Asymmetric Two-Stream Architecture for Accurate RGB-D Saliency Detection

Miao Zhang[1], Sun Xiao Fei[1⋆], Jie Liu[1⋆], Shuang Xu[1], Yongri Piao[1✉], and Huchuan Lu[1,2]

[1] Dalian University of Technology, Dalian, China
[2] Pengcheng Lab, Shenzhen, China
{xiaofeisun, 1605721375, sxu1997}@mail.dlut.edu.cn,
{miaozhang, yrpiao, lhchuan}@dlut.edu.cn
https://github.com/OIPLab-DUT/ATSA

Abstract. Most existing RGB-D saliency detection methods adopt symmetric two-stream architectures for learning discriminative RGB and depth representations. In fact, there is another level of ambiguity that is often overlooked: if RGB and depth data are necessary to fit into the same network. In this paper, we propose an asymmetric two-stream architecture taking account of the inherent differences between RGB and depth data for saliency detection. First, we design a flow ladder module (FLM) for the RGB stream to fully extract global and local information while maintaining the saliency details. This is achieved by constructing four detail-transfer branches, each of which preserves the detail information and receives global location information from representations of other vertical parallel branches in an evolutionary way. Second, we propose a novel depth attention module (DAM) to ensure depth features with high discriminative power in location and spatial structure being effectively utilized when combined with RGB features in challenging scenes. The depth features can also discriminatively guide the RGB features via our proposed DAM to precisely locate the salient objects. Extensive experiments demonstrate that our method achieves superior performance over 13 state-of-the-art RGB-D approaches on the 7 datasets. Our code will be publicly available.

Keywords: Saliency detection · Flow ladder · Depth attention

1 Introduction

Salient object detection, which involves identifying the visually interesting regions, is a well-researched domain of computer vision. It serves as an essential pre-processing step for various visual tasks such as image retrieval [7,15,17,28], visual tracking [2,20,38], object segmentation [12,39,40,43,42], object recognition [10,36,37], and therefore makes an important contribution towards sustainable development.

⋆ denotes equal contributions.

A majority of existing works [21,26] for saliency detection focus on operating RGB images. While RGB-based saliency detection methods have achieved great success, appearance features in RGB data are less predictive to some challenging scenes, such as multiple or transparent objects, similar foreground and background, complex background, low-intensity environment, etc.

The depth cue has the preponderance of discriminative power in location and spatial structure, which has been proved beneficial to accurate saliency prediction [35]. Moreover, the paired depth data for RGB natural images are widely available with the advent of depth sensors, e.g., Kinect and Lytro Illum. Consequently, using depth information gains growing interests in saliency detection.

Most RGB-D-based methods utilize symmetric two-stream architectures for extracting RGB and depth features [4,6,18,32]. However, we observe that while RGB data contain more information such as color, texture, contour, as well as limited location, grayscale depth data provide more information such as spatial structure and 3D layout. In consequence, a symmetric two-stream network may overlook the inherent differences of RGB and depth data. Asymmetric architectures have been adopted in few works to extract RGB and depth features, taking the differences between two modalities into account. Zhu *et al.* [48] present an architecture composed of a master network for processing RGB values, and a sub-network making full use of depth cues, which incorporates depth-based features into the master network via direct concatenation. Zhao *et al.* [46] incorporate the contrast prior to enhance the depth maps and then integrate them into the RGB stream for saliency detection. However, simple fusion strategies like direct concatenation or summation are less adaptive to locate the salient objects due to myriad possibilities of salient objects positions in the real world. Overall, these above methods overlook the fact that depth cue contributes differently to the salient object prediction in various scenes. Furthermore, existing RGB-D methods inevitably suffer from detail information loss [41,16] for adopting strides and pooling operations in the RGB and depth streams. An intuitive solution is to use skip-connections [22] or short-connections [21] for reconstructing the detail information. Although these strategies have brought satisfactory improvements, they remain restrictive to predict the complete structures with fine details.

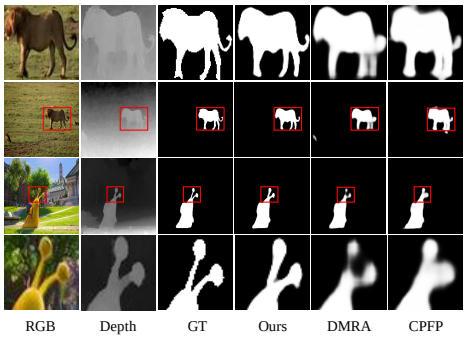

Fig. 1. The comparison of predicted maps between our method and two top-ranking RGB-D-based methods on salient objects details, i.e., DMRA [32], CPFP [46]. The 1^{st} row and the 4^{th} row are the enlarged images of the red box area of the middle two rows, which show superior performance of our method on saliency details

Building on the above observation, we strive to take a further step towards the goal of accurate saliency detection with an asymmetric two-stream model. **The primary challenge** towards this goal is how to effectively extract rich global context information while preserving local saliency details. **The second challenge** is how to effectively utilize the discriminative power of depth features to guide the RGB features for locating salient objects accurately.

To confront these challenges, we propose an asymmetric two-stream architecture as illustrated in Fig. 2. Concretely, our contributions are:

• We design a flow ladder module (FLM) and a lightweight depth network (DepthNet) with a small model size of 6.7MB. Instead of adopting skip-connections or short-connections, our FLM can effectively extract local detail information (see Fig. 1) and global context information through a local-global evolutionary fusion flow for accurate saliency detection.

• We propose a novel depth attention module (DAM) to ensure that the depth features can effectively guide the RGB features by using the discriminative power of depth cues. Its effectiveness has been experimentally verified (see Table 4).

• Furthermore, we conduct extensive experiments on 7 datasets and demonstrate that our method achieves consistently superior performance over 13 state-of-the-art RGB-D approaches in terms of 4 evaluation metrics. Numerically, our approach reduces the MAE performance by nearly 33% on DUT-RGBD dataset. In addition, our method minimizes the model size by 33% compared with the existing minimum method (PDNet) and achieves Top-2 running speed of 46 FPS.

2 Related work

RGB-D saliency detection. Although many RGB-based saliency detection methods have achieved appealing performance [16,29,33,44,45,47], they may not accurately detect the salient area because the appearance features in RGB data are less predictive when encountering with some complex scenes, such as low-contrast scenes, transparent objects, foreground sharing similar contents with background, multiple objects, and complex backgrounds. With the advent of consumer-grade depth cameras such as Kinect cameras, light field cameras and lidars, depth cues with a wealth of geometric and structural information is widely used in salient object detection (SOD).

Existing RGB-D saliency detection methods can be generally classified into two categories: **Traditional methods** [49,50,9,31,8,35,24,11]. Ren *et al.* [35] propose a two-stage RGB-D saliency detection framework using the validity of global priors. Lang *et al.* [24] introduce the depth prior into the saliency detection model to improve detection performance. Desingh *et al.* [11] use a non-linear regression to combine the RGB-D saliency detection model with the RGB model to measure the saliency values. **CNNs-based methods** [46,32,6,5,48,4,18,34]. To better mine salient information in challenging scenes, some CNNs-based methods combine depth information with RGB information for more accurate results. Practices and theories that lead to symmetric two-stream architectures which extract RGB and depth representations equally have been studied for a long

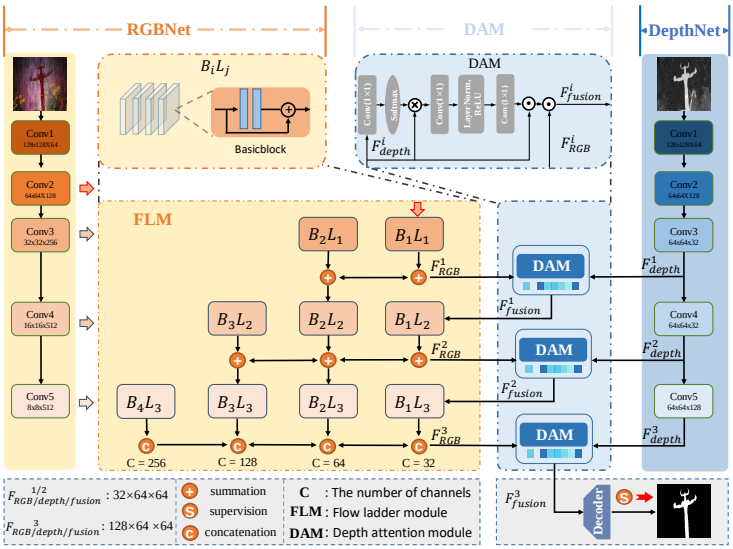

Fig. 2. The overall architecture of our proposed approach. Our asymmetric architecture consists of three parts, i.e., the RGBNet, the DAM and the lightweight DepthNet. The RGBNet includes a VGG-19 backbone and a flow ladder module. For depth stream, we also employ the same backbone of the RGBNet. The black arrows represent the information flows

time [32,18,6,5,4]. Han *et al.* [18] design a symmetric architecture for fusing the deep representations of depth and RGB views automatically to obtain the final saliency map. Chen *et al.* [6] utilize two-stream CNNs-based models for introducing cross-model interactions in multiple layers by direct summation. Recently, several asymmetric architectures are proposed for processing different data types [46,48]. Zhao *et al.* [46] use the enhanced depth information as an auxiliary cue and adopt a pyramid decoding structure to obtain more accurate salient regions.

Because of the inherent differences between RGB and depth information, classic symmetric two-stream architectures and simple fusion strategies may work their ways down to inaccurate prediction. Besides, the strides and pooling operations adopted in existing RGB-D-based methods for downsampling inevitably result in information loss. To address the above-mentioned issues, in this work, we design an asymmetric network and ably fuse RGB and depth information by a depth attention mechanism for precise saliency detection.

3 The proposed method

The overall architecture of our proposed method is shown in Fig. 2. In this section, we begin with describing the overall architecture in Section 3.1, then introduce the DepthNet in Section 3.2, the flow ladder module in Section 3.3, and finally the proposed depth attention module in Section 3.4.

Table 1. Details of our DepthNet architecture, k represents the kernel, s represents the stride, chns represents the number of input/output channels for each layer, p represents the padding, in and out represent the input and output feature size

Name	Layer	k	s	p	chns	in	out
Conv1	$*1$	3	1	1	3/64	256*256	256*256
	Maxpool	-	2	-	64/64	256*256	128*128
Conv2	$*1$	3	1	1	64/128	128*128	128*128
	Maxpool	-	2	-	128/128	128*128	64*64
	Transition1	3	1	1	128/32	64*64	64*64
Conv3	$*4$	3	1	1	32/32	64*64	64*64
		3	1	1	32/32	64*64	64*64
Conv4	$*4$	3	1	1	32/32	64*64	64*64
		3	1	1	32/32	64*64	64*64
	Transition2	3	1	1	32/128	64*64	64*64
Conv5	$*4$	3	1	1	128/128	64*64	64*64
		3	1	1	128/128	64*64	64*64

3.1 The overall architecture

Considering that most RGB-D-based methods utilizing symmetric two-stream architectures overlook the inherent differences between RGB and depth data, we propose an asymmetric two-stream architecture, as illustrated in Fig. 2. Our two-stream architecture includes a lightweight depth stream and a RGB stream with a flow ladder module, namely DepthNet and RGBNet, respectively. As for the depth stream, we design a lightweight architecture as shown in Table 1. Then the extracted depth features are fed into the RGB stream through a depth attention mechanism (DAM, see Fig. 3) to generate complementary features with affluent information of location and spatial structure. For the RGB stream, we adopt the commonly used architecture VGG-19 as our baseline. Based on this baseline, we propose a novel flow ladder module (FLM) to preserve the detail information as well as receive global location information from representations of other vertical parallel branches in an evolutionary way, which benefits locating salient regions and achieves considerable performance gains.

3.2 DepthNet

Compared with RGB data which contains richer color and texture information, depth cues focus on spatial location information. A large number of parameters in a complex depth extraction network are redundant, thus we consider that is unnecessary to process depth data with a complex network as large as RGBNet. In addition, the ablation experiments on symmetric and asymmetric architectures in Section 4.3 also confirm our claim. As illustrated in Fig. 2, we adopt a

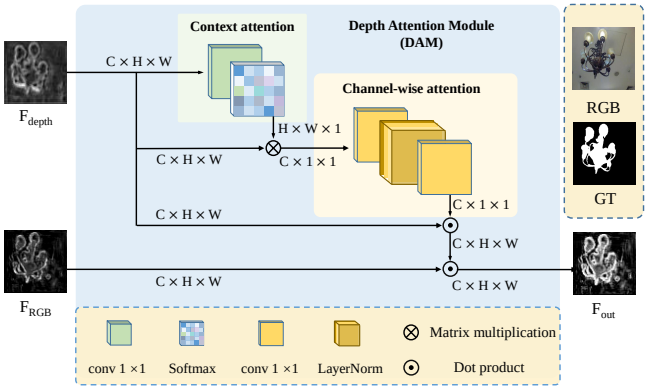

Fig. 3. Illustration of the depth attention module (DAM). The images above F_{out} are the corresponding original RGB image and ground truth

detail-transfer architecture for the depth stream (see Table 1 for detailed specification) and take the original depth maps as input. Our DepthNet transfers detail information in the whole architecture to capture fine spatial details. Considering the differences between RGB and depth data, numerous redundant channels of depth features are unnecessary. Therefore, we prune the number of feature channels to 32 in Conv3, 4 and 128 in final Conv, which further achieves a more lightweight DepthNet with a model size of 6.7MB.

3.3 RGBNet

Deeper networks are able to extract richer high-level information such as location and semantic information, but strides and pooling operations widely used in existing RGB-D-based methods may cause detail information loss, such as boundary, small object, for saliency detection. A straightforward solution to this issue is combining the low-level features with the high-level features by skip-connections [22]. However, the low-level features take a less discriminative and predictive power for complex scenes, which has trouble contributing to accurate saliency detection. Hence, we design a novel RGBNet consisting of a VGG backbone for fair comparison and a flow ladder model to preserve the local detail information by constructing four detail-transfer branches and fuse the global location information in an evolutionary way. In order to fit our task, we truncate the last three fully-connected layers and maintain the five convolution blocks as well as all pooling operations of VGG-19. The FLM can preserve the resolution of representations in multiple scales and levels, ensuring that the local detail information and global location information contribute to the precision of saliency detection. More details are described as follows.

In order to alleviate the detail information loss, we design a flow ladder model (FLM). This module is applied in VGG-19 and integrates four detail-transfer branches in the way of local-global evolutionary fusion flow. We design

detail-transfer branches for preserving the saliency details. As shown in Fig. 2, the first two branches consist of 3 layers. The number of the layers in the 3^{rd} and 4^{th} branch is decreased to 2, 1, respectively. Specifically, we simply denote the j^{th} layer of the i^{th} branch as B_iL_j, $i\epsilon[1,4], j\epsilon[1,3]$. B_iL_j is composed of four *basicblocks* [19], each of which consists of two convolutional layers as shown in the top of Fig. 2. Our FLM consists of 4 evolved detail-transfer branches. Instead of adopting strides and pooling operations, our FLM preserves the resolution of representations with more details in each branch by employing convolutional operations with 1*1 stride.

We design a novel **local-global evolutionary fusion flow** for integrating multi-scale local and global features extracted from detail-transfer branches. Each branch receives rich information from other vertical parallel representations through our local-global evolutionary fusion flow. In this way, rich global context representations are generated while more local saliency details are preserved. Specifically, the representations of the deeper branches are fused into the shallower branches by upsampling and summation operations as well as the representations of the shallower branches are fused into the deeper branches by downsampling and summation operations as shown in the FLM of Fig. 2. Through the evolution between different branches (shown in Fig. 2), the local detail information and the global context information are effectively combined, which benefits the precision of saliency detection. The whole fusion process is described as the following equations:

$$B_iL_j = \begin{cases} trans(Conv2) & i=1, j=1 \\ trans(Conv(i+1)) & i=j+1, j\epsilon[1,3] \\ \sum_{n=1}^{j} f(B_nL_{j-1}) & i\epsilon[1,j], j\epsilon[2,3] \end{cases}, \tag{1}$$

$$F_{RGB}^{j} = \sum_{n=1}^{j+1} f(B_nL_j) \quad j\epsilon[1,2], \tag{2}$$

$$F_{RGB}^{3} = cat(f(B_nL_3)) \quad n\epsilon[1,4], \tag{3}$$

where B_i and L_j denote the i^{th} branch and j^{th} layer, respectively. $f(\cdot)$ denotes $n-i$ times up-sampled when $n>i$ and $i-n$ times down-sampled when $n<i$. And when n is equal to i, $f(\cdot)$ means no operation. $Conv(i)$ means the output features of the i^{th} Conv block in VGG-19 and $trans(\cdot)$ is operated by a convolutional layer to realize the transformation of the number of channels. $cat(\cdot)$ denotes concatenating all features together. The final output of our LFM namely F_{RGB}^{3} is a concatenation of multi-scale features extracted from four branches. In conclusion, the features with local and global information are transferred to the parallel branches in an evolutionary way. Our proposed LFM can not only alleviate the object detail information loss but also effectively integrate multi-scale and multi-level features for precise saliency prediction.

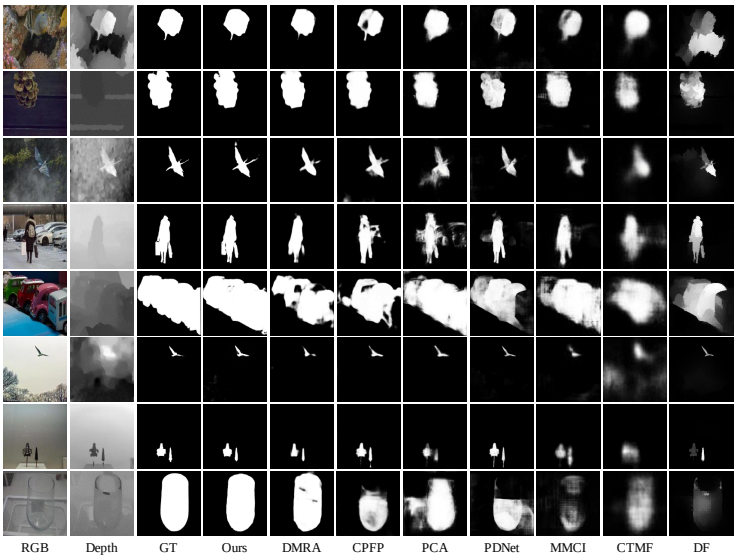

Fig. 4. Comparisons of ours with state-of-the-art CNNs-based methods. Those methods are top ranking ones in quantitative evaluation. Obviously, our results are more consistent with the ground truths (GT), especially in complex scenes

3.4 Depth Attention Module

Changes in statistics of object positions in the real world makes linear fusion strategies of RGB and depth data less adaptive to complex scenes. To take full advantage of the depth cues with the discriminative power in location and spatial structure, we design a depth attention module to adaptively fuse the RGB and depth representations as shown in Fig. 3. Firstly, the depth features contain abundant spatial and structural information. We utilize the context attention block which contains a 1*1 convolutional layer W_k and a softmax function for extracting the salient location cues more precisely, instead of applying simple fusion like summation or concatenation. Then a matrix multiplication operation is adopted to aggregate all location features together to generate attention weights of each channel i ($i.e.$, α_i) for capturing pixel-wise spatial dependencies. Moreover, the degree of response to the salient regions varies between features of different channels. Thus we adopt a channel-wise attention block which contains two 1*1 convolutional layers W_c and a LayerNorm function to capture the inter-dependencies between channels, and further achieves a weighted depth feature β. Then we adopt dot product operation to fuse β into the RGB stream, which helps guide the RGB features at pixel-level to distinguish the foreground and background thoroughly. Furthermore, the ablation experiments in Section 4.3 also verify the effectiveness of our DAM compared with simple fusion. And the visual results in Fig. 6 (b) also prove that the salient regions are emphasized through the attention mechanism.

The details of these three blocks can be formulated as the following equations:

$$\alpha_i = \sum_{j=1}^{N_p} \frac{e^{W_k F_d^j}}{\sum_{m=1}^{N_p} e^{W_k F_d^m}} F_d^j, \tag{4}$$

$$\beta_i = \varsigma(W_{c2} ReLU(LN(W_{c1}\alpha_i)) \odot F_d), \tag{5}$$

$$F_{fusion} = \varsigma(F_{RGB} \odot \beta), \tag{6}$$

where α_i denotes the weight of the i^{th} channel to obtain the global context features. F_d^j means the j^{th} position in depth feature F_{depth}. N_p is the number of positions in the depth feature map (e.g., $N_p = H \cdot W$). W_k, W_{c1} and W_{c2} denote 1*1 convolutional operations. *LN* denotes the *LayerNorm* operation after the convolution W_{c1}, and *ReLU* is an activate function. The $\varsigma(\cdot)$ and $\odot$ mean the sigmoid function and dot product operation, respectively. The β_i indicates the depth pixel-wise attention map of i^{th} channel of F_{RGB}. F_{RGB} and F_{fusion} represent the input RGB feature and the output feature of the DAM, respectively. The F_{fusion} can be calculated as a DAM output with much more effective depth-induced context-aware attention features. Furthermore, experiments in Section 4.3 show that our DAM is capable of fusing depth features discriminatively and filtering out features which are guided by depth cues in mistake.

As illustrated in Fig. 3, the inputs of our DAM are F_{RGB}^i and F_{Depth}^i extracted from our LFM and DepthNet, respectively, $i = 1, 2, 3$. At the end, a simple decoder is adopted for supervision. The decoder module contains two bilinear upsample functions, each of which is followed by 3 convolutional layers. The total loss L can be represented as:

$$L = l_f\{Decoder(F_{fusion}^3); gt\}, \tag{7}$$

where F_{fusion}^3 represents the output fusion feature of the third DAM and gt means the ground-truth map. The cross-entropy loss l_f can be computed as:

$$l_f\{\hat{y}; y\} = ylog\hat{y} + (1-y)log(1-\hat{y}), \tag{8}$$

where y and $\hat{y}$ denote the saliency ground-truth map and the predicted map, respectively.

4 Experiments

4.1 Dataset

We perform our experiments on 7 public RGB-D datasets for fair comparisons, i.e., NJUD [23], NLPR [31], RGBD135 [8], STEREO [30], LFSD [27], DUT-RGBD [32], SSD [25]. We split those datasets as [4,6,18] to guarantee fair comparisons. We randomly select 800 samples from DUT-RGBD, 1485 samples from NJUD and 700 samples from NLPR for training. The remaining images in these 3 datasets and other 4 datasets are all for testing to verify the generalization ability of saliency models. To prevent overfitting, we additionally augment the training set by flipping, cropping and rotating those images.

Table 2. Quantitative comparisons of E-measure (E_γ), S-measure (S_λ), F-measure (F_β) and MAE (M) on 7 widely-used RGB-D datasets. The best three results are shown in **boldface**, *red*, *green* fonts respectively. From top to bottom: the latest CNNs-based RGB-D methods and traditional RGB-D methods

			DUT-RGBD				NJUD				NLPR				SSD	
Mehtod	Years	Backbone	$E_\gamma\uparrow$	$S_\lambda\uparrow$	$F_\beta\uparrow$	$M\downarrow$	$E_\gamma\uparrow$	$S_\lambda\uparrow$	$F_\beta\uparrow$	$M\downarrow$	$E_\gamma\uparrow$	$S_\lambda\uparrow$	$F_\beta\uparrow$	$M\downarrow$	$E_\gamma\uparrow$	$S_\lambda\uparrow$
Ours	-	VGG-19	**0.948**	**0.918**	**0.920**	**0.032**	**0.921**	**0.901**	**0.893**	**0.040**	**0.945**	**0.907**	**0.876**	**0.028**	**0.901**	**0.860**
CPFP[46]	CVPR19	VGG-16	0.814	0.749	0.736	0.099	0.906	0.878	0.877	0.053	0.924	0.888	0.822	0.036	0.832	0.807
DMRA[32]	ICCV19	VGG-19	0.927	0.888	0.883	0.048	0.908	0.886	0.872	0.051	0.942	0.899	0.855	0.031	0.892	0.857
MMCI[6]	PR19	VGG-16	0.855	0.791	0.753	0.113	0.878	0.859	0.813	0.079	0.871	0.855	0.729	0.059	0.860	0.814
TANet[5]	TIP19	VGG-16	0.866	0.808	0.779	0.093	0.893	0.878	0.844	0.061	0.916	0.886	0.795	0.041	0.879	0.839
PDNet[48]	ICME19	VGG-16	0.861	0.799	0.757	0.112	0.890	0.883	0.832	0.062	0.876	0.835	0.740	0.064	0.813	0.802
PCA[4]	CVPR18	VGG-16	0.858	0.801	0.760	0.100	0.896	0.877	0.844	0.059	0.916	0.873	0.794	0.044	0.883	0.843
CTMF [18]	TCyb17	VGG-16	0.884	0.834	0.792	0.097	0.864	0.849	0.788	0.085	0.869	0.860	0.723	0.056	0.837	0.776
DF [34]	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	0.802	0.742
MB [49]	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	0.633	0.499
CDCP [50]	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	0.714	0.604
DCMC [9]	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	0.790	0.706
NLPR [31]	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	0.726	0.562
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	0.383	0.341

Table 3. Continuation of Table 2

			SSD		STEREO				LFSD				RGBD135			
Method	Years	Backbone	$F_\beta\uparrow$	$M\downarrow$	$E_\gamma\uparrow$	$S_\lambda\uparrow$	$F_\beta\uparrow$	$M\downarrow$	$E_\gamma\uparrow$	$S_\lambda\uparrow$	$F_\beta\uparrow$	$M\downarrow$	$E_\gamma\uparrow$	$S_\lambda\uparrow$	$F_\beta\uparrow$	$M\downarrow$
Ours	-	VGG-19	**0.827**	**0.050**	**0.921**	**0.897**	**0.884**	**0.039**	**0.905**	**0.865**	**0.862**	**0.064**	**0.952**	**0.907**	**0.885**	**0.024**
CPFP [46]	CVPR19	VGG-16	0.725	0.082	0.897	0.871	0.827	0.054	0.867	0.828	0.813	0.088	0.927	0.874	0.819	0.037
DMRA [32]	ICCV19	VGG-19	0.821	0.058	0.920	0.886	0.868	0.047	0.899	0.847	0.849	0.075	0.945	0.901	0.857	0.029
MMCI [6]	PR19	VGG-16	0.748	0.082	0.890	0.856	0.812	0.080	0.840	0.787	0.779	0.132	0.899	0.847	0.750	0.064
TANet [5]	TIP19	VGG-16	0.767	0.063	0.911	0.877	0.849	0.060	0.845	0.801	0.794	0.111	0.916	0.858	0.782	0.045
PDNet [48]	ICME19	VGG-16	0.716	0.115	0.903	0.874	0.833	0.064	0.872	0.845	0.824	0.109	0.915	0.868	0.800	0.050
PCA[4]	CVPR18	VGG-16	0.786	0.064	0.905	0.880	0.845	0.061	0.846	0.800	0.794	0.112	0.909	0.845	0.763	0.049
CTMF [18]	TCyb17	VGG-16	0.709	0.100	0.870	0.853	0.786	0.087	0.851	0.796	0.781	0.120	0.907	0.863	0.765	0.055
DF [34]	TIP17	-	0.709	0.151	0.844	0.763	0.761	0.142	0.841	0.786	0.810	0.142	0.801	0.685	0.566	0.130
MB [49]	CAIP17	-	0.414	0.219	0.693	0.579	0.572	0.178	0.631	0.538	0.543	0.218	0.798	0.661	0.588	0.102
CDCP [50]	ICCVW17	-	0.524	0.219	0.801	0.727	0.680	0.149	0.737	0.658	0.634	0.199	0.806	0.706	0.583	0.119
DCMC [9]	SPL16	-	0.551	0.200	0.838	0.745	0.761	0.150	0.842	0.754	0.815	0.155	0.674	0.470	0.228	0.194
NLPR [31]	ECCV14	-	0.073	0.500	0.781	0.567	0.716	0.179	0.742	0.558	0.708	0.211	0.850	0.577	0.857	0.097
DES [8]	ICIMCS14	-	0.684	0.168	0.451	0.473	0.223	0.417	0.475	0.440	0.228	0.415	0.786	0.627	0.689	0.289

4.2 Experimental setup

Evaluation Metrics. To comprehensively evaluate various methods, we adopt 4 evaluation metrics including F-measure (F_β) [1], mean absolute error (M) [3], S-measure (S_λ) [13], E-measure (E_γ) [14]. Specifically, the F-measure can evaluate the performance integrally. The M represents the average absolute difference between the saliency map and ground truth. The S-measure which is recently proposed can evaluate the structural similarities. The E-measure can jointly capture image level statistics and local pixel matching information.

Implementation details. Our method is implemented with pytorch toolbox and trained on a PC with GTX 2080Ti GPU and 16 GB memory. The input images are uniformly resized to 256*256. The momentum, weight decay, batch-size

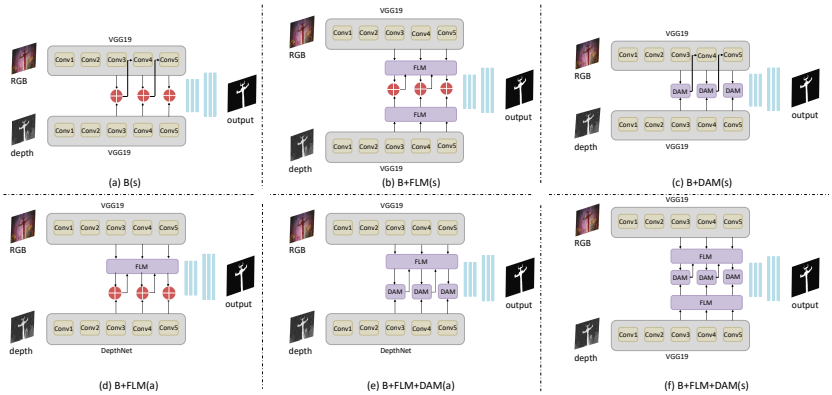

Fig. 5. Illustration of the six ablation experiments

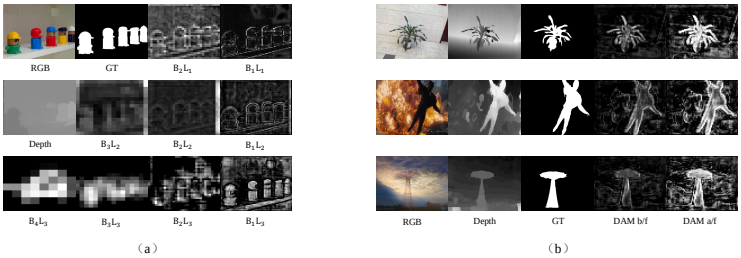

Fig. 6. (a) The visualization of the feature maps in FLM. The B_iL_j presents the output features of corresponding block in Fig. 2. (b) Visualization of the effectiveness of the DAM. The 4^{th} (DAM b/f) and the 5^{th} columns (DAM a/f) show the feature maps before and after adopting DAM, respectively

and learning rate of our network are set as 0.9, 0.0005, 2 and 1e-10, respectively. During training, we use softmax entropy loss described in Section 3.4 and the network converges after 60 epochs with mini-batch size 2.

4.3 Ablation analysis

Effect of FLM. We adopt the commonly two-stream VGG-19 network fused by direct summation as our baseline(denoted as 'B[s]' shown in Fig. 5 (a)). In order to verify the effectiveness of FLM, we employ the FLM in both the RGB and depth streams ('B+FLM[s]' shown in Fig. 5 (b)). The experimental results of (a) and (b) in Table 4 clearly demonstrate that our FLM obtains impressive performance gains. Moreover, as shown in Fig. 7, we can note that after employing FLM, the saliency maps achieve sharper boundaries as well as finer structures. Furthermore, for effectively analyzing the working mechanism of FLM module, we visualized the output features of each block in FLM. As shown in Fig. 6 (a), we can see that the branch 4 and branch 3 extract the

Table 4. Ablation analysis on 7 datasets. The [s] and [a] following the modules represent the symmetric and asymmetric architectures, respectively. Obviously, each component of our architecture can achieve considerable accuracy gains. (a), (b), (c), (d), (e), (f) represent the modules indexed by the corresponding letters in Fig. 5

Components	Index	Modules	DUT-RGBD		NJUD		NLPR		STEREO		LFSD		RGBD135		SSD	
			$F_\beta\uparrow$	$M\downarrow$	$F_\beta\uparrow$	$M\downarrow$	$F_\beta\uparrow$	$M\downarrow$	$F_\beta\uparrow$	$M\downarrow$	$F_\beta\uparrow$	$M\downarrow$	$F_\beta\uparrow$	$M\downarrow$	$F_\beta\uparrow$	$M\downarrow$
FLM	(a)	B[s]	0.822	0.068	0.810	0.071	0.766	0.050	0.816	0.068	0.812	0.088	0.794	0.044	0.749	0.080
	(b)	B+FLM[s]	0.911	0.035	0.893	0.040	0.872	0.029	0.881	0.041	0.858	0.064	0.882	0.025	0.836	0.048
DAM	(a)	B[s]	0.822	0.068	0.810	0.071	0.766	0.050	0.816	0.068	0.812	0.088	0.794	0.044	0.749	0.080
	(c)	B+DAM[s]	0.839	0.059	0.811	0.064	0.799	0.041	0.818	0.061	0.817	0.087	0.822	0.037	0.738	0.082
	(d)	B+FLM[a]	0.909	0.034	0.886	0.041	0.870	0.029	0.879	0.041	0.870	0.060	0.882	0.025	0.825	0.052
	(e)	B+FLM+DAM[a]	0.920	0.032	0.893	0.040	0.876	0.028	0.884	0.039	0.862	0.064	0.885	0.024	0.827	0.050
Asymmetric	(b)	B+FLM[s]	0.911	0.035	0.893	0.040	0.872	0.029	0.881	0.041	0.858	0.064	0.882	0.025	0.836	0.048
	(d)	B+FLM[a]	0.909	0.034	0.886	0.041	0.870	0.029	0.879	0.041	0.870	0.060	0.882	0.025	0.825	0.052
	(f)	B+FLM+DAM[s]	0.920	0.033	0.895	0.040	0.890	0.025	0.891	0.038	0.863	0.066	0.876	0.026	0.834	0.052
	(e)	B+FLM+DAM[a]	0.920	0.032	0.893	0.040	0.876	0.028	0.884	0.039	0.862	0.064	0.885	0.024	0.827	0.050

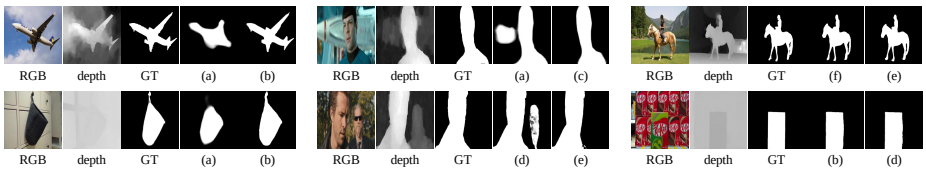

Fig. 7. Visual comparisons of ablation analyses. (a), (b), (c), (d), (e), (f) represent the visual results of the experiments indexed by the corresponding letters in Fig. 5

global location information while the branch 2 and branch 1 preserve more local detail information. This benefits from the evolutionary process of salient regions with finer details.

Effect of DAM. We conduct contrast experiments for verifying the effectiveness of our DAM on both symmetric and asymmetric architectures. In terms of symmetric architecture, we replace simple summation with DAM on our baseline (denoted as'B+DAM[s]', as shown in Fig. 5 (c)). From the results of (a) and (c) in Table 4 we can see that the MAE is reduced by 18% on NLPR dataset after employing DAM, which intuitively verifies the effect of DAM. Meanwhile, the corresponding visual results in Fig. 7 also illustrate that our DAM can fuse depth features discriminatively and filter out features which are guided by depth cues in mistake. On the other hand, we employ FLM in the RGB stream and replace the VGG-19 backbone with DepthNet in the depth stream (denoted as 'B+FLM[a]', as shown in Fig. 5 (d)). And we adopt DAM on 'B+FLM[a]' for verifying the effect of DAM on asymmetric architecture (denoted as 'B+FLM+DAM[a]' shown in Fig. 5 (e)). The comparison results of (d) and (e) in Table 4 demonstrate the effectiveness of DAM on asymmetric architecture over all datasets. Additional-

ly, we visualized the feature maps of our two-stream asymmetric architecture (before and after adopting DAM). As shown in Fig. 6 (b), we can see that the salient regions are emphasized after adopting DAM, which significantly improves our detection accuracy.

Effect of asymmetric architecture. In order to illustrate the effectiveness of adopting asymmetric architecture, we compare the results of (b) and (d) in Fig. 5. Furthermore, for fair comparison, we adopt our FLM and DAM on the two-stream symmetric network (denoted as 'B+FLM+DAM[s]', as shown in Fig. 5 (f)). As we can see from Table 4 (Asymmetric), the asymmetric architecture achieves considerable performance compared with symmetrical architecture, but has a small size. Specifically, the asymmetric architecture tremendously minimizes the model size by 47% (128.9MB vs. 244.4MB). Based on the above observation, we consider that is unnecessary to utilize large network as RGBNet for extracting features and we can replace it with a more lightweight network.

4.4 Comparison with state-of-the-art

Considering that most of the existing approaches are based on VGG network, we adopt VGG as our backbone for fair comparisons. And We compare our model with 13 RGB-D based salient object detection models including 8 CNNs-based methods: CPFP [46], DMRA [32], MMCI [6], TANet[5], PDNet [48], PCA [4], CTMF [18], DF [34], and 5 traditional methods: MB [49], CDCP [50], DCMC [9], NLPR [31], DES [8]. For fair comparisons, the results of the competing methods are generated by authorized codes or directly provided by authors.

Quantitative Evaluation. Table 2 and 3 show the validation results in terms of 4 evaluation metrics on 7 datasets. As we can see, our model achieves significant outperformance over all other methods. It is noted that our approach outperforms all other methods by a dramatic margin on datasets DUT-RGBD, NJUD and RGBD135, which are considered as more challenging datasets due to the large number of complex scenes like similar foreground and background, low-contrast and transparent object. It further indicates that our model can be generalized to various challenging scenes.

Qualitative Evaluation. We also visually compare our method with the most representative methods as shown in Fig. 4. From those results, we can observe that our saliency maps are closer to the ground truths. For instance, other methods have trouble in distinguishing salient objects in complex environments such as cluttered background (see the 1^{th} row), while ours can precisely identify the whole object and exquisite details. And our model can locate and detect the entire salient object with sharp details more accurately than others in more challenging scenes such as low-contrast (see the 2^{nd} - 3^{rd} rows), transparent object (see the 8^{th} row), multiple objects and small object (see the 5^{th} - 7^{th} rows). Those results further verify the effectiveness and robustness of our proposed model.

Table 5. Complexity comparisons on two datasets. The best three results are shown in **boldface**, red, green fonts respectively

			DUT-RGBD		NLPR	
Methods	$Size \downarrow$	$FPS \uparrow$	$F_\beta \uparrow$	$M \downarrow$	$F_\beta \uparrow$	$M \downarrow$
PCA	533.6MB	15	0.760	0.100	0.794	0.044
TANet	951.9MB	14	0.779	0.093	0.795	0.041
MMCI	929.7MB	19	0.753	0.113	0.729	0.059
PDNet	192MB	19	0.757	0.112	0.740	0.064
CPFP	278MB	6	0.736	0.099	0.822	0.036
CTMF	826MB	**50**	0.792	0.097	0.723	0.056
DMRA	238.8MB	22	0.883	0.048	0.855	0.031
Ours	**128.9MB**	46	**0.920**	**0.032**	**0.876**	**0.028**

Complexity Evaluation. We compare the model size and execution time of our method with other 7 representative models, as shown in Table 5. It can be seen that our method achieves Top-1 model size and Top-2 FPS. To be specific, the model size of our architecture is only 128.9MB which is 2/3 of the existing minimum model size (PDNet). Compared with the best performing method DMRA, our architecture tremendously minimizes the model size by 46% and boosts the FPS by 109%. Besides, we achieve a high running speed with 46 Frame Per Second (FPS) compared with the representative approaches.

5 Conclusion

In this paper, we propose an asymmetric two-stream architecture taking account of the inherent differences between RGB and depth data for saliency detection. For the RGB stream, we introduce a flow ladder module (FLM) for effectively extracting rich global context information while preserving local saliency details. And we design a lightweight DepthNet for depth stream with a small model size of 6.7MB. Besides, we propose a depth attention module (DAM) ensuring that the depth cues can discriminatively guide the RGB features for precisely locating salient objects. Our approach significantly advances the state-of-the-art methods over the widely used datasets and is capable of precisely capturing salient regions in challenging scenes.

Acknowledgement. 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).

References

1. Achanta, R., Hemami, S.S., Estrada, F.J., Ssstrunk, S.: Frequency-tuned salient region detection. In: CVPR. pp. 1597–1604 (2009) 4.2
2. Borji, A., Frintrop, S., Sihite, D.N., Itti, L.: Adaptive object tracking by learning background context. In: CVPR. pp. 23–30 (2012), https://academic.microsoft.com/paper/2158535435 1
3. Borji, A., Sihite, D.N., Itti, L.: Salient object detection: a benchmark. In: ECCV. pp. 414–429 (2012) 4.2
4. Chen, H., Li, Y.: Progressively complementarity-aware fusion network for rgb-d salient object detection. In: CVPR. pp. 3051–3060 (2018) 1, 2, 4.1, 2, 3, 4.4
5. Chen, H., Li, Y.: Three-stream attention-aware network for rgb-d salient object detection. TIP **28**(6), 2825–2835 (2019) 2, 2, 3, 4.4
6. Chen, H., Li, Y., Su, D.: Multi-modal fusion network with multi-scale multi-path and cross-modal interactions for rgb-d salient object detection. PR **86**, 376–385 (2019) 1, 2, 4.1, 2, 3, 4.4
7. Cheng, M.M., Hou, Q.B., Zhang, S.H., Rosin, P.L.: Intelligent visual media processing: When graphics meets vision. JCST **32**(1), 110–121 (2017), https://academic.microsoft.com/paper/2571295082 1
8. Cheng, Y., Fu, H., Wei, X., Xiao, J., Cao, X.: Depth enhanced saliency detection method. In: ICIMCS. pp. 23–27 (2014) 2, 4.1, 2, 3, 4.4
9. 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. SPL **23**(6), 819–823 (2016) 2, 2, 3, 4.4
10. Dai, J., Li, Y., He, K., Sun, J.: R-fcn: object detection via region-based fully convolutional networks. In: NIPS. pp. 379–387 (2016), https://academic.microsoft.com/paper/2407521645 1
11. Desingh, K., K, M.K., Rajan, D., Jawahar, C.V.: Depth really matters: Improving visual salient region detection with depth. In: BMVC (2013) 2
12. Donoser, M., Urschler, M., Hirzer, M., Bischof, H.: Saliency driven total variation segmentation. In: ICCV. pp. 817–824 (2009), https://academic.microsoft.com/paper/2546160422 1
13. Fan, D.P., Cheng, M.M., Liu, Y., Li, T., Borji, A.: Structure-measure: A new way to evaluate foreground maps. In: ICCV. pp. 4558–4567 (2017), https://academic.microsoft.com/paper/2963868681 4.2
14. Fan, D.P., Gong, C., Cao, Y., Ren, B., Cheng, M.M., Borji, A.: Enhanced-alignment measure for binary foreground map evaluation. In: IJCAI. pp. 698–704 (2018) 4.2
15. Fan, D.P., Wang, J., Liang, X.M.: Improving image retrieval using the context-aware saliency areas. AMM **734**, 596–599 (2015), https://academic.microsoft.com/paper/2090323693 1
16. Feng, M., Lu, H., Ding, E.: Attentive feedback network for boundary-aware salient object detection. In: CVPR. pp. 1623–1632 (2019), https://academic.microsoft.com/paper/2948510860 1, 2
17. Gao, Y., Wang, M., Tao, D., Ji, R., Dai, Q.: 3-d object retrieval and recognition with hypergraph analysis. TIP **21**(9), 4290–4303 (2012), https://academic.microsoft.com/paper/2068078373 1
18. Han, J., Chen, H., Liu, N., Yan, C., Li, X.: Cnns-based rgb-d saliency detection via cross-view transfer and multiview fusion. TSMC **48**(11), 3171–3183 (2018) 1, 2, 4.1, 2, 3, 4.4

19. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR). pp. 770–778 (2016) 3.3
20. Hong, S., You, T., Kwak, S., Han, B.: Online tracking by learning discriminative saliency map with convolutional neural network. In: ICML. pp. 597–606 (2015), https://academic.microsoft.com/paper/1854404533 1
21. Hou, Q., Cheng, M.M., Hu, X., Borji, A., Tu, Z., Torr, P.H.S.: Deeply supervised salient object detection with short connections. In: CVPR. vol. 41, pp. 815–828 (2017) 1, 1
22. Hou, Q., Cheng, M.M., Hu, X., Borji, A., Tu, Z., Torr, P.H.S.: Deeply supervised salient object detection with short connections. TPAMI **41**(4), 815–828 (2019), https://academic.microsoft.com/paper/2569272946 1, 3.3
23. Ju, R., Ge, L., Geng, W., Ren, T., Wu, G.: Depth saliency based on anisotropic center-surround difference. In: ICIP. pp. 1115–1119 (2014) 4.1
24. Lang, C., Nguyen, T.V., Katti, H., Yadati, K., Kankanhalli, M.S., Yan, S.: Depth matters: influence of depth cues on visual saliency. In: ECCV. pp. 101–115 (2012) 2
25. Li, G., Zhu, C.: A three-pathway psychobiological framework of salient object detection using stereoscopic technology. In: ICCVW. pp. 3008–3014 (2017), https://academic.microsoft.com/paper/2766315367 4.1
26. Li, G., Yu, Y.: Deep contrast learning for salient object detection. In: CVPR. pp. 478–487 (2016) 1
27. Li, N., Ye, J., Ji, Y., Ling, H., Yu, J.: Saliency detection on light field. PAMI **39**(8), 1605–1616 (2017) 4.1
28. Liu, G., Fan, D.: A model of visual attention for natural image retrieval. In: ISCC-C. pp. 728–733 (2013), https://academic.microsoft.com/paper/2314707829 1
29. Liu, N., Han, J.: Dhsnet: Deep hierarchical saliency network for salient object detection. In: CVPR. pp. 678–686 (2016) 2
30. Niu, Y., Geng, Y., Li, X., Liu, F.: Leveraging stereopsis for saliency analysis. In: CVPR. pp. 454–461 (2012) 4.1
31. Peng, H., Li, B., Xiong, W., Hu, W., Ji, R.: Rgbd salient object detection: A benchmark and algorithms. In: ECCV. pp. 92–109 (2014) 2, 4.1, 2, 3, 4.4
32. Piao, Y., Ji, W., Li, J., Zhang, M., Lu, H.: Depth-induced multi-scale recurrent attention network for saliency detection. In: ICCV (2019) 1, 1, 2, 4.1, 2, 3, 4.4
33. Qin, X., Zhang, Z., Huang, C., Gao, C., Dehghan, M., Jagersand, M.: Basnet: Boundary-aware salient object detection. In: CVPR. pp. 7479–7489 (2019), https://academic.microsoft.com/paper/2961348656 2
34. Qu, L., He, S., Zhang, J., Tian, J., Tang, Y., Yang, Q.: Rgbd salient object detection via deep fusion. TIP **26**(5), 2274–2285 (2017) 2, 2, 3, 4.4
35. Ren, J., Gong, X., Yu, L., Zhou, W., Yang, M.Y.: Exploiting global priors for rgb-d saliency detection. In: CVPRW. pp. 25–32 (2015) 1, 2
36. Ren, S., He, K., Girshick, R., Sun, J.: Faster r-cnn: Towards real-time object detection with region proposal networks. TPAMI **39**(6), 1137–1149 (2017), https://academic.microsoft.com/paper/639708223 1
37. Ren, Z., Gao, S., Chia, L.T., Tsang, I.W.H.: Region-based saliency detection and its application in object recognition. TCSVT **24**(5), 769–779 (2014), https://academic.microsoft.com/paper/2055180303 1
38. Smeulders, A.W.M., Chu, D.M., Cucchiara, R., Calderara, S., Dehghan, A., Shah, M.: Visual tracking: An experimental survey. TPAMI **36**(7), 1442–1468 (2014), https://academic.microsoft.com/paper/2126302311 1
39. Wang, W., Shen, J., Porikli, F.: Saliency-aware geodesic video object segmentation. In: CVPR. pp. 3395–3402 (2015) 1

40. Wang, W., Shen, J., Sun, H., Shao, L.: Video co-saliency guided co-segmentation. TCSVT **28**(8), 1727–1736 (2018), https://academic.microsoft.com/paper/2887503470 1
41. Wu, R., Feng, M., Guan, W., Wang, D., Lu, H., Ding, E.: A mutual learning method for salient object detection with intertwined multi-supervision. In: CVPR. pp. 8150–8159 (2019), https://academic.microsoft.com/paper/2962680827 1
42. Zhang, M., Ji, W., Piao, Y., Li, J., Zhang, Y., Xu, S., Lu, H.: Lfnet: Light field fusion network for salient object detection. IEEE Transactions on Image Processing **29**, 6276–6287 (2020) 1
43. Zhang, M., Li, J., Ji, W., Piao, Y., Lu, H.: Memory-oriented decoder for light field salient object detection. In: NeurIPS 2019 : Thirty-third Conference on Neural Information Processing Systems. pp. 898–908 (2019) 1
44. Zhang, P., Wang, D., Lu, H., Wang, H., Ruan, X.: Amulet: Aggregating multi-level convolutional features for salient object detection. In: ICCV. pp. 202–211 (2017), https://academic.microsoft.com/paper/2963032190 2
45. Zhang, X., Wang, T., Qi, J., Lu, H., Wang, G.: Progressive attention guided recurrent network for salient object detection. In: CVPR (2018) 2
46. Zhao, J.X., Cao, Y., Fan, D.P., Cheng, M.M., Li, X.Y., Zhang, L.: Contrast prior and fluid pyramid integration for rgbd salient object detection. In: CVPR. pp. 3927–3936 (2019) 1, 2, 2, 3, 4.4
47. Zhao, R., Ouyang, W., Li, H., Wang, X.: Saliency detection by multi-context deep learning. In: CVPR. pp. 1265–1274 (2015) 2
48. Zhu, C., Cai, X., Huang, K., Li, T.H., Li, G.: Pdnet: Prior-model guided depth-enhanced network for salient object detection. In: ICME (2019) 1, 2, 2, 3, 4.4
49. Zhu, C., Li, G., Guo, X., Wang, W., Wang, R.: A multilayer backpropagation saliency detection algorithm based on depth mining. In: CAIP. pp. 14–23 (2017) 2, 2, 3, 4.4
50. Zhu, C., Li, G., Wang, W., Wang, R.: An innovative salient object detection using center-dark channel prior. In: ICCVW. pp. 1509–1515 (2017) 2, 2, 3, 4.4