DAGNet: Exploring the Structure of Objects for Saliency Detection

Haobo Rao, Zhiheng Zhou, Bo Li*
School of Electronic and Information Engineering
South China University of Technology
Guangzhou, China
eerhb@mail.scut.edu.cn, {zhouzh, leebo}@scut.edu.cn

Xin Shu
College of Information Science and Technology
Nanjing Agricultural University
Nanjing, China
xinshu@njau.edu.cn

Abstract—**Fully Convolutional Neural Networks (FCNs) greatly promote the development of saliency detection. However, most of the FCN-based models have suffered from the structure of salient objects challenges. The extracted multi-scale features by previous models could help locate the objects with various scales, but they cannot contribute to effectively locating the objects with complex shapes, especially the salient regions that might intertwine with non-salient regions. Moreover, the style of decoder in previous models cannot adequately filter out the disturbance in low-level features, which is sub-optimal to sharpen the boundary of salient objects. In this paper, we propose DAGNet that explores the structure of salient objects from multi-level features to precisely detect salient objects. Firstly, the new dense multi-scale context extraction modules (DMCEMs) are implemented to transmit the rich structural information flow of salient objects from shallower layers to deeper layers, by which our model can locate the objects with complex shapes. Secondly, attention-based deeply refining modules (ADRMs) are designed in an effective attention-based style to effectively restore the boundary of objects stage-by-stage. In the style, the semantic information of high-level features is utilized to guide the shallow layer to filter out disturbance and refine the high-level features. Considering the salient objects surrounded by a cluttered scene, we propose a global context extraction module (GCEM) that can sufficiently understand the cluttered scene of an image from a global view. Comprehensive experiments indicate that our model is superior to 13 state-of-the-art models on 5 benchmark datasets under different evaluation metrics.**

Index Terms—**Fully convolutional neural network, Saliency detection, Structure of salient objects, Deep learning**

I. Introduction

Salient object detection aims to locate and segment the most visually distinctive regions in an image. It is widely served as the first step for many computer vision tasks, such as in [1], discriminant track is framed as a saliency problem, and solved based on discriminant center-surround saliency. Moreover, saliency detection is also applied to image editing [2], image captioning [3], [4], question answering [5].

Recently, convolutional neural networks (CNNs) have a strong ability to extract multi-level features from the initial image, which greatly facilitate the development of saliency detection. Due to rich high-level semantic information can be obtained by CNN, Some CNN-based models [6]–[8] achieve more remarkable performance than the traditional models. However, these CNN-based models extract features by stacking a series of stride convolutional and pooling layers, resulting in a significant decrease in the resolution of the initial image. Besides, each super-pixel is fed into the deep network, which is time-consuming. To address the problem, several effective saliency models [9]–[11] have been proposed based on the success of fully convolutional neural networks (FCNs) in other dense prediction tasks [12], [13]. Although these FCN-based models achieve excellent performance, there is still exist two challenges to tackle:

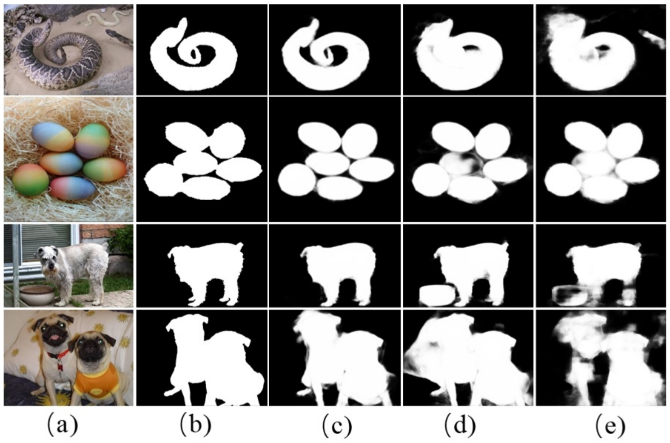

Fig. 1. Visual examples of our method and other state-of-the-art models. From left to right: (a) input image, (b) ground truth, (c) our method, (d) BMPM [14] (e) Amulet [15]

a) Due to salient objects have large variations in scale, shape, and location, it brings a challenge to precisely locate and highlight the entire salient objects. Subjected to the limitation of the receptive fields, the models of saliency detection are unable to learn rich contextual features [14], [15]. Hence, implementing modules to extract multi-scale features is an effective way to gain different receptive fields [14]–[16]. The captured different receptive fields information enables models to sense the salient objects with various scales. However, as the model deepens, the deeper layers encode a large amount of semantic information while the detailed structural information of the salient objects is lost. These models may fail to locate those salient objects with irregular and complex shapes. For

Bo Li (leebo@scut.edu.cn) is the corresponding author.

978-1-7281-6926-2/20/$31.00 ©2020 IEEE

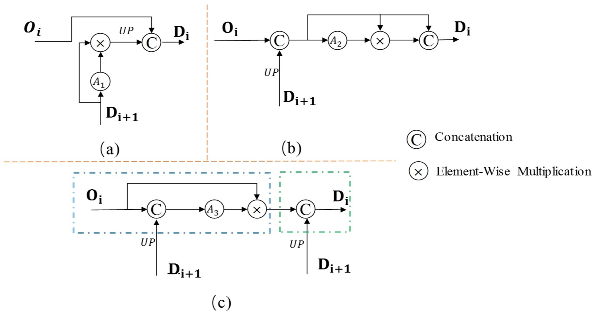

Fig. 2. Simply description of different styles for refining high-level features by low-level features. The style (c) is the simple description of the red box in Fig. 5. $\mathbf{O}_i$ is the low-level features that are fine but contain a large amount of disturbance. $\mathbf{D}_{i+1}$ is the high-level features that are rough-edged and almost noiseless. A_1, A_2, A_3 represent the attention modules in the corresponding paper. "*UP*" is an up-sampling layer and its factor is set to 2.

example, the salient regions might intertwine with non-salient regions (like row 1-2 in Fig. 1). The empirical receptive fields are much smaller than the theoretical receptive fields [17], by which the extracted features are short of the global receptive fields, and the models are unable to sufficiently understand the cluttered scene (like row 3-4 in Fig. 1).

In this paper, we solve the problems mentioned above from two aspects: **i)** we design the dense multi-scale context extraction modules (DMCEMs) to capture more effective multi-scale features. Different from the above model, DMCEMs pass the messages from shallow layers to deep layers, of which the low-level features can enhance the capability of high-level features to precisely locate salient with irregular and complex shapes. **ii)** we propose a global context extraction module (GCEM). GCEM further captures global receptive fields and makes each pixel can sense the salient objects from a global view, which effectively releases the disturbance from backgrounds.

b) The saliency map generated by the deepest layer is a low-resolution map with blurry boundaries of the salient object. How to restore it to a high-resolution map with sharp boundaries is another challenge. Previous methods [14], [18] use the low-level features of the shallower layer to refine high-level features of the deeper layer by direct concatenation [18] or element-wise addition [14] operations. Moreover, considering the different contributions of the features of each level for saliency detection, other works [19], [20] utilize attention mechanisms to alleviate it. As shown in Fig. 2, their attention modules are used to filter out the distractions from high-level features (style (a)) or the combinative features of low-level features and high-level features (style (b)). The high-level features are coarse and noiseless (high-level features aim to locate the salient objects). If the location information is inaccurate, it will be hardly rectified in the next stages. Correspondingly, the low-level features are fine-grained but contain a large amount of disturbance (redundant background regions information that is bad for the process of restoring sharp boundaries). The two styles of (a) and (b) in Fig. 2 are unable to specifically filter out the disturbance in low-level features. Hence, the residual disturbance obstructs the refining process of high-level features, causing the final prediction map cannot obtain sharp boundaries.

Based on the above analysis, we propose the attention-based deeply refining modules (ADRMs) to specifically suppress the disturbance of low-level features for restoring a high-resolution saliency map with sharp boundaries. ADRMs utilize semantic information of high-level features to guide the selection of low-level to obtain the fine-grained and noiseless features (the blue dotted box of style (c) in Fig. 2), then the features can effectively refine high-level features (the green dotted box of style (c) in Fig. 2). Besides, we embed an effective channel-wise and spatial attention module (the green and blue solid boxes in Fig. 5) into the guidance process to distinguish the contribution of each spatial position and different feature channels.

Our main contributions are summarized as three folds:

- We design the new dense multi-scale context extraction modules to make the model more robust for the salient objects which have large variations in scale and shape. Moreover, the global context extraction modules are proposed to gain global receptive fields so that the model can precisely locate the salient objects from a cluttered backgrounds.
- We propose the effective attention-based deeply refining modules to specifically suppress the disturbance of low-level features, which is used to restore the low-resolution saliency map with blurry boundaries to the high-resolution map with sharp boundaries.
- We compare our model with 13 state-of-the-art methods on five datasets. Comprehensive experiment results indicate that our proposed model performs favorably against state-of-the-art models under different evaluation metrics.

II. Related Work

Over the past two decades, a large number of salient object detection methods have been proposed. Traditional models extract low-level hand-crafted features using heuristic saliency priors [21]–[23]. These models simply utilize low-level features without the assistance of high-level semantic information, which make the model cannot precisely highlight the entire salient objects.

Due to the superior feature extraction ability of CNN, several early deep algorithms [6]–[8] predict the salient score of super-pixel. Despite the excellent performance achieved by these approaches, each super-pixel is fed into the deep network, which is time-consuming. Besides, these CNN-based models also decrease the resolution of the original image. Therefore, many effective FCN-based models [11], [24] are developed.

Recently, based on the following observations: High-level features aim to provide the location information, and low-level features containing rich spatial details that can refine the boundaries of salient objects. Many works [14]–[16], [18], [25] integrate multi-level features to enhance the performance of saliency detection. Li *et al.* [25] directly aggregate the features of each level of the backbone to obtain an effective feature presentation. Hou *et al.* [16] embed the short connection into the skip-layer structure within the HED [13]. Luo *et al.* [18] design a multi-resolution 4×5 grid network to extract global information and local contrast feature. Zhang *et al.* [15] aggregate multi-level features into five different resolutions, then the model predicts the saliency maps in each resolution and fuses them to produce a more fine saliency map. Although the above works obtain remarkable results, they integrate multi-level features without distinction. Therefore, several models [14], [19], [20] utilize attention mechanisms to alleviate the problem. Such as Lu *et al.* [14] design gate function to selectively pass the message between the shallow layer and deep layer. Zhang *et al.* [19] employ channel-wise and spatial attention modules to filter out disturbance of the high-level features of deeper layers recursively, then the low-level features are utilized to refine the coarse high-level features. Liu *et al.* [20] generate global and local attention maps to select effective features for refining coarse high-level features.

In this paper, their models may not be robust to the salient objects with various shapes or the salient objects surrounded by a cluttered background. And their attention-based style cannot specifically suppress the distractions in low-level features, obstructing the process of sharpening the boundaries of saliency objects. Different from the models above, our model passes low-level features containing rich structural information of salient objects to deeper layers, instead of transmitting high-level features containing rich semantic information to shallower layers. we also design a module to extract global context for precisely locating the entire salient objects surrounded by a cluttered background. Besides, we propose a new deeply refining style to effectively refine the coarse high-level features from deeper layer, and an effective attention module is embedded into the style to weigh each spatial position and different feature channels.

III. Proposed Method

In this paper, the DAGNet is proposed to explore the structure of salient objects for accurate saliency detection. In Sec. III-A, the overall architecture of the proposed network will be described. Then Sec. III-B provides the detailed principle of dense multi-scale context extraction modules (DMCEMs). Next, the global context extraction module (GCEM) will be given in Sec. III-C. At last, We introduce the attention-based deeply refining modules (ADRMs).

A. Overall Architecture

Fig. 3 shows the details of the overall architecture. The proposed model is an encoder-decoder fashion with VGG-16 [26] network as the pre-trained model. The VGG-16 network is modified to fit the saliency detection. All the fully connected layers and the last pooling layer of the VGG-16 network are removed to focus on pixel-wise prediction and maintain more details of the deepest layer, respectively. For the input image with size $H \times W$, the revised VGG-16 network extracts multi-level features at five stages, which denoted as $\mathbf{F}_i, i = 1, 2, 3, 4, 5$ with resolution $[\frac{H}{2^{i-1}}, \frac{W}{2^{i-1}}]$. Then these multi-resolution features are fed into the DMCEMs to capture more effective multi-scale features representation. Besides, GCEM is added on the deepest layer of the VGG-16 network to further gain global receptive fields so that the model can precisely locate the salient objects in a cluttered background. Finally, ADRMs are designed to deeply refine coarse high-level features by fine-grained low-level features, and generate a series of prediction maps $\mathbf{S}_i, i = 1, 2, 3, 4, 5$. $\mathbf{S}_1$ is the final saliency map with sharp boundaries.

B. Dense multi-scale context extraction module

The salient objects have large variations in scale and shape. Previous methods [14]–[16] extract multi-scale features from each side output of the backbone. These multi-scale features are robust to the salient objects with various scales, but not to the objects with various shapes. In this paper, we propose the novel dense multi-scale context extraction modules (DMCEMs) to capture more effective multi-scale features representation.

Fig. 3 shows the details of the DMCEM-i. To handle various scales of objects, we make DMCEM-i to capture different receptive fields by stacking a series of pooling layers and up-sampling layers in parallel. After pooling and up-sampling layers, DMCEM-i generates the original multi-scale features $\mathbf{M}_i = \{\boldsymbol{m}_i^k, k = 1, 2, 3, 4, 5\}$.

The original multi-scale features $\mathbf{M}_i$ is robust to the salient objects with various scales, but it may not effectively deal with those salient objects with complex shapes. Low-level features of shallow layers contain rich structural information of salient objects, which can be utilized to improve the capability of locating those salient objects with complex shapes. Hence, DMCEM-i utilizes dense connection to deliver the low-level features $\mathbf{H}_j = \{\boldsymbol{h}_j^k, j = 1, \cdots, i-1\}$ to the current level who are at the deeper layer. Then, high-level features $\mathbf{M}_i$ can concatenate resolution-matching features of $\mathbf{H}_j$ across channels to enhance original multi-scale $\mathbf{M}_i$. Finally, DMCEM-i generates the more effective multi-scale features $\mathbf{O}_i$. The $\mathbf{O}_i$ not only adapts to the objects with various scales but also can locate those objects with complex shapes. The whole process can be formulated as follows:

$$\boldsymbol{h}_i^k = \psi(\phi(Cat(\boldsymbol{m}_i^k, \mathbf{h}_1^k, \cdots, \mathbf{h}_{i-1}^k), \theta_k)) \quad (1)$$

$$\mathbf{O}_i = \psi(\phi(Cat(Down(\boldsymbol{h}_i^{1\sim(i-1)}), \boldsymbol{h}_i^i, Up(\boldsymbol{h}_i^{(i+1)\sim 5})), \omega) \quad (2)$$

Where $i \in \{1, 2, 3, 4, 5\}$, $k \in \{1, 2, 3, 4, 5\}$. $\phi(\cdot, \theta)$ is a 1×1 convolutional layer with parameter θ, $Cat(\cdot)$ is the cross-channel concatenation operation. $\psi(\cdot)$ is ReLU activate function. $Down(\cdot)$ and $Up(\cdot)$ represents pooling layers and up-sampling layers, respectively. $Down(\boldsymbol{h}_i^{1\sim(i-1)})$ denotes the set of $Down(\boldsymbol{h}_i^1), \cdots Down(\boldsymbol{h}_i^{i-1})$, $Up(\boldsymbol{h}_i^{(i+1\sim 5)})$ denotes the set of $Up(\boldsymbol{h}_i^{i+1}), \cdots Up(\boldsymbol{h}_i^5)$.

DMCEMs passe the low-level features to deeper layers by dense connection, which has two advantages: **i**) high-level features fuse the rich structural information from low-level features, which makes the model can precisely locate the salient objects with complex shapes. **ii**) Because of the

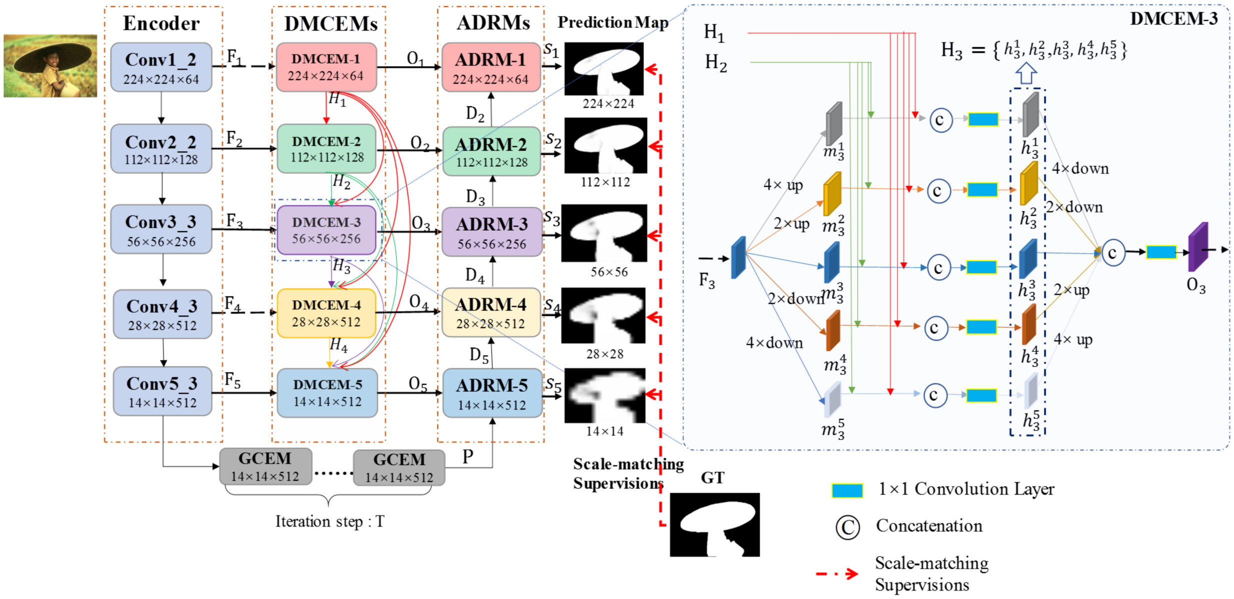

Fig. 3. The overall network architecture is on the left. The input image is fed to encoder to extract multi-level features $\mathbf{F}_i$. Then a series of GCEM are built on the deepest layer $\mathbf{F}_5$ to further capture global contextual information $\mathbf{P}$. We pass $\mathbf{F}_i$ into the DMCEM-i (the illustration is on the right). DMCEM-i generates the $\mathbf{H}_i$ and the more effective multi-scale features $\mathbf{O}_i$. Low-level features $\mathbf{H}_i$ is passed to other high-level by dense connection. $\mathbf{O}_i$ is fed into the ADRM-i to refine the coarse features of the deeper layer.

gradient backpropagation, the low-level can learn richer spatial details, especially the structural information of salient objects. After the DMCEMs, model gains a series of more effective multi-scale features, which can effectively locate the salient objects with various scales and shapes.

C. Global context extraction module

To obtaining global receptive fields, Wang *et al.* [27] apply a PPM to extract global contextual information. However, PPM utilizes multiple pooling layers with different kernel sizes, which loses abundant spatial details. Other works utilize convolutional layer with larger kernel size, which is time-consuming and increases the amount of memory and calculation. In this subsection, we propose the global context extraction module (GCEM) which combines different convolutional layers without pooling layers for retaining the rich spatial details in the deepest layer.

Fig. 4 shows the details of GCEM. GCEM separates the $\mathbf{F}_5$ into four equal parts $\{\mathbf{X}_1, \mathbf{X}_2, \mathbf{X}_3, \mathbf{X}_4\}$ along the channels by a 1×1 convolutional layer, which can reduce the computational overhead and the number of parameters of the module. GCEM employs the combined convolutional layers ($1\times k + k\times 1$, $k\times 1 + 1\times k$ and 3×3) on every $\mathbf{X}_i$ without pooling layers. Cascading the output of combined convolutional layers from $\mathbf{C}_1$ to $\mathbf{C}_4$ makes GCEM to obtain global receptive fields without any pooling layers. For $\mathbf{X}_i$, GCEM concatenates $\mathbf{X}_i$ and $\mathbf{C}_{i-1}$ across channels to generate the $\mathbf{K}_i$ and utilizes the combined convolutional layers to further enlarge receptive fields. Furthermore, we cascade a series of GCEM to make sure that each pixel can sense the whole image from a global view. The whole process is performed as follows.

$$\mathbf{K}_i = \phi((\mathit{Cat}(\mathbf{X}_i, \mathbf{C}_{i-1}), \kappa_i) \quad i = 2,3,4 \tag{3}$$

$$\mathbf{C}_i = \begin{cases} \psi_1(\mathbf{X}_i, \omega_i) + \psi_2(\mathbf{X}_i, \mu_i) + \psi_{3\times3}(\mathbf{X}_i, \theta_i) & i = 1 \\ \psi_1(\mathbf{K}_i, \omega_i) + \psi_2(\mathbf{K}_i, \mu_i) + \psi_{3\times3}(\mathbf{K}_i, \theta_i) & i = 2,3,4 \end{cases} \tag{4}$$

$$\mathbf{P} = \phi(\mathit{Cat}(\mathbf{C}_1, \mathbf{C}_2, \mathbf{C}_3, \mathbf{C}_4), \nu) + \mathbf{F}_5 \tag{5}$$

Where $\phi(\cdot, \kappa_i)$ is a 1×1 convolutional layer with the parameter κ_i, and $\psi_1(\cdot, \omega_i)$ represents the combination of $1\times k + k\times 1$ convolutional layer with parameter ω_i. $\psi_2(\cdot, \mu_i)$ represents the combination of $k\times1+1\times k$ convolutional layer with parameter μ_i. $\psi_{3\times3}(\cdot, \theta_i)$ is a 3×3 convolutional layer with parameter θ_i. $\mathbf{P}$ is the final output of the GCEM. $\phi(\cdot, \nu)$ is a 1×1 convolutional layer with the parameter ν.

D. Attention-based deeply refining module

The saliency map generated by the deepest layer is low-resolution with fuzzy boundaries. Recent works [19], [20] utilize attention mechanism (the style (a) and (b) in Fig. 2) to integrate multi-level features to restore a high-resolution saliency map with sharp boundaries. However, the two styles are unable to specifically filter out the distractions in low-level features, which is sub-optimal to fully refine high-level features. Different from them, the attention-based deeply refining modules (ADRMs) is designed in a new attention-based style to effectively restore the boundary of objects stage-by-stage. Moreover, motivated by [28], ADRMs embed channel-wise and spatial attention modules into the style to weigh the contribution of each spatial position and different feature channels.

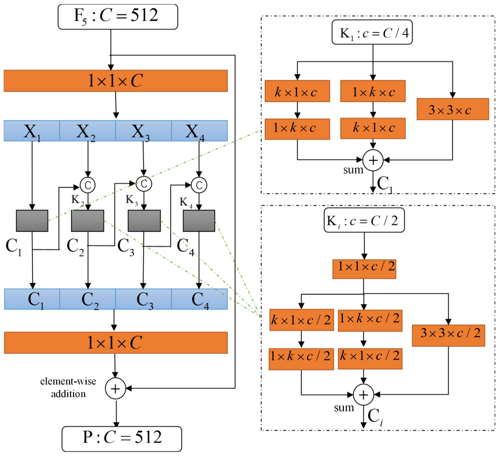

Fig. 4. The details of GCEM. The orange box represents the convolutional layer and the "$1 \times k \times C$" denotes the convolution kernel and channel number of the convolutional layer is $1 \times k$ and C. In this paper, we set the k equal to 7. "sum" denotes the element-wise addition operation.

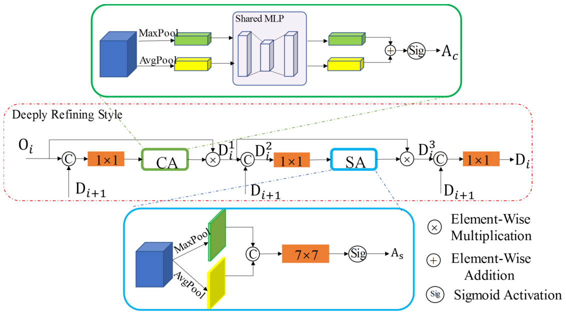

Fig. 5. The details of ADRM-i. "CA" represents the channel-wise attention module. "SA" represents the spatial attention module. The orange box represents the convolutional layer and the "7×7" denotes the convolution kernel size of convolutional layer is 7×7. The "1×1" denotes the convolution kernel size of convolutional layer is 1×1.

Fig. 5 shows the details of the proposed ADRM-i. A new deeply refining style (the red dotted box) is designed in the ADRM-i. Each channel of low-level features encodes different features about the image. Some channels encode effective features about salient objects while others have a strong response to the background regions. Instead of directly adding channel-wise attention module on low-level features, ADRM-i utilizes the strong semantic information of high-level features $\mathbf{D}_{i+1}$ to guide low-level features $\mathbf{O}_i$ by a channel-wise attention module. After the channel-wise guiding process, the disturbance of the $\mathbf{O}_i$ can be filtered out along the channel axis and the output feature $\mathbf{D}_i^1$ more focus on the channels that have a high response to the foreground. Similarly, not each spatial pixel contributes to saliency detection. Some background pixels may cause serious distractions. Therefore, ADRM-i also utilizes $\mathbf{D}_{i+1}$ to guide $\mathbf{D}_i^1$ by a spatial attention module. After the spatial guiding process, $\mathbf{D}_i^1$ can further highlight the salient regions and suppress background pixels distractions. Finally, the fine-grained and noiseless $\mathbf{D}_i^3$ deeply refines the high-level features $\mathbf{D}_{i+1}$. The whole process of the style is formulated as follows:

$$\begin{aligned}\mathbf{A}_c = \sigma(&\mathit{MLP}(\mathit{AP}(\phi(\mathit{Cat}(\mathbf{O}_i, \mathbf{D}_{i+1})), \tau), w)\\ &+\mathit{MLP}(\mathit{MP}(\phi(\mathit{Cat}(\mathbf{O}_i, \mathbf{D}_{i+1})), \tau), w))\end{aligned} \tag{6}$$

$$\mathbf{D}_i^1 = \mathbf{O}_i \otimes \mathbf{A}_c \tag{7}$$

$$\mathbf{D}_i^2 = \phi(\mathit{Cat}(\mathbf{D}_i^1, \mathbf{D}_{i+1}), \epsilon) \tag{8}$$

$$\mathbf{A}_s = \sigma(\phi_{7\times7}(\mathit{Cat}(\mathit{AP}(\mathbf{D}_i^2), \mathit{MP}(\mathbf{D}_i^2)), \alpha) \tag{9}$$

$$\mathbf{D}_i = \phi(\mathit{Cat}((\mathbf{D}_i^1 \otimes \mathbf{A}_s), \mathbf{D}_{i+1}), \lambda) \tag{10}$$

Where $\sigma(\cdot)$ is the sigmoid function. $\mathit{MLP}(\cdot, w)$ is the multi-layer perceptron with the shared parameter w. $\mathit{AP}(\cdot)$ and $\mathit{MP}(\cdot)$ denotes the average-pooling and max-pooling layer respectively. $\otimes$ is the element-wise multiplication. $\phi(\cdot)$ is a 1×1 convolutional layer to reduce the number of channels to the same as $\mathbf{O}_i$. $\phi_{7\times7}(\cdot, \lambda)$ is a 7×7 convolutional layer with parameter α.

We adopt deep supervision to facilitate the model training. To be specific, for each ADRM-i, we add a 1×1 convolutional layer with sigmoid activation on $\mathbf{D}_i$ to obtain the saliency map $\mathbf{S}_i$, then we use cross-entropy loss between the $\mathbf{S}_i$ and corresponding ground truth.

IV. Experiments

A. Experiment Setup

Implementation Details. VGG-16 network initializes the parameters of the first 13 convolutional layers and the rest one are initialized by Xavier [29]. Our model is trained by the Adam [30] optimizer with an initial learning rate of 1e-5 which is divided by 5 after 31 epochs. The reduction ratio r is set to 16. Meanwhile, similarly to [20], the weight of the cross-entropy loss in ADRM-i, $i = 5, 4, 3, 2, 1$ are set to 0.5, 0.5, 0.5, 0.8, 1 respectively. During training, we use horizontal flipping for data augmentation. Then each image is resized to 256×256 and randomly cropped to 224×224. During testing, we directly resize each image to 224×224, and these images are fed into our model to generate corresponding saliency maps. The batch size is set to 4. Our model is trained on the DUTS-TR dataset. It is trained for 50 epochs in total and takes about 50 hours on a GTX Titan XP GPU. The code can be found at https://github.com/CVisionProcessing.

Datasets. Our model is compared with other methods on five benchmark datasets: ECSSD [31], DUTS-TE [32], HKU-IS [6], PASCAL-S [33], SOD [34].

Evaluation metrics. Precision-recall (PR) curves, mean absolute error (MAE) and F-measure score (F_β) are used to evaluate the performance of the proposed model and other models. PR curve is a popular way to evaluate the predicted saliency map. The saliency map is binarized by a threshold that slides from 0 to 255. Then the binarized saliency map is compared with the ground truth to calculate the value of precision and recall. The F_β is a comprehensive performance measurement.

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

Where β^2 is set to 0.3 to emphasize that precision is more important than recall. The MAE calculates the average difference between the predicted map and the corresponding ground truth.

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

Where S and G are predicted saliency map and corresponding ground truth, respectively.

B. Ablation Studies

In this section, a series of ablation experiments are conducted to verify the effectiveness of each module employed in our model. We also investigate the performance of the component of DMCEMs, GCEM, and ADRMs.

TABLE I
ABLATION EXPERIMENTS ON THREE DATASETS. THE BEST RESULTS ARE HIGHLIGHTED IN RED.

No.	Modules			ECSSD		PASCAL-S		DUTS-TE	
	DMCEMs	GCEM	ADRMs	MaxF	MAE	MaxF	MAE	MaxF	MAE
1				0.903	0.061	0.844	0.087	0.806	0.065
2	√			0.919	0.052	0.854	0.079	0.912	0.042
3		√		0.931	0.043	0.864	0.073	0.850	0.046
4			√	0.932	0.045	0.864	0.075	0.850	0.048
5	√	√		0.932	0.044	0.869	0.072	0.857	0.046
6	√	√	√	0.939	0.040	0.874	0.069	0.866	0.043

The Effectiveness of DMCEMs. A series of experiments are conducted to invalidate the effectiveness of DMCEMs. First, we only add DMCEMs in the baseline (No.2 in Tab. I). The score of both F-measure and MAE all obviously surpass the FCN baseline (No.1 in Tab. I) on three datasets. Second, to investigate the performance of the component in DMCEMs, we remove dense connection from DMCEMs (No.1 in Tab. II). We also replace DMCEMs with MCFEM (No.2 in Tab. II). Compared to the whole model (No.9 in Tab. II), the score of the two experiments drop significantly. These results demonstrate two things: **i**) transmitting rich structural information of salient objects from low-level to high-level effectively locates the salient objects with complex shapes. **ii**) DMCEMs can extract more effective multi-scale features than MCFEM. The multi-scale features not only are robust to the salient objects with various scales, but also to the objects with complex shapes.

The Effectiveness of GCEM. First, we perform a series of experiments (No.3, No.4, and No.9 in Tab. II) to find the optimal value of T. The results indicate the performance is best when T = 2. We only add GCEM on baseline (No.3 in Tab. I) or replace it with PPM (No.5 in Tab. II). The numerical results verify that GCEM can effectively obtain global receptive fields without losing spatial details in the deepest layer, which makes our model can precisely locate the salient objects surrounded by a cluttered background. Besides, we embed both DMCEMs and GCEM into the baseline (No.5 in Tab. I). The result of further improvement reflects that the two modules work harmoniously, effectively alleviating the challenge that salient objects have large variations in scale, shape and location.

TABLE II
THE PERFORMANCE OF COMPONENT OF EVERY MODULE. THE BEST RESULTS ARE HIGHLIGHTED IN RED.OURS REPRESENTS COMBINATION OF DMCEMS, GCEM(T=2), ADRMS.

No.	model setting	ECSSD		PASCAL-S		DUTS-TE	
		MaxF	MAE	MaxF	MAE	MaxF	MAE
1	w/o dense connection	0.932	0.042	0.870	0.071	0.855	0.046
2	MCFEM [14]	0.933	0.044	0.874	0.073	0.854	0.051
3	GCEM with T=1	0.933	0.043	0.873	0.073	0.856	0.050
4	GCEM with T=3	0.936	0.043	0.876	0.072	0.863	0.046
5	PPM [11]	0.935	0.041	0.873	0.070	0.853	0.049
6	w/o attention	0.932	0.044	0.869	0.072	0.857	0.046
7	Style (a)	0.916	0.050	0.855	0.077	0.829	0.052
8	Style (b)	0.934	0.043	0.876	0.069	0.860	0.044
9	Ours	0.939	0.041	0.874	0.069	0.866	0.043

The Effectiveness of ADRMs. We only equip ADRMs at baseline (No.4 in Tab. I). The score increased markedly compared to the baseline. Then, we remove both channel-wise and spatial attention modules from ADRMs (No.6 in Tab. II). It proves that the attention modules can effectively filter out distractions and capture distinguishing features in low-level features. Besides, to investigate the performance of the proposed style of ADRMs (style (c) in Fig. 2), we replace it with style (a) and style (b) in Fig. 2, respectively (No.7 and No.8 in Tab. II). The proposed style achieves best performance (No.9 in Tab. II). It proves that the proposed style can effectively refine the high-level features and restore a high-resolution saliency map with sharp boundaries. Finally, we equip ADRM at the combination of DMCEMs and GCEM (No.6 of Tab. I). The score of both F-measure and MAE significantly improves. It demonstrates that the proposed model works collaboratively for accurate saliency detection.

C. Comparisons to State-of-the-Art Methods

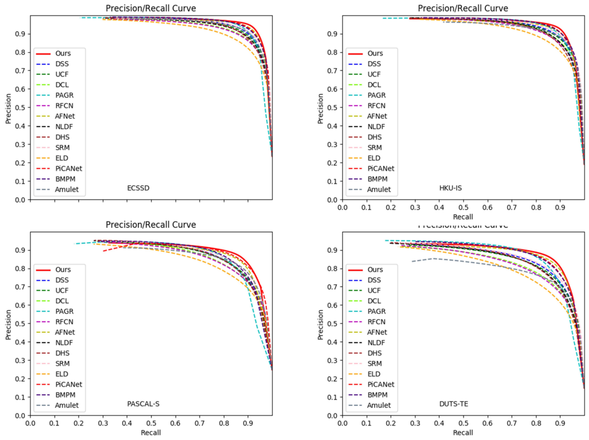

Fig. 6. Precision-Recall curves of our model and 13 state-of-the-art methods on four datasets

Our proposed model is compared with other 13 state-of-the-art models, including AFNet [36], PiCANet [20], PAGR [19],

TABLE III
QUANTITATIVE COMPARISONS OF DIFFERENT MODELS ON 5 BENCHMARK DATASETS. THE BEST THREE RESULTS ARE HIGHLIGHTED IN RED, GREEN AND BLUE.

Method	ECSSD		HKU-IS		SOD		PASCAL-S		DUTS-TE	
	MaxF	MAE	MaxF	MAE	MaxF	MAE	MaxF	MAE	MaxF	MAE
UCF [35]	0.911	0.078	0.886	0.074	0.803	0.169	0.846	0.128	0.771	0.117
SRM [11]	0.917	0.054	0.906	0.046	0.845	0.132	0.847	0.085	0.827	0.059
RFCN [11]	0.898	0.095	0.898	0.080	0.807	0.166	0.850	0.132	0.783	0.090
NLDF [18]	0.905	0.063	0.902	0.048	0.837	0.123	0.845	0.112	0.812	0.066
ELD [9]	0.867	0.079	0.839	0.074	0.760	0.154	0.773	0.123	0.738	0.093
DSS [16]	0.916	0.053	0.911	0.040	0.846	0.126	0.846	0.112	0.825	0.057
DHS [10]	0.907	0.060	0.902	0.054	0.827	0.133	0.841	0.111	0.829	0.065
DCL [25]	0.901	0.075	0.885	0.137	0.825	0.198	0.823	0.189	0.782	0.150
Amulet [15]	0.915	0.059	0.896	0.052	0.808	0.145	0.858	0.103	0.778	0.085
BMPM [14]	0.928	0.044	0.920	0.038	0.851	0.106	0.862	0.076	0.850	0.049
PAGR [19]	0.901	0.075	0.885	0.137	0.825	0.198	0.823	0.189	0.782	0.150
PiCANet [20]	0.931	0.047	0.921	0.042	0.855	0.108	0.880	0.088	0.851	0.054
AFNet [36]	0.935	0.042	0.923	0.036	-	-	0.868	0.071	0.862	0.046
Ours	0.939	0.040	0.926	0.036	0.853	0.105	0.874	0.069	0.866	0.043

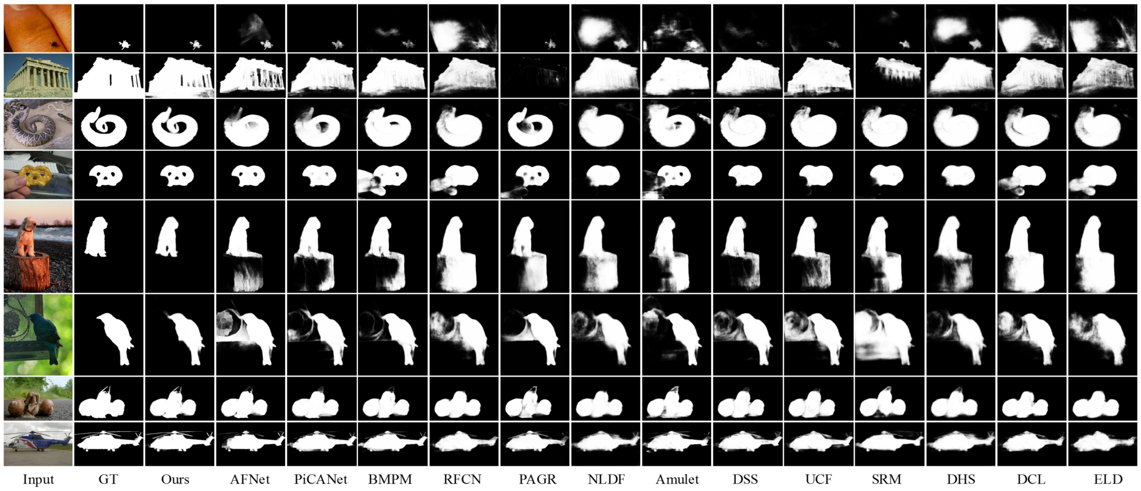

Fig. 7. Visual comparisons of our model and 13 state-of-the-art methods.

BMPM [14], Amulet [15], DCL [25], DSS [16], DHS [10], ELD [9], NLDF [18], RFCN [11], SRM [11], UCF [35], MDF [6]. For fair comparison, saliency maps are generated with running original codes with recommended parameters setting or provided by the authors.

Quantitative Comparisons. Tab. III shows the quantitative results of the proposed model and other 13 state-of-the-art models on five datasets. For different evaluation metrics, our model is superior to other models in most datasets, which indicates the effectiveness of our model.

PR Curves. We also draw the PR curves on four datasets as shown in Fig. 6. The PR curves of our model surpass other methods on four datasets. It verifies our method is more effective than other methods.

Visual Comparisons. Fig. 7 shows some saliency maps generated by our model and other state-of-the-art methods. These saliency maps are from test datasets. The visual results indicate that our model can precisely locate salient objects and restore a high-resolution saliency map with sharp boundaries. For example, the small objects and large objects (like row 1-2 in Fig. 7) are all located. It means that our model is more robust to the salient objects with various scales. Also, other models are unable to precisely detect those salient objects with irregular and complex shapes (like row 3-4 in Fig. 7). Due to the foreground regions intertwined with some background regions, the background regions of the eggs and snake are possibly misidentified as foreground. But our model can precisely locate the objects. These visual results also reflect the effectiveness of the proposed DMCEMs. The images with low contrast between foreground and background (like row 5-6 in Fig. 7) can also be precisely detected, which indicates that our model can sufficiently understand the cluttered scene of the image. Hence, our model is more robust to the salient objects surrounded by clutter background than other models, which verifies the effectiveness of the proposed GCEM. Besides, the saliency map of the snail and helicopter (row 7-8 in Fig. 7) generated by our model have more sharp boundaries than other models. It verifies that the ADRMs can effectively restore a saliency map with sharp boundaries.

V. Conclusion

In this paper, we propose DAGNet for accurate saliency detection. Firstly, the proposed dense multi-scale context extraction modules (DMCEMs), in addition to extracting original multi-scale features through a series of pooling layers, is more important in transmitting the low-level features containing rich structural information of salient objects to deeper layers to obtain more effective multi-scale features. DMCEMs effectively alleviate the challenge, of which salient objects have large variation in shapes. Secondly, the global context extraction module (GMCE) is designed to further extract global receptive fields so that the model can precisely locate the salient objects from a cluttered background. Thirdly, the attention-based deeply refining modules (ADRMs) are learning to gradually recover the low-resolution saliency map with blurry boundaries to the high-resolution map with sharp boundaries. The comprehensive experiments indicate that our model is superior to other state-of-the-art models on five benchmark datasets.

VI. Acknowledgments

This research was supported by National Key R&D Program of China (No. 2017YFC0806000), National Natural Science Foundation of China (No. 11627802, 51678249, 61871188), State Key Lab of Subtropical Building Science, South China University of Technology (2018ZB33), and the State Scholarship Fund of China Scholarship Council (201806155022).

References

[1] V. Mahadevan and N. Vasconcelos, "Saliency-based discriminant tracking," in *IEEE Conference on Computer Vision & Pattern Recognition*, 2009.

[2] M. M. Cheng, F. L. Zhang, N. J. Mitra, X. Huang, and S. M. Hu, "Repfinder: Finding approximately repeated scene elements for image editing," *Acm Transactions on Graphics*, vol. 29, no. 4, pp. 157–166, 2010.

[3] A. Das, H. Agrawal, L. Zitnick, D. Parikh, and D. Batra, "Human attention in visual question answering: Do humans and deep networks look at the same regions?" *Computer Vision & Image Understanding*, p. S1077314217301649, 2016.

[4] H. Fang, S. Gupta, F. N. Iandola, R. K. Srivastava, L. Deng, P. Dollr, J. Gao, X. He, M. Mitchell, and J. C. Platt, "From captions to visual concepts and back." 2015.

[5] Y. Lin, Z. Pang, D. Wang, and Y. Zhuang, "Task-driven visual saliency and attention-based visual question answering," *arXiv preprint arXiv:1702.06700*, 2017.

[6] G. Li and Y. Yu, "Visual saliency based on multiscale deep features," in *Computer Vision & Pattern Recognition*, 2015.

[7] L. Wang, H. Lu, R. Xiang, and M. H. Yang, "Deep networks for saliency detection via local estimation and global search," in *IEEE Conference on Computer Vision & Pattern Recognition*, 2015.

[8] Z. Rui, W. Ouyang, H. Li, and X. Wang, "Saliency detection by multi-context deep learning," in *Computer Vision & Pattern Recognition*, 2015.

[9] 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*, 2016, pp. 660–668.

[10] N. Liu and J. Han, "Dhsnet: Deep hierarchical saliency network for salient object detection," in *Computer Vision & Pattern Recognition*, 2016.

[11] L. Wang, L. Wang, H. Lu, P. Zhang, and R. Xiang, "Salient object detection with recurrent fully convolutional networks," *IEEE Transactions on Pattern Analysis & Machine Intelligence*, vol. PP, no. 99, pp. 1–1, 2018.

[12] 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*, 2015, pp. 3431–3440.

[13] S. Xie and Z. Tu, "Holistically-nested edge detection," *International Journal of Computer Vision*, 2017.

[14] Z. Lu, D. Ju, H. Lu, H. You, and W. Gang, "A bi-directional message passing model for salient object detection," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[15] P. Zhang, W. Dong, H. Lu, H. Wang, and R. Xiang, "Amulet: Aggregating multi-level convolutional features for salient object detection," in *IEEE International Conference on Computer Vision*, 2017.

[16] Q. Hou, M. M. Cheng, A. Borji, Z. Tu, and P. Torr, "Deeply supervised salient object detection with short connections," in *Computer Vision & Pattern Recognition*, 2017.

[17] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, "Object detectors emerge in deep scene cnns," *Computer Science*, 2014.

[18] Z. Luo, A. Mishra, A. Achkar, J. Eichel, and P. M. Jodoin, "Non-local deep features for salient object detection," in *Computer Vision & Pattern Recognition*, 2017.

[19] 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*, 2018, pp. 714–722.

[20] N. Liu, J. Han, and 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.

[21] R. Achantay, S. Hemamiz, F. Estraday, and S. Ssstrunky, "Frequency-tuned salient region detection," in *IEEE Conference on Computer Vision & Pattern Recognition*, 2009.

[22] W. Zhu, L. Shuang, Y. Wei, and S. Jian, "Saliency optimization from robust background detection," in *IEEE Conference on Computer Vision & Pattern Recognition*, 2014.

[23] Z. Jiang and L. S. Davis, "Submodular salient region detection," in *Computer Vision & Pattern Recognition*, 2013.

[24] X. Li, L. Zhao, L. Wei, M. Yang, F. Wu, Y. Zhuang, H. Ling, and J. Wang, "Deepsaliency: Multi-task deep neural network model for salient object detection," *IEEE TIP*, 2016.

[25] G. Li and Y. Yu, "Deep contrast learning for salient object detection," in *Computer Vision & Pattern Recognition*, 2016.

[26] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *Computer Science*, 2014.

[27] 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*, 2017.

[28] S. Woo, J. Park, J. Y. Lee, and I. S. Kweon, "Cbam: Convolutional block attention module," *ECCV*, 2018.

[29] X. Glorot and Y. Bengio, "Understanding the difficulty of training deep feedforward neural networks," in *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, 2010, pp. 249–256.

[30] D. Kingma and J. Ba, "Adam: A method for stochastic optimization," *Computer Science*, 2014.

[31] Q. Yan, L. Xu, J. Shi, and J. Jia, "Hierarchical saliency detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2013, pp. 1155–1162.

[32] L. Wang, H. Lu, Y. Wang, M. Feng, and R. Xiang, "Learning to detect salient objects with image-level supervision," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[33] L. Yin, X. Hou, C. Koch, J. M. Rehg, and A. L. Yuille, "The secrets of salient object segmentation," in *IEEE Conference on Computer Vision & Pattern Recognition*, 2014.

[34] V. Movahedi and J. H. Elder, "Design and perceptual validation of performance measures for salient object segmentation," in *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition-Workshops*. IEEE, 2010, pp. 49–56.

[35] P. Zhang, D. Wang, H. Lu, H. Wang, and B. Yin, "Learning uncertain convolutional features for accurate saliency detection," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 212–221.

[36] M. Feng, H. Lu, and E. Ding, "Attentive feedback network for boundary-aware salient object detection," in *Computer Vision & Pattern Recognition*, 2019.