Select, Supplement and Focus for RGB-D Saliency Detection

Miao Zhang[1,2*] Weisong Ren[1*] Yongri Piao[1†] Zhengkun Rong[1] Huchuan Lu[1,3]
[1]Dalian University of Technology, China
[2]Key Lab for Ubiquitous Network and Service Software of Liaoning Province,
Dalian University of Technology, China
[3]Pengcheng Lab
{miaozhang, yrpiao, lhchuan}@dlut.edu.cn {beatlescoco, rzk911113}@mail.dlut.edu.cn

Abstract

Depth data containing a preponderance of discriminative power in location have been proven beneficial for accurate saliency prediction. However, RGB-D saliency detection methods are also negatively influenced by randomly distributed erroneous or missing regions on the depth map or along the object boundaries. This offers the possibility of achieving more effective inference by well designed models. In this paper, we propose a new framework for accurate RGB-D saliency detection taking account of global location and local detail complementarities from two modalities. This is achieved by designing a complimentary interaction module (CIM) to discriminatively select useful representation from the RGB and depth data, and effectively integrate cross-modal features. Benefiting from the proposed CIM, the fused features can accurately locate salient objects with fine edge details. Moreover, we propose a compensation-aware loss to improve the network's confidence in detecting hard samples. Comprehensive experiments on six public datasets demonstrate that our method outperforms 18 state-of-the-art methods.

1. Introduction

Salient object detection (SOD) aims to distinguish the most attractive object in a scene. This fundamental task plays an important role in a wide range of computer vision and robotic vision tasks [3], such as video/image segmentation [11], visual tracking [18], image captioning [9]. Many previous works in SOD focus on RGB images. Despite great progress RGB SOD methods have achieved, they may remain sensitive when it comes to challenging scenes. This is mainly because without accurate spatial constraints, appearance features alone are often intrinsically less predictive for saliency detection when the color contrasts between foreground and background are quite low, or the background is cluttered.

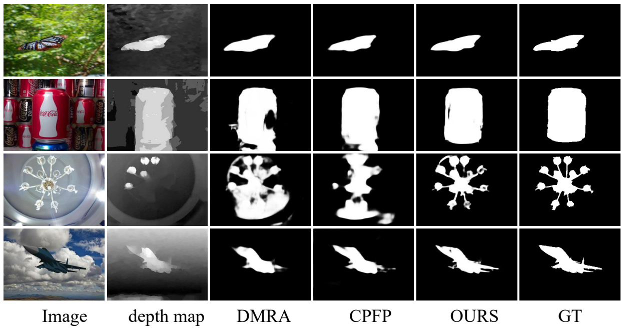

Figure 1. Samples from RGB-D saliency datasets. First row shows that ideal depth maps can significantly help the detection task. However, as shown in the other rows, undesirable depth maps can also significantly affect the prediction effect.

Depth maps with affluent spatial structure information have been proven beneficial for accurate saliency prediction. Within the last few years, tremendous efforts have been made towards RGB-D saliency detection [27, 37, 6, 5, 1]. The accuracy in RGB-D saliency detection is highly rely on the quality of the depth maps, which can be easily influenced by a variety of noise, such as the temperature of the camera, background illumination, and distance and reflectivity of the observed objects. Therefore, depth maps captured in real-life scenarios pose huge challenges to accurate RGB-D saliency detection in terms of two aspects. First, randomly distributed erroneous or missing regions are introduced on the depth map [33]. This is usually produced from sensors, absorption, or poor reflection, for example, a part of the object appears at the incorrect depth as shown in the 3_{rd} row of Figure 1. We also demonstrate in Figure 2 that the latest representative RGB-D methods are gradually losing the battle against the top ranking RGB method as similarity errors in the depth map increase. Second, erroneous depth measurements occur predominantly near object

[*] Equal Contributions
[†] Corresponding Author

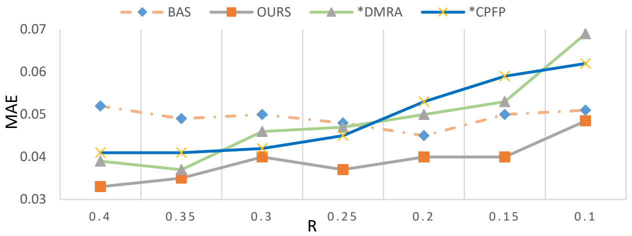

Figure 2. We show the performance comparisons of MAE regarding R. R represents the depth contrast value between the saliency region and the background of the input depth map, formulated as $R = |D_s - D_{ns}|$. D_s is the average depth value of the saliency region, and D_{ns} is the average depth value of the background region. We test two state-of-the-art RGB-D methods [27, 37] (remarked with $*$), one top ranking RGB method [32] as well as our proposed method in the NJUD+LFSD+NLPR dataset.

boundaries [37]. This is usually caused by the imaging principles, for example, the missing regions present along the depth boundaries as shown in the 4_{th} row of Figure 1. Unreliable boundary information in the depth map can also significantly affect subsequent performance.

In this work, we strive to embrace challenges towards accurate RGB-D saliency detection. The primary challenge towards this goal is in the design of a model that is discriminative enough to simultaneously reason about useful representation from the RGB-D data for cross-modal complements. The second challenge is in the design of the loss that has high confidence in the hard samples of the unreliable depth maps, leading to inaccurate and blurry predictions.

Our core insight is that we leverage the depth information despite the fact that the quality of some depth maps is far from perfection, to address the aforementioned challenges. Our approach focuses on effectively exploring and establishing complementarity and cooperation of cross-model features and meanwhile avoiding negative influence introduced by erroneous depth maps. The source code is released [1]. Concretely, our contributions are:

- We design a complimentary interaction module (CIM) for discriminatively exploring cross-modal complementarities, and effectively fusing cross-modal features. Our CIM associates the two modalities through a region-wise attention, and enhances each modality by supplementing rich boundary information.
- We introduce a compensation-aware loss to improve the networks's confidence for hard samples. To this end, the proposed loss further helps our network mine the structure information contained in the cross-modal features ensuring high stability for saliency detection in the challenging scenes.
- Our model outperforms 18 state-of-the-art SOD methods, including 9 RGB methods and 9 RGB-D methods, over 6 benchmark datasets.

2. Related Work

RGB-D Saliency Detection: A great number of RGB salient object detection (SOD) methods [19, 35, 36, 10, 23] have achieved outstanding performance. However, they may potentially appear fragile in some complex scenarios, such as similar foreground and background, complex background, transparent objects and low illumination. Therefore, additional auxiliary information should be exploited to assist the SOD task. Some works [37, 5] focus on the RGB-D saliency detection which uses depth cues to improve the performance in those complex scenes.

Traditional RGB-D saliency detection approaches mostly focus on introducing more effective cross-modal fusion methods, which can be divided into three categories: (a) [26, 31] concatenate the input depth map and the RGB image. (b) [13, 14] individually produce predictions from both RGB images and depth maps, and then integrate the results. (c) [15, 29] combine handcrafted RGB and depth saliency features to infer the final result.

Recently, CNNs are adopted in the RGB-D saliency detection to learn more discriminative RGB-D features. [28] feed handcrafted RGB-D features to a CNN for deep representations. [39] use a CNN-based network to process RGB information and a contrast-enhanced net to extract depth cues. [17, 5, 7] propose multi-modal multi-level fusion strategies to capture complementary information from RGB-D features. [6] propose a three-stream architecture to augment the RGB-D representation capacity in a bottom-up way, and introduce a top-down inference way to combine the cross-modal information. [37] enhance depth maps to work with RGB features, and design a fluid pyramid integration method to make better use of multi-scale cross-modal features. [27] use a depth included multi-scale weighting module to locate and identify salient objects and progressively generate more accurate saliency results through a recurrent attention model.

Different from the aforementioned methods, our work takes negative impacts caused by unreliable depth maps into account, and strives to exploit useful and precise information for cross-modal fusion.

3. The Proposed Framework

Overview. In this section, we describe the details of our proposed framework for the RGB-D saliency detection. Figure 3 shows an overview of the proposed network which consists of three main parts. First, two VGG-16 [30] based encoders learn feature representation from the RGB and depth images, respectively. Then we propose a cross modal attention unit (CAU) and a boundary refine unit (BSU) in Sec.3.1 to generate effective features of salient object location and boundary details. In the decoder part, we adopt a partial decoder to effectively fuse the extracted features.

[1] https://github.com/OIPLab-DUT/CVPR_SSF-RGBD

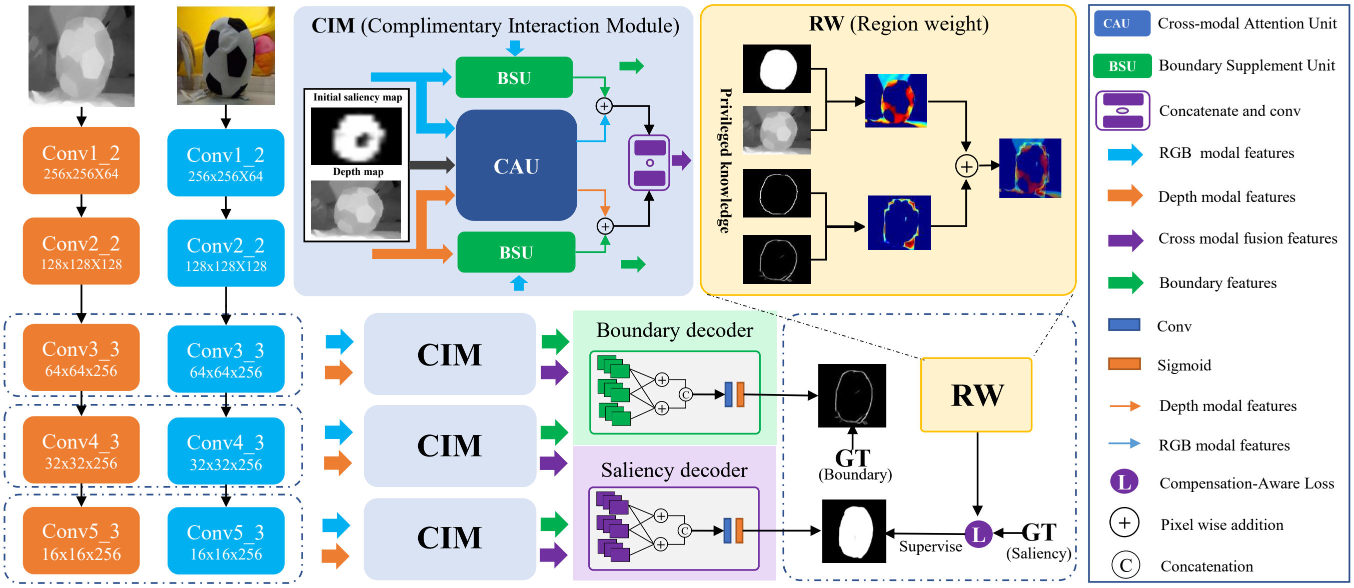

Figure 3. The overall architecture of our proposed network.

Details of this part are introduced in Sec.3.2. To guide the network to further learn from challenging scenes, we introduce a compensation-aware loss in Sec.3.3.

3.1. Complimentary Interaction Module

3.1.1 Cross-modal Attention Unit

The gap between different modalities makes linear fusion strategies of multi-modal features less adaptive to complex scenes. To tackle this dilemma, [37][5] propose fusion methods to better fuse cross-modal features. These methods are based on inputting depth maps with high contrast between the foreground and background objects. However, the ideal depth maps can not be always guaranteed during training and testing. Thus, these methods are easy to be negatively affected by erroneous depth maps.

To solve this problem, we propose a cross-modal attention unit, as shown in Figure 4(a). It aims to effectively select useful information from the RGB and depth modal features. First, we divide the depth map (0-1) into m binary mask maps to help find useful regions for salient object detection. The binary mask maps share the same spatial resolution with the depth map. To be specific, for the depth map, we set pixels in the depth region $(\frac{i-1}{m}, \frac{i}{m})$ as 1, and other pixels as 0 to generate the i_{th} mask map. For the RGB modal, we firstly generate a rough saliency map, namely the initial saliency map from the 5_{th} level of the RGB encoder through a 1×1 convolution ($S^{pr} = Conv(F_r^5)$). S^{pr} is supervised by the saliency ground truth. The initial saliency map guides the choice of depth layers, which is formulated as:

$$\tilde{F}_r^{[l]} = CA(F_r^{[l]}\sum_{i=1}^{m} d_i \frac{\sum\limits_{j=1}^{N}(d_{i,j}S_j^{pr})}{\sum\limits_{j=1}^{N} d_{i,j}}), \qquad (1)$$

where $F_r^{[l]}$ is the l_{th} level RGB features generated from the RFB block[34]. d_i represents the i_{th} binary mask map. S^{pr} represents the initial saliency map. N is the total number of pixels in the image. m is the number of the binary mask maps. CA denotes the channel-wise attention. $\tilde{F}_r^{[l]}$ is the output features generated from the channel-wise attention at the l_{th} level. When it comes to unreliable depth maps, it is hard to provide effective complementary from depth maps. To alleviate the above issue, we introduce another RGB attention block to work with the cross-modal attention block, denoted as: $CA(S^* * F_r^{[l]})$. If l=5, $S^* = S^{pr}$, else S^*=$(1 - S^{pr})$. This block maintains the high-level semantic information with the contribution of S^{pr} and residual details with the help of $(1 - S^{pr})$. In this case, our network can maintain the reliable information learned from the RGB modal. And the final output RGB features $\bar{F}_r^{[l]}$ is formulated as:

$$\bar{F}_r^{[l]} = Conv(Cat(\tilde{F}_r^{[l]}, CA(S^* * F_r^{[l]}))), \qquad (2)$$

where $Conv()$ represents a convolutional layer with 64 input channels and 32 output channels. For the depth modal, we also use the initial saliency map to weight the depth layers according to the proportion of saliency area. This process can be formulated as:

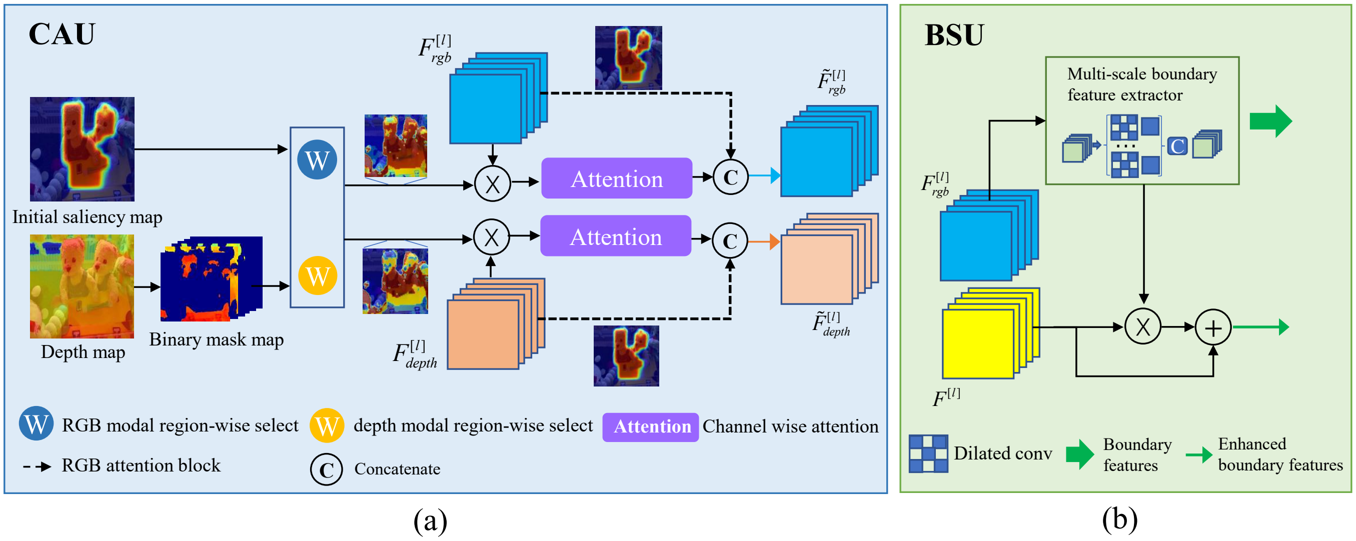

Figure 4. Detailed diagram of sub-units of the complimentary interaction model. (a) is the details of Cross-modal Attention Unit (CAU), (b) is the details of Boundary Supplement Unit(BSU).

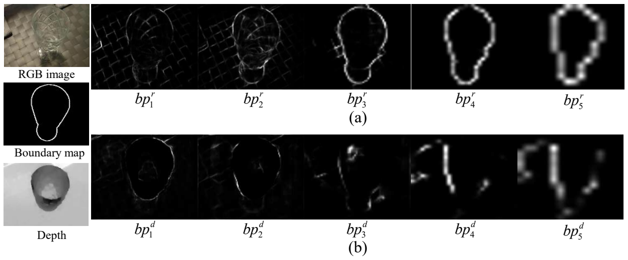

Figure 5. We generate the saliency edge prediction from each level of the VGG-16. (a): bp_i^r represents the side out edge prediction from the i_{th} level when inputting the RGB image. (b): bp_i^d represents the side out edge prediction from the i_{th} level when inputting the depth map. It is seen that bp_5^r, bp_4^r, bp_3^r can maintain the pure and complete saliency edge information.

$$\tilde{F}_d^{[l]} = CA(F_d^{[l]}\sum_{i=1}^{m} d_i \frac{\sum\limits_{j=1}^{N}(d_{i,j}S_j^{pr})}{\sum\limits_{j=1}^{N} S_j^{pr}}), \tag{3}$$

where $F_d^{[l]}$ represents RGB modal features from l_{th} level after the RFB block. This selection step can help our network focus on the important regions as well as channels of cross-modal features.

3.1.2 Boundary Supplement Unit

Existing RGB-D saliency detection methods still hold the problem of blurred boundaries due the the pooling operations. Specifically, as shown in Figure 5, it's hard to extract boundary details from the depth stream which leads to the blurred predictions. Based on this observation, we propose the boundary supplement unit, as shown in Figure 4(b). Different from previous methods which extract local edge information from low-level features [38], this unit aims to effectively explore edge details from high-levels (VGG16: l_3, l_4, l_5) of the RGB modal encoder. This is motivated by high level features of the RGB modal containing purer boundary information, as shown in Figure 5. Specifically, we design a multi-scale boundary feature extractor, which contains four parallel dilated convolutional blocks with different receptive fields (1, 3, 5, 7). Then we fuse the obtained complementary salient edge features and the saliency features at each level, shown as:

$$\tilde{F}_b = F_b \otimes F^{[l]} \oplus F^{[l]}, \tag{4}$$

where $\otimes$ denotes the element-wise multiplication, and $\oplus$ represents the element-wise addition. F_b denotes the extracted boundary features from RGB modal after the multi-scale boundary feature extractor. $\tilde{F}_b$ represents the boundary enhanced features. We decode F_b to the boundary prediction, supervised by the boundary ground truth generated by the saliency ground truth, to encourage F_b obtaining discriminative boundary inference. Then we concatenate features after the CAU and the BSU in each modal, and generate the enhanced feature $F_{dout}^{[l]}$, $F_{rout}^{[l]}$ from the depth modal and the RGB modal, respectively. Finally, we fuse cross-modal features, which is shown as:

$$F_f^{[l]} = Conv(cat(F_{dout}^{[l]}, F_{rout}^{[l]})), \tag{5}$$

where $Conv(\cdot)$ represents a convolutional layer with 64 input channels and 32 output channels.

Table 1. Quantitative comparisons of S-measure, F-measure and MAE scores on six RGB-D datasets. Methods with/without $*$ represent RGB-D methods and RGB methods respectively. The best three results are shown in red, green, and blue.

Methods	Year	DUT-RGBD			NJUD			NLPR			STEREO			LFSD			RGBD135		
		$S_\alpha \uparrow$	$F_\beta \uparrow$	MAE$\downarrow$	$S_\alpha \uparrow$	$F_\beta \uparrow$	MAE$\downarrow$	$S_\alpha \uparrow$	$F_\beta \uparrow$	MAE$\downarrow$	$S_\alpha \uparrow$	$F_\beta \uparrow$	MAE$\downarrow$	$S_\alpha \uparrow$	$F_\beta \uparrow$	MAE$\downarrow$	$S_\alpha \uparrow$	$F_\beta \uparrow$	MAE$\downarrow$
DSS	CVPR17	.767	.732	.127	.807	.776	.108	.816	.755	.076	.841	.814	.087	.718	.694	.166	.763	.697	.098
Amulet	ICCV17	.846	.803	.083	.843	.798	.085	.848	.722	.062	.881	.842	.062	.827	.817	.101	.842	.725	.070
*CTMF	Tcyb17	.833	.792	.097	.849	.788	.085	.860	.723	.056	.853	.786	.087	.796	.781	.120	.863	.765	.055
*DF	TIP17	.730	.748	.145	.735	.744	.151	.769	.682	.099	.763	.761	.142	.685	.566	.130	.685	.566	.130
*CDCP	ICCV17	.687	.633	.159	.673	.618	.181	.724	.591	.114	.727	.680	.149	.658	.634	.199	.706	.583	.119
PiCAN	CVPR18	.832	.826	.080	.847	.806	.071	.834	.761	.053	.868	.835	.062	.761	.730	.134	.854	.797	.042
PAGRN	CVPR18	.831	.836	.079	.829	.827	.081	.844	.795	.051	.851	.856	.067	.779	.786	.117	.858	.834	.044
R[3]Net	IJCAI18	.819	.781	.113	.837	.775	.092	.798	.649	.101	.855	.800	.084	.797	.791	.141	.847	.728	.066
*PCA	CVPR18	.801	.760	.100	.877	.844	.059	.873	.794	.044	.880	.845	.061	.800	.794	.112	.845	.763	.049
*MMCI	PR19	.791	.753	.113	.859	.813	.079	.855	.729	.059	.856	.812	.080	.787	.779	.132	.847	.750	.064
*TANet	TIP19	.808	.779	.093	.878	.844	.061	.886	.795	.041	.877	.849	.060	.801	.794	.111	.858	.782	.045
*PDNet	ICME19	.799	.757	.112	**.883**	.832	.062	.835	.740	.064	.874	.833	.064	**.845**	.824	.109	.868	.800	.050
*CPFP	CVPR19	.749	.736	.099	—	—	—	.888	.822	**.036**	—	—	—	.828	.813	.088	.874	.819	.037
PoolNet	CVPR19	**.892**	.871	.049	.872	.850	.057	.867	.791	.046	**.898**	**.877**	**.045**	.826	**.830**	.094	.888	.852	.031
BASNet	CVPR19	**.900**	.881	**.042**	.872	.841	**.055**	**.890**	.838	**.036**	**.896**	.865	**.042**	.823	.825	**.086**	.889	**.861**	.030
CPD	CVPR19	.875	**.872**	.055	.862	**.853**	.059	.885	**.840**	.037	.885	**.880**	.046	.806	.808	.097	**.893**	**.860**	**.028**
EGNet	ICCV19	.872	.866	.059	.869	.846	.060	.867	.800	.047	.889	**.876**	.049	.818	.812	.101	.878	.831	.035
*DMRA	ICCV19	.888	**.883**	**.048**	**.886**	**.872**	**.051**	**.899**	**.855**	**.031**	.886	.868	.047	**.847**	**.849**	**.075**	**.901**	.857	**.029**
Ours	-	**.915**	**.915**	**.033**	**.899**	**.886**	**.043**	**.914**	**.875**	**.026**	**.893**	**.880**	**.044**	**.859**	**.867**	**.066**	**.905**	**.876**	**.025**

3.2. Decoder

For the m_{th} layer, we first adopt a backward dense connection to skip-connect features of all deeper layers. Considering that the m_{th} layer only learns the level-specific representations, we use deeper features to complement context information for the m_{th} layer. Then, we upsample the multi-layer features to the spatial resolution with 128×128, and concatenate them. The final results can be generated using a 1×1 convolution, which is defined as:

$$\tilde{F}_f^{[l]} = Conv(Cat(F_f^{[l]}, \sum_{i=l+1}^{n} Conv(up(F_f^{[i]})))), \quad (6)$$

where $F_f^{[l]}$ represents the fused features of the l_{th} level. $\tilde{F}_f^{[l]}$ represents the updated feature of the l_{th} level. $up(\cdot)$ is the upsample operation. n is the total number of levels ($n = 5$). We achieve the final results S_f^{pr} from $\tilde{F}_f^{[3]}$.

3.3. Compensation-Aware Loss

The proposed CIM can effectively enhance the extracted features from location and boundary details. However, for some hard samples, the extracted cross-modal compensation and the boundary details still remain unreliable.

Thus, we involve a tailor-made loss function to pay more attention to those hard samples. Specifically, we mine these samples from two aspects: (1) samples with challenging boundary information. (2) samples with unreliable depth information. First, we use the boundary predictions as the privileged information to mine challenging boundary regions of RGB images. After generating the boundary predictions, we use the following operation to generate a weight map of the challenging region w_b:

$$w_b = Max(p_k^{\max}(b_l^{gt}), p_k^{\max}(b_l^{pr})) - p_k^{\max}(b_l^{gt}) * p_k^{\max}(b_l^{pr}), \quad (7)$$

where $p_k^{\max}$ represents the max-pooling operation with kernel size k. The max-pooling is used to enlarge the coverage area of the boundary. We set $k = 8$ in our paper. $Max()$ means the max operation. b^{gt} is the ground truth of the saliency edge and b^{pr} is the predicted saliency edge.

For those unreliable depth samples, the depth value of saliency regions is similar to that of the background. Thus, we calculate the average depth value of the saliency region D_s, and non-saliency region D_{ns}. The sample weight is defined as: $w_{sample} = 1 - |D_s - D_{ns}|$. Then, we weight depth samples by region to further evaluate those samples. The region weight map is defined as follows:

$$w_d = \sum_{i=1}^{m_f} d_i(1 - (\frac{\sum\limits_{k=1}^{N} (d_i^k g_s^k)}{\sum\limits_{k=1}^{N} d_i^k})), \quad (8)$$

where m_f represents a collection of the binary mask maps

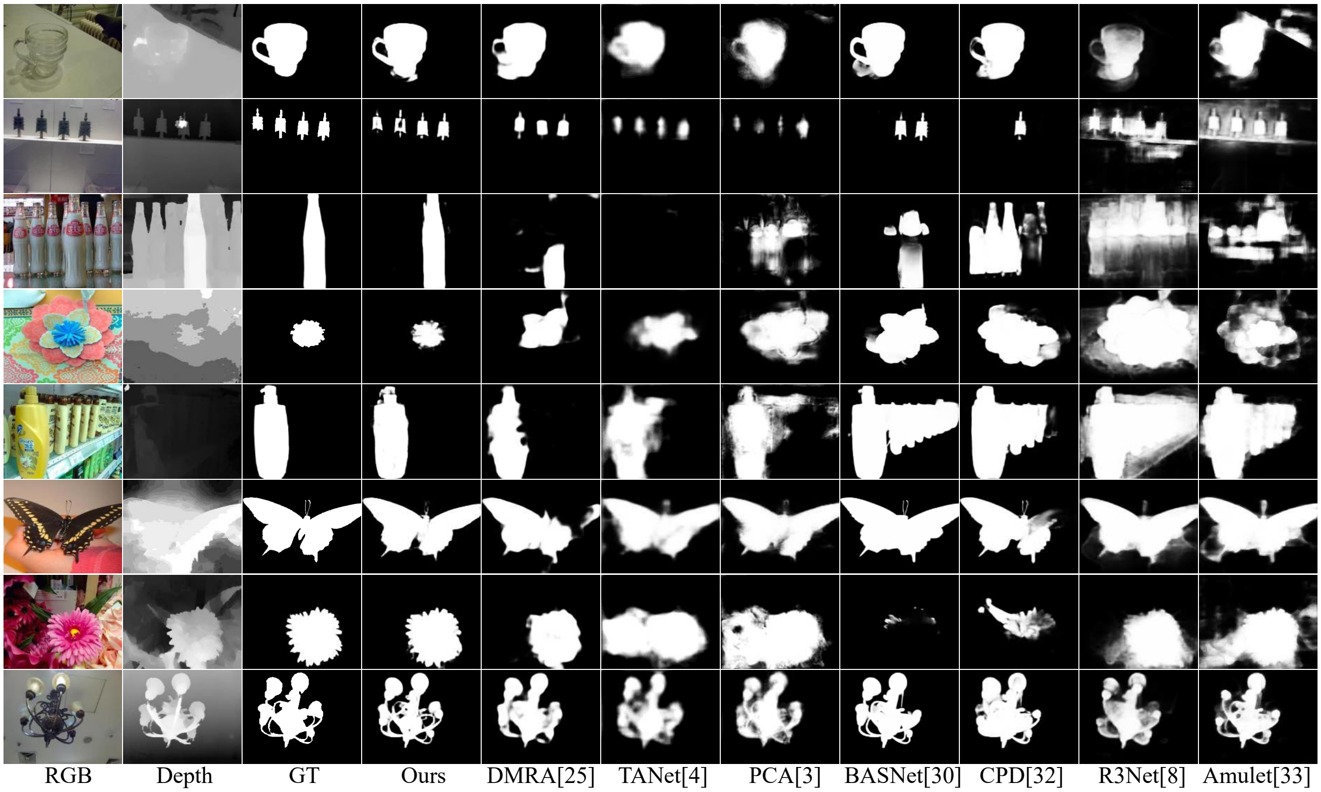

Figure 6. Visual comparisons of the proposed method and the state-of-the-art algorithms.

(d_i) that contain the saliency region. N is the total number of pixels. g_s is the saliency ground truth. We use the w_{sample}, w_d, w_s to work with the cross-entropy loss, and our compensation-aware loss can be given as:

$$L_{cl} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{c\in\{0,1\}} w^i(y(v_i)=c)(\log(\hat{y}(v_i)=c)), \quad (9)$$

where $w^i = \lambda_1 w_b^i + \lambda_2 w_d^i + \lambda_3 w_{sample}$. We set $\lambda_1 = 1$ $\lambda_2 = 1$ and $\lambda_3 = 0.5$. y represents the saliency ground truth. $\hat{y}$ is the saliency prediction. Our final loss that combines the BCE loss and the compensation-aware loss is given as:

$$L = l_{bce}(S_f^{pr}, g_s) + l_{bce}(S^{pr}, g_s) + l_{bce}(B^{pr}, g_b) + L_{cl}, \quad (10)$$

where l_{bce} represents the BCE loss. S_f^{pr} is the final prediction. S^{pr} is the initial saliency map generated by the RGB modal. B^{pr} is the boundary prediction. g_b is the boundary ground truth of salient objects, which is generated by the saliency ground truth through the Prewitt operator.

4. Experiments

4.1. Experimental Setup

Implementation details. We implement the proposed model based on the Pytorch toolbox with a Nvidia RTX 2080Ti GPU. The parameters of the backbone network are initialized by the VGG-16 [30]. Other convolutional parameters are randomly assigned. All the training and test images are resized to 256×256. The batch size is set as 20. The proposed model is trained by the Adam optimizer [21] with the initial learning rate of 1e-4 which is divided by 10 after 35 epochs. Our network is trained for 40 epochs in total.

Table 2. Ablation analyses on DUT-RGBD, LFSD and STEREO. Baseline denotes the baseline architecture shown in Figure 8. CAU and BSU are introduced in the Sec.3.1. closs represents the proposed compensation-aware loss introduced in the Sec.3.3.

Methods	DUT-RGBD			LFSD			STEREO		
	S_α ↑	F_β ↑	MAE↓	S_α ↑	F_β ↑	MAE↓	S_α ↑	F_β ↑	MAE↓
Baseline	.869	.876	.053	.775	.808	.103	.791	.810	.085
Baseline+CAU	.894	.897	.042	.824	.823	.084	.867	.859	.056
Baseline+BSU	.900	.904	.039	.823	.843	.082	.869	.870	.055
Baseline+CAU+BSU	.904	.904	.038	.845	.851	.074	.883	.874	.050
Baseline+CAU+BSU+closs	**.915**	**.915**	**.033**	**.859**	**.867**	**.066**	**.893**	**.880**	**.044**

Evaluation Metrics. We adopt 3 commonly used metrics, namely mean F-measure [2], mean absolute error (MAE) [4], and recently released structure measure (S-measure) [12], to evaluate the performance of each method.

4.2. Datasets

We conduct our experiments on six widely used RGB-D benchmark datasets. **DUT-RGBD** [27]: contains 1200

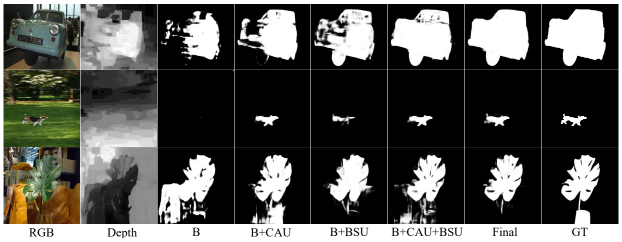

Figure 7. Visual comparisons of ablation analysis. The meaning of indexes has been explained in the caption of Table 2.

images captured by Lytro camera in real life scenes. **NJUD** [20]: includes 1985 RGB-D stereo images, in which the stereo images are collected from the Internet, 3D movies and photographs taken by a Fuji W3 stereo camera. **NLPR** [26]: contains 1000 image pairs captured by Kinect under different illumination conditions. **LFSD** [22]: contains 100 images captured by the Lytro camera. **STEREO** [25]: contains 797 stereoscopic images downloaded from the Internet. **RGBD-135** [8]: contains 135 indoor images collected by Microsoft Kinect. As the same splitting way in [27], we split 800 samples from DUT-RGBD 1485 samples from NJUD and 700 samples from NLPR for training. The remaining images in these three datasets and other three datasets are all for testing.

4.3. Comparison With the State-of-the-art

We compare our model with 18 salient object detection models including 9 latest CNNs-based RGB-D methods (remarked with $*$): *DMRA [27], *CPFP [37], *PDNet [39], *TANet [6], *MMCI [7], *PCA [5], *CDCP [40], *DF [28], *CTMF [17]; 9 top ranking CNNs-based RGB methods: EGNet [38], CPD [34], BASNet [32], PoolNet [23], R3Net [10], PAGRN [36], Amulet [35],PiCAN [32], DSS [19]. For fair comparisons, we use the released code and their default parameters to reproduce those methods. In terms of methods without released source code, we use their published results for comparisons.

Quantitative Evaluation. Table 1 shows the validation results in terms of three metrics including Mean F-measure, MAE and S-measure on six datasets. As can be seen in Table 1, our method significantly outperforms the existing methods, improving the MAE by 15.6% on the NJUD dataset. The improvement is consistently observed on other two metrics. Especially, benefitting from the proposed complimentary interaction module (CIM) and the helpful compensation-aware loss, ours results outperform all other methods on the STEREO and NLPR, where the scenes are considered to be relatively complicated.

Qualitative Evaluation. For a more intuitive view, we show some visualization results to exhibit the superiority of the proposed approach in Figure 6. The first 4 rows show the challenging scenes, including transparent object (Row 1), multiple objects (Row 2), low contrast scene (Row 3), and small object (Row 4). These results indicate that our network is capable of accurately capturing salient regions under these challenging situations. Furthermore, Row 5-6 demonstrate the superiority of our method when it comes to the unreliable depth maps. In these scenes, the existing RGB-D methods fail to detect the saliency part, misled by undesirable depth maps. On the other hand, our network can mine useful information to cope with these scenes by the proposed cross-modal attention unit (CAU). Additionally, we select two examples both with complex salient objects boundaries (Row 7-8) to show that our model not only locates the salient object but also segments objects with more accurate boundary details.

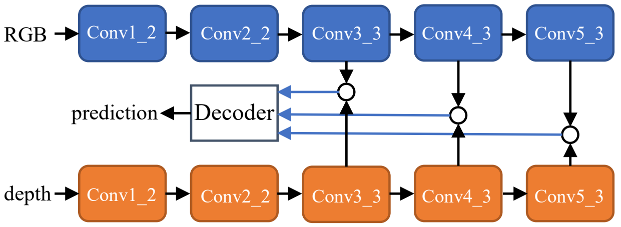

Figure 8. The baseline of our proposed network. ○ represents concatenating features from RGB and depth modalities.

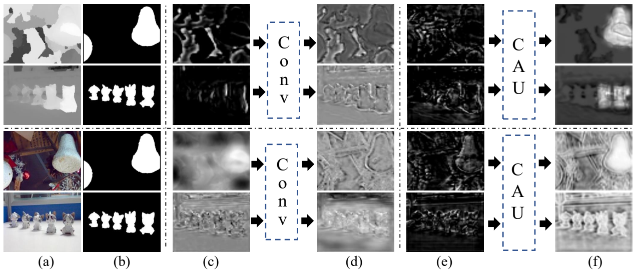

Figure 9. Visualizing feature maps around the CAU. The top two rows show the features from the depth modal, and the 3 and 4 rows show the features from the RGB modal. (a): RGB and depth inputs. (b): saliency ground truths. (c-f): visualization feature maps at different places. CAU: the proposed cross-modal attention unit. Conv: two convolution layers.

4.4. Ablation Studies

In this section, we perform ablation analysis to demonstrate the effect of each component on three testing datasets in Table 2. The baseline is the VGG16-based architecture shown in Figure 8.

Effect of the Cross-modal Attention Unit (CAU). To verify the effectiveness of the CAU, we analyse the performance of enabling our CAU as shown in Table 2. It is seen that our CAU improves the baseline across three datasets. Intuitively, we visualize the results before/after employing the CAU as shown in Figure 7. We observe that the predictions produced by our CAU can better locate the salient object. Furthermore, we visualize the feature maps before/after employing the CAU to validate its ability of se-

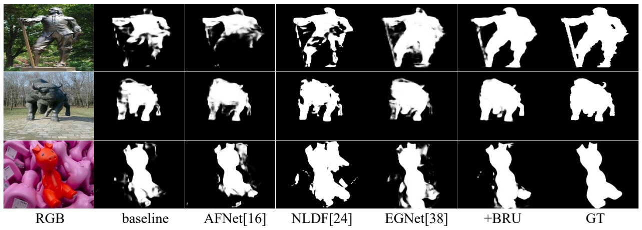

Figure 10. Visual comparisons of different methods for boundary refinement of the salient object.

Table 3. Ablation analyses on different mechanisms of using edge cues. Baseline+AFNet_edge, Baseline+NLDF_edge and Baseline+EGNet_edge are introduced in the Sec.4.4.

Methods	DUT-RGBD			LFSD			STEREO		
	S_α ↑	F_β ↑	MAE↓	S_α ↑	F_β ↑	MAE↓	S_α ↑	F_β ↑	MAE↓
Baseline	.869	.876	.053	.775	.808	.103	.791	.810	.085
Baseline+AFNet_edge	.873	.879	.051	.784	.814	.099	.796	.809	.083
Baseline+NLDF_edge	.875	.882	.050	.793	.815	.093	.815	.826	.072
Baseline+EGNet_edge	.882	.882	.048	.792	.823	.098	.806	.820	.081
Baseline+BSU	**.900**	**.904**	**.039**	**.823**	**.843**	**.082**	**.869**	**.870**	**.055**

lecting useful information, shown in Figure 9. Evidently, feature maps after the CAU show more precisely extracted the location information of salient objects (Column f), compared to those after two convolution layers (Column d).

Effect of the Boundary Supplement Unit (BSU). We compare our BSU with some other designs using edge information [24, 16, 38] to evaluate its effectiveness. The results are shown in Table 3. NLDF_edge: we add the same IoU loss as [24] to the baseline for minimizing the error of edges. AFNet_edge: we add the same CE loss as [16] to the baseline. EGNet_edge: based on [38], we adjust our RGB stream into the EGNet fashion, and extract local boundary features from the Conv2-2. Edge supplement is added to the saliency features through the O2OGM. As shown in Table 3, considerable performance boosts are achieved on three datasets. These improvements are logical since our BSU extract purer boundary details from the high levels, as shown in Figure 5. Meanwhile, the visual effects of enabling our BSU in Figure 10 illustrate the ability of capturing the edge of the salient objects.

Effect of the Compensation-Aware Loss. We verify the strength of the compensation-aware loss by employing the loss to the baseline+CAU+BSU, as shown in Table 2. It is seen that our compensation-aware loss improves the baseline+CAU+BSU across three datasets. Furthermore, we visualize the predictions before/after adding the compensation-aware loss to demonstrate its ability. As shown in Figure 11(a), for challenging samples which are hard to supplement boundary details, our loss helps our network pay more attention to hard pixels in the training stage for more accurate predictions. Moreover, the proposed loss can also help samples with unreliable depth maps to generate useful knowledge from the RGB images, shown in Figure 11(b).

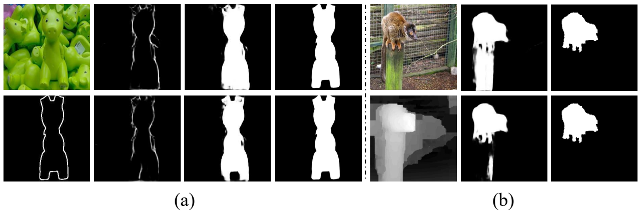

Figure 11. Visual comparisons of the results with/without the compensation-aware loss. Row 2 and Row 1 show the boundary and saliency predictions with/without the compensation-aware loss, respectively. Column 2 show the edge predictions. Column 3 and 6 show the saliency predictions. Column 4 and 7 represent the saliency ground truths.

Table 4. The effect of the number of binary maps (m).

Methods	DUT-RGBD			LFSD			STEREO		
	S_α ↑	F_β ↑	MAE↓	S_α ↑	F_β ↑	MAE↓	S_α ↑	F_β ↑	MAE↓
m=2	.898	.898	.040	.810	.841	.088	.835	.842	.069
m=5	.911	.914	.035	.853	.866	.071	.869	.869	.055
m=10	.915	.915	.033	.859	.867	.066	.893	.880	.044
m=20	.910	.917	.035	.842	.853	.070	.890	.882	.045

Hyperparameters Setting. m represents the number of the binary masks in our CAU. We increase m from 2 to 20 and measure the corresponding scores, shown in Table 4. As m increases, the depth map is divided more precisely for helping select accurate cross-modal information. However, when m is greater than 10, the accuracy gain is not significant, but with more computation costs. In our experiment, m is set to 10.

5. Conclusion

In this paper, we strive to embrace challenges towards accurate RGB-D saliency detection. We propose a new framework for accurate RGB-D saliency detection taking account of local and global complementarities from two modalities. It includes a complimentary interaction model, which consists of a cross-modal attention unit and a boundary supplement unit to capture effective features for salient object location and boundary detail refinement. Moreover, we propose a compensation-aware loss to improve the networks confidence in detecting hard samples. Experimental results demonstrate that the proposed method achieves state-of-the-art performance on 6 public saliency benchmarks.

Acknowledgements. This work was supported by the Science and Technology Innovation Foundation of Dalian (2019J12GX034), the National Natural Science Foundation of China (61976035, 61725202, U1903215, U1708263, 61829102, 91538201 and 61751212), and the Fundamental Research Funds for the Central Universities (DUT19JC58).

References

[1] An effective graph and depth layer based rgb-d image foreground object extraction method. *Computational visual media*, (4):85–91.

[2] Radhakrishna Achanta, Sheila Hemami, Francisco Estrada, and Sabine Süsstrunk. Frequency-tuned salient region detection. In *IEEE International Conference on Computer Vision and Pattern Recognition (CVPR 2009)*, number CONF, pages 1597–1604, 2009.

[3] Ali Borji, Ming Ming Cheng, Qibin Hou, Huaizu Jiang, and Jia Li. Salient object detection: A survey. *Eprint Arxiv*, 16(7):3118, 2014.

[4] Ali Borji, Ming-Ming Cheng, Huaizu Jiang, and Jia Li. Salient object detection: A benchmark. *IEEE transactions on image processing*, 24(12):5706–5722, 2015.

[5] Hao Chen and Youfu Li. Progressively complementarity-aware fusion network for rgb-d salient object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3051–3060, 2018.

[6] Hao Chen and Youfu Li. Three-stream attention-aware network for rgb-d salient object detection. *IEEE Transactions on Image Processing*, PP(99):1–1, 2019.

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

[8] Yupeng Cheng, Huazhu Fu, Xingxing Wei, Jiangjian Xiao, and Xiaochun Cao. Depth enhanced saliency detection method. In *Proceedings of international conference on internet multimedia computing and service*, page 23. ACM, 2014.

[9] Abhishek Das, Harsh Agrawal, Larry Zitnick, Devi Parikh, and Dhruv Batra. Human attention in visual question answering: Do humans and deep networks look at the same regions? *Computer Vision and Image Understanding*, 163:90–100, 2017.

[10] Zijun Deng, Xiaowei Hu, Lei Zhu, Xuemiao Xu, Jing Qin, Guoqiang Han, and Pheng-Ann Heng. R3net: Recurrent residual refinement network for saliency detection. In *Proceedings of the 27th International Joint Conference on Artificial Intelligence*, pages 684–690. AAAI Press, 2018.

[11] Michael Donoser, Martin Urschler, Martin Hirzer, and Horst Bischof. Saliency driven total variation segmentation. In *2009 IEEE 12th International Conference on Computer Vision*, pages 817–824. IEEE, 2009.

[12] Deng-Ping Fan, Ming-Ming Cheng, Yun Liu, Tao Li, and Ali Borji. Structure-measure: A new way to evaluate foreground maps. In *Proceedings of the IEEE international conference on computer vision*, pages 4548–4557, 2017.

[13] Xingxing Fan, Zhi Liu, and Guangling Sun. Salient region detection for stereoscopic images. In *2014 19th International Conference on Digital Signal Processing*, pages 454–458. IEEE, 2014.

[14] Yuming Fang, Junle Wang, Manish Narwaria, Patrick Le Callet, and Weisi Lin. Saliency detection for stereoscopic images. *IEEE Transactions on Image Processing*, 23(6):2625–2636, 2014.

[15] David Feng, Nick Barnes, Shaodi You, and Chris McCarthy. Local background enclosure for rgb-d salient object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2343–2350, 2016.

[16] Mengyang Feng, Huchuan Lu, and Errui Ding. Attentive feedback network for boundary-aware salient object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1623–1632, 2019.

[17] Junwei Han, Hao Chen, Nian Liu, Chenggang Yan, and Xuelong Li. Cnns-based rgb-d saliency detection via cross-view transfer and multiview fusion. *IEEE transactions on cybernetics*, 48(11):3171–3183, 2017.

[18] Seunghoon Hong, Tackgeun You, Suha Kwak, and Bohyung Han. Online tracking by learning discriminative saliency map with convolutional neural network. *Computer Science*, pages 597–606, 2015.

[19] Qibin Hou, Ming-Ming Cheng, Xiaowei Hu, Ali Borji, Zhuowen Tu, and Philip HS Torr. Deeply supervised salient object detection with short connections. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3203–3212, 2017.

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

[21] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *Computer Science*, 2014.

[22] Nianyi Li, Jinwei Ye, Yu Ji, Haibin Ling, and Jingyi Yu. Saliency detection on light field. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2806–2813, 2014.

[23] Jiang-Jiang Liu, Qibin Hou, Ming-Ming Cheng, Jiashi Feng, and Jianmin Jiang. A simple pooling-based design for real-time salient object detection. *arXiv preprint arXiv:1904.09569*, 2019.

[24] Zhiming Luo, Akshaya Mishra, Andrew Achkar, Justin Eichel, Shaozi Li, and Pierre-Marc Jodoin. Non-local deep features for salient object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6609–6617, 2017.

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

[26] Houwen Peng, Bing Li, Weihua Xiong, Weiming Hu, and Rongrong Ji. Rgbd salient object detection: A benchmark and algorithms. In *European conference on computer vision*, pages 92–109. Springer, 2014.

[27] Yongri Piao, Wei Ji, Jingjing Li, Miao Zhang, and Huchuan Lu. Depth-induced multi-scale recurrent attention network for saliency detection. In *ICCV*, 2019.

[28] Liangqiong Qu, Shengfeng He, Jiawei Zhang, Jiandong Tian, Yandong Tang, and Qingxiong Yang. Rgbd salient object detection via deep fusion. *IEEE Transactions on Image Processing*, 26(5):2274–2285, 2017.

[29] Riku Shigematsu, David Feng, Shaodi You, and Nick Barnes. Learning rgb-d salient object detection using background enclosure, depth contrast, and top-down features.

In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2749–2757, 2017.

[30] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *Computer Science*, 2014.

[31] Hangke Song, Zhi Liu, Huan Du, Guangling Sun, Olivier Le Meur, and Tongwei Ren. Depth-aware salient object detection and segmentation via multiscale discriminative saliency fusion and bootstrap learning. *IEEE Transactions on Image Processing*, 26(9):4204–4216, 2017.

[32] Jinming Su, Jia Li, Changqun Xia, and Yonghong Tian. Selectivity or invariance: Boundary-aware salient object detection. *arXiv preprint arXiv:1812.10066*, 2018.

[33] Guijin Wang, Cairong Zhang, Xinghao Chen, Xiangyang Ji, Jing-Hao Xue, and Hang Wang. Bi-stream pose guided region ensemble network for fingertip localization from stereo images. *arXiv preprint arXiv:1902.09795*, 2019.

[34] Zhe Wu, Li Su, and Qingming Huang. Cascaded partial decoder for fast and accurate salient object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3907–3916, 2019.

[35] Pingping Zhang, Dong Wang, Huchuan Lu, Hongyu Wang, and Xiang Ruan. Amulet: Aggregating multi-level convolutional features for salient object detection. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 202–211, 2017.

[36] Xiaoning Zhang, Tiantian Wang, Jinqing Qi, Huchuan Lu, and Gang Wang. Progressive attention guided recurrent network for salient object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 714–722, 2018.

[37] Jia-Xing Zhao, Yang Cao, Deng-Ping Fan, Ming-Ming Cheng, Xuan-Yi Li, and Le Zhang. Contrast prior and fluid pyramid integration for rgbd salient object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[38] Jia-Xing Zhao, Jiangjiang Liu, Den-Ping Fan, Yang Cao, Jufeng Yang, and Ming-Ming Cheng. Egnet: Edge guidance network for salient object detection. *arXiv preprint arXiv:1908.08297*, 2019.

[39] Chunbiao Zhu, Xing Cai, Kan Huang, Thomas H Li, and Ge Li. Pdnet: Prior-model guided depth-enhanced network for salient object detection. 2018.

[40] Chunbiao Zhu, Ge Li, Wenmin Wang, and Ronggang Wang. An innovative salient object detection using center-dark channel prior. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1509–1515, 2017.