Feature Reintegration over Differential Treatment: A Top-down and Adaptive Fusion Network for RGB-D Salient Object Detection

Miao Zhang[1,2], Yu Zhang[1], Yongri Piao[1,*], Beiqi Hu[1], Huchuan Lu[1,3]
[1]Dalian University of Technology, Dalian, China
[2]Key Lab for Ubiquitous Network and Service Software of Liaoning Province, Dalian University of Technology, China
[3] Pengcheng Lab, Shenzhen, China
{miaozhang,yrpiao,lhchuan}@dlut.edu.cn,{zhangyu4195,hbq1211}@mail.dlut.edu.cn

ABSTRACT

Most methods for RGB-D salient object detection (SOD) utilize the same fusion strategy to explore the cross-modal complementary information at each level. However, this may ignore different feature contributions from two modalities on different levels towards prediction. In this paper, we propose a novel top-down multi-level fusion structure where different fusion strategies are utilized to effectively explore the low-level and high-level features. This is achieved by designing the interweave fusion module (IFM) to effectively integrate the global information and designing the gated select fusion module (GSFM) to discriminatively select useful local information by filtering out the unnecessary one from RGB and depth data. Moreover, we propose an adaptive fusion module (AFM) to reintegrate the fused cross-modal features of each level to predict a more accurate result. Comprehensive experiments on 7 challenging benchmark datasets demonstrate that our method achieves the competitive performance over 14 state-of-the-art RGB-D alternative methods.

CCS CONCEPTS

• **Computing methodologies** → *Interest point and salient region detections.*

KEYWORDS

Salient object detection; Interweave fusion; Gated select fusion; Adaptive fusion

ACM Reference Format:
Miao Zhang[1,2], Yu Zhang[1], Yongri Piao[1,*], Beiqi Hu[1], Huchuan Lu[1,3] . 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 (MM '20), October 12–16, 2020, Seattle, WA, USA.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3394171.3413969

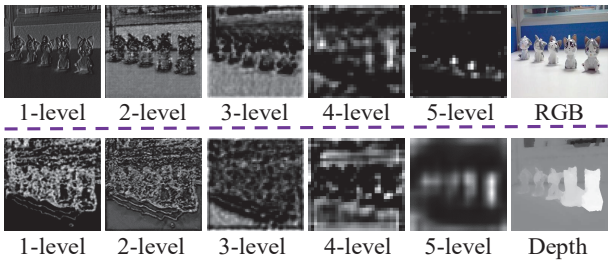

Figure 1: The feature maps visualized from the first level to the last level of the RGB and depth streams, respectively.

1 INTRODUCTION

Salient object detection (SOD) aims to identify the most distinctive objects or regions in a scene [17]. This fundamental task plays an important role in various computer vision applications, including image segmentation [24], object tracking [2, 16] and pose estimation [9].

Earlier saliency detection methods mainly focus on extracting hand-crafted features [26, 30, 36, 45]. Based on limited knowledge which lacks high-level contexts representation, these methods may have less robustness in different scenes. Recently, benefiting from the powerful ability of CNNs [20] in feature extraction, CNNs-based methods have been designed and shown outstanding performance in salient object detection. Many works [14, 23, 37, 39, 40, 42] focus on distinguishing the saliency region based on RGB images and have achieved spectacular performance. But these methods might be sensitive to some complex scenes, e.g., similar foreground and background, multiple objects or complex background, due to lack of accurate spatial constraints.

Depth data containing 3D layout information and spatial structure have been introduced to overcome the above issues in SOD. Many RGB-D methods have been explored and have achieved significant performance. Yet there still is large room for further improvement in two aspects: (1) In RGB-D saliency detection tasks, the contribution of two modalities is different at each level of the network, as exemplified in Figure 1. In high levels, the depth features typically carry more global contextual information than RGB features. On the other hand, in low levels, the RGB features contain more local information than depth features. Moreover, some cluttered and distractive information is inevitably blended in shallow features of two modalities, which may negatively influence

* Corresponding author
This work was supported by the Science and Technology Innovation Foundation of Dalian (2019J12GX034), the National Natural Science Foundation of China (61976035), and the Fundamental Research Funds for the Central Universities (DUT19JC58, DUT20JC42).

MM '20, October 12–16, 2020, Seattle, WA, USA

ACM ISBN 978-1-4503-7988-5/20/10…$15.00
https://doi.org/10.1145/3394171.3413969

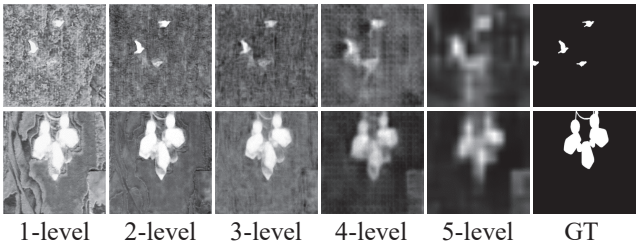

Figure 2: The fused cross-modal feature maps visualized from all levels of the top-down fusion model.

the discrimination ability of networks by indiscriminately fusing two modal features. Many RGB-D methods [4–6, 27] adopt the same fusion manner in all levels. This ignores the contribution of the local and global feature variation in different levels. How to differentially fuse the high-level and low-level features from two modalities should be explored. (2) The top-down inference is widely adopted in SOD [4, 5], where high-level features are gradually integrated with low-level features to obtain a fine-grained result. In this process, the network may achieve suboptimal performance as high-level features including position information are diluted in the flow, albeit simple in structure, as shown in Figure 2. Therefore, how to reintegrate features in different levels to obtain a superior result should be considered.

Our core insight is that the different fusion strategies can be leveraged, targeted at low-level and high-level features. In high levels, our model focuses on effective cross-modal fusion of global and contextual information for locating salient objects correctly. In low levels, our model focuses on fusing useful information by filtering out the distractive one. In addition, we take a further step toward the top-down inference while being free of global information degradation. Our main contributions are as follows:

- We propose a novel top-down multi-level fusion structure which adopts diffrernent fusion strategies in high and low levels, considering the distinction of RGB and depth features in different levels. A simple but effective interweave fusion module (IFM) is designed to fully extract and fuse global information in high levels, while the gated select fusion module (GSFM) is utilized to selectively process the useful information from two modal features in low levels.
- We design an adaptive fusion module (AFM) to effectively reintegrate the fused cross-modal features based on the top-down fusion structure. The module fully explores the contributions of the fused features in different levels and learns the exclusive weights to predict a more accurate result.
- We demonstrate that the proposed model can accurately reliably locate the salient objects, outperforming 14 state-of-the-art RGB-D methods on seven widely used benchmark datasets.

2 RELATED WORK

The primary challenge associated with RGB saliency detection is that they are sensitive when it comes to complex scenes. These include complex background, similar foreground and background, low-contrast environment and multiple objects. There have been many attempts to boost the performance of RGB methods [23, 29, 37–39].

Depth contains structural information and 3D layout information, which have been introduced to SOD [8, 10, 26, 32, 34, 44, 45]. Decent progress has been made by RGB-D saliency detection methods, especially in complex scenes. Ren et al. [33] explore the validity of global priors for SOD. Feng et al. [13] propose a SOD method based on local background enclosure. These methods mainly relied on the hand-crafted features are difficult to understand the global context, for lacking high-level semantic information.

Recently, CNNs have been adopted in RGB-D SOD to learn high-level representations and more discriminative features, having achieved significant performance. Qu et al. [31] fuse hand-crafted features from RGB and depth images before feeding these features to a CNN to learn deep representations and make inference. This method achieves great improvement comparing to some methods based on hand-crafted RGB-D features. However, in this method, designd low-level features are fused via a simple network, and the high-level features are not well integrated. Besides, this network can not be trained in an end-to-end manner. Han et al. [15] extract the two modal features by a two-stream network, then fuse RGB-D deep features to obtain final saliency maps. However, this fusion manner only focuses on fusing the high-level features, while the complementary information in low levels is ignored. Different from these methods which combine RGB-D features in a certain point (i.e, early or late), some methods [4–6, 27] try to explore a new fusion manner by which cross-modal features at each level are combined and cross-level features are fused progressively to make joint decisions. Chen et al. [4] exploit the level-wise cross-modal complementarity and propose a top-down progressive fusion network to fuse the two modal features. His another work [5] designs a three-stream network which combines the cross-modal information of each level in a cooperative top-down and bottom-up inference way. Piao et al. [27] fuse the cross-modal and cross-level features in a bottom-up way, then further extract and refine the information to predict a more accurate result. Chen et al. [6] fuse the deep and shallow cross-modal complements by cross-modal cross-level fusion strategies. These methods mainly utilize the same fusion operation in all levels, few consider the different contributions of RGB and depth in different levels. In addition, some works [18, 28, 41, 43] explore the asymmetric architecture for processing different data types. Zhao et al. [41] enhance the depth map as an attention map and design a fluid pyramid integration method to obtain a more accureate saliency map. Piao et al. [28] propose an adaptive and attentive depth distiller to transfer the depth knowledge from the depth stream to the RGB stream, and achieve a lightweight architecture without the use of depth data at test time.

Our work has several key differences with the aforementioned works: Firstly, we design two different fusion modules for effectively fusing low-level and high-level features from two modalities, respectively. Secondly, we reintegrate the fused cross-modal features of each level based on the level-specific contributions to make effective combinations. Our model brings a fresh perspective for RGB-D saliency detection and achieves better results. The source code is released [1].

[1] https://github.com/OIPLab-DUT/ACM-MM-FRDT

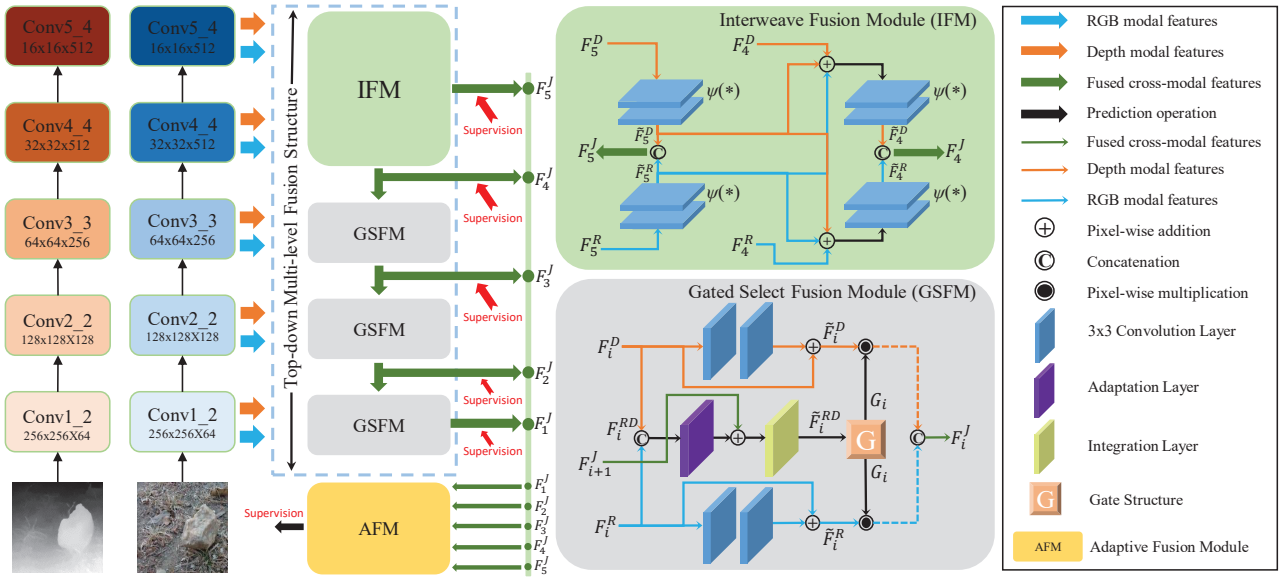

Figure 3: The overall architecture of our proposed network.

3 THE PROPOSED METHOD

3.1 The overall architecture

We adopt VGG-19 [35] as our basic architecture for both RGB and depth streams, discarding the last pooling and fully-connected layers, as shown in Figure 3. We collect all side-out features in all levels extracted from RGB and depth streams, which are denoted as $\{F_1^R, F_2^R, F_3^R, F_4^R, F_5^R\}$ and $\{F_1^D, F_2^D, F_3^D, F_4^D, F_5^D\}$, respectively. We refine and fuse paired side-out features at each level by employing the proposed top-down multi-level fusion structure which contains the interweave fusion module (IFM) and gated select fusion module (GSFM). The five fused cross-modal features are generated, denoted as $\left\{F_1^J, F_2^J, F_3^J, F_4^J, F_5^J\right\}$. Then the fused features are reintegrated by the proposed adaptive fusion module (AFM) to predict the final saliency map.

3.2 Top-down Multi-level Fusion Structure

In RGB-D salient object detection tasks, the top-down fusion strategy is widely used in fusing cross-modal and cross-level features. Many previous works utilize the same fusion manner in all levels. However, this manner ignores different contributions of two modalities in different levels. To address this problem, we propose a novel top-down multi-level fusion structure where different fusion strategies are utilized to effectively combine the low-level and high-level features. This structure contains two tailored fusion modules: the interweave fusion module (IFM) and gated select fusion module (GSFM). Specifically, as illustrated in Figure 3, the IFM focuses on combining global information from two modalities in high levels (the 4^{th} and 5^{th} levels), and the GSFM aims to fuse useful local information by filtering out the unnecessary one in low levels (the 1^{st} - 3^{rd} levels). Next, we will introduce the two modules in detail.

3.2.1 Interweave Fusion Module. The proposed IFM aims at fully fusing the available information from high levels to locate the salient objects more accurately. Since the high-level feature maps have low resolution and contain more location information, as shown in Figure 1, it is not necessary to utilize complex operations to extract features. To this end, we design a simple but effective module to fuse the high-level features from two modalities. Specifically, as shown in Figure 3, we design a light component $\psi(*)$, which contains two convolution layers and two ReLU activation functions. In the 5^{th} level, we fuse the F_5^R and the F_5^D to generate a fused cross-modal feature F_5^J:

$$\widetilde{F}_5^R = \psi(F_5^R), \quad \widetilde{F}_5^D = \psi(F_5^D), \quad F_5^J = \widetilde{F}_5^R \copyright \widetilde{F}_5^D \tag{1}$$

where $\copyright$ denotes the concatenation operation. In the 4^{th} level, the cross-modal cross-level global features are further combined by:

$$\begin{aligned} \widetilde{F}_4^R &= \psi(Up(\widetilde{F}_5^R) + Up(\widetilde{F}_5^D) + F_4^R) \\ \widetilde{F}_4^D &= \psi(Up(\widetilde{F}_5^R) + Up(\widetilde{F}_5^D) + F_4^D) \end{aligned} \tag{2}$$

$$F_4^J = \widetilde{F}_4^R \copyright \widetilde{F}_4^D \tag{3}$$

where $Up(*)$ is the 2× upsample operation with bilinear interpolation. In this way, our IFM can effectively fuse cross-modal features on high levels and help the network to locate the salient objects more accurately.

3.2.2 Gated Select Fusion Module. The proposed GSFM aims to fully extract and fuse the useful local information from two modalities in low levels. A straightforward solution is to simply concatenate or summate the two modal features. However, this direct fusion manner introduces redundant information which exist in low-level features, as shown in Figure 1. These redundant information may negatively influence the discrimination ability of the network. An effective fusion manner which can filter out these redundant information should be considered. Inspired by the gate mechanism [7] which aims to selectively control the data flow, we propose an efficient gated select fusion module (GSFM) to overcome this issue.

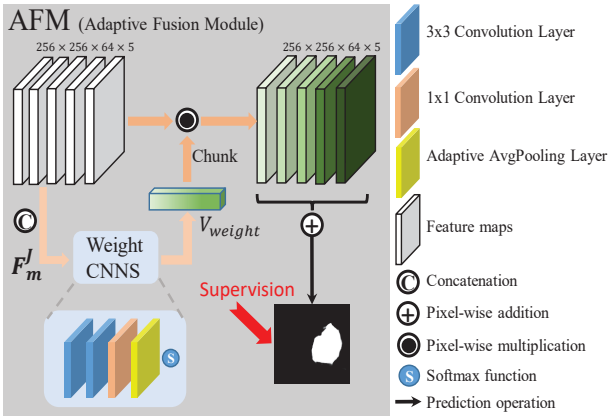

Figure 4: Details of adaptive fusion module.

Specifically, for the $i^{th}(i = 1, 2, 3)$ level, the inputs of the GSFM are the RGB feature F_i^R, the depth feature F_i^D and the combined RGB-D feature F_{i+1}^J from high levels. Our GSFM consists of three units, one gated select unit (GSU) and two residual units (RU), as shown in Figure 3. Detailed description for each unit is given below.

In the GSU, we aim to compute a gate map G_i based on the information of two modalities. We compute the joint representation F_i^{RD} by concatenating the F_i^R and F_i^D, then feed it into an adaptation layer. The fused RGB-D feature F_{i+1}^J from the $(i+1)^{th}$ level is added to the processed feature with a residual connection, followed by an integration layer to further combine the cross-level complementary information. In this way, the cross-modal information from the i^{th} level and the $(i+1)^{th}$ level can be fully fused as:

$$\widetilde{F}_i^{RD} = W^I * (W^A * F_i^{RD} + F_{i+1}^j) \tag{4}$$

where W^I and W^A represent the corresponding parameters of integration and adaptation layers, respectively (Details of these two layers of each level are shown in Table 1). Then the $\widetilde{F}_i^{RD}$ is fed into a core component gate structure containing two 3x3 convolution layers and a sigmod function which squashes values to [0,1] range. The gate map G_i is generated by:

$$G_i = \theta(W_1^G * (W_0^G * \widetilde{F}_i^{RD})) \tag{5}$$

where W_1^G and W_0^G are the corresponding parameters of two convolution operations, and θ represents the sigmod function. The gate map G_i will help the GSFM selectively combine the useful features from two modalities in the i^{th} level.

In the RU, we aim to further extract more useful infromation from the unimodal features. For the RGB feature F_i^R, we feed it into two successive 3x3 convolution layers to enlarge the receptive field and extract more useful unimodal features. Then, the original feature F_i^R is added by a residual connection to learn a more refined feature $\widetilde{F}_i^R$. Similarly, the same operation are also introduced for F_i^D to generate an enchanced depth feature $\widetilde{F}_i^D$. In the end, the final feature of the i^{th} level is generated by:

$$F_i^J = cat(\widetilde{F}_i^R \times G_i, \widetilde{F}_i^D \times G_i) \tag{6}$$

where the $cat(*)$ and $\times$ denote the concatenation operation and pixel-wise multiplication, respectively.

Table 1: Illustration of the parameters of the intra-level adaptation layer and intergration layer inside the gated select fusion module (GSFM), the transition layer between two neighboring levels

Level	Adaptation Layer kernel in/out	Integration Layer kernel in/out	Transition Layer kernel in/out
IFM	-	-	1×1, 1024/256
GSFM-3	1×1, 512/256	1×1, 256/256	1×1, 512/128
GSFM-2	1×1, 256/128	1×1, 128/128	1×1, 256/64
GSFM-1	1×1, 128/64	1×1, 64/64	-

In addition, we add a transition layer and an upsample operation to adapt the transferance between two adjacent levels. The detailed parameters of transition layers are shown in Table 1. Moreover, we add intermediate supervisions on all outputs of the IFM and GSFMs to guarantee that the most useful information can be fused explicitly for accurately identifying salient objects.

3.3 Adaptive Fusion Module

Though we adopt different strategies to deal with low-level and high-level features from two modalities, the cross-level features are still combined in a top-down manner. This may cause high-level features to be diluted as they are transmitted to the lower levels. To better reintegrate the fused cross-modal features of each level, we propose an adaptive fusion module (AFM) to emphasize the useful features and suppress unnecessary ones by learning the exclusive weights of fused features at each level, as shown in Figure 4.

We first reshape $\left\{F_i^J\right\}_{i=1}^5$ to the same resolution utilizing an upsample operation with bilinear interpolation and a 1x1 convolution operation. These reshaped features are denoted as $\left\{\widetilde{F}_i^J\right\}_{i=1}^5$. Then $\left\{\widetilde{F}_i^J\right\}_{i=1}^5$ are concatenated as F_m^J. We adopt several operations (two 3x3 convolution layers, a 1x1 convolution layer, a global average pooling layer, and a softmax function) to learn a feature-wise attention vector $V_{weight} \in R^{1\times1\times5}$. This procedure can be defined as:

$$V_{weight} = \delta(Avgpooling(conv(F_m^J))) \tag{7}$$

where the $conv$ represents the successive convolution operations in which parameters can be learned and δ denotes the softmax function. The concatenated $\left\{\widetilde{F}_i^J\right\}_{i=1}^5$ is weighted according to the V_{weight}, and then the weighted features are added together in a feature-wise manner. A saliency map is predicted by:

$$\begin{aligned} w_1, w_2, w_3, w_4, w_5 &= chunck(V_{weight}) \\ sal_f &= PRE(w_1 * F_1^J + w_2 * F_2^J \\ &+ w_3 * F_3^J + w_4 * F_4^J + w_5 * F_5^J) \end{aligned} \tag{8}$$

where w_1, w_2, w_3, w_4, w_5 are the weights of five feature blocks. $chunck$ represents the splitting operation to V_{weight} and PRE is the prediction operation by one 1x1 convolution operation. We take the sal_f as the final prediction. In addition, we also add a supervision to encourage the AFM can learn the most discriminative information for saliency detection.

Table 2: Quantitative comparisions of E-measure, S-measure, F-measure and MAE on seven widely-used RGB-D datasets. The top three scores in each column are marked in boldface, red, and green fonts, respectively. From top to bottom: CNNs-based RGB-D methods and traditional RGB-D methods.

		DUT-RGBD [27]				NJUD [19]				NLPR [26]				STEREO [25]	
Methods	Years	$E_\gamma\uparrow$	$S_\lambda\uparrow$	$F_\beta\uparrow$	$MAE\downarrow$	$E_\gamma\uparrow$	$S_\lambda\uparrow$	$F_\beta\uparrow$	$MAE\downarrow$	$E_\gamma\uparrow$	$S_\lambda\uparrow$	$F_\beta\uparrow$	$MAE\downarrow$	$E_\gamma\uparrow$	$S_\lambda\uparrow$
Ours	-	**0.941**	**0.910**	**0.903**	**0.039**	**0.917**	**0.898**	**0.879**	**0.048**	**0.945**	**0.914**	0.867	**0.029**	**0.927**	**0.901**
A2dele [28]	CVPR20	0.924	0.885	0.892	0.042	0.897	0.869	0.874	0.051	0.944	0.898	**0.872**	**0.029**	0.916	0.885
DMRA [27]	ICCV19	0.927	0.888	0.883	0.048	0.908	0.886	0.872	0.051	0.942	0.899	0.855	0.031	0.920	0.886
CPFP [41]	CVPR19	0.814	0.749	0.736	0.099	-	0.878	0.850	0.053	0.923	0.888	0.822	0.036	0.897	0.871
MMCI [6]	PR19	0.855	0.791	0.753	0.113	0.878	0.859	0.813	0.079	0.871	0.855	0.729	0.059	0.890	0.856
TANet [5]	TIP19	0.866	0.808	0.779	0.093	0.893	0.878	0.844	0.061	0.916	0.886	0.795	0.041	0.911	0.877
PDNet [43]	ICME19	0.861	0.799	0.757	0.112	0.890	0.883	0.832	0.062	0.876	0.835	0.740	0.064	0.903	0.874
PCA [4]	CVPR18	0.858	0.801	0.760	0.100	0.896	0.877	0.844	0.059	0.916	0.873	0.794	0.044	0.905	0.880
CTMF [15]	TCYB17	0.884	0.834	0.792	0.097	0.864	0.849	0.788	0.085	0.869	0.860	0.723	0.056	0.870	0.853
DF [31]	TIP17	0.842	0.730	0.748	0.145	0.818	0.735	0.744	0.151	0.838	0.769	0.682	0.099	0.844	0.763
MB [44]	CAIP17	0.691	0.607	0.577	0.156	0.643	0.534	0.492	0.202	0.814	0.714	0.637	0.089	0.693	0.579
CDCP [45]	ICCVW17	0.794	0.687	0.633	0.159	0.751	0.673	0.618	0.181	0.785	0.724	0.591	0.114	0.801	0.727
NLPR [26]	ECCV14	0.767	0.568	0.659	0.174	0.722	0.530	0.625	0.201	0.772	0.591	0.520	0.119	0.781	0.567
DES [8]	ICIMCS14	0.733	0.659	0.668	0.280	0.421	0.413	0.165	0.448	0.735	0.582	0.583	0.301	0.451	0.473
DCMC [10]	SPL16	0.712	0.499	0.406	0.243	0.796	0.703	0.715	0.167	0.684	0.550	0.328	0.196	0.838	0.745

Table 3: Continuation of Table 2

		STEREO [25]		LFSD [22]				RGBD135 [8]				SSD [21]			
Methods	Years	$F_\beta\uparrow$	$MAE\downarrow$	$E_\gamma\uparrow$	$S_\lambda\uparrow$	$F_\beta\uparrow$	$MAE\downarrow$	$E_\gamma\uparrow$	$S_\lambda\uparrow$	$F_\beta\uparrow$	$MAE\downarrow$	$E_\gamma\uparrow$	$S_\lambda\uparrow$	$F_\beta\uparrow$	$MAE\downarrow$
Ours	-	0.880	**0.043**	**0.899**	**0.857**	**0.855**	**0.073**	0.942	**0.902**	**0.868**	**0.028**	**0.905**	**0.872**	**0.827**	**0.053**
A2dele [28]	CVPR20	**0.884**	**0.043**	0.870	0.837	0.835	0.074	0.922	0.885	0.865	**0.028**	0.862	0.807	0.791	0.069
DMRA [27]	ICCV19	0.868	0.047	**0.899**	0.847	0.849	0.075	**0.945**	0.901	0.857	0.029	0.892	0.857	0.821	0.058
CPFP [41]	CVPR19	0.827	0.054	0.867	0.828	0.813	0.088	0.927	0.874	0.819	0.037	0.832	0.807	0.725	0.082
MMCI [6]	PR19	0.812	0.080	0.840	0.787	0.779	0.132	0.899	0.847	0.750	0.064	0.860	0.814	0.748	0.082
TANet [5]	TIP19	0.849	0.060	0.845	0.801	0.794	0.111	0.916	0.858	0.782	0.045	0.879	0.839	0.767	0.064
PDNet [43]	ICME19	0.833	0.064	0.872	0.845	0.824	0.109	0.915	0.868	0.800	0.050	0.813	0.802	0.716	0.115
PCA [4]	CVPR18	0.845	0.061	0.846	0.800	0.794	0.112	0.909	0.845	0.763	0.049	0.883	0.843	0.786	0.064
CTMF [15]	TCyb17	0.786	0.087	0.851	0.796	0.781	0.120	0.907	0.863	0.765	0.055	0.837	0.776	0.709	0.100
DF [31]	TIP17	0.761	0.142	0.841	0.796	0.810	0.142	0.801	0.685	0.566	0.130	0.802	0.742	0.709	0.151
MB [44]	CAIP17	0.572	0.178	0.631	0.538	0.543	0.218	0.798	0.661	0.588	0.102	0.633	0.499	0.414	0.219
CDCP [45]	ICCVW17	0.680	0.149	0.737	0.658	0.634	0.199	0.806	0.706	0.583	0.119	0.714	0.604	0.524	0.219
NLPR [26]	ECCV14	0.716	0.179	0.742	0.558	0.708	0.211	0.850	0.577	0.857	0.097	0.726	0.562	0.551	0.200
DES [8]	ICIMCS14	0.223	0.417	0.475	0.440	0.228	0.415	0.786	0.627	0.689	0.289	0.383	0.341	0.073	0.500
DCMC [10]	SPL16	0.761	0.150	0.842	0.754	0.815	0.155	0.674	0.470	0.228	0.194	0.790	0.706	0.684	0.168

4 EXPERIMENTS

4.1 Dataset

To evaluate the performance of our proposed method, we conduct comprehensive experiments on seven widely-used RGB-D datasets.
DUT-RGBD [27]: contains 1200 RGB-D scenes paired with corresponding depth maps and ground truths which are captured by a Lytro2 camera.
NJUD [19]: contains 1985 image pairs which are collected from indoor/outdoor environments and stereo movies. And the depth maps are estimated from the stereo images.
NLPR [26]: includes 1000 RGB images and corresponding high-quality depth maps which are captured by Kinect in both indoor and outdoor scenarios. Moreover, many images which contain multiple and small salient objects are included in this dataset.
STEREO [25]: includes 797 pairs of binocular images downloaded from the Internet. In some methods, it is also named SSB.
LFSD [22]: contains 100 image pairs captured by a Lytro camera.
RGBD135 [8]: includes 135 indoor images collected by the Microsoft Kinect. It is also named DES in some papers.
SSD [21]: contains 80 stereo images collected from three stereo movies.

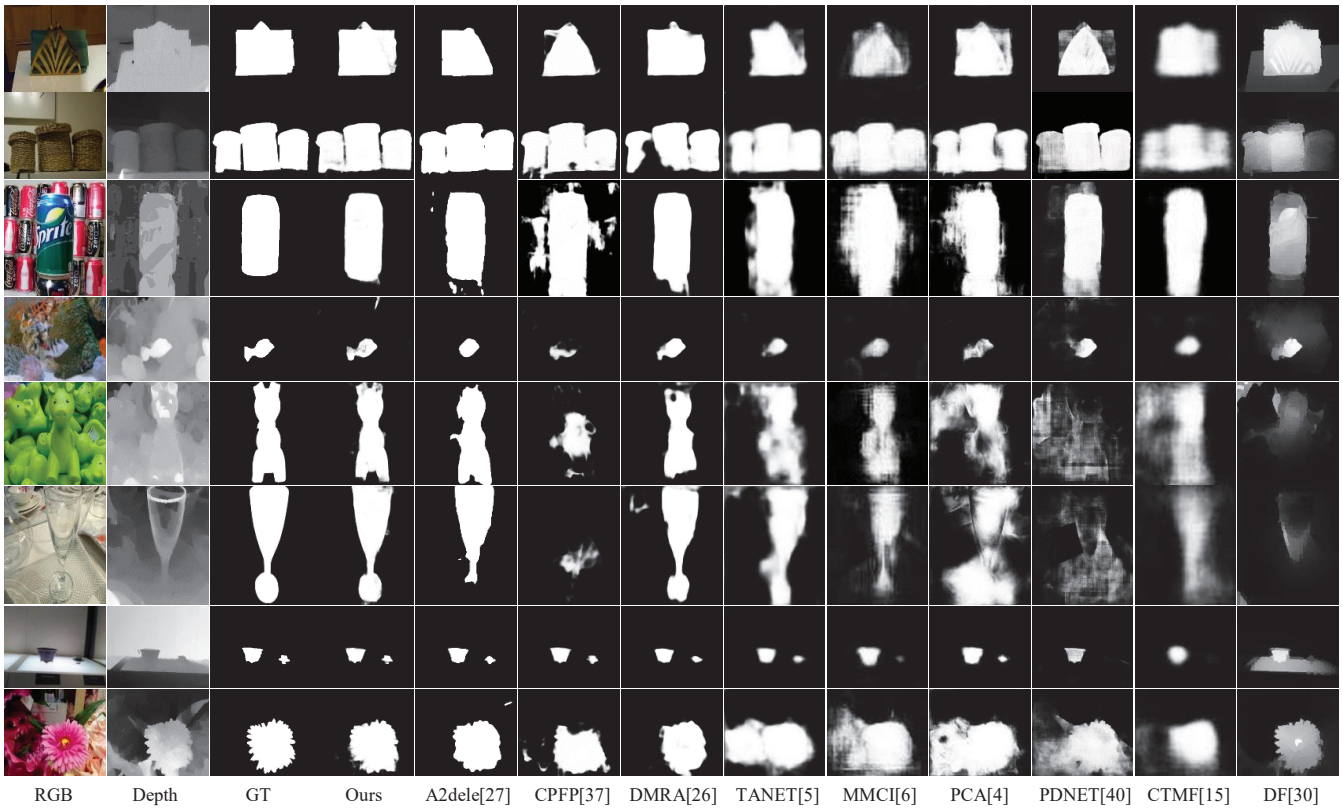

Figure 5: Visual comparisons of our method with latest CNNs-based approaches in challenging scenes, such as low contrast environments, complex background, multiple objects, transparent objects, and so on. Those methods are the top ranking ones in quantitative table shown in Table 1.

To guarantee a fair comparison, we adopt the similar splitting way as [4, 6, 41]. Specifically, we choose 800 samples from DUT-RGBD dataset, 700 samples from NLPR and 1485 samples from NJUD for training. The remaining images of these three datasets and the other four datasets are used for testing. In addition, we augment the training set by flipping, cropping and rotating to avoid overfitting.

4.2 Experimental Setup

Evaluation Metrics. For comprehensively evaluating various methods, we adopt five commonly used metrics, including S-measure (S_λ) [11], F-measure (F_β) [1], E-measure (E_γ) [12], mean absolute error (*MAE*) [3] and precision-recall (PR) curve. Specifically, the S-measure is a structure metric which can evaluate the structural similarities, the F-measure can evaluate the average precision and average recall, the E-measure can jointly capture image level statistics and local pixel matching information and the MAE is used to evaluate the average absolute difference between the prediction map and ground truth. In addition, the PR curve describes the different combination of precision and recall scores computed by comparing the binarized saliency map with the ground truth.

Implementation details. We choose the Pytorch toolbox to implement our method, trained on a PC with RTX 2080Ti GPU and 16 GB memory. The training and testing images are uniformly resized to 256×256. During training, we use the standard SGD optimizer. The momentum, weight decay and learning rate are set as 0.9, 0.0005 and 1e-10, respectively. The cross entropy loss is adopted to train our network, converging after 15 epochs with batch size of 2.

Baseline. Our baseline is shown in Figure 7. To fully extract the useful features from origial RGB-D paired images, we adopt VGG19 for both RGB and depth streams. Simple concatenation is employed to fuse two-modal features. Additionally, we take the supervisions on each level.

4.3 Comparison with State-of-the-arts

We compare our method with 14 state-of-the-art RGB-D salient object detection methods, including 9 latest CNNs-based methods: A2dele [28], DMRA [27], CPFP [41], PDNet [43], PCA [4], CTMF [15], MMCI [6], DF [31], TANet [5]; and 5 traditional methods: DES [8], NLPR [26], DCMC [10], MB [44], CDCP [45]. For fair comparisons, We implement those methods with the released code and their default parameters. In terms of methods without the released source code, the results are directly provided by authors.

Quantitative Evaluation. Table 2 and 3 show the validation results in terms of four evaluation metrics on seven public datasets. It can be seen that our method can outperform across the existing methods, except the second-best E_γ score on RGBD135, and the second-best F_β scores on NLPR and STEREO. Moreover, except the

Table 4: Ablation analysis on seven datasets. We report the performance of BS as the baseline. Row (b), (c) and (d) show the influence of each individual component by adding it to the baseline, respectively. Row (e) illustrates the performance of the GSFM acting on each level of the baseline. Row (b),(f) and (g) show the influence of each individual component by successively adding it to the baseline. Please see Section 4.4 for more detailed analysis. In addition, H represents the last two levels, L represents the first three levels and A represents all levels.

						DUT-RGBD		NJUD		NLPR		STEREO		LFSD		RGBD135		SSD	
index	BS	IFM (H)	GSFM (L)	AFM	GSFM (A)	$F_\beta \uparrow$	$MAE \downarrow$	$F_\beta \uparrow$	$MAE \downarrow$	$F_\beta \uparrow$	$MAE \downarrow$	$F_\beta \uparrow$	$MAE \downarrow$	$F_\beta \uparrow$	$MAE \downarrow$	$F_\beta \uparrow$	$MAE \downarrow$	$F_\beta \uparrow$	$MAE \downarrow$
(a)	√					0.843	0.063	0.834	0.057	0.795	0.040	0.833	0.054	0.823	0.088	0.815	0.036	0.783	0.067
(b)	√	√				0.854	0.048	0.843	0.054	0.810	0.037	0.842	0.050	0.826	0.085	0.834	0.031	0.800	0.061
(c)	√		√			0.870	0.048	0.854	0.052	0.830	0.037	0.853	0.050	0.835	0.083	0.842	0.032	0.801	0.063
(d)	√			√		0.855	0.047	0.846	0.053	0.818	0.035	0.850	0.049	0.835	0.082	0.840	0.031	0.799	0.059
(e)	√				√	0.877	0.043	0.858	0.052	0.841	0.033	0.856	0.048	0.842	0.081	0.849	0.030	0.806	0.058
(f)	√	√	√			0.888	0.042	0.866	0.050	0.852	0.033	0.868	0.047	0.848	0.079	0.857	0.029	0.815	0.056
(g)	√	√	√	√		**0.903**	**0.039**	**0.879**	**0.048**	**0.867**	**0.029**	**0.880**	**0.043**	**0.855**	**0.073**	**0.868**	**0.028**	**0.827**	**0.053**

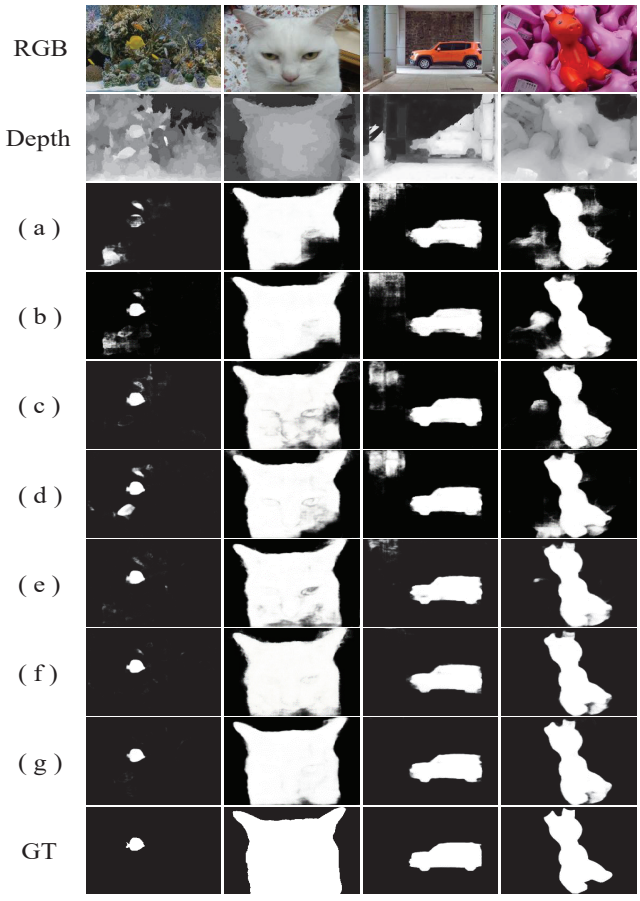

Figure 6: The visual results of ablation analysis. The meaning of indexes has been shown in Table 4.

RGBD135, our method improves the S_λ by a large margin on the other six datasets. This indicates that our method is more powerful to detect complete salient objects in terms of region-aware and object-aware structural similarity between the saliency map and ground truth. The PR curves in Figure 8 also consistently demonstrate the superior performance of our method.

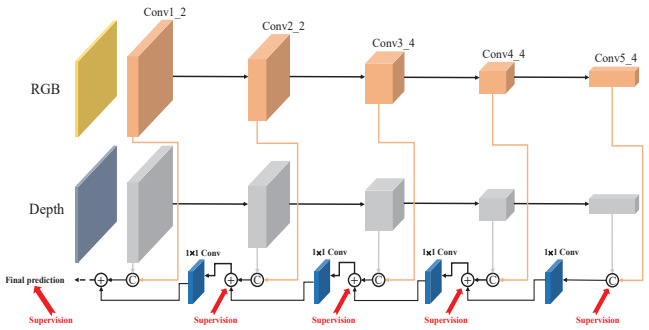

Figure 7: The baseline of our method.

Qualitative Evaluation. For a more intuitive view, we visually compare our method with the most representative methods in Figure 5. As we can see, our method is capable of accurately capturing salient regions in some challenging scenes, including cluttered background (Row 3), similar foreground and background (Row 5), transparent object (Row 6), small object (Row 4 and 7) and multiple objects (Row 2 and 7). This indicates that our method can effectively filter out the redundant information to predict the results accurately. Moreover, compared to some methods which utilize the top-down or bottom-up cross-modal cross-level fusion manner (DMRA [27], PCA [4], MMCI [6], TANet [5]), our method locates and detects the salient objects more accurately. It further demonstrates the superiority of our method benefiting from feature reintegration.

4.4 Ablation Studies

Effect of IFM. In order to verify the effectiveness of the proposed IFM, we replace the simple concatenation operation with the IFM in the last two levels of the baseline. As shown in Table 4 (b), our IFM improves the performance of the baseline across all datasets. Intuitively, as shown in Figure 6 (b), the predictions produced by adding the IFM can better locate the salient object. This advance confirms the superiority of our IFM in effectively extracting and fusing paired complementary information from high levels.

Effect of GSFM. To give evidence for the effectiveness of the GSFM, we replace the simple concatenation operation with the GSFM in the first three (1-3) levels of the baseline. By comparing the results in Table 4 (c) and (a), we observe that our proposed GSFM achieves significant improvement than the baseline. Meanwhile, as shown

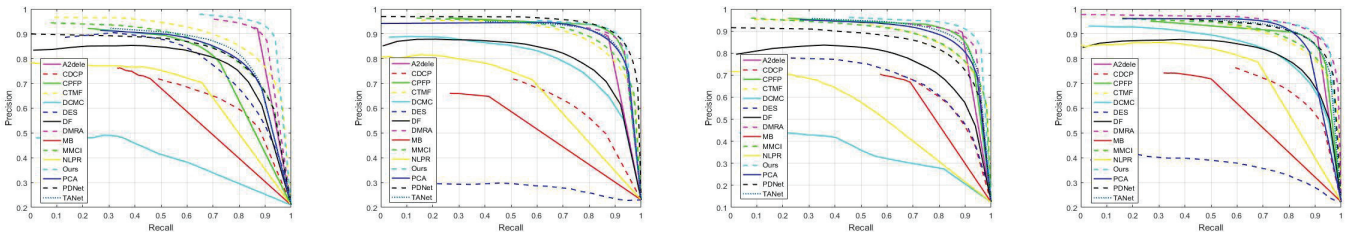

Figure 8: The PR curves of the proposed method and other state-of-the-art approaches over DUT-RGBD, NJUD, NLPR and STEREO datasets.

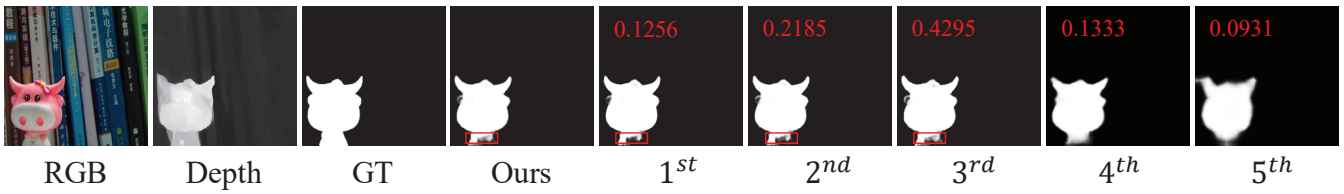

Figure 9: From left to right (Column 5-9): the predictions generated from the first level to the last level of the proposed model, respectively. Numbers on each prediction map indicate the corresponding attention weights learned from the AFM.

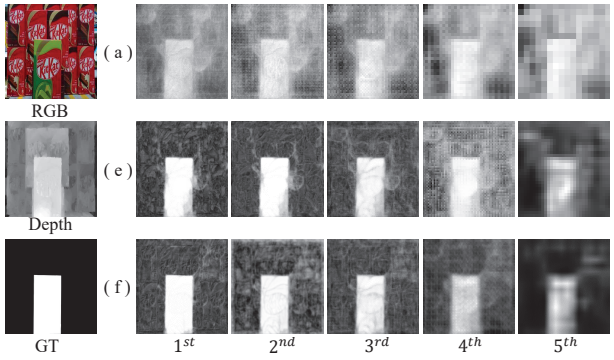

Figure 10: The feature maps visualized from the first level to the last level of the baseline (a), the baseline+GSFM (all levels) (e) and the proposed model (f), respectively. The meaning of indexes is the same as the ones in Table 4.

in Figure 6 (c), the results produced by adding the GSFM contain more useful and accurate details than the baseline. Moreover, we visualize the feature maps generated from each level of (a), (e) and (f), same as the indexes in Table 4, as shown in Figure 10, it is seen that the sharper salient object standing out from the background in the first three levels of (e) and (f). This further confirms that the redundant features can be filtered out by the proposed GSFM in low levels.

Effect of AFM. We compare the baseline network with it adding the AFM to prove the effectiveness of the AFM. As shown in Table 4 (d), the improved performances across all datasets are achieved by adding our proposed AFM. By comparing the visual results in Figure 6 (d) and (a), the predictions produced by AFM contain more complete information than the baseline. Furthermore, as shown in Table 4 (g) and (f), the final results of our method are improved by the AFM. The visual results in Figure 6 (g) and (f) further prove the effectiveness of our proposed AFM. To better understand the difference of the features learned by the IFM and GSFM, we visualize the predictions and calculate the corresponding attention weights for each level by the proposed AFM, as shown in Figure 9. As we can see, the AFM can give exclusive weights of the fused cross-modal features of each level to emphasize the useful features and suppress unnecessary ones.

Effect of our architecture. To illustrate the effectiveness of our different strategies for processing low-level and high-level features, we replace the concatenation operation with the GSFM at each level of the baseline. Comparing the results in Table 4 (f) with (e), we observe that our different strategies achieve better performance. As shown in Figure 6 (f), we can see that our structure can predict more accurate and complete salient objects. On the other hand, we can observe from Figure 10 (e) and (f) that the IFM can effectively and efficiently explore the high-level features (the last two levels) with a simpler structure than the GSFM. This indicates that it is not necessary to utilize complex operations to deal with high-level features. It further demonstrates that the different fusion strategies can be leveraged, targeted at low-level and high-level features.

5 CONCLUSION

In this work, we adopt different fusion strategies in high and low levels. Taking account of global and local complementarities from two modalities, we propose a novel top-down multi-level fusion structure. It includes an interweave fusion module (IFM) which can fully extract and fuse global information in high levels, and three gated select fusion modules (GSFM) which can selectively process the useful information from two modal features in low levels. Moreover, considering that the high-level features will be diluted when transmitted to the lower levels, we design an adaptive fusion module (AFM) to reintegrate the fused cross-modal features based on the top-down fusion structure. Experimental results demonstrate that our method achieves state-of-the-art performance on 7 public RGB-D datasets.

REFERENCES

[1] Radhakrishna Achanta, Sheila Hemami, Francisco Estrada, and Sabine Susstrunk. 2009. Frequency-tuned salient region detection. In *IEEE Conference on Computer Vision and Pattern Recognition*. 1597–1604.

[2] Ali Borji, Simone Frintrop, Dicky N. Sihite, and Laurent Itti. 2012. Adaptive object tracking by learning background context. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*. 23–30.

[3] Ali Borji, Dicky N. Sihite, and Laurent Itti. 2012. Salient object detection: a benchmark. In *European Conference on Computer Vision*. 414–429.

[4] Hao Chen and Youfu Li. 2018. Progressively Complementarity-Aware Fusion Network for RGB-D Salient Object Detection. In *IEEE Conference on Computer Vision and Pattern Recognition*. 3051–3060.

[5] Hao Chen and Youfu Li. 2019. Three-Stream Attention-Aware Network for RGB-D Salient Object Detection. *IEEE Transactions on Image Processing* 28, 6, 2825–2835.

[6] Hao Chen, Youfu Li, and Dan Su. 2019. Multi-modal fusion network with multi-scale multi-path and cross-modal interactions for RGB-D salient object detection. *Pattern Recognition* 86, 376–385.

[7] Yanhua Cheng, Rui Cai, Zhiwei Li, Xin Zhao, and Kaiqi Huang. 2017. Locality-Sensitive Deconvolution Networks with Gated Fusion for RGB-D Indoor Semantic Segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition*. 1475–1483.

[8] Yupeng Cheng, Huazhu Fu, Xingxing Wei, Jiangjian Xiao, and Xiaochun Cao. 2014. Depth Enhanced Saliency Detection Method. In *International Conference on Internet Multimedia Computing and Service*. 23–27.

[9] Xiao Chu, Wei Yang, Wanli Ouyang, Cheng Ma, Alan Loddon Yuille, and Xiaogang Wang. 2017. Multi-context Attention for Human Pose Estimation. In *IEEE Conference on Computer Vision and Pattern Recognition*. 5669–5678.

[10] Runmin Cong, Jianjun Lei, Changqing Zhang, Qingming Huang, Xiaochun Cao, and Chunping Hou. 2016. Saliency Detection for Stereoscopic Images Based on Depth Confidence Analysis and Multiple Cues Fusion. *IEEE Signal Processing Letters* 23, 6, 819–823.

[11] Deng-Ping Fan, Ming-Ming Cheng, Yun Liu, Tao Li, and Ali Borji. 2017. Structure-Measure: A New Way to Evaluate Foreground Maps. In *IEEE International Conference on Computer Vision*. 4558–4567.

[12] Deng-Ping Fan, Cheng Gong, Yang Cao, Bo Ren, Ming-Ming Cheng, and Ali Borji. 2018. Enhanced-alignment Measure for Binary Foreground Map Evaluation. In *International Joint Conference on Artificial Intelligence*. 698–704.

[13] David Feng, Nick Barnes, Shaodi You, and Chris McCarthy. 2016. Local Background Enclosure for RGB-D Salient Object Detection. In *IEEE Conference on Computer Vision and Pattern Recognition*. 2343–2350.

[14] Mengyang Feng, Huchuan Lu, and Errui Ding. 2019. Attentive Feedback Network for Boundary-Aware Salient Object Detection. In *IEEE Conference on Computer Vision and Pattern Recognition*. 1623–1632.

[15] Junwei Han, Hao Chen, Nian Liu, Chenggang Yan, and Xuelong Li. 2018. CNNs-Based RGB-D Saliency Detection via Cross-View Transfer and Multiview Fusion. *IEEE Transactions on Cybernetics* 48, 11, 3171–3183.

[16] Seunghoon Hong, Tackgeun You, Suha Kwak, and Bohyung Han. 2015. Online Tracking by Learning Discriminative Saliency Map with Convolutional Neural Network. In *International Conference on International Conference on Machine Learning*. 597–606.

[17] Laurent Itti, Christof Koch, and Ernst Niebur. 1998. A model of saliency-based visual attention for rapid scene analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 20, 11, 1254–1259.

[18] Wei Ji, Jingjing Li, Miao Zhang, Yongri Piao, and Huchuan Lu. 2020. Accurate RGB-D Salient Object Detection via Collaborative Learning. *European Conference on Computer Vision*.

[19] Ran Ju, Ling Ge, Wenjing Geng, Tongwei Ren, and Gangshan Wu. 2014. Depth saliency based on anisotropic center-surround difference. In *IEEE International Conference on Image Processing*. 1115–1119.

[20] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. 2017. ImageNet classification with deep convolutional neural networks. *Communications of The ACM* 60, 6, 84–90.

[21] Ge Li and Chunbiao Zhu. 2017. A Three-Pathway Psychobiological Framework of Salient Object Detection Using Stereoscopic Technology. In *IEEE International Conference on Computer Vision Workshops*. 3008–3014.

[22] Nianyi Li, Jinwei Ye, Yu Ji, Haibin Ling, and Jingyi Yu. 2017. Saliency Detection on Light Field. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 39, 8, 1605–1616.

[23] Nian Liu, Junwei Han, and Ming-Hsuan Yang. 2018. PiCANet: Learning Pixel-Wise Contextual Attention for Saliency Detection. In *IEEE Conference on Computer Vision and Pattern Recognition*. 3089–3098.

[24] Jonathan Long, Evan Shelhamer, and Trevor Darrell. 2015. Fully convolutional networks for semantic segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition*. 3431–3440.

[25] Yuzhen Niu, Yujie Geng, Xueqing Li, and Feng Liu. 2012. Leveraging stereopsis for saliency analysis. In *IEEE Conference on Computer Vision and Pattern Recognition*. 454–461.

[26] Houwen Peng, Bing Li, Weihua Xiong, Weiming Hu, and Rongrong Ji. 2014. RGBD Salient Object Detection: A Benchmark and Algorithms. In *European Conference on Computer Vision*. 92–109.

[27] Yongri Piao, Wei Ji, Jingjing Li, Miao Zhang, and Huchuan Lu. 2019. Depth-Induced Multi-Scale Recurrent Attention Network for Saliency Detection. In *IEEE International Conference on Computer Vision*. 7253–7262.

[28] Yongri Piao, Zhengkun Rong, Miao Zhang, Weisong Ren, and Huchuan Lu. 2020. A2dele: Adaptive and Attentive Depth Distiller for Efficient RGB-D Salient Object Detection. In *IEEE Conference on Computer Vision and Pattern Recognition*. 9060–9069.

[29] Xuebin Qin, Zichen Zhang, Chenyang Huang, Chao Gao, Masood Dehghan, and Martin Jagersand. 2019. BASNet: Boundary-Aware Salient Object Detection. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 7479–7489.

[30] Yao Qin, Huchuan Lu, Yiqun Xu, and He Wang. 2015. Saliency detection via Cellular Automata. In *IEEE Conference on Computer Vision and Pattern Recognition*. 110–119.

[31] Liangqiong Qu, Shengfeng He, Jiawei Zhang, Jiandong Tian, Yandong Tang, and Qingxiong Yang. 2017. RGBD Salient Object Detection via Deep Fusion. *IEEE Transactions on Image Processing* 26, 5, 2274–2285.

[32] Jingfan Quo, Tongwei Ren, and Jia Bei. 2016. Salient object detection for RGB-D image via saliency evolution. In *IEEE International Conference on Multimedia and Expo*. 1–6.

[33] Jianqiang Ren, Xiaojin Gong, Lu Yu, Wenhui Zhou, and Michael Ying Yang. 2015. Exploiting global priors for RGB-D saliency detection. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops*. 25–32.

[34] Tongwei Ren and Ao Zhang. 2019. RGB-D Salient Object Detection: A Review. 203–220.

[35] Karen Simonyan and Andrew Zisserman. 2015. Very Deep Convolutional Networks for Large-Scale Image Recognition. In *International Conference on Learning Representations*.

[36] Wei-Chih Tu, Shengfeng He, Qingxiong Yang, and Shao-Yi Chien. 2016. Real-Time Salient Object Detection with a Minimum Spanning Tree. In *IEEE Conference on Computer Vision and Pattern Recognition*. 2334–2342.

[37] Zhe Wu, Li Su, and Qingming Huang. 2019. Cascaded Partial Decoder for Fast and Accurate Salient Object Detection. In *IEEE Conference on Computer Vision and Pattern Recognition*. 3907–3916.

[38] Yu Zeng, Huchuan Lu, Lihe Zhang, Mengyang Feng, and Ali Borji. 2018. Learning to Promote Saliency Detectors. In *IEEE Conference on Computer Vision and Pattern Recognition*. 1644–1653.

[39] Pingping Zhang, Dong Wang, Huchuan Lu, Hongyu Wang, and Xiang Ruan. 2017. Amulet: Aggregating Multi-level Convolutional Features for Salient Object Detection. In *IEEE International Conference on Computer Vision*. 202–211.

[40] Jiaxing Zhao, Jiang-Jiang Liu, Deng-Ping Fan, Yang Cao, Jufeng Yang, and Ming-Ming Cheng. 2019. EGNet: Edge Guidance Network for Salient Object Detection. In *IEEE International Conference on Computer Vision*. 8778–8787.

[41] Jia-Xing Zhao, Yang Cao, Deng-Ping Fan, Ming-Ming Cheng, Xuan-Yi Li, and Le Zhang. 2019. Contrast Prior and Fluid Pyramid Integration for RGBD Salient Object Detection. In *IEEE Conference on Computer Vision and Pattern Recognition*. 3927–3936.

[42] Ting Zhao and Xiangqian Wu. 2019. Pyramid Feature Attention Network for Saliency Detection. In *IEEE Conference on Computer Vision and Pattern Recognition*. 3085–3094.

[43] Chunbiao Zhu, Xing Cai, Kan Huang, Thomas H. Li, and Ge Li. 2019. PDNet: Prior-Model Guided Depth-Enhanced Network for Salient Object Detection. In *IEEE International Conference on Multimedia and Expo*. 199–204.

[44] Chunbiao Zhu, Ge Li, Xiaoqiang Guo, Wenmin Wang, and Ronggang Wang. 2017. A Multilayer Backpropagation Saliency Detection Algorithm Based on Depth Mining. In *International Conference on Computer Analysis of Images and Patterns*. 14–23.

[45] Chunbiao Zhu, Ge Li, Wenmin Wang, and Ronggang Wang. 2017. An Innovative Salient Object Detection Using Center-Dark Channel Prior. In *IEEE International Conference on Computer Vision Workshops*. 1509–1515.