A Single Stream Network for Robust and Real-Time RGB-D Salient Object Detection

Xiaoqi Zhao[1], Lihe Zhang[1(✉)], Youwei Pang[1], Huchuan Lu[1,2], and Lei Zhang[3,4]

[1] Dalian University of Technology, Dalian, China
{zxq,lartpang}@mail.dlut.edu.cn, zhanglihe@dlut.edu.cn
[2] Peng Cheng Laboratory, Shenzhen, China
[3] Department of Computing, The Hong Kong Polytechnic University, Hung Hom, China
cslzhang@comppolyu.edu.hk, lhchuan@dlut.edu.cn
[4] DAMO Academy, Alibaba Group, Hangzhou, China

Abstract. Existing RGB-D salient object detection (SOD) approaches concentrate on the cross-modal fusion between the RGB stream and the depth stream. They do not deeply explore the effect of the depth map itself. In this work, we design a single stream network to directly use the depth map to guide early fusion and middle fusion between RGB and depth, which saves the feature encoder of the depth stream and achieves a lightweight and real-time model. We tactfully utilize depth information from two perspectives: (1) Overcoming the incompatibility problem caused by the great difference between modalities, we build a single stream encoder to achieve the early fusion, which can take full advantage of ImageNet pre-trained backbone model to extract rich and discriminative features. (2) We design a novel depth-enhanced dual attention module (DEDA) to efficiently provide the fore-/back-ground branches with the spatially filtered features, which enables the decoder to optimally perform the middle fusion. Besides, we put forward a pyramidally attended feature extraction module (PAFE) to accurately localize the objects of different scales. Extensive experiments demonstrate that the proposed model performs favorably against most state-of-the-art methods under different evaluation metrics. Furthermore, this model is 55.5% lighter than the current lightest model and runs at a real-time speed of 32 FPS when processing a 384×384 image.

Keywords: RGB-D salient object detection · Single stream · Depth-enhanced dual attention · Lightweight · Real-time

1 Introduction

Salient object detection (SOD) aims to estimate visual significance of image regions and then segment salient targets out. It has been widely used in

many fields, *e.g.*, scene classification [29], visual tracking [21], person re-identification [30], foreground maps evaluation [10], content-aware image editing [51], light field image segmentation [36] and image captioning [14], etc.
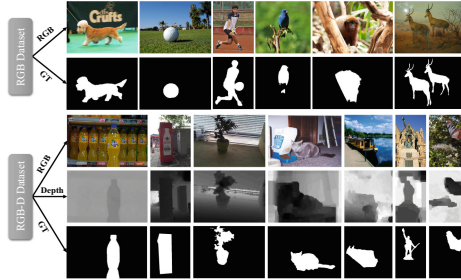

Fig. 1. Visual comparison of RGB and RGB-D SOD datasets.

With the development of deep convolutional neural networks (CNNs), a large number of CNN-based methods [6,24,27,33,35,37–39,41–44,47] have been proposed for RGB salient object detection and they achieve satisfactory performance. However, some complex scenarios are still unresolved, such as salient objects share similar appearances to the background or the contrast among different objects is extremely low. Under these circumstances, only using the information provided by the RGB image is not sufficient to predict saliency map well. Recently, benefiting from Microsoft Kinect and Intel RealSense devices, depth information can be conveniently obtained. Moreover, the stable geometric structures depicted in the depth map are robust against the changes of illumination and texture, which can provide important supplement information for handling complex environments, as shown in Fig. 1. These examples in the RGB-D dataset have more stereoscopic viewing angles and more severe interference from the background than ones in the RGB dataset.

For the RGB-D SOD task, many CNN-based methods [2–4,23,26,45] are proposed, but more efforts need be paid to achieve a robust, real-time and small-scale model. We analyze their restrictions here: (1) Most methods [2,4,16,26,34,48] use the two-stream structure to separately extract features from RGB and depth, which greatly increases the number of parameters in the network. In addition, due to small scale of existing RGB-D datasets and great difference between RGB and depth modalities, the deep network (e.g., VGG, ResNet) is very difficult to be trained from scratch if the RGB and depth channels are concatenated and fed into the network. To this end, we construct a single stream encoder, which can borrow the generalization ability of ImageNet pre-trained backbone to extract discriminative features from the RGB-D input and achieve SOD-oriented RGB-depth early fusion. (2) The depth map can naturally depict contrast cues at different positions, which provides important guidance for the fore-/back-ground segmentation. However, this observation has never been investigated in the existing literature. In this work, we introduce a spatial filtering mechanism between

the encoder and the decoder, which explicitly utilizes the depth map to guide the computation of dual attention, thereby promoting feature discrimination in the fore-/back-ground decoding branches. (3) Since the size of objects is various, the effective utilization of multi-scale contextual information is very key to accurately localize objects. Previous methods [9,26,35,42,46] do not explore the internal relationships between the parallel features of different receptive fields in the multi-scale feature extraction module (e.g. ASPP [5]). We think that each position in the feature map responds differently to objects and a strong activation area can better perceive the semantic cues of objects.

To address these above problems, we propose a single stream network with the novel depth-enhanced attention (DANet) for RGB-D saliency detection. First, we design a single stream encoder with a 4-channel input. It can not only save many parameters compared to previous two-stream methods, but also promote the regional discrimination of the low-level features because this encoder can effectively utilize the ImageNet pre-trained model to extract powerful features with the help of the proposed initialization strategy. Second, we build a depth-enhanced dual attention module (DEDA) between the encoder and the decoder. This module sequentially leverages both the mask-guided strategy and the depth-guided strategy to filter the mutual interference between depth prior and appearance prior, thereby enhancing the overall contrast between foreground and background. In addition, we present a pyramidal attention mechanism to promote the representation ability of the top-layer features. It calculates the spatial correlation among different scales and obtains efficient context guidance for the decoder.

Our main contributions are summarized as follows.

– We propose a single stream network to achieve both early fusion and middle fusion, which implicitly formulates the cross-modal information interaction in the encoder and further explicitly enhances this effect in the decoder.
– We design a novel depth-enhanced dual attention mechanism, which exploits the depth map to strengthen the mask-guided attention and computes fore-/back-ground attended features for the encoder.
– Through using a self-attention mechanism, we propose a pyramidally attended feature extraction module, which can depict spatial dependencies between any two positions in feature map.
– We compare the proposed model with ten state-of-the-art RGB-D SOD methods on six challenging datasets. The results show that our method performs much better than other competitors. Meanwhile, the proposed model is much lighter than others and achieves a real-time speed of 32 FPS.

2   Related Work

Generally speaking, the depth map can be utilized in three ways: early fusion [25, 32], middle fusion [15] and late fusion [13]. It is worth noting that the early fusion technique has not been explored in existing deep learning based saliency methods. Most of them use two streams to respectively handle RGB and depth

information. They achieve the cross-modal fusion only at a specific stage, which limits the usage of the depth-related prior knowledge. This issue motivates some efforts [2,4] to examine the multi-level fusion between the two streams. However, the two-stream design significantly increases the number of parameters in the network [2,4,16,34]. And, restricted by the scale of existing RGB-D datasets, the depth stream is hardly effectively trained and does not comprehensively capture depth cues to guide salient object detection. To this end, Zhao *et al.* [45] propose a trade-off method, which only feeds the RGB images into the encoder network and inserts a shallow convolutional subnet between adjacent encoder blocks to extract the guidance information from the depth map.

In this work, we integrate the depth map and the RGB image from starting to build a real single-stream network. This network can fully use the advantage of the ImageNet pre-trained model to extract color and depth features and remedy the deficiencies of individual grouping cues in color space and depth space. And we also show the effectiveness of the proposed early fusion strategy in the encoder through quantitative and qualitative analysis. Recently, Zhao *et al.* [45] exploit the depth map to compute a contrast prior and then use this prior to enhance the encoder features. Their contrast loss actually enforces the network to learn saliency cues from the depth map in a brute-force manner. Although the resulted attention map can coarsely distinguish the foreground from the background, it greatly reduces the ability of providing accurate depth prior for some easily-confused regions, thereby weakening the discrimination of the encoder feature in these regions. We think that the depth map is more suitable to play a guiding role because the grouping cues in depth space are very incompatible with those in color space. In this work, we combine the depth guidance and the mask guidance to explicitly formulate their complementary relation. Thus, we can effectively take advantage of the useful depth cues to assist in segmenting salient objects and weaken their incompatibility.

3   Proposed Method

We adopt the feature pyramid network [19] (FPN) as the basic structure and the overall architecture is shown in Fig. 2, in which encoder blocks, transition layers, saliency layers and background layers are denoted as $\mathbf{E}^i$, $\mathbf{T}^i$, $\mathbf{S}^i$ and $\mathbf{B}^i$, respectively. Here, $i \in \{1, 2, 3, 4, 5\}$ indexes different levels. And their output feature maps are denoted as E^i, T^i, S^i and B^i, respectively. Each transition layer uses a 3×3 convolution operation to process the features maps from each encoder block for matching the number of channels. The saliency layers and background layers compose the decoder. The final output is generated by integrating the predictions of the two branches using a residual connection. In this section, we first describe the encoder network in Sect. 3.1, then give the details of the proposed modules, including depth-enhanced dual attention module (DEDA) in Sect. 3.2 and pyramidally attended feature extraction module (PAFE) in Sect. 3.3.
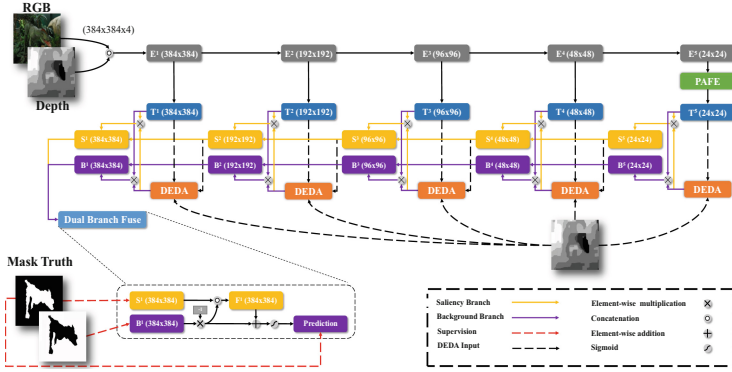
Fig. 2. Network pipeline. It consists of the VGG-16 ($\mathbf{E}^1 \sim \mathbf{E}^5$), five transition layers ($\mathbf{T}^1 \sim \mathbf{T}^5$), five saliency layers ($\mathbf{S}^1 \sim \mathbf{S}^5$), five background layers ($\mathbf{B}^1 \sim \mathbf{B}^5$), the pyramidally attended feature extraction module (PAFE) and the depth-enhanced dual attention module (DEDA). The final prediction is generated by using residual connections to fuse the outputs from $\mathbf{S}^1$ and $\mathbf{B}^1$.

3.1   Single Stream Encoder Network

In our model, the encoder is a single stream with a FCN structure. We take the VGG-16 [31] network as the backbone, which contains 13 Conv layers, 5 max-pooling layers and 2 fully connected layers. First, we concatenate the depth map with the RGB image as the 4-channel RGB-D input. We initialize the parameters of the first convolutional layer in block $\mathbf{E}^1$ using the He's method [17] and output a 64-channel feature. The other layers adopt the ImageNet pre-trained parameters. In this way, the two-modality information can be fused in the input stage and make the low-level features have a more powerful discriminant ability, which is conducive to extracting effective features for salient regions. Moreover, because four input channels are parallel in the channel direction, the network can easily learn to suppress the feature response of the depth channel when the quality of the depth map is poor and does not affect feature computation of the color channels. To demonstrate the effectiveness of this design, we compare two other schemes. Both of them combine the color channels with the depth channel by element-wise addition. One is to directly load the pre-trained parameters. The other is to use the above-mentioned parameter setting. When the depth map has a negative impact, the first layer simultaneously suppresses the color response and the depth response. The quantitative results in Table 3 show that our early fusion strategy performs better than other schemes.

Similar to most previous methods [2,3,26,34,45,48], we cast away all the fully-connected layers of the VGG-16 net and remove the last pooling layer to retain the details of the top-layer features.
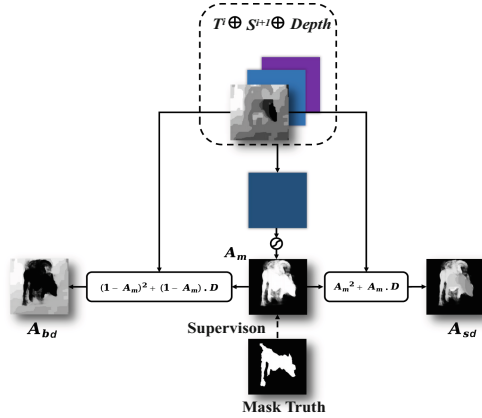
Fig. 3. Detailed diagram of depth-enhanced dual attention module.

3.2 Depth-Enhanced Dual Attention Module

Considering that the depth map can naturally describe contrast information in different depth positions, we utilize it to generate contrasted features for the decoder, thereby strengthening the detection ability for hard examples. In particular, we propose a depth-enhanced attention module and its detailed structure is shown in Fig. 3.

When the region of object has a large span at depth or the background and foreground areas are at the same depth, only depending on the depth map does not provide accurate grouping cues for saliency detection. Therefore, we adopt the mask supervision and depth guidance mechanism to filter the misleading information. We first combine the features from the current transition layer and the previous decoder block with the depth map to compute a mask-guided attention A_m, which is supervised by the saliency ground truth. The whole process is written as follows:

$$A_m = \begin{cases} \delta(Conv(T^i + S^{i+1} + D)) & \text{if } i = 1,2,3,4 \\ \delta(Conv(T^i + D)) & \text{if } i = 5, \end{cases} \tag{1}$$

where $\delta(\cdot)$ is an element-wise sigmoid function, $Conv(\cdot)$ refers to the convolution layer and D denotes the depth map. Although the resulted A_m shows high contrast between the foreground and the background under binary supervision, it inevitably exists two drawbacks: (1) Some background regions are wrongly classified to be salient. (2) Some salient regions are mislabelled as the background. To solve the first issue, we introduce the depth information to refine A_m:

$$A_{sd} = A_m \cdot A_m + A_m \cdot D, \tag{2}$$

where A_{sd} denotes the depth-enhanced attention of the saliency branch. It can provide additional contrast guidance for the misjudged regions in A_m and main-

tain high contrast between foreground and background, thereby enhancing mask-guided attention. To resolve the second issue, we design the depth-enhanced attention A_{bd} for the background branch as follows:

$$A_{bd} = (1 - A_m) \cdot (1 - A_m) + (1 - A_m) \cdot D. \tag{3}$$

We combine A_m and D by the above formulas to construct foreground attention A_{sd} and background attention A_{bd}. There are three benefits: (1) When the depth value is very small or even zero, the attention still work because the first terms in Equ. (2) and Equ. (3) are independent of D. (2) The depth map does not have the semantic distinction between foreground and background, which may introduce noise and interference when segmenting salient object. However, the DEDA can still preserve high contrast between the foreground and the background while introducing depth information in Equ. (2) and Equ. (3). Becasue, the A_m usually shows high contrast between the foreground and the background under binary supervision. $A_m \cdot D$ or $1 - A_m \cdot D$ can limit D to only optimize the foreground or the background. (3) During the back-propagation process of gradient, A_{sd} and A_{bd} can obtain dynamic gradients, which help the network learn the optimal parameters. Taking A_{sd} for example, its derivation with respect to A_m is calculated as:

$$\frac{\mathrm{d}A_{sd}}{\mathrm{d}A_m} = 2 \cdot A_m + D, \tag{4}$$

from where it can be seen that the gradient changes with A_m although the depth D is fixed.
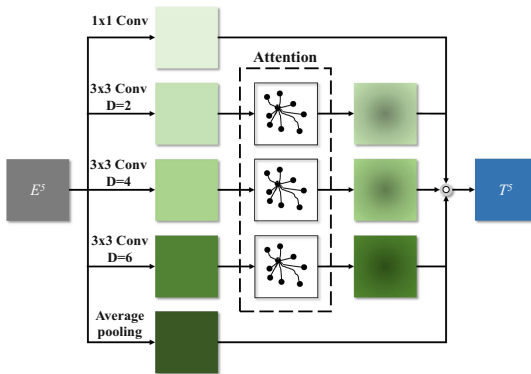

Fig. 4. Illustration of pyramidally attended feature extraction.

3.3   Pyramidally Attended Feature Extraction

The scale of objects is various in images. The single-scale features can not capture the multi-scale context for different objects. Benefiting from the ASPP in

semantic segmentation [5], some SOD networks [9,42,46] also equip it. However, directly aggregating features at different scales may weaken the representation ability for salient areas because of the distraction of non-salient regions. Instead of equally treating all spatial positions, we respectively apply spatial attention to the features of different scales in order to focus more on the visually important regions. By integrating the attention-enhanced multi-scale features, we build a pyramidally attended feature extraction module (PAFE). Its detailed structure is shown in Fig. 4.

We first load in parallel several dilated convolutional layers with different dilation rates on the top-layer $\mathbf{E}^5$ to extract high-level and multi-scale features. Then, an attention module is followed in individual branch. Our attention design is inspired by the non-local idea [40]. We consider the pairwise relationship at any point in feature map to calculate the attention weight. Let $F_{in} \in \mathbb{R}^{C \times H \times W}$ and $F_{out} \in \mathbb{R}^{C \times H \times W}$ represent the input and the output of the attention module, respectively. The attention map A is computed as follows:

$$
\begin{aligned}
A = softmax(R_1(Conv(F_{in}))^\top \\
\times R_1(Conv(F_{in}))),
\end{aligned}
\tag{5}
$$

where $softmax(\cdot)$ is an element-wise softmax function and $R_1(\cdot)$ reshapes the input feature to $\mathbb{R}^{C \times N}$. $N = H \times W$ is the number of features.

Next, we combine A with F_{in} to yield the attention-enhanced feature map and then add the input F_{in} to obtain the output F_{out} as follows:

$$
F_{out} = F_{in} + R_2(R_1(Conv(F_{in})) \times A^\top),
\tag{6}
$$

where $R_2(\cdot)$ reshapes the input feature to $\mathbb{R}^{C \times H \times W}$. In particular, the 1×1 convolution branch and the global average pooling branch aim to maintaining the inherent properties of the input by respectively using the minimal and maximum receptive field. Therefore, we do not apply the attention module to the two branches.

4   Experiments

4.1   Dataset

We evaluate the proposed model on six public RGB-D SOD datasets which are **NJUD** [18], **RGBD135** [7] **NLPR** [25], **SSD** [49], **DUTLF-D** [26] and **SIP** [12]. On the DUTLF-D, we adopt the same way as the DMRA [26] to use 800 images for training and the rest 400 for testing. Following most state-of-the-art methods [2,4,16,45], we randomly select 1400 samples from the NJUD dataset and 650 samples from the NLPR dataset for training. Their remaining images and other three datasets are used for testing.

4.2   Evaluation Metrics

We adopt several widely used metrics for quantitative evaluation: precision-recall (PR) curves, F-measure score, mean absolute error (MAE, $\mathcal{M}$), the recently released S-measure (S_m) and E-measure (E_m) scores. The lower value is better for the MAE and higher is better for others. **Precision-Recall curve**: The pairs of precision and recall are calculated by comparing the binary saliency maps with the ground truth to plot the PR curve, where the threshold for binarizing slides from 0 to 255. **F-measure**: It is a metric that comprehensively considers both precision and recall:

$$\mathrm{F}_\beta = \frac{\left(1 + \beta^2\right) \cdot \text{precision} \cdot \text{recall}}{\beta^2 \cdot \text{precision} + \text{recall}}, \tag{7}$$

where β^2 is set to 0.3 as suggested in [1] to emphasize the precision. In this paper, we report the maximum F-measure (F_β^{max}) score across the binary maps of different thresholds, the mean F-measure (F_β^{mean}) socre across an adaptive threshold and the weighted F-measure (F_β^w) [22]. **Mean Absolute Error**: It is a complement to the PR curve and measures the average absolute difference between the prediction and the ground truth pixel by pixel. **S-measure**: It evaluates the spatial structure similarity by combining the region-aware structural similarity S_r and the object-aware structural similarity S_o:

$$\mathrm{S}_m = \alpha * S_o + (1 - \alpha) * S_r, \tag{8}$$

where α is set to 0.5 [10]. **E-measure**: The enhanced alignment measure [11] can jointly capture image level statistics and local pixel matching information.

4.3   Implementation Details

Our model is implemented based on the Pytorch toolbox and trained on a PC with GTX 1080Ti GPU for 40 epochs with mini-batch size 4. The input RGB image and depth map are both resized to 384×384. For the RGB image, we use some data augmentation techniques to avoid overfitting: random horizontally flip, random rotate, random brightness, saturation and contrast. For the optimizer, we adopt the stochastic gradient descent (SGD) with a momentum of 0.9 and a weight decay of 0.0005. The learning rate is set to 0.001 and later use the "poly" policy [20] with the power of 0.9 as a mean of adjustment. In this paper, we use the binary cross-entropy loss as supervision. The source code will be publicly available at https://github.com/Xiaoqi-Zhao-DLUT/DANet-RGBD-Saliency.

4.4   Comparison with State-of-the-Art Results

The performance of the proposed model is compared with ten state-of-the-art approaches on six benchmark datasets, including the DES [7], DCMC [8],

CDCP [50], DF [28], CTMF [16], PCA [2], MMCI [4], TANet [3], CPFP [45] and DMRA [26]. For fair comparisons, all the saliency maps of these methods are directly provided by authors or computed by their released codes.

Table 1. Quantitative comparison. ↑ and ↓ indicate that the larger and smaller scores are better, respectively. Among the CNN-based methods, the best results are shown in **red**. The subscript in each model name is the publication year.

Metric	Traditional Methods			VGG-16							VGG-19	
	DES_{14}	$DCMC_{16}$	$CDCP_{17}$	DF_{17}	$CTMF_{18}$	$PCANet_{18}$	$MMCI_{19}$	$TANet_{19}$	$CPFP_{19}$	DANet	$DMRA_{19}$	DANet
	[7]	[8]	[50]	[28]	[16]	[2]	[4]	[3]	[45]	Ours	[26]	Ours
SSD [49] F_β^{max} ↑	0.260	0.750	0.576	0.763	0.755	0.844	0.823	0.835	0.801	**0.888**	0.858	**0.866**
F_β^{mean} ↑	0.073	0.684	0.524	0.709	0.709	0.786	0.748	0.767	0.726	**0.831**	0.821	**0.827**
F_β^{w} ↑	0.172	0.480	0.429	0.536	0.622	0.733	0.662	0.727	0.709	**0.798**	0.787	**0.795**
S_m ↑	0.341	0.706	0.603	0.741	0.776	0.842	0.813	0.839	0.807	**0.869**	0.856	**0.864**
E_m ↑	0.475	0.790	0.714	0.801	0.838	0.890	0.860	0.886	0.832	**0.909**	0.898	**0.911**
$\mathcal{M}$ ↓	0.500	0.168	0.219	0.151	0.100	0.063	0.082	0.063	0.082	**0.050**	0.059	**0.050**
NJUD [18] F_β^{max} ↑	0.328	0.769	0.661	0.789	0.857	0.888	0.868	0.888	0.890	**0.905**	0.896	**0.910**
F_β^{mean} ↑	0.165	0.715	0.618	0.744	0.788	0.844	0.813	0.844	0.837	**0.877**	**0.871**	**0.871**
F_β^{w} ↑	0.234	0.497	0.510	0.545	0.720	0.803	0.739	0.805	0.828	**0.853**	0.847	**0.857**
S_m ↑	0.413	0.703	0.672	0.735	0.849	0.877	0.859	0.878	0.878	**0.897**	0.885	**0.899**
E_m ↑	0.491	0.796	0.751	0.818	0.866	0.909	0.882	0.909	0.900	**0.926**	0.920	**0.922**
$\mathcal{M}$ ↓	0.448	0.167	0.182	0.151	0.085	0.059	0.079	0.061	0.053	**0.046**	0.051	**0.045**
RGBD135 [7] F_β^{max} ↑	0.800	0.311	0.651	0.625	0.865	0.842	0.839	0.853	0.882	**0.916**	0.906	**0.928**
F_β^{mean} ↑	0.695	0.234	0.594	0.573	0.778	0.774	0.762	0.795	0.829	**0.891**	0.867	**0.899**
F_β^{w} ↑	0.301	0.169	0.478	0.392	0.687	0.711	0.650	0.740	0.787	**0.848**	0.843	**0.877**
S_m ↑	0.632	0.469	0.709	0.685	0.863	0.843	0.848	0.858	0.872	**0.905**	0.899	**0.924**
E_m ↑	0.817	0.676	0.810	0.806	0.911	0.912	0.904	0.919	0.927	**0.961**	0.944	**0.968**
$\mathcal{M}$ ↓	0.289	0.196	0.120	0.131	0.055	0.050	0.065	0.046	0.038	**0.028**	0.030	**0.023**
DUTLF-D [30] F_β^{max} ↑	0.770	0.444	0.658	0.774	0.842	0.809	0.804	0.823	0.787	**0.911**	0.908	**0.918**
F_β^{mean} ↑	0.667	0.405	0.633	0.747	0.792	0.760	0.753	0.778	0.735	**0.884**	0.883	**0.889**
F_β^{w} ↑	0.380	0.284	0.521	0.536	0.682	0.688	0.628	0.705	0.638	**0.847**	0.852	**0.860**
S_m ↑	0.659	0.499	0.687	0.729	0.831	0.801	0.791	0.808	0.749	**0.889**	0.887	**0.899**
E_m ↑	0.751	0.712	0.794	0.842	0.883	0.863	0.856	0.871	0.815	**0.929**	0.930	**0.937**
$\mathcal{M}$ ↓	0.280	0.243	0.159	0.145	0.097	0.100	0.112	0.093	0.100	**0.047**	0.048	**0.043**
NLPR [23] F_β^{max} ↑	0.695	0.413	0.687	0.752	0.841	0.864	0.841	0.876	0.884	**0.908**	0.888	**0.916**
F_β^{mean} ↑	0.583	0.328	0.592	0.683	0.724	0.795	0.730	0.796	0.818	**0.865**	0.855	**0.870**
F_β^{w} ↑	0.254	0.259	0.501	0.516	0.679	0.762	0.676	0.780	0.807	**0.850**	0.840	**0.862**
S_m ↑	0.582	0.550	0.724	0.769	0.860	0.874	0.856	0.886	0.884	**0.908**	0.898	**0.915**
E_m ↑	0.760	0.685	0.786	0.840	0.869	0.916	0.872	0.916	0.920	**0.945**	0.942	**0.949**
$\mathcal{M}$ ↓	0.301	0.196	0.115	0.100	0.056	0.044	0.059	0.041	0.038	**0.031**	0.031	**0.028**
SIP [12] F_β^{max} ↑	0.720	0.680	0.544	0.704	0.720	0.861	0.840	0.851	0.870	**0.901**	0.847	**0.892**
F_β^{mean} ↑	0.644	0.645	0.495	0.673	0.684	0.825	0.795	0.809	0.819	**0.864**	0.815	**0.855**
F_β^{w} ↑	0.342	0.414	0.397	0.406	0.535	0.768	0.712	0.748	0.788	**0.829**	0.734	**0.822**
S_m ↑	0.616	0.683	0.595	0.653	0.716	0.842	0.833	0.835	0.850	**0.878**	0.800	**0.875**
E_m ↑	0.751	0.787	0.722	0.794	0.824	0.900	0.886	0.894	0.899	**0.914**	0.858	**0.915**
$\mathcal{M}$ ↓	0.298	0.186	0.224	0.185	0.139	0.071	0.086	0.075	0.064	**0.054**	0.088	**0.054**

Quantitative Evaluation. 1) Table 1 shows performance comparisons in terms of the maximum F-measure, mean F-measure, weighted F-measure, S-measure, E-measure and MAE scores. It can be seen that our DANet achieves the best results on all six datasets under all six metrics. 2) Table 2 lists the model sizes and average speed of different methods in detail. Our model is the smallest and the fastest among these state-of-art methods and saves 55.5% of the parameters compared to the second lightest method DMRA [26]. 3) Figure 5 shows the

Table 2. The model sizes and average speed of different methods.

Model Name	PCANet [2]	MMCI [4]	TANet [3]	CPFP [45]	DMRA [26]	OURS(VGG-19)	OURS(VGG-16)
Model Size	533.6 (MB)	951.9 (MB)	929.7 (MB)	278 (MB)	238.8 (MB)	128.1 (MB)	106.7 (MB)
Average speed	17 (FPS)	20 (FPS)	14(FPS)	6 (FPS)	22 (FPS)	30 (FPS)	32 (FPS)

PR curves of different algorithms. We can see that the curves of the proposed method are significantly higher than those of other methods, especially on the NJUD, NLPR and RGBD135 datasets which contain plenty of relatively complex images. Through detailed quantitative comparisons, it can be seen that our method has significant advantages in accuracy and model size, which indicates it is necessary to further explore how to better utilize depth information.
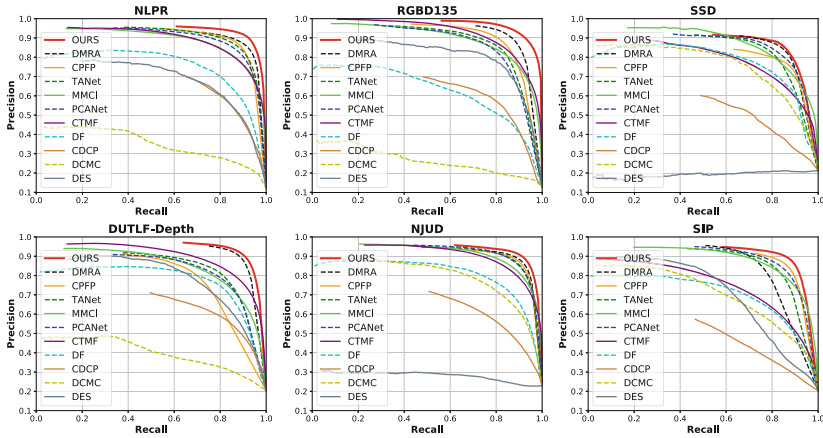

Fig. 5. Precision (vertical axis) recall (horizontal axis) curves on six RGB-D salient object detection datasets.

Qualitative Evaluation. Figure 6 illustrates the visual comparison with other approaches. Our method yields the results more close to the ground truth in various challenging scenarios. For example, for the images having multiple objects or the objects having slender parts, our method can accurately locate objects and capture more details (see the 1st–3th rows). In complex environments, with the guidance of the depth maps, the proposed method can precisely identify the whole object, while other methods fail (see the 4th–6th rows). Even when the depth information performs badly in separating the foreground from the background, our network still significantly outperforms other methods (see the 7th–9th rows).

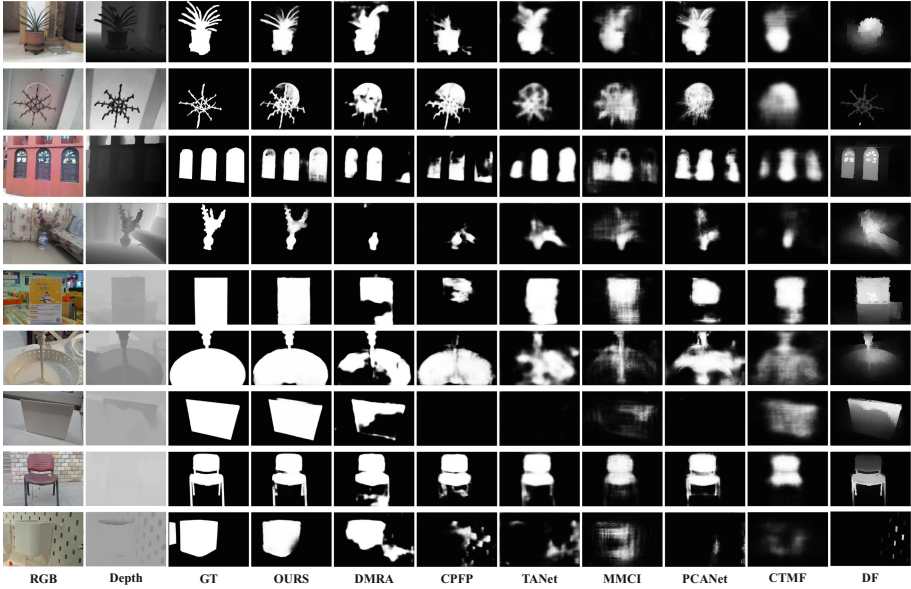| RGB | Depth | GT | OURS | DMRA | CPFP | TANet | MMCI | PCANet | CTMF | DF |

Fig. 6. Visual comparison between our results and the state-of-the-art methods.

4.5 Ablation Studies

We take the FPN network of the VGG-16 backbone as the baseline to analyze the contribution of each component. To verify their generalization abilities, we demonstrate the experimental results on five datasets.

Effectiveness of Depth Fusion in Encoder Network. We evaluate three early fusion strategies. The results are shown in Table 3. Add_p denotes the fusion by using element-wise addition and the ImageNet pre-trained first-layer convolution. Add_{He} and Cat_{He} use the He's initialization [17] instead of the pre-trained parameters in the first layer, and the latter adopts the 4-channel concatenation rather than element-wise addition. We can see that Cat_{He} is significantly better than the baseline and other early fusion methods across five datasets. In particular, it respectively achieves the gain of 4.53%, 5.44%, 5.25% and 16.36% in terms of the F_β^{max}, F_β^{mean}, F_β^w and MAE on the RGBD135 dataset. Furthermore, we visualize the features of different levels in Fig. 7. With the aid of the contrast prior provided the depth map, salient objects and their surrounding backgrounds can be clearly distinguished starting from the lowest level ($\mathbf{E}^1$). At the highest level ($\mathbf{E}^5$), the encoder feature is more concentrated on the salient regions, thereby providing the decoder with effective contextual guidance.

Effectiveness of Depth-Enhanced Dual Attention Module. We compare three attention modules based on the 'Cat_{He}' model. The results are shown in Table 3. We try to directly use the depth map as the attention between the encoder and decoder. Since the depth value often varies widely inside the

Table 3. Ablation analysis on five datasets.

Metric		**Baseline**	Add$_p$	Add$_{He}$	**Cat$_{He}$**	DA	MGA	DEFA	**DEDA**	ASPP	**PAFE**
SSD [49]	F_β^{max} ↑	0.799	0.812	0.817	**0.845**	0.837	0.843	0.858	**0.860**	0.879	**0.888**
	F_β^{mean} ↑	0.745	0.743	0.734	**0.758**	0.754	0.794	0.806	**0.810**	0.830	**0.831**
	F_β^{w} ↑	0.700	0.705	0.677	**0.710**	0.697	0.745	0.757	**0.761**	0.784	**0.798**
	S_m ↑	0.813	0.825	0.811	**0.835**	0.829	0.841	0.846	**0.847**	0.855	**0.869**
	E_m ↑	0.862	**0.857**	0.833	0.849	0.847	0.883	0.886	**0.887**	0.905	**0.909**
	$\mathcal{M}$ ↓	0.080	0.077	0.092	**0.076**	0.078	0.064	0.060	**0.062**	0.056	**0.050**
NJUD [18]	F_β^{max} ↑	0.855	0.861	0.857	**0.869**	0.865	0.882	0.889	**0.889**	0.896	**0.905**
	F_β^{mean} ↑	0.781	0.784	0.798	**0.815**	0.813	0.832	0.842	**0.849**	0.862	**0.877**
	F_β^{w} ↑	0.748	0.757	0.744	**0.770**	0.763	0.815	0.823	**0.826**	0.843	**0.853**
	S_m ↑	0.848	0.854	0.847	**0.860**	0.856	0.878	**0.881**	0.880	0.890	**0.897**
	E_m ↑	0.863	0.866	0.872	**0.880**	0.880	0.896	0.904	**0.907**	0.915	**0.926**
	$\mathcal{M}$ ↓	0.079	0.076	0.081	**0.073**	0.076	0.059	0.056	**0.055**	0.049	**0.046**
RGBD135 [7]	F_β^{max} ↑	0.839	0.860	0.865	**0.877**	0.881	0.897	0.904	**0.913**	0.907	**0.916**
	F_β^{mean} ↑	0.772	0.792	0.802	**0.814**	0.812	0.850	0.868	**0.876**	**0.894**	0.891
	F_β^{w} ↑	0.705	0.732	0.740	**0.742**	0.751	0.823	0.831	**0.846**	**0.860**	0.848
	S_m ↑	0.847	0.863	**0.867**	0.864	0.871	0.906	0.903	**0.907**	**0.915**	0.905
	E_m ↑	0.904	0.910	0.922	**0.922**	0.923	0.943	0.952	**0.954**	**0.966**	0.961
	$\mathcal{M}$ ↓	0.055	0.050	0.051	**0.046**	0.044	0.032	0.033	**0.029**	**0.026**	0.028
NLPR [25]	F_β^{max} ↑	0.852	0.852	0.860	**0.862**	0.859	**0.887**	0.886	0.880	0.903	**0.908**
	F_β^{mean} ↑	0.772	0.772	0.773	**0.774**	0.773	0.821	0.826	**0.832**	0.857	**0.865**
	F_β^{w} ↑	0.741	0.741	0.743	**0.743**	0.734	0.809	0.813	**0.815**	0.846	**0.850**
	S_m ↑	0.862	0.863	0.866	**0.868**	0.865	**0.893**	0.893	0.889	0.907	**0.908**
	E_m ↑	0.898	**0.900**	0.898	0.892	0.894	0.920	0.923	**0.926**	0.939	**0.945**
	$\mathcal{M}$ ↓	0.052	0.053	0.053	**0.052**	0.055	0.040	0.040	**0.038**	0.032	**0.031**
SIP [12]	F_β^{max} ↑	0.838	0.851	0.836	**0.849**	0.835	0.864	0.873	**0.876**	0.885	**0.901**
	F_β^{mean} ↑	0.780	0.784	0.758	**0.787**	0.771	0.804	0.830	**0.833**	0.847	**0.864**
	F_β^{w} ↑	0.716	0.721	0.692	**0.722**	0.699	0.767	0.791	**0.798**	0.813	**0.829**
	S_m ↑	0.833	0.840	0.824	**0.841**	0.833	0.854	0.863	**0.865**	0.871	**0.878**
	E_m ↑	0.882	**0.881**	0.867	0.880	0.868	0.889	**0.907**	**0.907**	0.909	**0.917**
	$\mathcal{M}$ ↓	0.085	**0.082**	0.095	0.083	0.092	0.070	0.062	**0.061**	0.057	**0.054**

foreground or the background, it easily misleads salient object segmentation and performs badly, even worse than the Cat$_{He}$ model. To this end, we use the mask-guided attention (MGA) and the performance is indeed improved. Based on it, we further introduce the depth guidance and build two attended branches to form the depth-enhanced dual attention module (DEDA). It can be seen that the DEFA and DEDA achieve significant performance improvement compared to the MGA. And, the gap between the DEFA and DEDA indicates that the background branch has important supplement to the final prediction. I should note is that we do not deeply consider the two-branch fusion. Since the output of each branch is only a single-channel map, it might not produce too much performance improvement no matter what fusion is used. In addition, we qualitatively show the benefits of the DEDA in Fig 8. It can be seen that the mask-guided attention wrongly classifies some salient regions as the background (see the 1st–3th columns) and some background regions to be salient (see the 4th–6th columns). By introducing extra contrast cues provided by the depth map for these regions, the decoder can very well correct some mistakes in the final predictions.
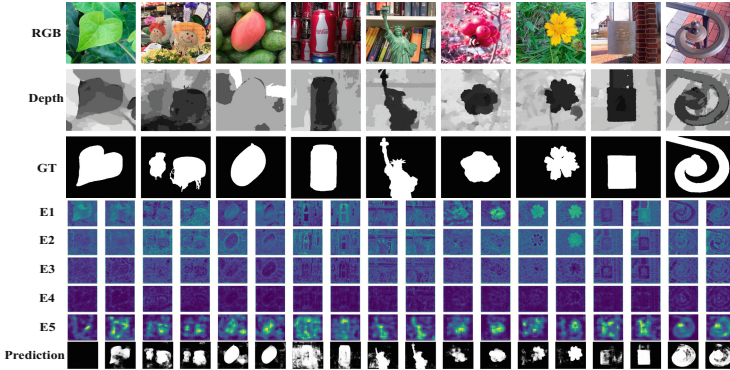
Fig. 7. Visual comparison between the 4-channel RGB-D FPN and the 3-channel RGB FPN (baseline). Each input image corresponds to two columns of feature maps ($\mathbf{E}^1 \sim \mathbf{E}^5$) and prediction. The left is the results of the 3-channel baseline, while the right is those of the 4-channel baseline.
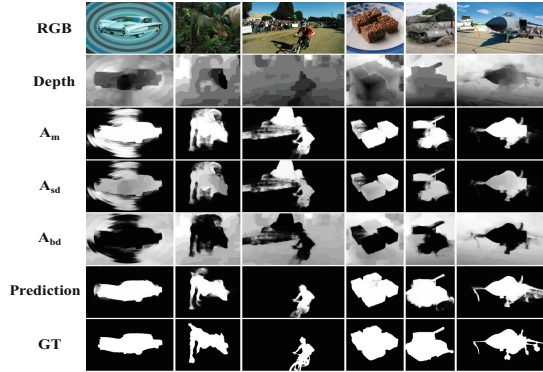

Fig. 8. Visual results of using the DEDA. A_m, A_{sd} and A_{bd} are calculated by Eq. 1, Eq. 2 and Eq. 3, respectively.

Effectiveness of Pyramidally Attended Feature Extraction. To be fair, we compare the PAFE with the ASPP which also uses the same convolution operations. That is, both the two modules equip a 1×1 convolution, three 3×3 atrous convolution with dilation rates of $[2, 4, 6]$ and a global average pooling. The results in Table 3 indicate that the proposed PAFE is more competitive than the ASPP. In addition, we also compare them in terms of Flops (4.00G vs. 3.86G) and Params (7.07M vs. 6.82M). Our PAFE does not increase much more computation cost.

5 Conclusions

In this paper, a more efficient way of using depth information is proposed. We build a single-stream network with the novel depth-enhanced dual attention for real-time and robust salient object detection. We first abandon the routines of the two-stream cross-modal fusion and design a single stream encoder to make full use of the representation ability of the pre-trained network. Next, we use the depth-enhanced dual attention module to make the decoder jointly optimize the fore-/back-ground predictions. Benefiting from the above two ingenious designs, the saliency detection performance is greatly improved while almost no parameters are increased. In addition, we introduce the self-attention mechanism to pyramidally weight multi-scale features, thereby obtaining accurate contextual information to guide salient object segmentation. Extensive experimental results demonstrate that the proposed model notably outperforms ten state-of-the-art methods under different evaluation metrics. Moreover, our model size is only 106.7 MB with the VGG-16 backbone and runs a real-time speed of 32 FPS.

References

1. Achanta, R., Hemami, S., Estrada, F., Süsstrunk, S.: Frequency-tuned salient region detection. In: CVPR, pp. 1597–1604 (2009)
2. Chen, H., Li, Y.: Progressively complementarity-aware fusion network for RGB-D salient object detection. In: CVPR, pp. 3051–3060 (2018)
3. Chen, H., Li, Y.: Three-stream attention-aware network for RGB-D salient object detection. IEEE TIP **28**(6), 2825–2835 (2019)
4. 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. Pattern Recog. **86**, 376–385 (2019)
5. Chen, L.C., Papandreou, G., Kokkinos, I., Murphy, K., Yuille, A.L.: Deeplab: semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs. IEEE TPAMI **40**(4), 834–848 (2017)
6. Chen, S., Tan, X., Wang, B., Hu, X.: Reverse attention for salient object detection. In: Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y. (eds.) ECCV 2018. LNCS, vol. 11213, pp. 236–252. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-01240-3_15
7. Cheng, Y., Fu, H., Wei, X., Xiao, J., Cao, X.: Depth enhanced saliency detection method. In: International Conference on Internet Multimedia Computing and Service, p. 23 (2014)
8. Cong, R., Lei, J., Zhang, C., Huang, Q., Cao, X., Hou, C.: Saliency detection for stereoscopic images based on depth confidence analysis and multiple cues fusion. IEEE SPL **23**(6), 819–823 (2016)

9. Deng, Z., et al.: R3Net: recurrent residual refinement network for saliency detection. In: IJCAI, pp. 684–690 (2018)

10. Fan, D.P., Cheng, M.M., Liu, Y., Li, T., Borji, A.: Structure-measure: a new way to evaluate foreground maps. In: ICCV, pp. 4548–4557 (2017)

11. Fan, D.P., Gong, C., Cao, Y., Ren, B., Cheng, M.M., Borji, A.: Enhanced-alignment measure for binary foreground map evaluation. arXiv preprint arXiv:1805.10421 (2018) 9

12. Fan, D.P., et al.: Rethinking RGB-D salient object detection: Models, datasets, and large-scale benchmarks. arXiv preprint arXiv:1907.06781 (2019)

13. Fan, X., Liu, Z., Sun, G.: Salient region detection for stereoscopic images. In: International Conference on Digital Signal Processing, pp. 454–458 (2014)

14. Fang, H., et al.: From captions to visual concepts and back. In: CVPR, pp. 1473–1482 (2015)

15. Feng, D., Barnes, N., You, S., McCarthy, C.: Local background enclosure for RGB-D salient object detection. In: CVPR, pp. 2343–2350 (2016)

16. Han, J., Chen, H., Liu, N., Yan, C., Li, X.: CNNs-based RGB-D saliency detection via cross-view transfer and multiview fusion. IEEE Trans. Cyber. **48**(11), 3171–3183 (2017)

17. He, K., Zhang, X., Ren, S., Sun, J.: Delving deep into rectifiers: surpassing human-level performance on imagenet classification. In: ICCV, pp. 1026–1034 (2015)

18. Ju, R., Ge, L., Geng, W., Ren, T., Wu, G.: Depth saliency based on anisotropic center-surround difference. In: ICIP, pp. 1115–1119 (2014)

19. Lin, T.Y., Dollár, P., Girshick, R., He, K., Hariharan, B., Belongie, S.: Feature pyramid networks for object detection. In: CVPR, pp. 2117–2125 (2017)

20. Liu, W., Rabinovich, A., Berg, A.C.: Parsenet: Looking wider to see better. arXiv preprint arXiv:1506.04579 (2015)

21. Mahadevan, V., Vasconcelos, N.: Saliency-based discriminant tracking. In: CVPR (2009)

22. Margolin, R., Zelnik-Manor, L., Tal, A.: How to evaluate foreground maps? In: CVPR, pp. 248–255 (2014)

23. Pang, Y., Zhang, L., Zhao, X., Lu, H.: Hierarchical dynamic filtering network for RGB-D salient object detection. In: ECCV (2020)

24. Pang, Y., Zhao, X., Zhang, L., Lu, H.: Multi-scale interactive network for salient object detection. In: CVPR, pp. 9413–9422 (2020)

25. Peng, H., Li, B., Xiong, W., Hu, W., Ji, R.: RGBD salient object detection: a benchmark and algorithms. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) ECCV 2014. LNCS, vol. 8691, pp. 92–109. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-10578-9_7

26. Piao, Y., Ji, W., Li, J., Zhang, M., Lu, H.: Depth-induced multi-scale recurrent attention network for saliency detection. In: ICCV, pp. 7254–7263 (2019)

27. Qin, X., Zhang, Z., Huang, C., Gao, C., Dehghan, M., Jagersand, M.: Basnet: boundary-aware salient object detection. In: CVPR, pp. 7479–7489 (2019)

28. Qu, L., He, S., Zhang, J., Tian, J., Tang, Y., Yang, Q.: RGBD salient object detection via deep fusion. IEEE TIP **26**(5), 2274–2285 (2017)

29. Ren, Z., Gao, S., Chia, L.T., Tsang, I.W.H.: Region-based saliency detection and its application in object recognition. IEEE TCSVT **24**(5), 769–779 (2013)

30. Rui, Z., Ouyang, W., Wang, X.: Unsupervised salience learning for person re-identification. In: CVPR (2013)

31. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014)

32. Song, H., Liu, Z., Du, H., Sun, G., Le Meur, O., Ren, T.: Depth-aware salient object detection and segmentation via multiscale discriminative saliency fusion and bootstrap learning. IEEE TIP **26**(9), 4204–4216 (2017)
33. Wang, L., Wang, L., Lu, H., Zhang, P., Ruan, X.: Saliency detection with recurrent fully convolutional networks. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016. LNCS, vol. 9908, pp. 825–841. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46493-0_50
34. Wang, N., Gong, X.: Adaptive fusion for RGB-D salient object detection. IEEE Access **7**, 55277–55284 (2019)
35. Wang, T., Borji, A., Zhang, L., Zhang, P., Lu, H.: A stagewise refinement model for detecting salient objects in images. In: ICCV, pp. 4019–4028 (2017)
36. Wang, T., Piao, Y., Li, X., Zhang, L., Lu, H.: Deep learning for light field saliency detection. In: ICCV, pp. 8838–8848 (2019)
37. Wang, T., et al.: Detect globally, refine locally: a novel approach to saliency detection. In: CVPR, pp. 3127–3135 (2018)
38. Wang, W., Shen, J., Cheng, M.M., Shao, L.: An iterative and cooperative top-down and bottom-up inference network for salient object detection. In: CVPR, pp. 5968–5977 (2019)
39. Wang, W., Zhao, S., Shen, J., Hoi, S.C., Borji, A.: Salient object detection with pyramid attention and salient edges. In: CVPR, pp. 1448–1457 (2019)
40. Wang, X., Girshick, R., Gupta, A., He, K.: Non-local neural networks. In: CVPR, pp. 7794–7803 (2018)
41. Zeng, Y., Zhang, P., Zhang, J., Lin, Z., Lu, H.: Towards high-resolution salient object detection. arXiv preprint arXiv:1908.07274 (2019)
42. Zhang, L., Dai, J., Lu, H., He, Y., Wang, G.: A bi-directional message passing model for salient object detection. In: CVPR, pp. 1741–1750 (2018)
43. 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)
44. Zhang, X., Wang, T., Qi, J., Lu, H., Wang, G.: Progressive attention guided recurrent network for salient object detection. In: CVPR, pp. 714–722 (2018)
45. 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 (2019)
46. Zhao, T., Wu, X.: Pyramid feature attention network for saliency detection. In: CVPR, pp. 3085–3094 (2019)
47. Zhao, X., Pang, Y., Zhang, L., Lu, H., Zhang, L.: Suppress and balance: a simple gated network for salient object detection. In: ECCV (2020)
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, pp. 199–204 (2019)
49. Zhu, C., Li, G.: A three-pathway psychobiological framework of salient object detection using stereoscopic technology. In: ICCV, pp. 3008–3014 (2017)
50. Zhu, C., Li, G., Wang, W., Wang, R.: An innovative salient object detection using center-dark channel prior. In: ICCV, pp. 1509–1515 (2017)
51. Zhu, J.Y., Wu, J., Xu, Y., Chang, E., Tu, Z.: Unsupervised object class discovery via saliency-guided multiple class learning. IEEE TPAMI **37**(4), 862–875 (2014)