM* represents our full model without the cmFM module (*i.e.*, no modulated features); and the baselines *w/cmFA* and *w/cmFC* refer to that the cmFM module is replaced by the cmFA or cmFC module (*i.e.*, the depth and RGB features are integrated by the element-wise addition or concatenation). **(2)** The baseline *w/o AFS* represents our full model without the AFS module (*i.e.*, the features after cmFM module are directly concatenated with the up-sampled saliency-related features); the baselines *w/o GFF* and *w/o CACA* correspond to removing the fused spatial features and the channel attention-on-channel attention features, respectively; and the baseline *w/CA* refers to that the AFS module is replaced by the conventional channel attention module [20]. **(3)** The baselines *w/o PEA*, *w/o PA*, and *w/o PE* correspond to our full model without the sg-PEA module, the position attention unit, and the edge attention unit, respectively.

detected salient boundaries are sharper, and the structure is more complete. In the fifth image illustrating a case of complex background, our method can still completely detect a small salient object (*i.e.*, the human).

The P-R curves of different methods are shown in Fig. 8. Our method (*i.e.*, the red solid line) achieves the highest precision compared to other methods on all datasets. The numerical results are reported in Table 1. Our method achieves the best quantitative results across all metrics, outperforming all competing methods. Compared with the **second best method** on the NJUD-Test dataset, the percentage gain reaches 1.6% for F-measure, 6.2% for MAE score, and 1.7% for S-measure. On the DUT-test dataset, the **minimum percentage gain** reaches 2.0% for F-measure, 14.1% for MAE score, and 2.5% for S-measure. All these measures demonstrate the superiority and effectiveness of our method.

4.4 Ablation Studies

To verify the impact of our key modules, we conduct experiments on the STEREO dataset and DUT-Test dataset. The quantitative results are shown in Table 2. An example of visual comparison is illustrated in Fig. 9.

Cross-Modality Feature Modulation (cmFM). We compare three variants: *w/o cmFM*, *w/cmFA*, and *w/cmFC*. In Fig. 9, the baseline *w/o cmFM* cannot effectively detect the salient object while the baselines *w/cmFA* and *w/cmFC* achieve the similar detection result. The same quantitative trend also reflects in Table 2. Compared with the full model, the results indicate that the proposed

Table 2. Quantitative comparisons of ablated models

Modules	Baselines	STEREO dataset			DUT-test dataset		
		$F_\beta \uparrow$	MAE $\downarrow$	$S_m \uparrow$	$F_\beta \uparrow$	MAE $\downarrow$	$S_m \uparrow$
	full model	**0.9084**	**0.0422**	**0.8895**	**0.9328**	**0.0366**	**0.8853**
cmFM	w/o cmFM	0.8727	0.0722	0.8573	0.8968	0.0616	0.8599
	w/cmFA	0.9020	0.0479	0.8820	0.9237	0.0429	0.8771
	w/cmFC	0.8995	0.0480	0.8825	0.9221	0.0617	0.8789
AFS	w/o AFS	0.8990	0.0546	0.8762	0.9165	0.0503	0.8666
	w/o GFF	0.9012	0.0690	0.8826	0.9212	0.0458	0.8777
	w/o CACA	0.9017	0.0517	0.8797	0.9276	0.0470	0.8742
	w/CA	0.9027	0.0503	0.8780	0.9216	0.0468	0.8747
sg-PEA	w/o PEA	0.9057	0.0450	0.8854	0.9205	0.0427	0.8796
	w/o PA	0.9064	0.0442	0.8857	0.9234	0.0409	0.8827
	w/o PE	0.9065	0.0481	0.8862	0.9296	0.0385	0.8806

cmFM module is important for improving the SOD performance. Besides, the simple addition and concatenation can only boost a little performance.

Adaptive Feature Selection (AFS). We compare with four baselines: *w/o AFS*, *w/o GFF*, *w/o CACA*, and *w/CA*. Observing Fig. 9 and Table 2, we found that the performance of the baseline *w/o AFS* is obviously worse than the baselines *w/o GFF*, *w/o CACA*, and *w/CA*. The visual results reflect that the baseline *w/o GFF* produces incomplete salient object while the baseline *w/o CACA* yields the result with an unclear boundary. Collectively, these results underscore the importance of progressive self-modality and cross-modality channel attention while fusing important spatial features of multi-modalities.

Saliency-guided Position-Edge Attention (sg-PEA). We compare with three baselines: *w/o PEA*, *w/o PA*, and *w/o EA*. In Fig. 9, the baseline *w/o PEA* fails to highlight the position and edge of salient object. The baseline *w/o PA* has a sharper boundary of partial complete object while the baseline *w/o PE* shows a more complete object but unclear boundary. In contrast, our full model achieves better performance than these three baselines as presented in Table 2.

In summary, the ablation studies demonstrate the effectiveness and advantages of the proposed three modules qualitatively and quantitatively. In addition, the ablation studies also demonstrate that careful feature modulation, selection, and refinement can effectively improve the performance of RGB-D SOD.

5 Conclusion

We propose an RGB-D SOD network equipped with cross-modality feature modulation and adaptive feature selection. The former effectively integrates the multi-modality complementarities while the latter adaptively highlights

saliency-related features. We demonstrate that both elaborate integration of cross-modality features and adaptive selection of multi-modality spatial and channel features can boost the performance of SOD. Experiment results also demonstrate that our method achieves new state-of-the-art performance on six benchmarks.

Acknowledgments. This research was supported by SenseTime-NTU Collaboration Project, Singapore MOE AcRF Tier 1 (2018-T1-002-056), NTU NAP, in part by the Fundamental Research Funds for the Central Universities under Grant 2019RC039, and in part by China Postdoctoral Science Foundation Grant 2019M660438.

References

1. Boer, P.T.D., Kroese, D.P., Mannor, S., Rubinstein, R.Y.: A tutorial on the cross-entropy method. Ann. Oper. Res. **134**(1), 19–67 (2005)
2. Borji, A., Cheng, M.M., Jiang, H., Li, J.: Salient object detection: a benchmark. IEEE Trans. Image Process. **24**(12), 5706–5722 (2015)
3. Chen, H., Li, Y.: Progressively complementarity-aware fusion network for RGB-D salient object detection. In: CVPR, pp. 3051–3060 (2018)
4. Chen, H., Li, Y.: Three-stream attention-aware network for RGB-D salient object detection. IEEE Trans. Image Process. **28**(6), 2825–2835 (2019)
5. Chen, H., Li, Y., Su, D.: Discriminative cross-modal transfer learning and densely cross-level feedback fusion for RGB-D salient object detection. IEEE Trans. Cybern., 1–13 (2019)
6. Chen, H., Li, Y., Su, D.: Multi-modal fusion network with multiscale multi-path and cross-modal interactions for RGB-D salient object detection. Pattern Recognit. **86**, 376–385 (2019)
7. Chen, L., et al.: SCA-CNN: spatial and channel-wise attention in convolutional networks for image captioning. In: CVPR, pp. 5659–5667 (2017)
8. Cong, R., Lei, J., Fu, H., Cheng, M.M., Lin, W., Huang, Q.: Review of visual saliency detection with comprehensive information. IEEE Trans. Circuits Syst. Video Technol. **29**(10), 2941–2959 (2019)
9. Cong, R., Lei, J., Fu, H., Hou, J., Huang, Q., Kwong, S.: Going from RGB to RGBD saliency: a depth-guided transformation model. IEEE Trans. Cybern. **50**(8), 3627–3639 (2020)
10. Cong, R., Lei, J., Zhang, C., Huang, Q., Cao, X., Hou, C.: Saliency detection for stereoscopic images based on depth confidence analysis and multiple cues fusion. IEEE Sig. Process. Lett. **23**(6), 819–823 (2016)
11. 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)
12. Fan, D.P., Zhai, Y., Borji, A., Yang, J., Shao, L.: BBS-Net: RGB-D salient object detection with a bifurcated backbone strategy network. In: Vedaldi, A., Bischof, H., Brox, T., Frahm, J.M. (eds.) ECCV 2020. LNCS, vol. 12357. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-58610-2_17
13. Feng, D., Barnes, N., You, S., McCarthy, C.: Local background enclosure for RGB-D salient object detection. In: CVPR, pp. 2343–2350 (2016)
14. Feng, M., Lu, H., Ding, E.: Attentive feedback network for boundary-aware salient object detection. In: CVPR, pp. 1623–1632 (2019)

15. Fu, J., Liu, J., Tian, H., Li, Y.: Dual attention network for scene segmentation. In: CVPR, pp. 3146–3154 (2019)
16. Fu, K.F., 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)
17. Guan, W., Wang, T., Qi, J., Zhang, L., Lu, H.: Edge-aware convolutional neural network based salient object detection. IEEE Sig. Process. Lett. **26**, 114–118 (2018)
18. Han, J., Chen, H., Liu, N., Yan, C., Li, X.: CNNs-based RGB-D saliency detection via cross-view transfer and multiview fusion. IEEE Trans. Cybern. **48**(11), 3171–3183 (2018)
19. Hou, Q., Cheng, M.M., Hu, X., Borji, A., Tu, Z., Torr, P.H.: Deeply supervised salient object detection with short connections. IEEE Trans. Pattern Anal. Mach. Intell. **41**(4), 815–828 (2019)
20. Hu, J., Shen, L., Sun, G.: Squeeze-and-excitation networks. In: CVPR, pp. 7132–7141 (2018)
21. Ju, R., Liu, Y., Ren, T., Ge, L., Wu, G.: Depth-aware salient object detection using anisotropic center-surround difference. Sig. Process. Image Commun. **38**, 115–126 (2015)
22. Li, C., et al.: ASIF-Net: attention steered interweave fusion network for RGBD salient object detection. IEEE Trans. Cybern., 1–13 (2020)
23. Li, G., Zhu, C.: A three-pathway psychobiological framework of salient object detection using stereoscopic technology. In: ICCVW, pp. 3008–3014 (2017)
24. Li, N., Ye, J., Ji, Y., Ling, H., Yu, J.: Saliency detection on light field. In: CVPR, pp. 2806–2813 (2014)
25. Li, X., Lu, H., Zhang, L., Ruan, X., Yang, M.H.: Saliency detection via dense and sparse reconstruction. In: ICCV, pp. 2976–2983 (2013)
26. Liu, J., 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)
27. Niu, Y., Geng, Y., Li, X., Liu, F.: Leveraging stereopsis for saliency analysis. In: CVPR, pp. 454–461 (2012)
28. Oreshkin, B.N., Rodriguez, P., Lacoste, A.: TADAM: task dependent adaptive metric for improved few-shot learning. In: NeurIPS, pp. 721–731 (2018)
29. Peng, H., Li, B., Ling, H., Hu, W., Xiong, W., Maybank, S.J.: Salient object detection via structured matrix decomposition. IEEE Trans. Pattern Anal. Mach. Intell. **39**(4), 818–832 (2017)
30. Peng, H., Li, B., Xiong, W., Hu, W., Ji, R.: RGBD salient object detection: a benchmark and algorithms. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) ECCV 2014. LNCS, vol. 8691. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-10578-9_7
31. Perez, E., Strub, F., de Vries, H., Dumoulin, V., Courville, A.: FiLM: Visual reasoning with a general conditioning layer. In: AAAI, pp. 3942–3951 (2018)
32. Piao, Y., Ji, W., Li, J., Zhang, M., Lu, H.: Depth-induced multi-scale recurrent attention network for saliency detection. In: ICCV, pp. 7254–7263 (2019)
33. 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)
34. Qin, X., Zhang, Z., Huang, C., Gao, C., Dehghan, M., Jagersand, M.: BASNet: boundary-aware salient object detection. In: CVPR, pp. 7479–7489 (2019)
35. Qu, L., He, S., Zhang, J., Tian, J., Tang, Y., Yang, Q.: RGBD salient object detection via deep fusion. IEEE Trans. Image Process. **26**(5), 2274–2285 (2017)

36. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014)
37. Song, H., Liu, Z., Du, H., Sun, G., Le Meur, O., Ren, T.: Depth-aware salient object detection and segmentation via multiscale discriminative saliency fusion and bootstrap learning. IEEE Trans. Image Process. **26**(9), 4204–4216 (2017)
38. Vaswani, A., et al.: Attention is all you need. In: NeurIPS, pp. 5998–6008 (2017)
39. Wang, W., Lai, Q., Fu, H., Shen, J., Ling, H.: Salient object detection in the deep learning era: An in-depth survey. arXiv preprint arXiv:1904.09146 (2019)
40. Wang, X., Yu, K., Dong, C., Loy, C.C.: Recovering realistic texture in image super-resolution by deep spatial feature transform. In: CVPR, pp. 606–615 (2018)
41. Yu, D., Fu, J., Mei, T., Rui, Y.: Multi-level attention networks for visual question answering. In: CVPR, pp. 4709–4717 (2017)
42. Yuan, Y., Li, C., Kim, J., Cai, W., Feng, D.D.: Reversion correction and regularized random walk ranking for saliency detection. IEEE Trans. Image Process. **27**(3), 1311–1322 (2018)
43. Zhang, J., et al.: UC-Net: uncertainty inspired RGB-D saliency detection via conditional variational autoencoders. In: CVPR, pp. 8582–8591 (2020)
44. 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)
45. Zhang, Y., Li, K., Li, K., Wang, L., Zhong, B., Fu, Y.: Image super-resolution using very deep residual channel attention networks. In: Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y. (eds.) ECCV 2018. LNCS, vol. 11211. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-01234-2_18
46. Zhao, J., 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: CVPR, pp. 3927–3936 (2019)
47. Zhao, J., Liu, J., Fan, D.P., Cao, Y., Yang, J., Cheng, M.M.: EGNet: edge guidance network for salient object detection. In: ICCV, pp. 8779–8788 (2019)
48. Zhu, C., Li, G.: A multilayer backpropagation saliency detection algorithm and its applications. Multimed. Tools Appl. **77**(19), 25181–25197 (2018). https://doi.org/10.1007/s11042-018-5780-4