Neurocomputing 372 (2020) 55–63

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom

Deep layer guided network for salient object detection

Zhengyi Liu, Quanlong Li, Wei Li*

School of Computer Science and Technology, Anhui University, Heifei, China

ARTICLE INFO

Article history:
Received 25 April 2019
Revised 24 August 2019
Accepted 11 September 2019
Available online 24 September 2019

Communicated by Dr. Shen Jianbing

Keywords:
Salient object detection
Fully convolution network
Deep layer guidance
Discriminative feature

ABSTRACT

Recently salient object detection with convolutional neural networks has made great progress. More and more methods design more complex networks to integrate the features of each stage from backbone extractor. Considering that global information and spatial details of an image are better captured by features of deep layers and shallow layers of CNN respectively, deep layer guided network in which global information from deep layers is progressively transmitted to shallow layers in a guided manner is proposed. Hybrid feature enhancement block receives the feature maps of adjacent stages, and outputs enhanced feature maps which reduce the loss of spatial details and reduce the impact of varying in shape, scale and position of object. Discriminative feature block highlights the consistency in different stages from channel and spatial dimensions. Optimized discriminative features focus more on channels which show higher response to salient objects and positions consistent with the foreground in the saliency map from higher stage. Saliency inference block optimizes feature by merging with adjacent layer feature further. Thus feature becomes more enhanced, discriminative and refined in a high-to-low manner. Experimental results on five public datasets show that our salient object detection method reaches state-of-the-art performance under evaluation metrics.

1. Introduction

Saliency detection includes eye-fixation prediction (FP) [1,2] which can predict scene locations where a human observer may fixate and salient object detection(SOD) which can capture most visually conspicuous objects or regions. The former is originated from cognitive and psychology research communities. The latter is driven by object-level vision applications, such as visual tracking [3–6], person re-identification [7,8], image retrieval [9,10], image resizing [11,12] and image segmentation [13]. The initial research of salient object detection focuses on salient object detection in an image, then it is extended to RGB-D saliency detection [14–20] in which depth information can be utilized, co-saliency detection [21–24] in which inner and inter saliency constraint need be considered simultaneously, and video saliency detection [25–29] in which temporal and spatial relationship are explored. Our work focuses on SOD in an image.

In the past two decades, many salient object detection methods have been proposed. They are divided into traditional methods based on low-level cues and deep learning methods based on high-level semantic features. Traditional methods are difficult to achieve the outstanding performance when dealing with images with complex backgrounds. However, deep learning methods, especially full convolutional neural networks (FCNs) can overcome this bottleneck. Recently, state-of-the-art methods using FCNs have focused on integrating the hierarchical features of feature extractor. In [30] the high-level features are transmitted to shallower side-output layers by short connections. In [31] the multi-level features are fused by resolution-based feature combination modules. In [32] the features of different layer pass messages to each other through a bi-directional message passing module. Although these methods have achieved excellent performance, their network structures are more and more complicated. Whether do we also need to design a more complex network to further improve performance?

We note that there are three problems when dealing with the salient object detection problem by FCN: (1) the inherent disadvantage of FCNs about the operation of the pooling layer will bring the loss of feature information, (2) the salient object is variable in shape, scale and position in natural images, (3) multi-level features of FCNs are lack of discriminative because these features also contain cluttered and noisy information. Therefore, in order to solve these problems, we propose a deep layer guided network. First, we use the modified VGG-16 network [33] as our basic feature extraction network to extract original features of six stages. Then the original features of adjacent stages are sent to our proposed hybrid feature enhancement block (HFEB) to obtain enhanced features. Because there is a max-pooling operation between adjacent

* Corresponding author.
E-mail address: liuzywen@ahu.edu.cn (W. Li).

https://doi.org/10.1016/j.neucom.2019.09.018
0925-2312/

stages, the features of the current stage are least lost compared to the features of the previous stage. Therefore, taking the features of adjacent stages as input helps to mitigate the information losses caused by FCNs. In HFEB, there is a set of parallel convolution operations of different numbers and different types to alleviate the effects of varied salient object. Enhanced features of different resolutions also have noise or redundant information, so the proposed discriminative feature block (DFB) is used to process enhanced features to get discriminative features. In the DFB, we obtain discriminative features by adjacent higher layer guidance again and enhance the discrimination of the feature in the channel and spatial dimensions. In order to further refine the discriminative features, adjacent higher layer feature is combined and transferred from deep to shallow in saliency inference block (SIB). Thus the original features are optimized by these three blocks into enhanced, discriminative and refined features.

In summary, our contributions can be summarized as follows:

(1) We propose a deep layer guided network to optimize the multi-level convolutional features. It consists of backbone network, hybrid feature enhancement block, discriminative feature block and saliency inference block. Rich semantic information of deep layer conveys to adjacent layers step by step in a guidance manner among the network. Optimized features become enhanced, discriminative and refined for salient object detection task.
(2) We propose hybrid feature enhancement block which receives the feature maps of adjacent stages, and outputs enhanced feature maps. Under the guidance of adjacent higher layer feature, each hierarchical enhanced feature reduces the loss of spatial details and the impact of varying in shape, scale and position of object.
(3) We propose discriminative feature block which highlights the consistency in different stages from channel and spatial dimensions. Under the supplementation of adjacent higher layer feature, optimized discriminative features focus more on channels which show higher response to salient objects and maintain the consistency with the saliency map from higher stage in spatial positions.
(4) The proposed method has achieved new state-of-the-art performance on the public five salient object detection datasets.

2. Related work

2.1. Salient object detection

Salient object detection goes through the process from traditional methods [34] including example-driven method [35], to deep learning methods [36]. Traditional methods are based on low-level hand-crafted features and rely on certain heuristics [37–40]. Wang et al. [35] propose a correspondence-based saliency transfer approach in which the nearest-support images or patches and their annotations are warped for inferring the saliency of the input image according to global and local correspondences. Deep learning methods design various network architectures to extract feature and fuse feature. Hu et al. [41] design VGG16-based backbone network equipped with a Guided Superpixel Filtering (GSF) layer which combines superpixels to further detect saliency within objects and suppress saliency outside the objects. Wang and Shen [1] design an encoder-decoder framework with skip layer architecture supervised in top three layers. More and more researches [42–44] emphasize on complex network combined with contour information to detect salient objects, while we aim at exploring more enhanced, discriminative and refined features for the better performance.

2.2. Enhanced features

By taking convolution or other operations behind the multi-layer features of basic network to enhance the representation of features have shown its efficiency in various salient object detection architectures [30,32,45] and face detection [46]. In [45], Luo et al. utilize a convolution and contrast features operations to obtain enhanced features. In [30], Hou et al. use a multi-features fusion method which high-level features are transformed to shallower side-output layers by introducing short connections to the skip-layer structure within the Holistically-Nested Edge Detector (HED) [47] architecture. Zhang et al. [32] use the multi-scale context-aware feature extraction module (MCFEM) which contains four dilated convolutions with different dilation rates to get the enhanced features of each layer. Li et al. [46] propose feature enhance module aiming to enhance the original features. The module takes the adjacent layer features of basic network as input and contains three different numbers of dilation convolutional layers. Inspired by these works, we propose HFEB which captures multi-scale context information to generate the enhanced features by deep layer guided network.

2.3. Discriminative features

Recently attention mechanics is explored to make the feature discriminative. Wang et al. [48] add extra supervision from the visual attention map which is acted as a strong prior in its attention box prediction network. Wang et al. [49] build a neural network called Attentive Saliency Network (ASNet) that learns to detect salient objects from fixation maps. ASNet first captures a global and high-level understanding of a scene in its higher layers, by learning to predict human fixations. Then, it uses a stack of convLSTMs to progressively infer object saliency from the fixation map in a top-down and coarse-to-fine manner. The whole network is simultaneously trained to predict fixation locations and detect salient objects in an end-to-end way. Wang et al. [25] design dynamic saliency network capturing temporal statistics from frame pairs and the output saliency map of static saliency network which is served as spatial attentive priors indicative of potential salient regions. Wang et al. [2] encode a supervised fixation prediction to capture static saliency information and helps CNN-LSTM network better capture dynamic saliency representations over successive frames. Feng et al. [43] utilize attentive feedback and boundary-enhanced loss to produce exquisite boundaries. Wang et al. [50] propose a global Recurrent Localization Network(RLN) which used to select distinctive features in spatial dimension. Chen et al. [51] propose reverse attention block based side-output residual features for localizing salient object regions progressively. In these methods, the attentional mechanism is used to select discriminative features. Inspired by these references, DFB we proposed uses the feature information and saliency maps of higher layer as the guidance to select discriminative features on the channel and spatial dimensions.

3. Proposed method

The framework we propose consists of four parts, as shown in Fig. 1. The first is backbone net which is modified from VGG-16 network [33]. The second is hybrid feature enhancement block which extracts more enhanced features. The third is discriminative feature block which is used to select discriminative features. The last is saliency inference block which generates final saliency map. Four parts are elaborated in Section 3.1–Section 3.4, respectively.

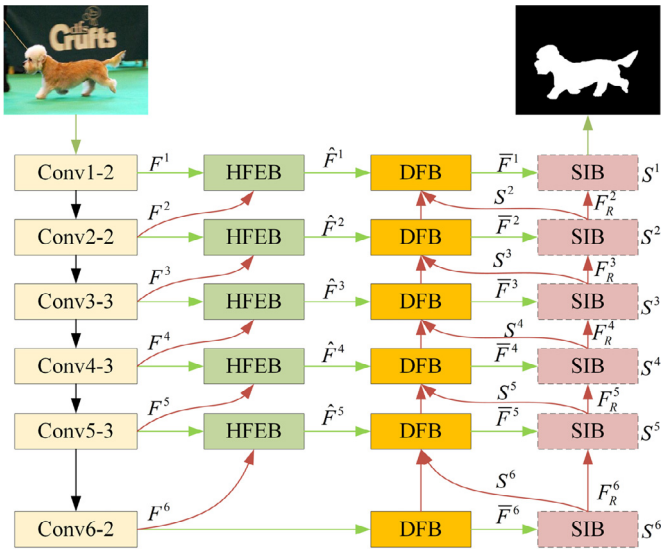


Fig. 1. Overall architecture of our proposed method. An image is sent to modified VGG-16 network [33] and multi-level convolutional features are generated. At first hybrid feature enhancement block (HFEB) attends to adjacent convolutional features for more enhanced features. Then discriminative feature block (DFB) refines the features more discriminative by attention mechanics. At last saliency inference block (SIB) outputs the saliency map in the deep-to-shallow manner. The red, green and black lines represent the upsample operation, information transfer and max-pooling operation respectively. Dotted box indicates supervised learning.

3.1. Backbone network

The VGG-16 model plays a backbone network role in a large number of computer vision tasks due to its simple framework and excellent performance. The VGG-16 model is used for image classification and is not suit for image-to-image density estimation. Therefore, it is revised and the last two fully connected layers are replaced by two convolution layers. As described in [52,53], global information can better help to recognize which objects or regions are salient, so increasing the receptive field of the network is beneficial for saliency detection. Therefore, the sizes of the last two convolutional layers are set as $5 \times 5 \times 1024$ and $3 \times 3 \times 1024$, respectively. The backbone network provides feature maps at six scales, which are denoted as $F^i(i = 1, \ldots, 6)$ from bottom to top. More specifically, for the input image I with $W \times H$ size, the resolution of feature maps F^i is $\frac{W}{2^{i-1}} \times \frac{H}{2^{i-1}} (i = 1, \ldots, 6)$.

3.2. Hybrid feature enhancement block

Hierarchical features $F^i(i = 1, \ldots, 6)$ in VGG-16 network have two drawbacks. At first, max-pooling operation is performed between the feature maps F^i and F^{i+1}, so F^{i+1} will lose the spatial details compared with F^i. Second, the shape, scale and position of the salient objects or regions in the image are varied, while the feature maps F^i can only observe the information of the fixed receptive field. Recent works [32,46] have proposed dilated convolution to enhance feature. Inspired by these works, hybrid feature enhancement block is proposed to solve the above two defects.

An overview of HFEB is illustrated in Fig. 2. First, convolution operation is performed to normalize the feature maps of adjacent stages F^i and F^{i+1}, and then F^{i+1} is upsampled as the same size as F^i by bilinear interpolation operation. Then, these two feature maps are concatenated, and followed by a convolutional layer and three dilation convolution layers with different numbers in cascade mode to capture richer feature information at multiple scales. At last, all the feature maps are concatenated and performed by two sequential convolution layers to obtain enhanced feature map $\hat{F}^i(i = 1, \ldots, 5)$. In order to clearly describe above process, it is expressed by the following formula:

$$\begin{aligned} G^i &= Cat(Conv(F^i), Up(Conv(F^{i+1}))) \\ Q^i &= Cat(Conv(G^i), 1 * DilationConv(G^i), \\ &\quad 2 * DilationConv(G^i), 3 * DilationConv(G^i)) \\ \hat{F}^i &= Conv(Conv(Q^i)) \end{aligned} \tag{1}$$

where $Conv(\cdot)$ denotes 3×3 convolutional layer with 96 channels accompanied with Relu activation function, $Up(\cdot)$ denotes the $2\times$ upsampling operation using bilinear interpolation, $Cat(\cdot)$ denotes the concatenation operation along channel axis, $n * DilationConv(\cdot)(n = 1, 2, 3)$ operation denotes performing dilation convolution n times, $DilationConv(\cdot)$ is a 3×3 dilation convolutional layer [54] with rate=3, n is the number of dilation convolution.

Each HFEB receives the feature maps of adjacent stages F^i and F^{i+1} as input, and output enhanced feature maps $\hat{F}^i(i = 1, \ldots, 5)$ which reduce the loss of spatial details caused by max-pooling operation by the help of the features of adjacent layers, and adapt to varying in shape, scale and position of object by the convolution and dilation convolution operations of different numbers in parallel. The proposed HFEB makes the original features more enhanced, and it is verified by the experiments in Table. 3.

3.3. Discriminative feature block

When the network combines the features of adjacent stages, it enhances the representation ability of the feature but ignores the consistency in different stages. There is no strong consistency constraint between these features $\hat{F}^i(i = 1, \ldots, 5)$. In order to remedy the drawback, we propose discriminative feature block which explores deep layer guidance. As shown in Fig. 3, strong consistency constraints are introduced in the channel and spatial dimensions respectively. In the channel axis, the channel attention block [55,56] is used to optimize enhanced features to focus more on channels which show higher response to salient objects. In the spatial axis, the saliency map S^{i+1} of higher stage is introduced into DFB as spatial attention guidance to assign larger weights to foreground regions inspired by Chen et al. [51].

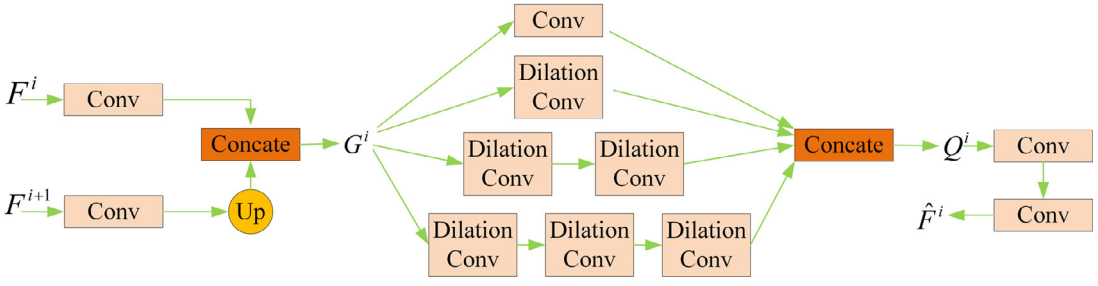


Fig. 2. HFEB: hybrid feature enhancement block.

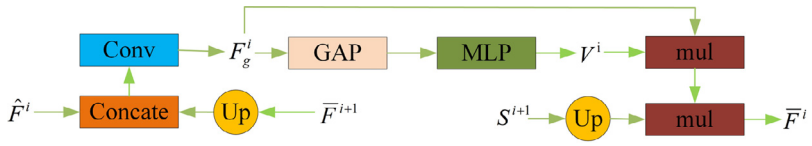


Fig. 3. DFB: discriminative feature block.

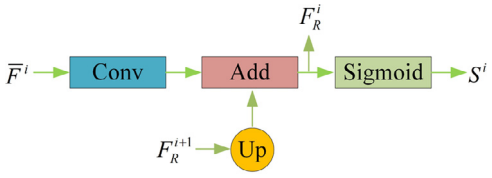


Fig. 4. SIB: saliency inference block.

DFB is executed from deep to shallow. The last DFB is unique and different from the others. Feature F^6 is first performed a convolution operation with 3×3 kernel size and 96 channels, and then convolution block attention module (CBAM) [57] is utilized to further get discriminative global features $\bar{F}^6$. Then DFB combines global features $\bar{F}^{i+1}$ and hierarchical enhanced feature $\hat{F}^i$ from deep to shallow for more discriminative feature $\bar{F}^i$ by attention mechanics. Its specific process is expressed by the following formula:

$$F_g^i = Conv(Cat(\hat{F}^i, Up(\bar{F}^{i+1}))) \tag{2}$$

where F_g^i is a refined feature by global information, and $Conv(\cdot)$ denotes 3×3 convolutional layer with 96 channels without activation function. Then channel and spatial attention are performed, and it is illustrated as:

$$\bar{F}^i = (F_g^i \otimes MLP(GAP(F_g^i))) \otimes Up(S^{i+1}), i \in \{1, 2, 3, 4, 5\} \tag{3}$$

where $\otimes$ denotes element-wise multiplication, *GAP* denotes global average pooling operation and *MLP* is single hidden layer fully connected neural network [55]. Thus discriminative and consistent features $\bar{F}^i$ can be achieved.

Discriminative feature block can exploit the consistency in different stages by channel and spatial attention, and utilize global information to make hierarchical feature more discriminative, as shown in Table 3. The proposed attention strategy is superior to CBAM [57] in our framework, as shown in Table 5.

3.4. Saliency inference block

In order to obtain more accurate saliency maps with different resolutions, the network is supervised at different stages. Global feature $\bar{F}^6$ is first performed a convolutional operation to generate refined hierarchical F_R^6. Then as shown in Fig. 4, refined hierarchical feature $F_R^{i+1}(i = 1, \ldots, 5)$ is $2 \times$ upsampling and fused with the feature $\bar{F}^i$ after a convolutional layer by a simple element-wise summation operation to generate refined hierarchical feature $F_R^i(i = 1, \ldots, 5)$. All the refined hierarchical F_R^i are performed sigmoid function to generate saliency map S^i which is both sent to DFB for spatial attention guidance and supervised by the ground truth. Above process can be expressed as the following formula:

$$F_R^i = \begin{cases} Conv(\bar{F}^i) + Up(F_R^{i+1}), & i = 1, 2, \ldots, 5 \\ Conv(\bar{F}^i), & i = 6 \end{cases} \tag{4}$$

where the $Conv(\cdot)$ denotes 3×3 convolutional layer with 1 channels without activation function. Finally,in order to get saliency map S^i, a sigmoid operation is applied as:

$$S^i = Sigmoid(F_R^i), i = 1, 2, \ldots, 6 \tag{5}$$

Thus the feature from deep layers are progressively transmitted to shallow layers, and S^1 is used as our final saliency map. The proposed model is end-to-end trained using cross-entropy loss function between the saliency maps of different resolution and the ground truth maps. The loss function is defined as:

$$L^m = -\sum_i y_i^m log(S_i^m) + (1 - y_i^m) log(1 - S_i^m), m = 1, 2, \ldots, 6 \tag{6}$$

where y_i^m is the label of the pixel i in mth ground truth map with the same resolution as S_i^m, and S_i^m is the saliency value of pixel i in the mth saliency map.

4. Experiments

4.1. Experimental setup

Datasets. The proposed method is evaluated on five benchmark datasets: ECSSD [58], DUT-OMRON [39], PASCAL-S [59], HKU-IS [60] and DUTS [61]. ECSSD [58] dataset contains 1000 natural images and PASCAL-S [59] dataset contains 850 natural images from the PASCAL VOC dataset [62]. HKU-IS [60] dataset contain multiple salient objects. DUT-OMRON [39] dataset which contain 5168 well annotated images is more challenging than the others. DUTS [61] dataset is a latest released large dataset which consists of 10,553 training images and 5019 test images.

Evaluation criteria. Our model and other state-of-the-art salient object detection models are evaluated in three main metrics: precision–recall (P–R) curves, max F-measure score and mean absolute error (MAE). The precision and recall are calculated by $|M \cap G|/|M|$ and $|M \cap G|/|G|$ respectively. M is a binary mask by thresholding the predicted saliency map and G is the corresponding ground truth. Operation $|\cdot|$ counts the number of non-zeros in a mask. P–R curve is generated by averaging the precision and recall values of all images in the dataset. F-measure score is an indicator of overall performance and is defined as:

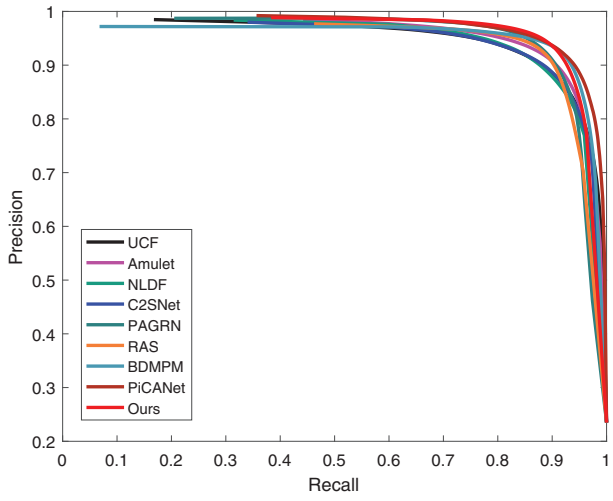
$$F_\beta = \frac{(1 + \beta^2 \times Precision \times Recall)}{\beta^2 \times Precision + Recall}, \tag{7}$$

where β^2 is usually set to 0.3 to emphasize that accuracy is more important than recall. MAE is the average pixel-wise absolute difference between the predicted saliency map S and the corresponding truth map G:

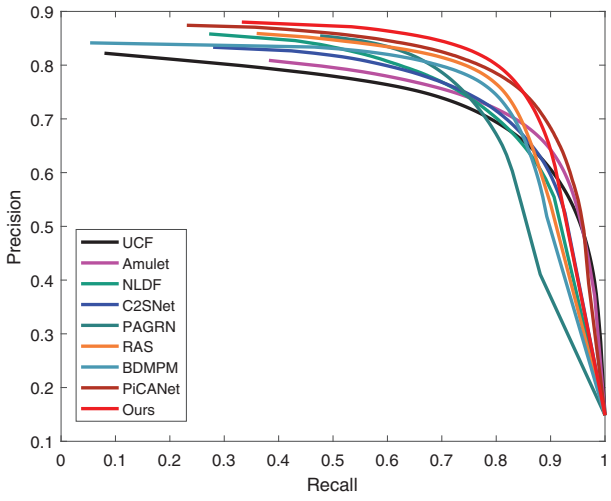
$$MAE = \frac{1}{W \times H} \sum_{i=1}^{W} \sum_{j=1}^{H} |S(i, j) - G(i, j)|, \tag{8}$$

where W and H represent width and height respectively.

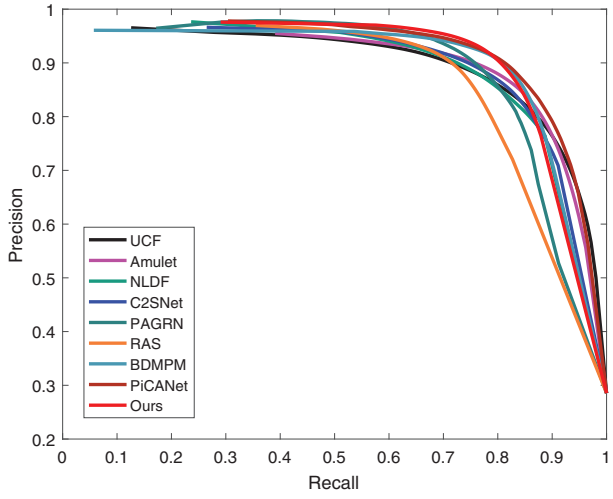
Implementation details. The proposed model is implemented in TensorFlow [63], and it is trained on the DUTS [61] training set. We train the model until training loss converges and no validation set are used, as suggested in [32]. To alleviate the over-fitting problem, we enhance the training set by flipping the images horizontally and vertically. The weights of backbone network are initialized with the pretrained weights of VGG-16 network [33]. And the other weights of convolutional layers are initialized with a truncated normal method. Upsampling is achieved by bilinear interpolation. We use the Adam [64] method to train our model with an initial learning rate about 1e−6. The model is trained by 10



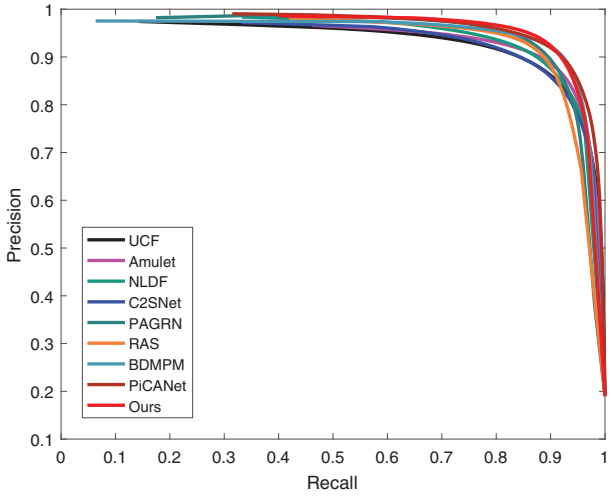
(a) ECSSD



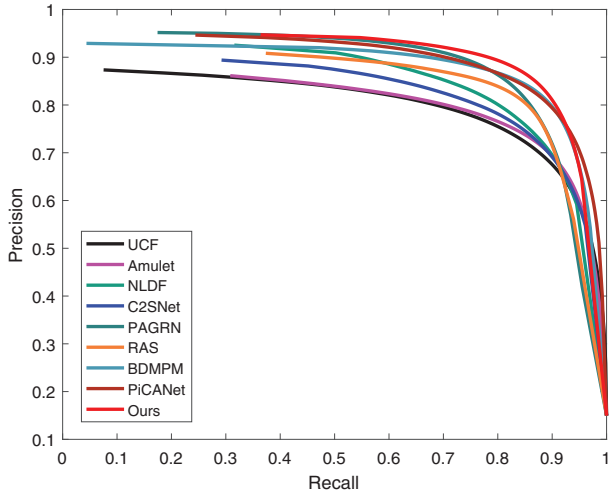
(b) DUT-OMRON



(c) PASCAL-S



(d) HKU-IS



(e) DUTS-TE

Fig. 5. Quantitative results of P–R curves for the proposed model and other 8 state-of-the-art methods.

Fig. 6. Visual comparison results of our method and other the state-of-the-art salient object detection models.

Table 1
Quantitative comparison of our proposed model with other 8 state-of-the-art methods in terms of max *F*-measure and MAE scores on five datasets. The larger max *F*-measure value is, the better the model is. The smaller MAE scores is, the better the model is. The best three models are highlighted in bold, italic and bolditalic, respectively.

	ECSSD		DUT-OMRON		PASCAL-S		HKU-IS		DUTS-TE	
	F-measure↑	MAE↓	*F*-measure↑	MAE↓	*F*-measure↑	MAE↓	*F*-measure↑	MAE↓	*F*-measure↑	MAE↓
UCF_{17}	0.903	0.068	0.729	0.120	0.825	0.114	0.887	0.061	0.772	0.112
$Amulet_{17}$	0.915	0.058	0.742	0.097	0.840	0.097	0.897	0.050	0.777	0.084
$NLDF_{17}$	0.905	0.062	0.753	0.079	0.833	0.099	0.901	0.047	0.812	0.065
$C2SNet_{18}$	0.903	0.057	0.751	0.073	0.844	0.086	0.888	0.049	0.793	0.067
$PAGRN_{18}$	0.926	0.060	0.770	0.070	0.857	0.093	0.917	0.047	*0.854*	0.055
RAS_{18}	0.921	0.056	***0.786***	*0.061*	0.838	0.104	0.912	0.0453	0.831	0.059
$BDMPM_{18}$	***0.928***	**0.044**	0.774	***0.063***	***0.863***	*0.074*	***0.920***	*0.038*	***0.851***	*0.049*
$PiCANet_{18}$	*0.931*	***0.046***	*0.794*	0.067	*0.871*	***0.076***	*0.920*	***0.041***	***0.851***	***0.054***
Ours	**0.934**	**0.044**	**0.809**	**0.055**	**0.880**	**0.071**	**0.927**	**0.035**	**0.870**	**0.043**

epochs with a single image batch size on a single NVIDIA 1080Ti GPU. When testing, the proposed model runs at about 21 fps with 256 × 256 resolution.

Table 2
Average runtime comparison on DUTS-TE dataset.

Method	NLDF	BDMPM	PiCANet	Ours
Runtime (s)	0.047	0.034	0.118	0.048

4.2. Performance comparison

We compare the proposed algorithm with the recent 8 state-of-the-art saliency models on five test datasets, including UCF [65], Amulet [31], NLDF [45], C2SNet [66], PAGRN [67], RAS [51], BDMPM [32] and PiCANet [68]. For fair comparison, we used saliency maps released by the author.

Quantitative comparison. We compare our model and other state-of-the-art saliency detection models based on P-R curves, maximum *F*-measure and MAE value. The results of the evaluation are shown in Table 1 and Fig. 5. As shown in Table 1, our model achieves the best score on all five datasets in terms of maximum *F*-measure and MAE criteria. According to maximum *F*-measure scores, our model outperforms the second best method by 0.3%, 1.9%, 1.0%, 0.8%, 2.2% over ECSSD, DUT-OMRON, PASCAL-S, HKU-IS, DUTS-TE respectively. And as for MAE, our model lowers the value by 17.9%, 4.1%, 7.9%, 12.2% on DUT-OMRON, PASCAL-S, HKU-IS, DUTS-TE respectively and wins the first with BDMPM [32] on ECSSD. From Fig. 5, the P–R curve is slightly higher than other methods at the highest point in all five datasets. Therefore, the results in Table 1 and Fig. 5 show that the proposed method is effective.

Table 3
The performance of different architectures on five datasets.

	ECSSD		DUT-OMRON		PASCAL-S		HKU-IS		DUTS-TE	
	F-measure↑	MAE↓	*F*-measure↑	MAE↓	*F*-measure↑	MAE↓	*F*-measure↑	MAE↓	*F*-measure↑	MAE↓
Backbone+SIB	0.917	0.052	0.781	0.059	0.853	0.082	0.907	0.042	0.842	0.049
Backbone+HFEB+SIB	0.933	0.045	0.801	0.057	0.872	0.076	0.922	0.037	0.862	0.046
Backbone+HFEB+DFB+SIB	**0.934**	**0.044**	**0.809**	**0.055**	**0.880**	**0.071**	**0.927**	**0.035**	**0.870**	**0.043**

Table 4
Comparison between DSS and our HFEB on five datasets.

	ECSSD		DUT-OMRON		PASCAL-S		HKU-IS		DUTS-TE	
	F-measure↑	MAE↓	*F*-measure↑	MAE↓	*F*-measure↑	MAE↓	*F*-measure↑	MAE↓	*F*-measure↑	MAE↓
DSS	0.926	0.051	0.792	0.061	0.867	0.079	0.918	0.040	0.853	0.050
Backbone+HFEB+SIB	**0.933**	**0.045**	**0.801**	**0.057**	**0.872**	**0.076**	**0.922**	**0.037**	**0.862**	**0.046**

Table 5
Comparison between CBAM and Ours on five datasets.

	ECSSD		DUT-OMRON		PASCAL-S		HKU-IS		DUTS-TE		Runtime (s)↓
	F-measure↑	MAE↓	*F*-measure↑	MAE↓	*F*-measure↑	MAE↓	*F*-measure↑	MAE↓	*F*-measure↑	MAE↓	
CBAM	**0.935**	**0.042**	0.802	0.056	0.877	**0.070**	0.925	0.036	0.862	0.044	0.049
Ours	0.934	0.044	**0.809**	**0.055**	**0.880**	0.071	**0.927**	**0.035**	**0.870**	**0.043**	**0.046**

Visual comparison. In order to qualitatively evaluate our method with other methods, a visual comparison of the results is shown in Fig. 6. The examples (rows 1–4) have single salient object that contains complex backgrounds in different scenarios. The example (row 3) has similar color appearance with the background. We show some examples including multiple salient objects (rows 5–7). As can be seen, our visual saliency map is closer to the ground truth than the other state-of-the-art methods.

Runtime comparison. We compare the test runtime of ours with some TensorFlow-based method (NLDF [45], BDMPM [32], PiCANet [68]) in the same GPU. As can be seen from Table 2, the test runtime of our method is at the middle level. It is neither the fastest nor the lowest. It can be applied in real-time applications.

4.3. Ablation analysis

In order to verify the validity of our proposed different block, ablation experiment is conducted. Modified VGG-16 network is encoder network, and it need SIB to complete decoder process. Thus two parts are grouped and served as baseline. Then HFEB and DFB are added step by step. From Table 3, we can see that HFEB and DFB contribute their abilities for the final results. The whole network achieves the best performance.

Meanwhile in order to compare DSS [30] with our model fairly, DSS [30] are retrained by the same training set as ours, DFB is removed from our model, and the comparison result is shown in Table 4. From the result, we can see that our fusion strategy of adjacent feature is better than short connection strategy in which large-rate upsampling and single-channel feature of higher layer give rise to information loss.

At last in order to verify our DFB, the comparison between the attention strategy in CBAM [57] and ours is conducted, and the result is shown in Table 5. From the result, we can see that major evaluation metrics of ours are superior to CBAM [57], and the runtime is relatively less than CBAM [57].

5. Conclusion

In the paper, we propose a novel saliency detection model named deep layer guided network. We first designed a hybrid feature enhancement block that consists of convolution and multiple different receptive field size dilation convolutions. It is used to enhance the side features of the backbone network. Then, we propose a discriminative feature block that further filters features in the channel and spatial dimensions. At last, we refine the feature maps gradually by using high-level features to guide adjacent low-level layer feature. The experiments have proved the effectiveness of our proposed method. In the future, we attempt to apply our idea of deep layer guidance to detect salient object in videos, and select the better features in spatial and temporal domain.

Declaration of Competing Interest

We declare that we have no financial and personal relationships with other people or organizations that can inappropriately influence our work, there is no professional or other personal interest of any nature or kind in any product, service or company that could be construed as the review of the manuscript.

Acknowledgment

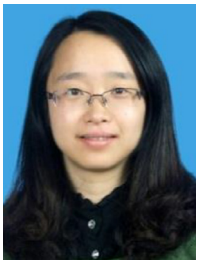
This research is supported by Natural Science Foundation of Anhui Province (1908085MF182) and Key Program of Natural Science Project of Educational Commission of Anhui Province (KJ2019A0034).

References

[1] W. Wang, J. Shen, Deep visual attention prediction, IEEE Trans. Image Process. PP (99) (2018) 1.
[2] W. Wang, J. Shen, F. Guo, M.-M. Cheng, A. Borji, Revisiting video saliency: a large-scale benchmark and a new model, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 4894–4903.
[3] A. Borji, S. Frintrop, D.N. Sihite, L. Itti, Adaptive object tracking by learning background context, in: Proceedings of the 2012 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops, IEEE, 2012, pp. 23–30.
[4] V. Mahadevan, N. Vasconcelos, Saliency-based discriminant tracking, in: Proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognition, IEEE, 2009, pp. 1007–1013.
[5] L. Bertinetto, J. Valmadre, J.F. Henriques, A. Vedaldi, P.H.S. Torr, Fully-convolutional siamese networks for object tracking, in: Proceedings of the IEEE European Conference on Computer Vision, Springer, 2016, pp. 850–865.
[6] X. Dong, J. Shen, D. Wu, K. Guo, X. Jin, F. Porikli, Quadruplet network with one-shot learning for fast visual object tracking, IEEE Trans. Image Process. 28 (7) (2019) 3516–3527.

[7] S. Bi, G. Li, Y. Yu, Person re-identification using multiple experts with random subspaces, J. Image Graph. 2 (2) (2014) 151–157.
[8] L. Wu, Y. Wang, J. Gao, X. Li, Deep adaptive feature embedding with local sample distributions for person re-identification, Pattern Recognit. 73 (2018) 275–288.
[9] M.-M. Cheng, Q.-B. Hou, S.-H. Zhang, P.L. Rosin, Intelligent visual media processing: when graphics meets vision, J. Comput. Sci. Technol. 32 (1) (2017) 110–121.
[10] J. He, J. Feng, X. Liu, T. Cheng, T.-H. Lin, H. Chung, S.-F. Chang, Mobile product search with bag of hash bits and boundary reranking, in: Proceedings of the 2012 IEEE Conference on Computer Vision and Pattern Recognition, IEEE, 2012, pp. 3005–3012.
[11] W. Wang, J. Shen, Deep cropping via attention box prediction and aesthetics assessment, in: Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 2186–2194.
[12] W. Wang, J. Shen, Y. Yu, K.-L. Ma, Stereoscopic thumbnail creation via efficient stereo saliency detection, IEEE Trans. Vis. Comput. Graph. 23 (8) (2017) 2014–2027.
[13] M. Donoser, M. Urschler, M. Hirzer, H. Bischof, Saliency driven total variation segmentation, in: Proceedings of the 12th IEEE International Conference on Computer Vision, IEEE, 2009, pp. 817–824.
[14] D.-P. Fan, Z. Lin, J.-X. Zhao, Y. Liu, Z. Zhang, Q. Hou, M. Zhu, M.-M. Cheng, Rethinking RGB-D Salient Object Detection: Models, Datasets, and Large-Scale Benchmarks, arXiv:1907.06781 (2019).
[15] J. Zhao, Y. Cao, D.-P. Fan, X.-Y. Li, L. Zhang, M.-M. Cheng, Contrast prior and fluid pyramid integration for RGBD salient object detection, in: Proceedings of the 2019 IEEE Computer Vision and Pattern Recognition, CVPR, 2019, pp. 1–10.
[16] Z. Liu, S. Shi, Q. Duan, W. Zhang, P. Zhao, Salient object detection for RGB-D image by single stream recurrent convolution neural network, Neurocomputing 363 (2019) 45–57.
[17] H. Chen, Y. Li, Three-stream attention-aware network for RGB-D salient object detection, IEEE Trans. Image Process. 28 (6) (2019) 2825–2835.
[18] H. Chen, Y. Li, D. Su, Multi-modal fusion network with multi-scale multi–path and cross-modal interactions for RGB-D salient object detection, Pattern Recognit. 86 (2019) 376–385.
[19] C. Zhu, X. Cai, K. Huang, T.H. Li, G. Li, PDNet: Prior-Model Guided Depth-Enhanced Network for salient Object Detection, arXiv:1803.08636 (2018).
[20] N. Wang, X. Gong, Adaptive Fusion for RGB-D Salient Object Detection, arXiv:1901.01369 (2019).
[21] L. Wei, S. Zhao, O.E.F. Bourahla, X. Li, F. Wu, Y. Zhuang, Deep group-wise fully convolutional network for co-saliency detection with graph propagation, IEEE Trans. Image Process. 28 (10) (2019) 2052–5063.
[22] K. Zhang, T. Li, B. Liu, Q. Liu, Co-saliency detection via mask-guided fully convolutional networks with multi-scale label smoothing, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 3095–3104.
[23] C. Wang, Z.-J. Zha, D. Liu, H. Xie, Robust deep co-saliency detection with group semantic, in: Proceedings of the AAAI Conference on Artificial Intelligence, 2019, pp. 8917–8924.
[24] M. Li, S. Dong, K. Zhang, Z. Gao, X. Wu, H. Zhang, G. Yang, S. Li, Deep learning intra-image and inter-images features for co-saliency detection, in: Proceedings of the 2018 British Machine Vision Conference, BMVC, 2018, p. 291.
[25] W. Wang, J. Shen, L. Shao, Video salient object detection via fully convolutional networks, IEEE Trans. Image Process. 27 (1) (2018) 38–49.
[26] D.-P. Fan, W. Wang, M.-M. Cheng, J. Shen, Shifting more attention to video salient object detection, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 8554–8564.
[27] H. Song, W. Wang, S. Zhao, J. Shen, K.-M. Lam, Pyramid dilated deeper ConvLSTM for video salient object detection, in: Proceedings of the 2018 European Conference on Computer Vision (ECCV), 2018, pp. 715–731.
[28] W. Wang, J. Shen, R. Yang, F. Porikli, Saliency-aware video object segmentation, IEEE Trans. Pattern Anal. Mach. Intell. 40 (1) (2017) 20–33.
[29] J. Shen, J. Peng, L. Shao, Submodular trajectories for better motion segmentation in videos, IEEE Trans. Image Process. 27 (6) (2018) 2688–2700.
[30] Q. Hou, M.-M. Cheng, X. Hu, A. Borji, Z. Tu, P.H.S. Torr, Deeply supervised salient object detection with short connections, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 3203–3212.
[31] P. Zhang, D. Wang, H. Lu, H. Wang, X. Ruan, Amulet: aggregating multi-level convolutional features for salient object detection, in: Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 202–211.
[32] L. Zhang, J. Dai, H. Lu, Y. He, G. Wang, A bi-directional message passing model for salient object detection, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 1741–1750.
[33] K. Simonyan, A. Zisserman, Very Deep Convolutional Networks for Large-Scale Image Recognition, arXiv:1409.1556 (2014).
[34] A. Borji, M.-M. Cheng, H. Jiang, J. Li, Salient object detection: a benchmark, IEEE Trans. Image Process. 24 (12) (2015) 5706–5722.
[35] W. Wang, Correspondence driven saliency transfer, IEEE Trans. Image Process. 25 (11) (2016) 5025–5034.
[36] W. Wang, Q. Lai, H. Fu, J. Shen, H. Ling, Salient Object Detection in the Deep Learning Era: An In-Depth Survey, arXiv:1904.09146 (2019).
[37] M.M. Cheng, G.X. Zhang, N.J. Mitra, X. Huang, S.M. Hu, Global contrast based salient region detection, in: Proceedings of the IEEE Computer Vision and Pattern Recognition, 2011, pp. 409–416.
[38] Y. Wei, F. Wen, W. Zhu, J. Sun, Geodesic saliency using background priors, in: Proceedings of the IEEE European conference on computer vision, Springer, 2012, pp. 29–42.
[39] C. Yang, L. Zhang, H. Lu, X. Ruan, M.-H. Yang, Saliency detection via graph-based manifold ranking, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2013, pp. 3166–3173.
[40] W. Zhu, S. Liang, Y. Wei, J. Sun, Saliency optimization from robust background detection, in: Proceedings of the IEEE Computer Vision and Pattern Recognition, 2014, pp. 2814–2821.
[41] P. Hu, B. Shuai, J. Liu, G. Wang, Deep level sets for salient object detection, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 2300–2309.
[42] S. Zhou, J. Wang, F. Wang, D. Huang, SE2Net: Siamese Edge-Enhancement Network for Salient Object Detection, arXiv:1904.00048 (2019).
[43] M. Feng, H. Lu, E. Ding, Attentive feedback network for boundary-aware salient object detection, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 1623–1632.
[44] Z. Chen, H. Zhou, X. Xie, J. Lai, Contour Loss: Boundary-Aware Learning for Salient Object Segmentation, arXiv:1908.01975 (2019).
[45] Z. Luo, A. Mishra, A. Achkar, J. Eichel, S. Li, P.M. Jodoin, Non-local deep features for salient object detection, in: Proceedings of the IEEE Computer Vision and Pattern Recognition, 2017, pp. 6593–6601.
[46] J. Li, Y. Wang, C. Wang, Y. Tai, J. Qian, J. Yang, C. Wang, J. Li, F. Huang, DSFD: Dual Shot Face Detector, arXiv:1810.10220 (2018).
[47] S. Xie, Z. Tu, Holistically-nested edge detection, in: Proceedings of the IEEE International Conference on Computer Vision, 2015, pp. 1395–1403.
[48] W. Wang, J. Shen, H. Ling, A deep network solution for attention and aesthetics aware photo cropping, IEEE Trans. Pattern Anal. Mach. Intell. 41.7 (2018) 1531–1544.
[49] W. Wang, J. Shen, X. Dong, A. Borji, Salient object detection driven by fixation prediction, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 1711–1720.
[50] T. Wang, L. Zhang, S. Wang, H. Lu, G. Yang, X. Ruan, A. Borji, Detect globally, refine locally: a novel approach to saliency detection, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 3127–3135.
[51] S. Chen, X. Tan, B. Wang, X. Hu, Reverse attention for salient object detection, in: Proceedings of the European Conference on Computer Vision (ECCV), 2018, pp. 234–250.
[52] M.-M. Cheng, N.J. Mitra, X. Huang, P.H.S. Torr, S.-M. Hu, Global contrast based salient region detection, IEEE Trans. Pattern Anal. Mach. Intell. 37 (3) (2015) 569–582.
[53] Y. Liu, Y. Qiu, L. Zhang, J. Bian, G.-Y. Nie, M.-M. Cheng, Salient Object Detection Via High-to-Low Hierarchical Context Aggregation, arXiv:1812.10956 (2018).
[54] F. Yu, V. Koltun, Multi-Scale Context Aggregation by Dilated Convolutions, arXiv:1511.07122 (2015).
[55] J. Hu, L. Shen, G. Sun, Squeeze-and-excitation networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 7132–7141.
[56] C. Yu, J. Wang, C. Peng, C. Gao, G. Yu, N. Sang, Learning a discriminative feature network for semantic segmentation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 1857–1866.
[57] S. Woo, J. Park, J.-Y. Lee, I. So Kweon, CBAM: Convolutional block attention module, in: Proceedings of the European Conference on Computer Vision (ECCV), 2018, pp. 3–19.
[58] Q. Yan, L. Xu, J. Shi, J. Jia, Hierarchical saliency detection, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2013, pp. 1155–1162.
[59] Y. Li, X. Hou, C. Koch, J.M. Rehg, A.L. Yuille, The secrets of salient object segmentation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2014, pp. 280–287.
[60] G. Li, Y. Yu, Visual saliency based on multiscale deep features, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 5455–5463.
[61] L. Wang, H. Lu, Y. Wang, M. Feng, D. Wang, B. Yin, X. Ruan, Learning to detect salient objects with image-level supervision, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 136–145.
[62] M. Everingham, L. Van Gool, C.K.I. Williams, J. Winn, A. Zisserman, The pascal visual object classes (VOC) challenge, Int. J. Comput. Vis. 88 (2) (2010) 303–338.
[63] M. Abadi, P. Barham, J. Chen, Z. Chen, A. Davis, J. Dean, M. Devin, S. Ghemawat, G. Irving, M. Isard, et al., Tensorflow: a system for large-scale machine learning, in: Proceedings of the 12th IEEE USENIX Symposium on Operating Systems Design and Implementation (OSDI 16), 2016, pp. 265–283.
[64] D.P. Kingma, J. Ba, Adam: A Method for Stochastic Optimization, arXiv:1412.6980 (2014).
[65] P. Zhang, D. Wang, H. Lu, H. Wang, B. Yin, Learning uncertain convolutional features for accurate saliency detection, in: Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 212–221.
[66] X. Li, F. Yang, H. Cheng, W. Liu, D. Shen, Contour knowledge transfer for salient object detection, in: Proceedings of the European Conference on Computer Vision (ECCV), 2018, pp. 355–370.
[67] X. Zhang, T. Wang, J. Qi, H. Lu, G. Wang, Progressive attention guided recurrent network for salient object detection, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 714–722.

[68] N. Liu, J. Han, M.-H. Yang, PiCANet: learning pixel-wise contextual attention for saliency detection, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 3089–3098.



Zhengyi Liu is an associate professor in School of Computer Science and Technology, Anhui University, China. She received her B.S., M.S., and Ph.D. from Anhui University, China in 2001, 2004 and 2007, respectively. Her research interests include image and video processing, computer vision and deep learning.

Quanlong Li is a M.S. Candidate of Anhui University. He received his B.S. from Fuyang Normal University, China in 2017. His research interests include image and video processing and computer vision.

Wei Li is a professor in School of Computer Science and Technology, Anhui University, China. She received her B.S., M.S., and Ph.D. from Anhui University, Hefei University of Technology and Anhui University, China in 1991,2001 and 2006, respectively. Her research interests include data mining, computer vision and deep learning.