# OGNet: Salient Object Detection with Output-guided Attention Module

Shiping Zhu, Member, IEEE, Lanyun Zhu

Abstract—Attention mechanisms are widely used in salient object detection models based on deep learning, which can effectively promote the extraction and utilization of useful information by neural networks. However, most of the existing attention modules used in salient object detection are input with the processed feature map itself, which easily leads to the problem of 'blind overconfidence'. In this paper, instead of applying the widely used self-attention module, we present an output-guided attention module built with multiscale outputs to overcome the problem of 'blind overconfidence'. We also construct a new loss function, the intractable area F-measure loss function, which is based on the F-measure of the hard-to-handle area to improve the detection effect of the model in the edge areas and confusing areas of an image. Extensive experiments and abundant ablation studies are conducted to evaluate the effect of our methods and to explore the most suitable structure for the model. Tests on several datasets show that our model performs very well, even though it is very lightweight.

Index Terms—Salient object detection, multi output neural network, attention mechansim

#### I. INTRODUCTION

ALIENT object detection aims to estimate the region of the most attractive object in an image, and it is an important research area in computer vision. It has great application value in the fields of scene classification [1], object detection [2], image retrieval [3], [4] and visual tracking [5]–[7]. Salient object detection is a very challenging problem because it requires both a correct identification of the salient object and an accurate display of the salient region. In recent decades, many algorithms for salient object detection have been proposed. Inspired by the human visual attention mechanism, traditional unsupervised methods [8]–[13] typically apply handcrafted features in images to determine the salient region. These methods do not perform well when the background of the image or the shape of the salient object is very complicated.

Recently, deep learning has made rapid development, and many methods based on deep learning [14]–[22] have greatly improved the accuracy of salient object detection. Models based on convolutional neural networks and recurrent neural networks have achieved remarkable performance in many tasks, such as image classification [23], [24], object detection

This work is supported by the National Natural Science Foundation of China (NSFC) under grant No. 61375025, No. 61075011 and No. 60675018, and also the Scientific Research Foundation for the Returned Overseas Chinese Scholars from the State Education Ministry of China.

Shiping Zhu and Lanyun Zhu are with the Department of Measurement Control and Information Technology, School of Instrumentation Science and Optoelectronics Engineering, Beihang University, 100191, HaiDian District, XueYuan Road No. 37., Beijing, China. (phone: +86-13391687912; e-mail: spzhu@163.com)

[25]–[27] and semantic segmentation [28], [29]. A deep neural network can effectively extract and fuse different levels of features in images, which effectively solves the insufficiency of image feature extraction and fusion in traditional methods. Many salient object detection models based on deep learning adopt the encoder-decoder as the basic structure of the neural network [21], [30], [31]. This structure, represented by FCN [32], reduces the image resolution by passing the encoder and extracting the image features from different levels; and then, it gradually recovers the image resolution by the decoder and finally gains the saliency map. The encoder-decoder structure is widely adopted since it can recover the contour shape of the salient objects well.

Since the encoder-decoder has a weaker extraction ability for semantic information, the simple utilization of the encoderdecoder cannot gain quite an outstanding performance. Hence, many research studies are committed to improving the original encoder-decoder structure. Two methods can upgrade the detection effects by increasing only a small memory footprint and the cost of computing. One is to use networks with multiscale outputs [33]. Different from most models that have only one output, the multiscale output structure, represented by deeply supervised net, obtains many outputs in various positions of the neural network. Such a structure can make the deeper parts of the neural network easier to train and lead the network to place more emphasis on the required tasks, avoiding information turbulence and mistakes. Many models based on multiscale output structures for image segmentation [34], object detection [35] and salient object detection [14], [15], [33], [36] have achieved good performance. The other method uses the attention mechanism. In recent years, the attention mechanism has become one of the most important research directions of deep learning [37] because it can significantly improve the effect of models by adding only a small amount of computation. Attention mechanisms can reinforce useful or key information and impair useless or incorrect information. Salient object detection is a task of classifying each pixel into two categories, and the introduction of an attention mechanism can enhance the confidence for the model's judgement.

In this paper, we make full use of the features of these two structures. Traditional attention modules used in segmentation and salient object detection tasks usually apply the processed feature maps themselves as the input of the attention modules. In such a structure, it is easy to realize two kinds of favorable operations, reinforcing 'true positive' information and weakening 'false negative' information, as well as generate two kinds of faulty operations, reinforcing 'false positive' information and weakening 'true negative' information. This is a problem

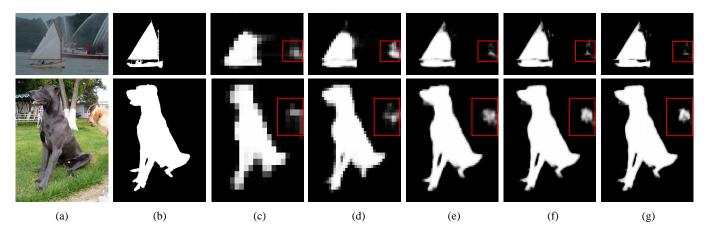


Fig. 1. Some examples of output in different layers. (a) Input image; (b) ground truth; (c) output of layer 1; (d) output of layer 2; (e) output of layer 3; (f) output of layer 4 (g) output of layer 5. The area in the red box is where the output is misjudged. For the first line, output of the shallower layers makes fewer mistakes. For the second layers, output of the shallower layers makes more mistakes.

we call the 'blind overconfidence' of the attention module judgement. To solve this problem, we take the deeper layer's low-resolution output as the input of the shallower layer's attention module to establish a new output-guided attention module. Compared with the ordinary self-attention module, taking the deeper layers' output as the input of the attention module can integrate the advantage of each layer of the neural network, thus preventing the attention module of decoder in one layer from enlarging the false information caused by this layer. Considering that different input feature maps have different importance in attention module processing for different input images, in order to reasonably select multiple input of the attention module, we make the network learn a set of weights. And all the input features maps are weighted before fed into the attention module. Based on the output-guided module, we built a new salient object detection model applying the classic encoder-decoder structure. Our model gains remarkable detection effects with a small memory footprint and fast detection speed. Moreover, with the enlightenment of the features of outputs in different layers, we propose a method to identify regions that are difficult to estimate in images of the training set. On this basis, we propose the intractable area F-measure loss function, which can pay more attention to the areas that are difficult to judge in the image. The main contributions of this paper can be summarized as follows:

- We propose a new output-guided attention module built with outputs in various positions of an neural network, which can overcome the shortcomings of many other self attention modules.
- We propose a new end-to-end neural network for salient object detection applying the output-guided attention module.
- We propose an intractable area loss function based on the features of the multi-output structure. The introduction of this loss function makes the model more effective facing complicated images.

The rest of this paper is organized as follows: in Section II, we make a conclusion about the existing classic salient object detection models and attention modules; in Section III,

we introduce the output-guided attention module, OGNet and intractable area F-measure loss; in Section IV, we demonstrate our experimental results during the research process. Finally, we conclude this paper in Section V.

#### II. RELATED WORK

## A. Salient Object Detection

In the early stages of development, salient object detection models were usually based on low-level hand-crafted features, such as color features [8] and textural features [11], [37], [38]. Although these models generated certain effects, their performances were not ideal for images with complicated backgrounds or complex salient objects. When making saliency judgments, the human eyes always confront complicated elements, while in traditional methods, fully considering and integrating various factors are difficult. Deep learning explores a new route for the research of salient object detection. Early salient object detection models based on deep learning usually select the convolutional layer - fully connected layer structure, which is the same as most image classification models. Wang et al. [16] proposed two neural networks to detect salient objects, one for learning local patch features to determine the saliency value of each pixel and the other for predicting the saliency score of each object region based on global features. Li and Yu [39] first segmented the image into several areas and then formulated a neural network with some branches to train these areas. Then, their method utilizes several convolutional layers to connect them together to achieve information integration among different layers. Zhao et al. [40] built a multicontext deep learning framework with two branches that extract global context and local context and then integrate them together. After the appearance of fully convolutional networks, many salient object detection models based on deep learning adopted the encoder-decoder structure represented by FCN and then made some adjustments to that structure. Liu and Han [14] made use of the hierarchical recurrent convolution to build up the decoder part of the neural network. Zhang et al. [17] applied the reformulated dropout to some convolutional layers on the strength of the basic

encoder-decoder structure to extract the salient information more conveniently. However, due to the inadequate use of different levels of information, it is difficult to achieve very good performance in a simple encoder-decoder. Hou *et al.* [33] utilized a large number of short connections to join the decoders in different layers together and drew on the idea of DenseNet [41], which worked to realize the full integration of information in different layers. Similarly, Zhang *et al.* [18] proposed the Amulet. Wang *et al.* [42] proposed a multistage structure and used pyramid pooling in the joint part to obtain and merge the information from different layers together. Such methods usually perform well. However, due to a demand to connect feature maps in different layers, they often need to consume a large amount of memory and require a large amount of computation.

#### B. Attention Mechansim

During deep learning, the attention mechanism was applied to the field of machine translation [43], [44] at the earliest stages and accomplished outstanding effects. Then, it is applied to the neural network models of computer vision. For the past few years, many models applying attention mechanisms have greatly improved the effects in image classification [45], semantic segmentation [46], [47], action recognition [48] and other fields. The core ideology of the attention mechanism is to selectively enhance or weaken the large amount of information constructed by neural networks. The attention module of SENet proposed by Hu et al. [45] includes two processes: squeeze and excitation. The squeeze process applies global average pooling to compress the feature maps, and the excitation process utilizes two fully connected layers to obtain a series of weights, which are used to weigh feature maps from channels. This method improves the accuracy of image classification models immensely. Furthermore, CBAM proposed by Woo et al. [46] expands the attention mechanisms treatment dimension from the channel dimension of SENet to two dimensions channel and spatial and selects both the average value and maximum value to compress the feature maps, which further increases the effects of the attention module. The structures of SENet and CBAM can expand to many other computer vision task models. In recent years, many salient object detection models have also utilized various kinds of attention modules. The module proposed by Zhang et al. builds two attention modules from the channel and spatial layers, which is similar to the establishment method of CBAM. Liu et al. [31] applied a convolution and bidirectional LSTM to formulate local pixelwise attention and global pixelwise attention, which enlarges the receptive field to reduce mistakes. All the attention modules mentioned above only use the processed feature maps themselves as the input, which is the main difference between the proposed method and other methds for salient object detection only using self attention [30], [31].

#### III. METHODOLOGY

In this section, we introduce our proposed methods. The output-guided attention module is introduced in Subsection

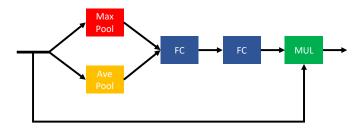


Fig. 2. The structure of channel attention. FC is the fully connected layer. MUL is a multiplication operation.

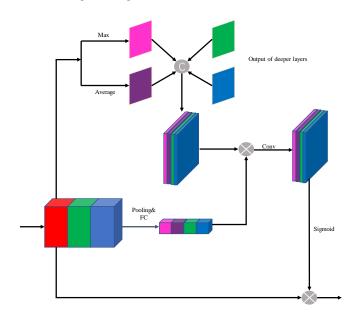


Fig. 3. The structure of spatial attention. CAT is the concatenation of some groups of feature maps from channels.

III-A and the complete structure of the output-guided network (OGNet) is shown in Subsection III-B. The intractable area F-measure loss and the training method are shown in Subsection III-C and Subsection III-D, respectively.

#### A. Output-guided Attention Module

1) Blind Overconfidence: At present, attention mechanisms composed of both spatial attention and channel attention, represented by CBAM [49], is one of the most popular attention modules used in various kinds of computer vision models. CBAM builds up two attention modules - channel attention and spatial attention, taking the processed feature maps themselves as input. The effects of models can be greatly improved through these two attention modules. Such a module is very suitable for image classification because image classification does not concern the shape and location of objects in an image; thus, the enhancement of incorrect information caused by attention modules will not have a great impact on the final judgment. However, we think that this kind of spatial attention, which only takes processed feature maps as input, faces some problems when applied to salient object detection. Salient object detection aims to classify each pixel in an image into two categories, which means that the final saliency map is

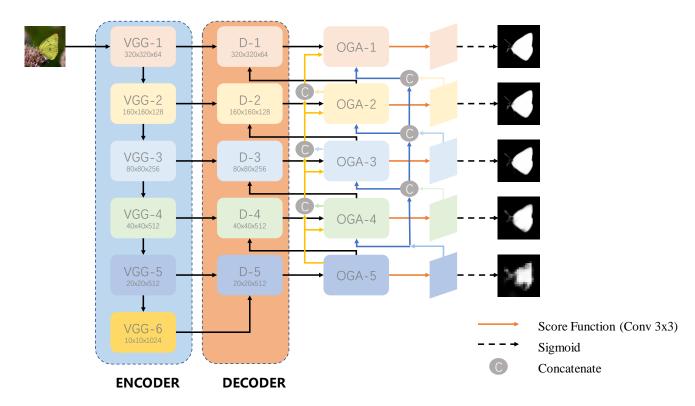


Fig. 4. The detailed structure of the output-guided model.

binary. Assuming that a pixel in an image is salient, the value of the corresponding position in the feature maps should be large. If a layer of the neural networks misjudges the pixel, the following attention module will greatly magnify the wrong information, which is the problem of 'blind overconfidence' in attention modules. An ideal attention module should magnify the correct information and avoid the enhancement of wrong information. A good resolution is to enlarge the receptive field of the attention module to capture more information. However, this requires a lot of computation. To solve the problem of blind overconfidence, we built a new attention module. Similar to CBAM, it consists of both channel attention and spatial attention. The structures of channel attention module and spatial attention module are shown in Fig. 2 and Fig. 3, respectively.

2) Channel Attention: In a layer of neural networks, not all feature maps have the same significance. The channel attention module is a feature detector that can enhance information in useful feature maps and reduce information in useless feature maps. If we adopt the whole feature map as the input of the attention module, the computation will be quite large, which violates the design principle that the attention module should be lightweight. Thus, we should find a method whose receptive field is large enough to express the global feature of a feature map. Similar to CBAM, we use both max pooling and average pooling to demonstrate the global feature of the input feature map  $\mathbf{F} \in \mathbb{R}^{C \times W \times H}$ . The channel attention map can be calculated as follows:

$$\mathbf{W_c} = Sigmoid(L_2(L_1(GMP(\mathbf{F}))) + L_2(L_1(GAP(\mathbf{F}))))$$
(1)

where GMP is global max pooling and GAP is global average pooling. The input size of  $L_1$  and the output size of  $L_2$  are C. The input size of  $L_2$  and the output size of  $L_1$  are C/4. This setup is designed to deepen the network to extract more information with reducing the additional memory footprints caused by these two fully connected layers.  $L_1$  is followed by a rectified linear unit (ReLU) [50]. Note that  $L_1$  and  $L_2$  are shared for feature maps after max pooling and average pooling. Sigmoid is a function used to get the attention map. For each resolution  $\mathbf{X}_{\mathbf{c}}^i$  in a feature map processed by two fully networks:

$$\mathbf{W}_{\mathbf{c}}^{i} = \frac{1}{1 + e^{-\mathbf{X}_{\mathbf{c}}^{i}}} \tag{2}$$

3) Spatial Attention: Spatial attention is used to enhance the confidence of the model on its judgement. In salient object detection, the use of spatial attention can also make a model focus on the foreground region, which is beneficial for saliency prediction. Unlike CBAM, apart from taking the average and maximum value from channels, outputs from other layers are also taken as the input. We think that taking the outputs of other layers into the attention module is a kind of balance and compensation, which can avoid the enhancement of wrong information caused by the attention module in one layer. Beyond that, this structure can also be regarded as a special form of short connection, which can make full use of information from different layers and make the deep neural network easier to train. For feature maps from the decoder in layer m, we obtain two feature maps,  $\mathbf{F}_m^{max}$  and  $\mathbf{F}_m^{min}$ , which express the comprehensive information of all layers by calculating the maximum and average values on the channel. For layer m, the input of the attention module is  $\{\mathbf{F}_m^{max}, \mathbf{F}_m^{min}, \mathbf{O}_{m+1}, ..., \mathbf{O}_M\}$ , where  $\mathbf{O}_i$  is the output of the i<sup>th</sup> layer.

A straightforward idea is to input these maps directly into some convolution layers to obtain spatial attention weight. However, this approach regards all maps as having the same importance and ignores the differences between them. As shown in Fig. 1, for the first column, output maps in shallower layers make fewer errors in saliency judgments for the area in the red box. Thus, when fed into the attention module, shallower output should be more important. However, for the second column where outputs in deeper layers judge better, deeper output should be emphasized. Thus, weighting these maps before feeding them into the spatial attention module is necessary. The weight is also obtained from the neural network. We first concatenate  $\mathbf{F}_m$  with output of output-guided attention modules of all deeper layers  $\{\mathbf{OG}_{m-1}, \mathbf{OG}_{m-2}, ..., \mathbf{OG}_{M}\}$ to get a feature map C. C passes through two fully connected layers which are similar to that in channel attention module and obtain a vetor V with M-m dimensions. Finally, the spatial attention map can be generated as follows:

$$\mathbf{W_s} = Sigmoid(f^{7\times7}(\mathbf{V}.CAT(\mathbf{F}_m^{max}, \mathbf{F}_{min}^m, \mathbf{O}_{m-1}, ..., \mathbf{O}_M)))$$
(3)

where  $f^{7\times7}$  is a  $7\times7$  convolution layer. The size of the convolution layer is bigger than the usual one which is  $3\times3$  because the receptive field should be large enough to fully extract pixel relationship for the sptial attention. CAT is the concatenation of feature maps from channels.

When utilizing the attention module, we let the processed feature maps pass the channel attention module first and then pass the spatial attention module to obtain the final output of the output-guided attention module. The output can be obtained as follows:

$$\mathbf{F}_{out} = \mathbf{W}_s.\mathbf{W}_c.\mathbf{F}_{in} \tag{4}$$

This arrangement that passes the channel attention module first is also inspired by [49], which argues that channel-first is slightly better than spatial-first.

#### B. OGNet

Based on the output-guided attention module introduced above, we propose a new model for the salient object detection: output-guided model (OGNet). Our model strengthens the basic encoder-decoder structure. To compare fairly with most salient object detection models, we choose the most commonly used VGG16 [51] as the backbone of the encoder. Note that, similar to most models, the backbone can be flexibly selected and can be replaced by other networks, such as ResNet [52] and Xception [53].

Our model's decoder contains five layers so that five saliency maps with different resolutions are gained. Each layer of the decoder has the same structure. The structure of the decoder is shown in Fig. 5 and details of each convolution are shown in Table I. The ith layer of the decoder takes the output of the encoder in the same layer and the output of the previous layer's decoder as input. First, the decoder feature maps are bilinearly upsampled by a factor of 2, and

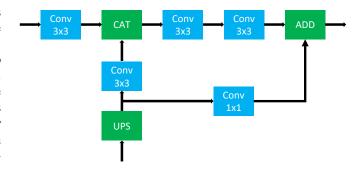


Fig. 5. The detailed structure of a layer of the decoder. Feature maps from encoder and decoder are input from the left and bottom, respectively. UPS is the bilinearly upsamping.

# TABLE I STRUCTURE OF EACH LAYER OF THE DECODER. CONV\_E, CONV\_D REPRESENT THE CONVOLUTION LAYERS PROCESSING INPUT FROM

ENCODER AND DECODER. CONV\_1 AND CONV\_2 REFER TO TWO CONVOLUTION LAYERS AFTER THE CONCATENATION.

No. Layer	conv_e	conv_d	conv_1	conv_2
1	$3 \times 3, 128 \\ 3 \times 3, 128$	$3 \times 3, 128 \\ 3 \times 3, 128$	$3 \times 3,256 \\ 3 \times 3,256$	$3 \times 3,256 \\ 3 \times 3,256$
3	$3 \times 3,128$ $3 \times 3,64$	$3 \times 3,128$ $3 \times 3,64$	$3 \times 3, 230$ $3 \times 3, 128$	$3 \times 3,230 \\ 3 \times 3,128$
4 5	$3 \times 3, 32 \\ 3 \times 3, 32$	$3 \times 3, 32 \\ 3 \times 3, 32$	$3 \times 3, 64 \\ 3 \times 3, 64$	$3 \times 3, 64 \\ 3 \times 3, 64$

then two  $3 \times 3$  convolutions are applied on feature maps from the encoder and decoder separately. Note that we do not use deconvolution directly because bilinearly upsampling performs slightly better than deconvolution. Inspired by [33], we tried to use a larger-sized convolution such as  $7 \times 7$ and  $5 \times 5$  to process feature maps from encoder but found that it could not improve performance but instead caused overfitting. We performed some experiments to find the most suitable convolution size, and the results are shown in Section 4.3. The encoder feature maps and decoder feature maps are concatenated, and another two  $3 \times 3$  convolutions are applied to further fuse and extract information from the feature maps. Inspired by the structure of ResNet [52], a residual block is applied to construct the decoder. For each layer of the decoder, we apply a  $1 \times 1$  convolution to convert the feature map which has been bilinearly upsampled to the same number of channels as the output of the decoder in this layer. Then this feature map is added to the output of the decoder to obtain the final output. Section 4.3 shows the comparison between the performance of models using residual blocks and not using residual blocks.

The output of every layer of the decoder passes an output-guided attention module, which is the input of the decoder in the next layer, as well as passes a  $3 \times 3$  convolution and Sigmoid function to obtain this layer's output saliency map. Note that the inputs of the output-guided attention module are the saliency maps that have not passed the Sigmoid function. All convolutions in the decoder are followed by a batch normalization and ReLU. The structure of the output-guided model is shown in Fig. 4.

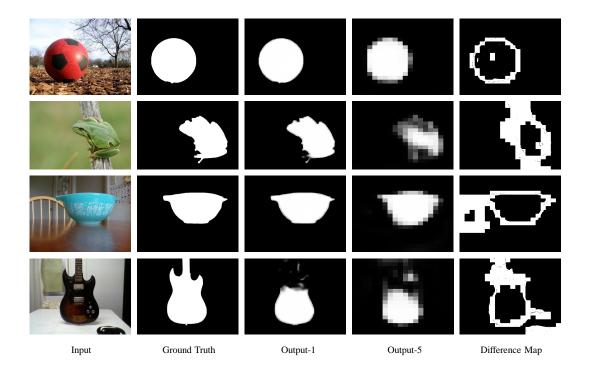


Fig. 6. Some examples of multi output and the difference map. The output of different layers in the network is different in some areas. From the difference maps, we can find that the different areas are usually the boundary of the objects or where the disturbing objects are located.

#### C. Intractable Area F-measure Loss

We observe that in multioutput encoder-decoder neural networks, outputs from different positions with different resolutions have different characteristics. Generally speaking, taking the deeply supervised multioutput network as an example, deeper outputs with low resolutions can capture semantic information better while shallower outputs with high resolutions concerns more on the spatial features. Some examples of outputs from different positions are shown in Fig. 1. As can be seen from Fig. 1, first, high-resolution output saliency maps are more precise than low-resolution maps at the boundary of objects: second, there are some interference objects that are easily misjudged and different outputs make different saliency judgments on them. Both the object boundary and interference objects are difficult points to improve the detection accuracy. The judgement ability in these areas is always a significant factor affecting the performance of a salient object detection model.

Thus, we propose a new loss function to promote the model's performance in these areas. We need to find the intractable areas of images in the training set. First, we apply another dataset with fewer images to train the model for fewer iterations, and the training result is rough. Then, we test images in the training set utilizing the roughly trained model, and some saliency maps with different resolutions are obtained. For input image I, there are five output saliency maps  $S_i, i \in \{1, 2, 3, 4, 5\}$ . We apply  $S_1$  with the largest resolution and  $S_5$  with the smallest resolution to calculate the difference map based on the observation that difference between high-resolution maps can only show the boundary area but fail to get the intractable area such as the disturbing objects. First, we

bilinearly upsampled these two saliency maps to the resolution of the original image and obtain  $S_1^{'}$  and  $S_5^{'}$ . Then the different areas can be obtained by the pixel-level comparison between  $S_1^{'}$  and  $S_5^{'}$  and the coordinate set C of the different areas can be calculated as follows, for all coordinates (i,j) in  $S_1^{'}$  and  $S_5^{'}$ :

$$\begin{cases} (i,j) \in C & \text{if} \quad S_{1}^{'}(i,j) - S_{5}^{'}(i,j) = 0\\ (i,j) \notin C & \text{if} \quad S_{1}^{'}(i,j) - S_{5}^{'}(i,j) \neq 0 \end{cases}$$
 (5)

After getting the different maps, we train the model for the second time. For the second training, the saliency score is binarized, and the intractable area F-measure loss is calculated as follows:

$$L_f = 1 - \frac{(1+\beta^2) \times P_c \times R_c}{\beta^2 \times P_c + R_c} \tag{6}$$

where  $P_c$  and  $R_c$  represent the precision and recall of area C. The formula of intractable area F-measure loss equals to 1 minus the F-measure of area C. The effectiveness can be understood from two aspects. On the one hand, the loss function is designed directly according to the evaluation metric, which is proved to be useful to promote the test results in a lot of computer vision tasks such as object detection [54] and semantic segmentation [55]; On the other hand, the IAF loss is only calculated on the intractable areas, thus promoting the model to process these areas more effectively and enhancing the generalization ability of the model in dealing with complex images.

Note that, the second training is **not** the fine-tuning of the model gained by the first training. The only purpose of the first training is to obtain the difference maps for the training

TABLE II

QUANTITATIVE COMPARISON OF MAE, F-MEASURE AND S-MEASURE WITH 15 METHODS ON 5 DATASETS. A HIGHER F-MEASURE SCORE, HIGHER S-MEASURE SCORE AND LOWER MAE SCORE REPRESENT BETTER PERFORMANCE. THE TOP THREE RESULTS ARE HIGHLIGHTED IN RED, GREEN AND BLUE, RESPECTIVELY.

	Datasets		HKU-IS			ECSSD			SOD		DU	JT-OMR	ON		DUTS-TE	Ξ
Methods		MAE	$\mathcal{F}$	$\mathcal{S}$	MAE	$\mathcal{F}$	S									
LEGS	CVPR2015	0.119	0.732	0.742	0.019	0.75	0.786	0.195	0.683	0.658	0.133	0.592	0.714	0.138	0.585	0.696
MDF	CVPR2015	0.096	0.801	0.810	0.105	0.807	0.776	0.164	0.721	0.674	0.092	0.611	0.721	0.092	0.644	0.728
ELD	CVPR2016	0.074	0.769	0.868	0.080	0.810	0.841	0.155	0.712	0.705	0.092	0.611	0.751	0.098	0.628	0.754
DCL	CVPR2016	0.075	0.820	0.877	0.137	0.736	0.868	0.198	0.641	0.747	0.157	0.575	0.771	0.150	0.606	0.796
KSR	ECCV2016	0.120	0.747	0752	0.135	0.782	0.763	-	-	-	0.131	0.591	0.722	0.121	0.602	0.715
RFCN	ECCV2016	0.089	0.835	0.859	0.107	0.834	0.852	0.169	0.743	0.794	0.111	0.627	0.764	0.090	0.712	0.859
DHS	ECCV2016	0.054	0.806	0.870	0.060	0.841	0.884	0.133	0.686	0.749	-	-	-	0.065	0.698	0.818
NLDF	CVPR2017	0.048	0.838	0.879	0.063	0.839	0.875	0.130	0.708	0.889	0.080	0.634	0.770	0.066	0.710	0.816
Amulet	ICCV2017	0.052	0.813	0.886	0.059	0.841	0.894	0.140	0.755	0.757	0.098	0.626	0.780	0.085	0.657	0.804
SRM	ICCV2017	0.046	0.874	0.887	0.056	0.892	0.895	0.132	0.671	0.741	0.069	0.707	0.798	0.059	0.678	0.836
UCF	ICCV2017	0.062	0.823	0.875	0.069	0.852	0.883	0.169	0.644	0.753	0.120	0.628	0.760	0.117	0.588	0.782
PAGRN	CVPR2018	0.048	0.886	0.887	0.064	0.891	0.889	-	-	-	0.072	0.711	0.775	0.055	0.788	0.838
PICA	CVPR2018	0.042	0.847	0.905	0.047	0.865	0.916	0.108	0.721	0.776	0.068	0.691	0.825	0.054	0.748	0.863
C2S	ECCV2018	0.046	0.848	0.889	0.057	0.860	0.896	0.122	0.702	0.760	0.072	0.698	0.799	0.062	0.686	0.831
RA	ECCV2018	0.045	0.913	0.887	0.059	0.896	0.893	0.124	0.709	0.764	0.062	0.701	0.814	0.059	0.723	0.839
Ours		0.041	0.916	0.909	0.047	0.916	0.903	0.114	0.863	0.815	0.066	0.743	0.833	0.047	0.807	0.884

set in the second training. When testing, only the model from the second training is applied to obtain saliency maps. Thus, our proposed method is end-to-end when testing though the training involves two processes.

# D. Training

Suppose that the multioutput neural networks can be divided into M layers and that every layer of the decoder generates an output. Every output can produce a loss term. The final loss function can be defined as:

$$L(I, G, W, w) = \beta l_f(I, G, W, w^{(1)}) + \sum_{m=1}^{M} \alpha_m l_{side}^m(I, G, W, w^{(M)})$$
(7)

where  $\alpha_m$  is the weight of the cross-entropy loss in the  $m^{th}$  layer and  $\beta$  is the weight of intractable area F-measure loss. I and G represent the input image and its ground truth. Each output is obtained by a separate score function  $\mathbf{w}^{(m)}$ , and  $\mathbf{w}$  refers to the set of all score fuctions:

$$\mathbf{w} = (\mathbf{w}^1, \mathbf{w}^2, ..., \mathbf{w}^M)$$
 (8)

Here,  $l_f(I,G,\mathbf{W},\mathbf{w}^{(1)})$  represents the intractable area F-measure loss function, and  $l^m_{side}(I,G,\mathbf{W},\mathbf{w}^{(M)})$  refers to the

cross-entropy loss function of the  $m^{th}$  output and can be calculated as follows:

$$l_{side}^{m}(I, G, \mathbf{W}, \mathbf{w}^{(m)}) = -\sum_{z=1}^{|I|} G(z) log P(G(z) = 1 | I(z), \mathbf{W}, \mathbf{w}^{m})$$
$$-\sum_{z=1}^{|I|} (1 - G(z)) log P(G(z) = 0 | I(z), \mathbf{W}, \mathbf{w}^{m})$$
(9)

In the output-guided network, M equals 5 so that 5 outputs are gained. Instead of fusing these outputs as in [33] by adding additional computing, we directly apply the output of the first layer, which has the highest resolution, as our final saliency score. Considering that the output of the first layer has the highest importance,  $\alpha_1$  is set higher than others, the weights of all the loss functions are:

$$\{\alpha_1, \alpha_2, \alpha_3, \alpha_4, \alpha_5, \beta\} = \{50, 4, 4, 4, 4, 25\}$$
 (10)

#### IV. EXPERIMENTAL RESULTS

### A. Implementation Details

We use the PyTorch framework to train and test our model. All images are resized to  $320 \times 320$  pixels for training and testing. We select SGD with a weight decay of 0.0005 and a momentum of 0.9 as the optimizer. Inspired by [56], we use the 'poly' policy to set the learning rate. For an iteration, its learning rate equals the initial learning rate multiplied by  $(1-\frac{iter}{maxiter})^{power}$ , where the initial learning rate is set to

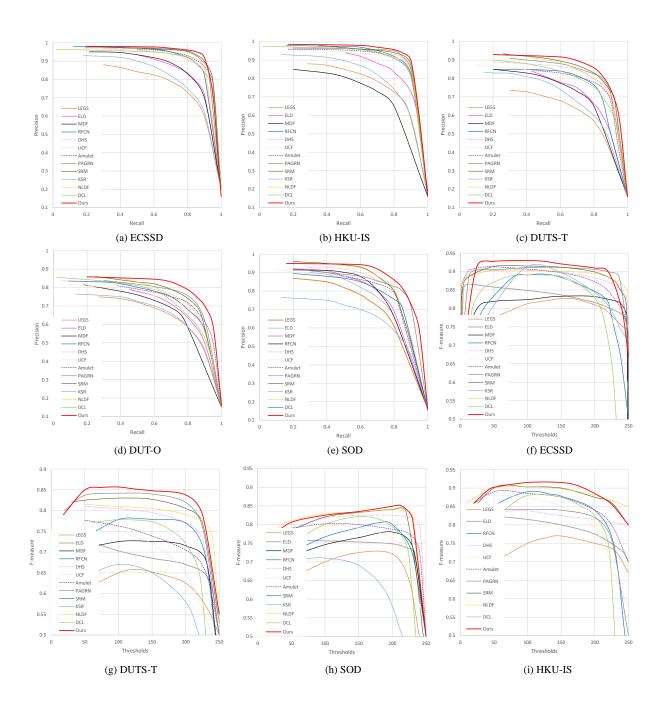


Fig. 7. (a)-(e) are P-R curves on various datasets, including ECSSD, HKU-IS, DUTS-T, DUT-O and SOD. (f)-(i) are F-measure curves on various datasets, including ECSSD, DUTS-T, SOD and HKU-IS.

0.0001 and power is set to 0.9. Due to the use of IAF loss, the model needs to go through two separate training processes, for which we used the same parameter configurations. It takes approximately 21 hours to train 40 epochs on a sever with an NVIDIA Titan X GPU (with 12G memory).

#### B. Datasets and Evaluation Metrics

1) Datasets: Six datasets are used to train and test our models: MSRB [58], DUTS [59], ECSSD [11], DUT-OMRON [60], HKU-IS [39], and SOD [61].

MSRB: This dataset contains 5000 high quality images with

high precision marks. These images are abundant in species, but their backgrounds are usually simple.

**DUTS**: This dataset includes 10553 images for training and 5019 images for testing. This datasets images are characterized by a large quantity of abundant species and high marked quality.

**ECSSD**: This dataset contains 1000 images with a complex background, and the ground truth of the image in the dataset usually contains very rich semantic information.

**DUT-OMRON**: This dataset contains 5168 high quality images. The images of this dataset include one or more salient

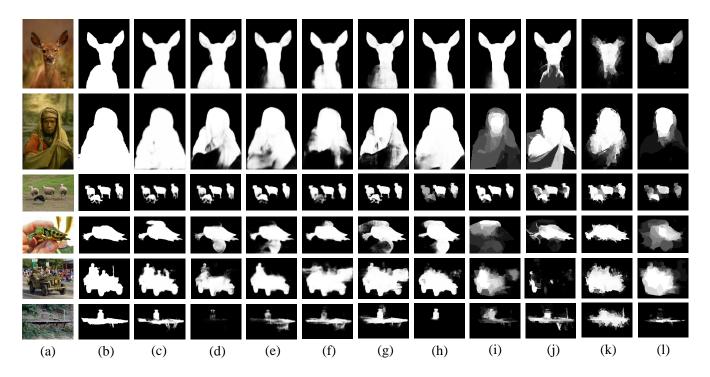


Fig. 8. Visual comparison with 9 state-of-the-art methods. (a) Input image; (b) ground truth; (c) ours; (d) PAGRN [30]; (e) SRM [42]; (f) Amulet [18]; (g) UCF [17]; (h) NLDF [19]; (i) KSR [57]; (j) MDF [39]; (k) ELD [20]; (l) LEGS [16]. Our method performs best for images with various characteristics.

objects and their backgrounds are very complicated. It is relatively more difficult to achieve salient object detection on these images. Hence, it is a significant dataset to determine whether a salient object detection model can perform well for complex images.

**HKU-IS**: This dataset contains 4447 high-precision labeled images. Images in the dataset are often equipped with many salient objects, and some of these salient objects are located at the edge of the images, which brings a great challenge to salient object detection.

**SOD**: This dataset contains 300 images. These images' background and the shape of the salient objects are quite complex. It is a very challenging dataset.

We use MSRB to train our model for the first time and then use this model to test the training set of DUTS and obtain the difference maps. Then, the training set of DUTS and the difference maps are used for the second training to obtain the final model. The test set of DUTS and other datasets are used to test the model.

2) Evaluation Metrics: We utilize four methods that are extensively applied in the salient object detection field to test our models performance on test sets: precision-recall (PR) curves, F-measure and mean absolute error (MAE) and S-measure. The saliency maps are binarized by varying the threshold from 0 to 255, and pairs of precision and recall under different thresholds are computed to plot the PR curve. Then, the saliency map is binarized with a fixed threshold, which is determined as twice the mean saliency value of the saliency map. The F-measure is calculated as follows:

$$F_{\beta} = \frac{(1+\beta^2) \times Precision \times Recall}{\beta^2 \times Precision + Recall}$$
 (11)

Similar to most other methods [17], [18], [42], we set  $\beta^2$  to 0.3, making the precision's influence factors larger than that of the recall.

Due to the binarization of the saliency map, the F-measure cannot directly measure the difference between the ground truth and the saliency map obtained by the model. Hence, we also apply MAE, which values the average pixelwise absolute difference between the saliency map and binary ground truth:

MAE = 
$$\frac{1}{W \times H} \sum_{x=1}^{W} \sum_{y=1}^{H} |S(x,y) - G(x,y)|$$
 (12)

where W and H are the width and height of the saliency map S, respectively.

Structure measure (S-measure) [62] is a new evaluation metric to evaluate region-aware and object-aware structural similarity between saliency maps and ground truth maps. It can be calculated as follows:

$$S = \alpha * S_o + (1 - \alpha) * S_r \tag{13}$$

where  $S_o$  and  $S_r$  represent object-aware and region-aware structural similarity, respectively.  $\alpha$  is set to 0.5. A model with a higher F-measure score ,lower MAE score and higher S-measure score has better performance.

#### C. Performance Comparison

We compare 15 state-of-the-art classic salient object detection methods, including LEGS [16], ELD [20], MDF [39], KSR [57], DCL [22], RFCN [21], NLDF [19], DHS [14], UCF [17], Amulet [18], PAGRN [30], SRM [42], C2S [63], RA [15] and PICA [31]. Most of these methods are based on deep learning.

- 1) Qualitative Evaluation: Fig. 8 displays the visual comparison between our method and the others. Our method can judge the salient object better and more accurately display the area of the salient object. Our method performs much better than the other methods in the following challenging situations:
- (1) When confronting multiple salient objects in an image, our method makes more accurate decisions on multiple salient objects. As shown in the third line, our method precisely judges all four salient objects.
- (2) When the shape of the salient object is complicated, our algorithm still demonstrates the shape of the salient object obviously and favorably. As shown in the fifth line, although the salient object's upper edge contour is quite complex and the rough sketch feature is quite blurry, our method precisely recovers the rough sketch of the salient object and do not generate an erroneous judgement.
- (3) Thanks to the introduction of the attention mechanism, when salient objects are surrounded by some interferential factors disturbing the salient judgement, our method is able to perform better and had strong antijamming ability. For example, in the first line, the lower left quarter of the salient object is highly similar to the surrounding areas, so it is quite easy to cause an erroneous judgement. Our method makes a very accurate judgement, while most of the other state-of-theart methods incorrectly judge that area as nonsalient.
- 2) Quantitative Evaluation: The PR curves and F-measure curves are shown in Fig. 7. For a PR curve, a higher precision and slower attenuation represents a better performance. Compared with the other methods, our method has the best performance on all the datasets.

In Table II, we also compare our method with the state-of-the-art methods in terms of MAE, F-measure and S-measure. For MAE score, we obtained the best performance on most of the datasets. Although we did not realize the best performance on SOD and DUT-OMRON on MAE, our method demonstrates high competition. For the F-measure score, our method performs the best on all the datasets. Compared with the second-ranked method, our method improves the F-measures score by 4.8%, 2.7%, 14.3%, 4.5% and 2.4% on HKU-IS, ECSSD, SOD, DUT-OMRON and DUTS-TE, respectively. For S-measure score, our method performs best on three datasets and and ranks second on another two datasets.

Based on the indexes being synthesized, in comparison to the other state-of-the-art methods, our method shows the best performance overall. The excellent execution on all datasets demonstrates that our method possesses stronger universality.

3) Memory Comparison: The algorithms based on deep learning usually require a large computation and memory footprint. In general, a deeper neural network can gain better performance, but it is also followed by a larger memory footprint and computation so it is difficult to apply the model to real-time detection and to use it on mobile terminals, which reduces the practicability. Hence, the size of the neural network model is also one of the significant factors when measuring a salient object detection algorithm based on deep learning. Fig. 9 shows some methods' model size and F-measure on ECSSD. The model size of many methods is very large, while those with smaller model sizes usually have a

TABLE III
ABLATION EXPERIMENTS ON DUTS AND SOD.

No.	settings	SOD	DUTS
(a)	Comparison of attention module		
1	basline	0.13082	0.05323
2	+SE	0.12433	0.05232
3	+CBAM	0.12234	0.05185
4	+OGAM	0.11619	0.04895
(b)	Comparison of IAF loss		
5	baseline(No.4 setting)	0.11619	0.04895
6	BCE loss+IAF loss	0.11362	0.04658
(c)	Comparison of residual blocks		
7	baseline(No.4 setting)	0.11619	0.04895
8	without residual block	0.11794	0.04978

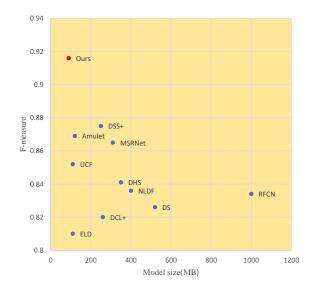


Fig. 9. Memory comparison with some methods, including Amulet[], DSS+[33], MSRNet [64], UCF [17], DHS [14], NLDF [19], RFCN [21], DS [65], DCL+ [22] and ELD [20].

general effect. Our method is the only one with a model size of less than 100 MB and an F-measure score higher than 0.9. Our model is lightweight but very effective.

#### D. Ablation Studies

1) Evaluation of output-guided attention: As shown in table III, to verify the effect of the output-guided attention module, we compare the effects of models with and without the output-guided attention module. The experimental results show that the output-guided attention module can greatly improve the model's effect. In addition, we also test the effects of some other types of attention modules on model improvement. Two attention modules are tested: SE [45] and CBAM [49]. Different from the output-guided attention module, SE only uses channel attention, and both SE and CBAM only take the processed feature maps themselves as input. The experimental results show that the effect of using SE alone is not significant enough and CBAM utilizing both channel attention and spatial attention can produce better effects. Our output-guided

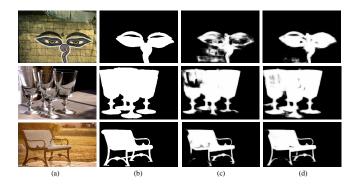


Fig. 10. Comparison between output obtained by models applying IAF loss and not applying IAF loss. (a) Input image; (b) ground truth; (c) output of model not applying IAF loss; (d) output of model applying IAF loss.

attention module performs best among these three kinds of attention modules.

- 2) Evaluation of intractable area F-measure loss: The intractable area F-measure loss is used to upgrade the model's judgement ability when encountering difficult areas. To test its effectiveness, we test the performance of models trained applying the intractable area F-measure loss and not applying the intractable F-measure loss on SOD and DUTS-TE, respectively. As shown in Fig. 10, comparing the test results of the two models, the model performs better in the marginal areas of the salient objects and makes a more precise judgment of the difficult areas after utilizing the intractable area F-measure loss. Quantitative analysis is shown in Table III. After utilizing the intractable F-measure loss, the MAE score on both datasets decline.
- 3) Ablation of residual block: Residual blocks can make a very deep neural network easier to train and improve the effect of the neural network. To test the residual blocks' influence on our model, we eliminate the original residual block of the model and then test its performance. The experimental results are shown in Table III. Observing the training process, we find that the model with the residual blocks converged faster and that the final loss value was smaller. The application of the residual blocks slightly raised the model's effects. Thus, we deemed that the utilization of residual blocks in our model causes overfitting.
- 4) Selection of convolution size: The choice of convolution size has a great influence on the performance of convolutional neural networks. DSS [33] uses a large convolution to process the feature maps extracted from the encoder in every layer. Theoretically, a large convolution can increase the receptive field and extract more semantic information, so it is used to process feature maps from encoders that do not sufficiently extract semantic information compared with those from decoders. Inspired by DSS, we first choose a convolution of size  $7 \times 7$  but find that the performance of the model unexpectedly became worse. To determine the most suitable convolution size, we test the convolution of four sizes:  $7 \times 7$ ,  $5 \times 5$ ,  $3 \times 3$ ,  $1 \times 1$ . The performance of these models on five datasets are shown in Fig. 11. The lowest MAE score on all five datasets is achieved by the model with a  $3 \times 3$  convolution. The receptive field of a  $1 \times 1$  convolution is too small to integrate

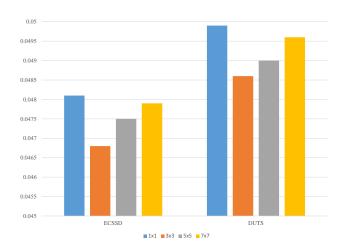


Fig. 11. Comparison of MAE score of four sizes of convolution on ECSSD and DUTS.

the information extracted from the encoder and overfitting is caused by a  $5 \times 5$  convolution and a  $7 \times 7$  convolution, which are too large. Feature maps extracted from encoders are mainly used to better restore the shape of salient objects, so spatial information is more important than semantic information. A large convolution may destroy the spatial information, which is harmful for the accurate display of salient objects.

5) Application of output-guided attention in the other models: The output-guided attention module proposed by us in this paper is a lightweight and universal module that can be used in all multioutput models. We test the effect of the output-guided attention module on some other multioutput models. DSS [33] is a classic salient object detection model with multiple outputs. The original DSS uses two convolutional layers to process each side output, and we add an output-guided attention module after the first convolutional layer. The experimental results are shown in Fig. 12, where the MAE score of the five datasets between the original DSS and the DSS with the output-guided attention module are compared. Compared with the original model, the MAE score of the five datasets after using the output-guided attention module decreases by 8.9%, 4.6%, 8.0%, 4.2%, and 5.1%, respectively.

# V. CONCLUSION

In this paper, we proposed a new output-guided attention module. Experimental results show that compared with other attention modules, the output-guided attention module constructed by the processed feature maps themselves and other resolution outputs can reduce errors and achieve better performance. Our proposed model, based on output-guided attention, showed outstanding performance on multiple datasets. Owing to the output-guided attention module, our model has stronger robustness. The proposed intractable area F-measure loss can effectively improve the performance of the model when facing images with complex backgrounds and salient objects with complicated shapes. The improvements of the output-guided attention module and intractable area F-measure loss on other multioutput methods demonstrate that these two methods are

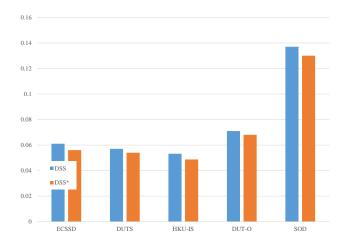


Fig. 12. Comparison of MAE scores on five datasets of the original DSS and DSS\*(DSS applying output-guided attention module and IAF loss.)

universal. We suggest that researchers try to use the outputguided attention module and intractable area F-measure loss when constructing other neural networks for salient object detection. We believe that blind overconfidence is a common problem faced by many attention modules in salient object detection and that the output-guided attention module provides a new way to solve this problem. In the future, we will further explore additional ways to solve the problem of 'blind overconfidence'.

### REFERENCES

- C. Siagian and L. Itti, "Rapid biologically-inspired scene classification using features shared with visual attention," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 29, no. 2, pp. 300–312, 2007.
- [2] Z. Ren, S. Gao, L.-T. Chia, and I. W.-H. Tsang, "Region-based saliency detection and its application in object recognition," *IEEE Transactions* on Circuits and Systems for Video Technology, vol. 24, no. 5, pp. 769– 779, 2014.
- [3] Y. Gao, M. Wang, D. Tao, R. Ji, and Q. Dai, "3-d object retrieval and recognition with hypergraph analysis," *IEEE Transactions on Image Processing*, vol. 21, no. 9, pp. 4290–4303, 2012.
- [4] J. He, J. Feng, X. Liu, T. Cheng, T.-H. Lin, H. Chung, and S.-F. Chang, "Mobile product search with bag of hash bits and boundary reranking," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3005–3012, 2012.
- [5] S. Hong, T. You, S. Kwak, and B. Han, "Online tracking by learning discriminative saliency map with convolutional neural network," in *International Conference on Machine Learning*, pp. 597–606, 2015.
- [6] V. Mahadevan and N. Vasconcelos, "Saliency-based discriminant tracking," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition(CVPR)*, pp. 1007–1013, 2009.
- [7] C. Ma, Z. Miao, X.-P. Zhang, and M. Li, "A saliency prior context model for real-time object tracking," *IEEE Transactions on Multimedia*, vol. 19, no. 11, pp. 2415–2424, 2017.
- [8] F. Perazzi, P. Krähenbühl, Y. Pritch, and A. Hornung, "Saliency filters: Contrast based filtering for salient region detection," in *Proceedings* of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 733–740, 2012.
- [9] J. Harel, C. Koch, and P. Perona, "Graph-based visual saliency," in Advances in Neural Information Processing Systems, pp. 545–552, 2007.
- [10] L. Itti, C. Koch, and E. Niebur, "A model of saliency-based visual attention for rapid scene analysis," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 20, no. 11, pp. 1254–1259, 1998.
- [11] Q. Yan, L. Xu, J. Shi, and J. Jia, "Hierarchical saliency detection," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition(CVPR), pp. 1155–1162, 2013.

- [12] H. Jiang, J. Wang, Z. Yuan, Y. Wu, N. Zheng, and S. Li, "Salient object detection: A discriminative regional feature integration approach," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition(CVPR), pp. 2083–2090, 2013.
- [13] Y. Liu, J. Han, Q. Zhang, and L. Wang, "Salient object detection via two-stage graphs," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 29, no. 4, pp. 1023–1037, 2019.
- [14] N. Liu and J. Han, "Dhsnet: Deep hierarchical saliency network for salient object detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition(CVPR)*, pp. 678–686, 2016.
- [15] S. Chen, X. Tan, B. Wang, and X. Hu, "Reverse attention for salient object detection," in *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 234–250, 2018.
- [16] L. Wang, H. Lu, X. Ruan, and M.-H. Yang, "Deep networks for saliency detection via local estimation and global search," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition(CVPR)*, pp. 3183–3192, 2015.
- [17] P. Zhang, D. Wang, H. Lu, H. Wang, and B. Yin, "Learning uncertain convolutional features for accurate saliency detection," in *IEEE Inter*national Conference on Computer Vision (ICCV), pp. 212–221, 2017.
- [18] P. Zhang, D. Wang, H. Lu, H. Wang, and X. Ruan, "Amulet: Aggregating multi-level convolutional features for salient object detection," in *Proceedings of the IEEE International Conference on Computer Vision(CVPR)*, pp. 202–211, 2017.
- [19] Z. Luo, A. Mishra, A. Achkar, J. Eichel, S. Li, and P.-M. Jodoin, "Non-local deep features for salient object detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition(CVPR)*, 2017.
- [20] G. Lee, Y.-W. Tai, and J. Kim, "Deep saliency with encoded low level distance map and high level features," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition(CVPR)*, pp. 660–668, 2016.
- [21] L. Wang, L. Wang, H. Lu, P. Zhang, and X. Ruan, "Saliency detection with recurrent fully convolutional networks," in *European Conference* on Computer Vision(ECCV), pp. 825–841, Springer, 2016.
- [22] G. Li and Y. Yu, "Deep contrast learning for salient object detection," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition(CVPR), pp. 478–487, 2016.
- [23] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in Neural Infor*mation Processing Systems, pp. 1097–1105, 2012.
- [24] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition(CVPR)*, pp. 1–9, 2015.
- [25] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," in *Advances in Neural Information Processing Systems*, pp. 91–99, 2015.
- [26] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition(CVPR)*, pp. 779–788, 2016.
- [27] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask r-cnn," in *IEEE International Conference on Computer Vision (ICCV)*, pp. 2980–2988, 2017
- [28] G. Lin, A. Milan, C. Shen, and I. Reid, "Refinenet: Multi-path refinement networks for high-resolution semantic segmentation," in *Proceedings* of the IEEE Conference on Computer Vision and Pattern recognition (CVPR), vol. 1, p. 3, 2017.
- [29] C. Peng, X. Zhang, G. Yu, G. Luo, and J. Sun, "Large kernel matter-simprove semantic segmentation by global convolutional network," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1743–1751, 2017.
- [30] X. Zhang, T. Wang, J. Qi, H. Lu, and G. Wang, "Progressive attention guided recurrent network for salient object detection," in *Proceedings* of the IEEE Conference on Computer Vision and Pattern Recognition(CVPR), pp. 714–722, 2018.
- [31] N. Liu, J. Han, and M.-H. Yang, "Picanet: Learning pixel-wise contextual attention for saliency detection," arXiv preprint arXiv:1708.06433, 2017.
- [32] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3431–3440, 2015.
- [33] Q. Hou, M.-M. Cheng, X. Hu, A. Borji, Z. Tu, and P. H. Torr, "Deeply supervised salient object detection with short connections," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition(CVPR), pp. 3203–3212, 2017.

- [34] H. Zhao, X. Qi, X. Shen, J. Shi, and J. Jia, "Icnet for real-time semantic segmentation on high-resolution images," in *Proceedings of* the European Conference on Computer Vision (ECCV), pp. 405–420, 2018.
- [35] S.-W. Kim, H.-K. Kook, J.-Y. Sun, M.-C. Kang, and S.-J. Ko, "Parallel feature pyramid network for object detection," in *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 234–250, 2018.
- [36] H. Xiao, J. Feng, Y. Wei, M. Zhang, and S. Yan, "Deep salient object detection with dense connections and distraction diagnosis," *IEEE Transactions on Multimedia*, vol. 20, no. 12, pp. 3239–3251, 2018.
- [37] C. Yang, L. Zhang, H. Lu, X. Ruan, and M.-H. Yang, "Saliency detection via graph-based manifold ranking," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition(CVPR)*, pp. 3166–3173, 2013.
- [38] A. Manno-Kovacs, "Direction selective contour detection for salient objects," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 29, no. 2, pp. 375–389, 2019.
- [39] G. Li and Y. Yu, "Visual saliency based on multiscale deep features," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition(CVPR), pp. 5455–5463, 2015.
- [40] R. Zhao, W. Ouyang, H. Li, and X. Wang, "Saliency detection by multi-context deep learning," in *Proceedings of the IEEE Conference* on Computer Vision and Pattern Recognition(CVPR), pp. 1265–1274, 2015.
- [41] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proceedings of the IEEE Confer*ence on Computer Vision and Pattern Recognition(CVPR), pp. 4700– 4708, 2017.
- [42] T. Wang, A. Borji, L. Zhang, P. Zhang, and H. Lu, "A stagewise refinement model for detecting salient objects in images," in *IEEE International Conference on Computer Vision(ICCV)*, pp. 4019–4028, 2017.
- [43] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems*, pp. 5998–6008, 2017.
- [44] M.-T. Luong, H. Pham, and C. D. Manning, "Effective approaches to attention-based neural machine translation," arXiv preprint arXiv:1508.04025, 2015.
- [45] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition(CVPR), pp. 7132–7141, 2018.
- [46] C. Yu, J. Wang, C. Peng, C. Gao, G. Yu, and N. Sang, "Learning a discriminative feature network for semantic segmentation," arXiv preprint arXiv:1804.09337, 2018.
- [47] J. Fu, J. Liu, H. Tian, Z. Fang, and H. Lu, "Dual attention network for scene segmentation," arXiv preprint arXiv:1809.02983, 2018.
- [48] Z. Fan, X. Zhao, T. Lin, and H. Su, "Attention-based multiview reobservation fusion network for skeletal action recognition," *IEEE Trans*actions on Multimedia, vol. 21, no. 2, pp. 363–374, 2019.
- [49] S. Woo, J. Park, J.-Y. Lee, and I. So Kweon, "Cbam: Convolutional block attention module," in *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 3–19, 2018.
- [50] V. Nair and G. E. Hinton, "Rectified linear units improve restricted boltzmann machines," in *Proceedings of the 27th International Conference* on Machine Learning (ICML-10), pp. 807–814, 2010.
- [51] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," arXiv preprint arXiv:1409.1556, 2014.
- [52] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision* and Pattern Recognition(CVPR), pp. 770–778, 2016.
- [53] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," arXiv preprint, pp. 1610–02357, 2017.
- [54] L. Tychsen-Smith and L. Petersson, "Improving object localization with fitness nms and bounded iou loss," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 6877–6885, 2018.
- [55] C. H. Sudre, W. Li, T. Vercauteren, S. Ourselin, and M. J. Cardoso, "Generalised dice overlap as a deep learning loss function for highly unbalanced segmentations," in *Deep learning in medical image analysis* and multimodal learning for clinical decision support, pp. 240–248, Springer, 2017.
- [56] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 40, no. 4, pp. 834–848, 2018.

- [57] T. Wang, L. Zhang, H. Lu, C. Sun, and J. Qi, "Kernelized subspace ranking for saliency detection," in *European Conference on Computer Vision(ECCV)*, pp. 450–466, Springer, 2016.
- [58] T. Liu, Z. Yuan, J. Sun, J. Wang, N. Zheng, X. Tang, and H.-Y. Shum, "Learning to detect a salient object," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 33, no. 2, pp. 353–367, 2011.
- [59] L. Wang, H. Lu, Y. Wang, M. Feng, D. Wang, B. Yin, and X. Ruan, "Learning to detect salient objects with image-level supervision," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition(CVPR), pp. 136–145, 2017.
- [60] C. Yang, L. Zhang, H. Lu, X. Ruan, and M.-H. Yang, "Saliency detection via graph-based manifold ranking," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition(CVPR)*, pp. 3166–3173, 2013.
- [61] D. Martin, C. Fowlkes, D. Tal, and J. Malik, "A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics," in *IEEE International Conference on Computer Vision(ICCV)*, vol. 2, pp. 416–423, 2001.
- [62] D.-P. Fan, M.-M. Cheng, Y. Liu, T. Li, and A. Borji, "Structure-measure: A new way to evaluate foreground maps," in *Proceedings of the IEEE international conference on computer vision*, pp. 4548–4557, 2017.
- [63] X. Li, F. Yang, H. Cheng, W. Liu, and D. Shen, "Contour knowledge transfer for salient object detection," in *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 355–370, 2018.
- [64] G. Li, Y. Xie, L. Lin, and Y. Yu, "Instance-level salient object segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition(CVPR)*, pp. 247–256, 2017.
- [65] X. Li, L. Zhao, L. Wei, M.-H. Yang, F. Wu, Y. Zhuang, H. Ling, and J. Wang, "Deepsaliency: Multi-task deep neural network model for salient object detection," *IEEE Transactions on Image Processing*, vol. 25, no. 8, pp. 3919–3930, 2016.



**Shiping Zhu** (M05) received the B.Sc. and M.Sc. degrees in measuring and testing technologies and instruments from Xian University of Technology, Xian, China, in 1991 and 1994, respectively, and the Ph.D. degree in precision instrument and machinery from Harbin Institute of Technology, Harbin, China, in 1997.

From 1997 to 1999, he was a Postdoctoral Fellow with Beihang University, Beijing, China. From 2000 to 2002, he was a Postdoctoral Fellow with the Brain and Cognition Research Center. Universit Paul

Sabatier, Toulouse, France. From 2002 to 2004, he was a Postdoctoral Fellow with the Department of Computer Science and Department of Electrical and Computer Engineering, Universit de Sherbrooke, Sherbrooke, QC, Canada. Since 2005, he has been an Associate Professor with the Department of Measurement Control and Information Technology, School of Instrumentation Science and Optoelectronics Engineering, Beihang University. He has authored or coauthored more than 80 journal and conference papers. He received the second prize of National Technological Invention Award in 2013. He is the holder of 50 China invention patents. His current research interests include image processing and video coding, stereo matching, saliency detection and image/video object segmentation.



Lanyun Zhu is currently pursuing the B.Sc. degree with Beihang University, Beijing, China. He is a research assistant with school of instrumentation and optoelectronic engineering, Beihang University. His currently research interests manly focus on computer vision, image processing and deep learning.