DAANet: Dual Attention Aggregating Network for Salient Object Detection*

Yijie Li, Hewei Wang, Zhenqi Li, Shaofan Wang, Soumyabrata Dev, Guoyu Zuo

***Abstract*— When the Convolutional neural network (CNN) has been introduced for computer vision, many researches use CNN-based model to perform salient object detection (SOD). In recent years, The feature pyramid network (FPN) based structure is more popular for salient object detection tasks. In this paper, to improve the overall performance of salient object detection tasks, we propose a dual attention aggregating network (DAANet), which is an FPN-based deep convolutional neural network with a dual attention aggregation module (DAAM) and dilated refinement block (DRB). The DRB module uses convolutions with different dilation rates to expand the receptive field. The DAAM considers the salient map prediction from the low-level output as pseudo-attention which can efficiently aggregate multi-scale information. The convolution block attention module (CBAM) in DAAM can refine the aggregation of pseudo-attention which enables better performance. We evaluate DAANet on six benchmark datasets that prove the effectiveness of DAANet and its components. Our implementation can be found at: `https://github.com/Att100/DAANet`.**

I. Introduction

Salient object detection refers to the identification of vital visual information analog to the human attention mechanism. In robotics and automation, the salient object detection model can be used to determine the objective target from background full of complex content and robots can use the visual information provided by salient object detection model to plan the following actions, such as catching or moving an object. Most of the proposed methods of this computer vision task have been widely used in film and television production, and image matting. The early approaches for salient object detection are usually based on traditional visual extraction methods. Some of those methods use handcraft features or filters to extract important regions, and others may adopt the graph-based method to transfer the computer vision task to a graph optimization problem. When convolutional neural networks have been widely used in all kinds of computer vision tasks [1]–[5], many research works on CNN-based salient object detection have been released. Before the introduction of classical image classification networks, such as VGG [6] and ResNet [7], many CNN-based approaches used a simple design with the combination of CNN and fully connected layers to generate salient masks. Those early methods usually yield limited performance. Since 2016, many classical pre-trained image classification models have been used as feature extraction backbone network in salient object detection which significantly improves the overall performance of prediction. In recent years, FPN [8] and U-Net [9] have become the most popular structure in SOD. The shortcut connection between the encoders and decoders enhances the ability of feature aggregation.

However, the encoder-decoder based approach with straightforward design in decoders is hard to capture the details of a specific salient region and the drawbacks of the existing SOD methods motivate us to use dual attention to remit these issues:

- Several methods are estimated based on an encoder-decoder structure with straightforward design in decoders, leading to a sub-optimal capacity of capturing the details in a specific salient region.
- Multiple existing salient object detection methods directly use the feature maps from the backbone encoder without further processing which results in the limitation of the receptive field.

To ameliorate the aforementioned issues, we proposed a dual attention aggregation network (DAANet) to enhance the feature aggregation in decoding stages. DAANet includes a backbone network as an encoder and several dual attention aggregating modules (DAAM) as decoders, while the DRB is between encoders and decoders. The DRB uses several convolution layers with increasing dilation rates to expand the receptive field. The DAAM has two attention mechanisms while one of which is pseudo-attention while it considers the salient map from the previous decoding stage as attention weights. Another attention is a modified version of the convolution block attention module (CBAM) [10] which refines the previous aggregation results to enable more accurate prediction. We also use the combined binary cross entropy loss and intersection of union loss to improve the supervision.

The main contributions of DAANet are threefold:

- We proposed a novel decoder module for salient object detection: a dual attention aggregation module (DAAM), which can achieve better performance than baseline models.
- We introduced the dilated refinement block (DRB) for salient object detection to expand the receptive field and refine the feature maps output by the backbone encoder.

*This work was supported by the National Nature Science Foundation of China (62373016) and the Open Projects Program of State Key Laboratory of Multimodal Artificial Intelligence Systems (MAIS-2023-22).

Yijie Li, Hewei Wang, and Zhenqi Li are with the Beijing-Dublin International College, Beijing University of Technology, Beijing 100124, China. Shaofan Wang is with the Faculty of Information Technology, Beijing University of Technology, Beijing 100124, China. Soumyabrata Dev is with the School of Computer Science, University College Dublin, Dublin, Ireland. Guoyu Zuo is with the Faculty of Information Technology, Beijing University of Technology, and also with the Beijing Key Laboratory of Computing Intelligence and Intelligent Systems, Beijing 100124, China.

Corresponding author: Guoyu Zuo (zuoguoyu@bjut.edu.cn)

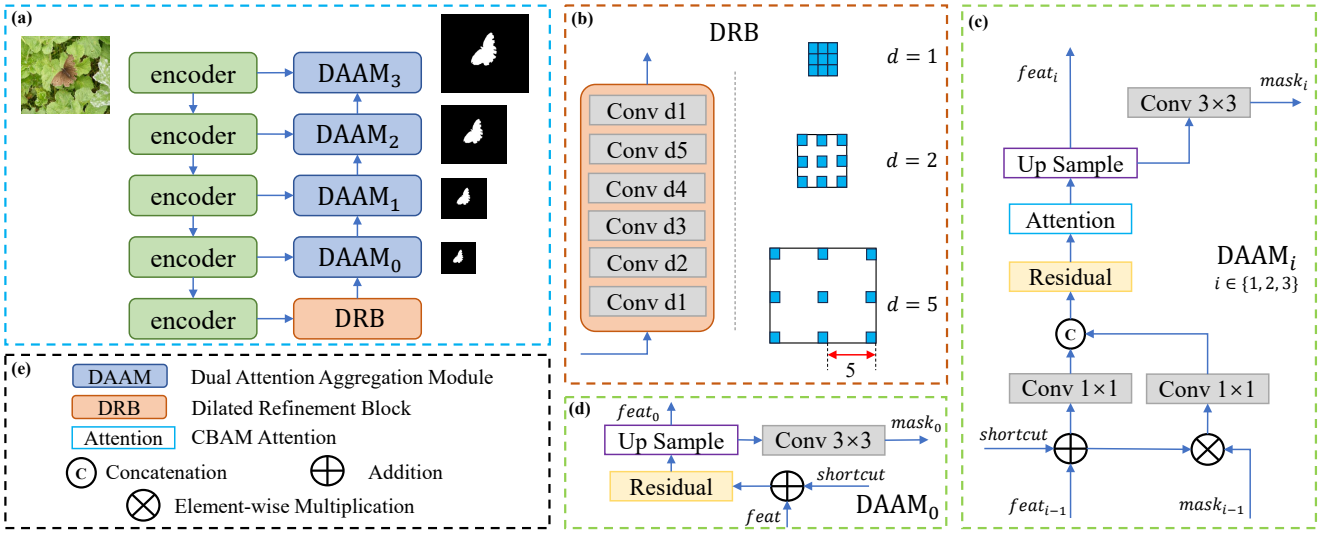

Fig. 1. (a) The overall pipeline of DAANet. (b) The structure of Dilated Refinement Block (DRB). (c) The architecture of Dual Attention Aggregation Module (DAAM) in stages 1, 2, and 3. (d) DAAM in stage 0. (e) Notation description

- We designed and conducted a thorough evaluation and comparison with twelve methods on six benchmark datasets, and DAANet can achieve advanced performance with a light-weighted configuration.

II. Related Works

Before the widely using of deep learning techniques, there already existed a great number of salient object detection methods such as [11]–[14], which usually use gradient variation and features to judge whether a pixel belongs to foreground or background, and some of these works use simple artificially designed features and algorithms to generate the estimation. Sai *et al.* [11] propose a foreground connectivity measurement to enhance the salient map retrieval. Their approach first builds an objectness map by utilizing the objectness proposals and capturing super-pixels containing the salient object and then uses their proposed foreground connectivity measure to assign weight to super-pixels. Finally, they apply a saliency optimization to combine the foreground weight and background to get the saliency map. Wei *et al.* [12] illustrate that most of the background area can easily connect to the boundary of the image. On the contrary, it is difficult for the area belonging to the salient object to link the image boundary. This motivated them to redefine the saliency of an image patch as the length of its shortest path to image boundaries. Yang *et al.* [13] proposed a graph-based approach that considers the image as a graph with super-pixel as nodes. They rank these nodes based on their similarity to the foreground and background queries to extract the background area and salient objects. In [14], Cheng *et al.* used global contrast as the core methodology to retrieve a salient map that defines the pixel saliency as the contrast ratio. They also introduce histogram-based contrast (HC) which uses color statistics of the input image to define saliency.

After the widely using of deep convolutional neural network (CNN) in the image classification area, VGG [6] and ResNet [7], for example, many saliency detection methods including [15]–[18] use more efficient pre-trained backbone networks to extract features. When FCN [19] has been introduced, the encoder-decoder structure has been widely used in salient object detection models. Zhang *et al.* [15] propose a bidirectional model with a contextual feature extraction module that allows the integration of multi-level features and the bi-directional structure to enhance the message passing between features at different levels. They also adopt a gate function to control the passing of information. Hu *et al.* [16] introduce a novel model to recurrently aggregate deep features which can effectively exploit the complementary information extracted by each layer. They compress the combination of outputs from each layer and merge the combination with the feature of each layer to enhance the discrimination of features and salient regions.

III. DAANet

DAANet uses an FPN-based encoder-decoder structure. The encoders support various backbone networks, including ResNet50, VGG16, and MobileNetV2. The decoders are our proposed DAAMs. When an image is fed into the network, the pre-trained backbone encoder will generate four intermediate feature maps and one output, and the output will then go through the DRB module. After that, it will be fed into a series of decoders, and each decoder will increase the resolution of the feature map and reduce the number of channels. As for the training technique, we apply multi-stage supervision to improve the converging speed and training stability. We use a hybrid loss to train our model which is a combination of IOU loss and BCE loss.

A. Encoder Networks

As mentioned in previous sections, DAANet is a FPN-liked encoder-decoder network. This fundamental structure allows us to support and switch different encode networks. In our implementation, we support ResNet50 [7], VGG16 [6], and MobileNetV2 [20] for training and inference. The outputs of each encoding stage will be fed into decoders.

The original ResNet-50 network that was designed for ImageNet has five down-sample stages which reduce the size of the feature map to 1/32 of the input image before fully-connected layers, but our DAANet only needs 4 down-sample stages to produce the four different outputs while the size of the smallest feature map is 16×16. To resolve this issue, we modify several parameters of the two convolution layers in the last encoding stage. The first convolution layer mentioned above refers to the second convolution layer in the first bottle-neck block of the last encoding stage whose new dilation is $(2,2)$, the new padding values are $(1,1)$, and the stride values are changes to $(1,1)$. The next convolution layer refers to the first convolution layer in the down-sample sub-block of the same bottle-neck block whose strides are changed to $(1,1)$. The VGG16 for our DAANet has the same problem and we solve this issue by removing the last max-pooling layer. As for the MobileNetV2, we reduce the stride of a convolution layer to $(1,1)$.

B. Dilated Refinement Block (DRB)

Most previous approaches with FPN structure directly use the feature maps from the backbone network without further processing which may have difficulty in extracting features from small objects, which inspired us to design the DRB, shown in Figure 1 (b). In our approach, we add six additional layers after the last encoder. increasing dilation rates and increasing kernel size can both improve the receptive field and the ability to extract features from small objects. In order to solve this problem and reduce the computation complexity, we use depth-wise convolution which means in this convolution layer, the number of input channels is equal to the group's number so that the parameters amount can be reduced. We also use dilation rates with 1, 2, 3, 4, 5, 1 for the six layers instead of using the large kernel size to further reduce the complexity.

C. Dual Attention Aggregation Module (DAAM)

The salient object detection task requires the model to segment the object with the highest probability to be the foreground object, whose mechanism is much similar to the attention technique which motivates us to integrate channels and spatial attention module (CBAM) [10], the structure as shown in Figure 2, into the dual attention aggregation module (DAAM). The architecture of the DAAM is shown in Figure 1 (c). To enhance the attention, we consider the output of a DAAM as a pseudo-attention which will be multiplied with the addition of a shortcut and input. Those two attention constructs are dual attention.

In each DAAM block, the shortcut, and the output together with the salient map of the previous layer will be accepted as inputs while its output will be sent to the next DAAM block. Firstly, we adopt an element-wise multiplication between the addition result and the saliency map of the previous DAAM block which constructed a rough weighted feature. Secondly, the additional results and the weighted feature will be concatenated after a Conv2d-Bn-ReLU group while this group of layers will reduce the channels to half of its inputs. After that, the features will go through a residual block which refers to the original bottle-neck design, and then it will be fed into CBAM to apply more accurate attention. Then an up-sample operation with a stride equal to 2 is applied to increase the resolution of the feature map. Finally, there will be two branches of outputs and one of which will go through a single convolution layer to build the prediction logits and another one will be fed into the next DAAM. These procedures can be formulated as,

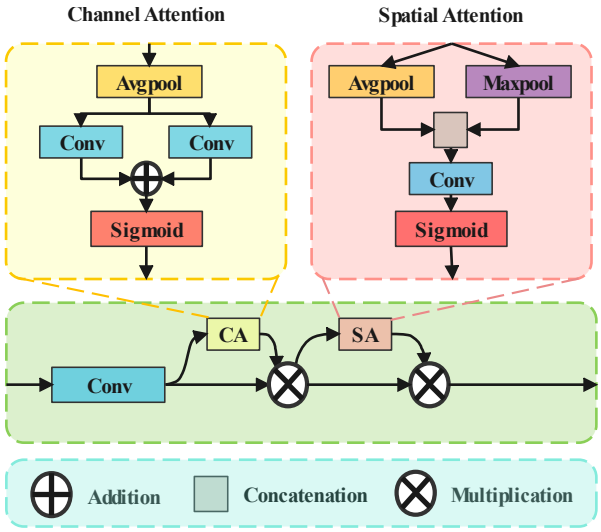

Fig. 2. The structure of modified CBAM [10]

$$a_i = feat_{i-1} + shortcut_i \quad (1)$$

$$z_i = \text{Concat}(\text{Conv}(a_i), \text{Conv}(a_i \odot mask_{i-1})) \quad (2)$$

$$feat_i = \text{UpSample}(\text{Attention}(\text{Residual}(z_i))) \quad (3)$$

$$mask_i = \text{Sigmoid}(\text{Conv}(feat_i)) \quad (4)$$

We also need to remind that the structure shown in Figure 1 (d) indicates the first DAAM decoder which only has a residual block, an up-sample layer, and a convolution layer which results from the low-resolution of features and it is hard to capture the spatial structure of the salient object and it is no need to apply the dual attention. Then the formulas can be simplified as,

$$a_i = feat + shortcut_i \quad (5)$$

$$feat_i = \text{UpSample}(\text{Residual}(a_i)) \quad (6)$$

$$mask_i = \text{Sigmoid}(\text{Conv}(feat_i)) \quad (7)$$

We observe that the original CBAM is designed for the backbone network and our design needs to aggregate it in decoders, the results show that the original CBAM may result in a significant decrease in MAE which motivates us to make some modifications. We modify it by removing the max-pooling layer in the channel attention module, as shown in Figure 2. The channel attention module of CBAM generates attention through a global average pooling followed by two paralleled convolution layers and a sigmoid function. The spatial attention module retrieves the attention map by passing the input feature into the max-pooling branch and average-pooling branch, concatenating them, and following by a convolution layer with sigmoid activation.

As for our lightweight approach with MobileNetV2 [20] backbone, we replace the original residual block with the inverted residual [20], which can significantly decrease the parameter amount by using a series of depth-wise convolution.

D. Loss function

The loss functions we used in training are binary cross entropy (BCE) and the intersection of union (IOU) loss. The BCE loss is widely used in regression and binary classification tasks and the IOU loss can add additional constrain to the supervision of overall structure, while those functions are shown below:

$$L_{\text{bce}} = \frac{1}{NM}\sum_{n=1}^{N}\sum_{i=1}^{M} -t_i \ln(p_i) + (1-t_i)\ln(1-p_i) \quad (8)$$

$$L_{\text{iou}} = 1 - \frac{1}{N}\sum_{n=1}^{N}\frac{\sum_{i=1}^{M}(t_i \times p_i)}{\sum_{i=1}^{M}(t_i + p_i - t_i p_i)} \quad (9)$$

in which, M is the total number of pixels, N equals to the batch size, while t_i and p_i indicate the label and prediction of ith pexel. To guarantee the convergence of loss and reduce the training iterations, we use multistage supervision for training and the total loss function can be given as:

$$L_{\text{total}} = \sum_{i=1}^{4}\beta_i(\alpha_1 L_{\text{bce}} + \alpha_2 L_{\text{iou}}) \quad (10)$$

where α is the weight to balance two different objective functions, while β is used to balance the contribution of outputs with different resolutions, 256×256, 128×128, 64×64, and 32×32. We empirically set β_1, β_2, β_3, and β_4 to 1, 0.8, 0.5, and 0.5. We let both α_1 and α_2 to 1.

E. Implementation Details

We perform the training on a single NVIDIA Tesla V100-SXM2 16GB GPU. We use the standard train-valid data split provided by original datasets. As for training and validation datasets, we train DAANet on DUTS-TR [21] and validate our DAANet with different configurations on six benchmark datasets, including DUTS-TE [21], SOD [22], HKU-IS [23], ECSSD [24], PASCAL-S [25], and DUT-OMRON [13]. We set the training batch size to 16 and trained 30 epochs for the model with ResNet50 and 25 epochs with other backbones. As for the optimizer, we use SGD with an initialized learning rate of $1e-2$, momentum is 0.9, and weight decay is $5e-4$. We also use exponential learning-rate decay with γ equal to 0.85 after each training epoch. We evaluate DAANet on the test set after every 5 epochs.

IV. Experiments

A. Datasets

We train our DAANet on DUTS-TR [21] dataset, and evaluate on DUTS-TE [21], SOD [22], HKU-IS [23], ECSSD [24], PASCAL-S [25], and DUT-OMRON [13]. DUTS-TR has 10,553 images pair in total which is widely used for training. DUTS-TE has 5,019 images for testing. SOD include only 300 images while each of them contains complex semantic information and it is the most difficult benchmark for validation. HKU-IS contains 4,447 image pairs. ECSSD contains 1,000 samples for testing. The PASCAL-S dataset has 850 images. DUT-OMRON has 5,168 images for testing.

B. Evaluation Metrics

We use three measurements to evaluate DAANet: mean absolute error (MAE) [26], mean F-measure, and PR curve. The MAE measures average pixel-wise differences between ground truth and prediction, given a salient ground truth t and a prediction p, the MAE can be defined as:

$$\text{MAE} = \frac{1}{NM}\sum_{n=1}^{N}\sum_{i=1}^{M}|t_i - p_i| \quad (11)$$

The mean F-measure is a weighted harmonic mean of precision and recall which can be represented as:

$$mF_\beta = \frac{(1+\beta^2)\times\text{Precision}\times\text{Recall}}{\beta^2\times\text{Precision}+\text{Recall}} \quad (12)$$

where β^2 is set to 0.3 which follows the widely used configuration. The PR curve is defined by a series of precision and recall pairs, each of which is calculated under a different threshold. We first cast the model output to 0-255 by multiplying the probability output by 255 and then set 256 threshold from 0 to 255, and then plot the PR curve with those 256 points.

C. Ablation Study

we conduct the ablation study, shown in Table I, on DAANet's module compositions, loss function, and backbone configurations to evaluate of our approach. We train the DAANet and baseline model on DUTS-TE under the setting described in section III-E.

Module composition and loss function: To show the influence of our proposed module on overall performance we perform experiments on DAAM, DRB, and IOU loss. We take FPN+VGG16 as our baseline model. Then we extend the baseline model with IOU loss, DAAM module, and DRB. As shown in Table I No. 1 and No. 2, the baseline approach can achieve an MAE of 0.053 and mF_β is 0.749, and ths MAE is

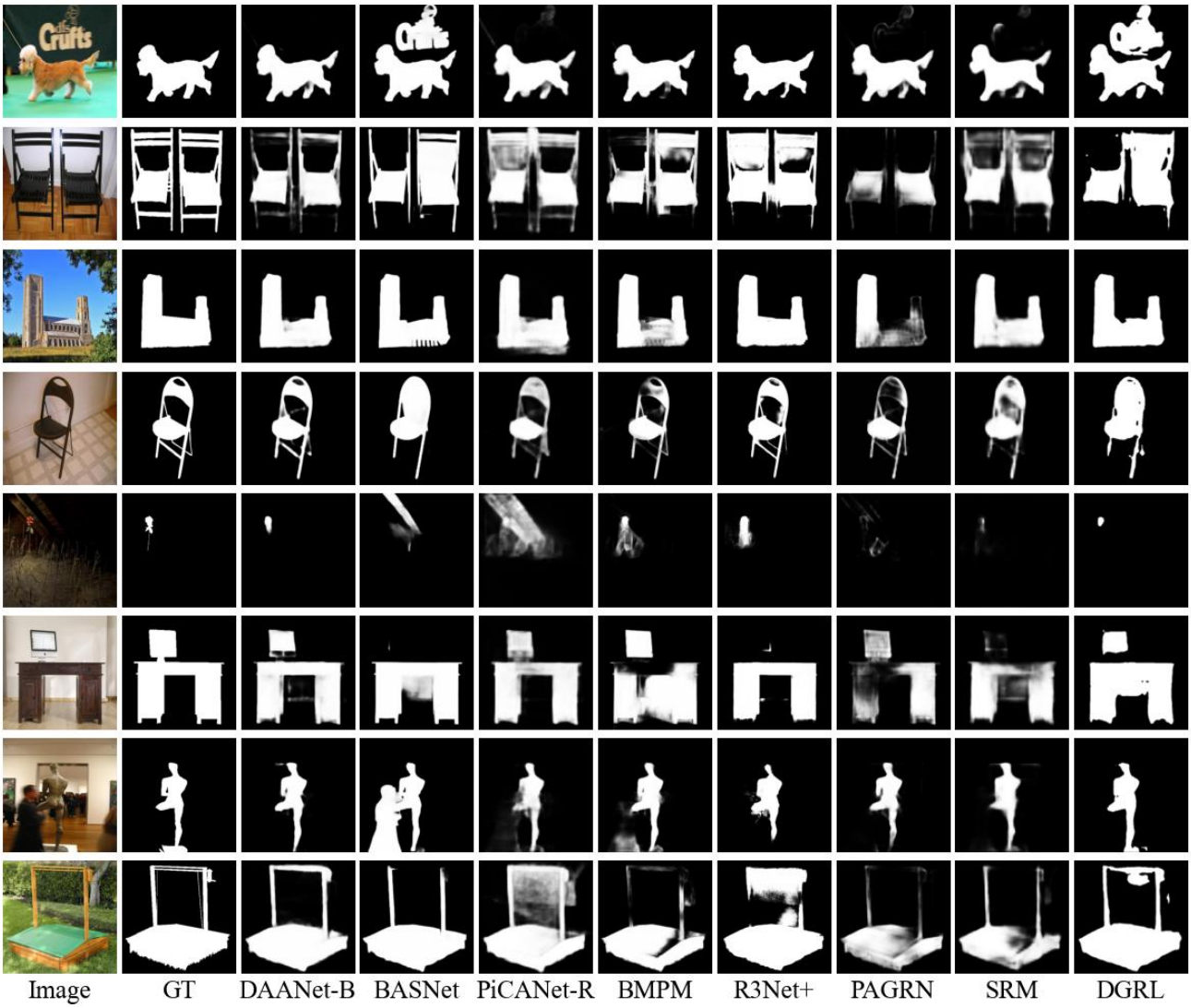

Fig. 3. Qualitative comparison on DUTS-TE between DAANet-B and previous approaches

TABLE I

ABLATION STUDY ON DIFFERENT MODULE COMPOSITIONS, LOSS FUNCTION, AND BACKBONE NETWORKS. IOU REPRESENTS IOU LOSS.

No.	Backbone	FPN configs			DUTS-TE	
		IOU	DAAM	DRB	mF_β	*MAE*
1	VGG16				0.749	0.053
2	VGG16	✓			0.773	0.049
3	VGG16	✓	✓		0.792	0.044
4	VGG16	✓	✓	✓	0.795	0.043
5	ResNet50	✓	✓		0.801	0.042
6	ResNet50	✓	✓	✓	**0.801**	**0.042**
7	MobileNetV2	✓	✓		0.767	0.051
8	MobileNetV2	✓	✓	✓	0.762	0.052

improved to 0.049 after adopting the IOU loss together with BCE loss. In experiment No.3, we extend No. 2 with the proposed DAAM and it can achieve the MAE of 0.044 and mF_β become 0.792. When applying the DRB module, the MAE and mF_β can be further improved to 0.043 and 0.795 which can prove the effectiveness of DAAM and DRB.

Experiments on different backbones: To analyze the performance of our proposed DAAM on different backbone networks, we perform another two experiments on ResNet50 and MobileNetV2. Table I shows that DAAM with ResNet50 can achieve the best performance on both two metrics. The configuration with MobileNetv2 can achieve an MAE of 0.051 with only 15.8 MBytes of parameters. When applying the ResNet50 backbone, the MAE can be improved to 0.042, and mF_β can reach 0.801. However, adding DRB to the DAANet with MobileNetV2 backbone leads to the decrease of overall performance which may results from the limited feature extraction ability of backbone network.

D. Qualitative Evaluation

In order to intuitively evaluate the performance of DAANet, we conduct the qualitative evaluation that comparing DAANet (ResNet50 backbone) with seven previous approaches, including BASNet [36], PiCANet [33], BMPM [15], R3Net+ [37], PAGRN [30], SRM [35] and DGRL

TABLE II

QUANTITATIVE COMPARISON WITH DAANETS (**DAANET-A**, **DAANET-B**, **DAANET-C**) AND 12 OTHER METHODS ON 6 BENCHMARK DATASETS. THE METRICS INCLUDING MAXIMUM F-MEASURE $maxF_\beta$, MEAN F-MEASURE mF_β, AND MAE. WE USE DIFFERENT COLORS TO LABEL THE TOP3 METHOD ON EACH BENCHMARK, RED, GREEN, BLUE INDICATE THE BEST, THE SECOND BEST, AND THE THIRD BEST MODEL RESULTS RESPECTIVELY.

Methods	**Size(MB)**	**Training Data**		**DUTS-TE** [21]			**HKU-IS** [23]			**SOD** [22]			**DUT-OMRON** [13]			**PASCAL-S** [25]			**ECSSD** [24]		
		Datasets	#Images	$maxF_\beta\uparrow$	$mF_\beta\uparrow$	$MAE\downarrow$	$maxF_\beta\uparrow$	$mF_\beta\uparrow$	$MAE\downarrow$	$maxF_\beta\uparrow$	$mF_\beta\uparrow$	$MAE\downarrow$	$maxF_\beta\uparrow$	$mF_\beta\uparrow$	$MAE\downarrow$	$maxF_\beta\uparrow$	$mF_\beta\uparrow$	$MAE\downarrow$	$maxF_\beta\uparrow$	$mF_\beta\uparrow$	$MAE\downarrow$
VGG backbone																					
UCF [27]	**117.9**	MSRA10K	10,000	0.771	0.629	0.117	0.886	0.808	0.074	0.803	0.699	0.164	0.734	0.613	0.132	0.828	0.706	0.126	0.911	0.840	0.078
Amulet [28]	132.6	MSRA10K	10,000	0.778	0.676	0.085	0.895	0.839	0.052	0.806	0.755	0.141	0.742	0.647	0.098	0.837	0.768	0.098	0.915	0.870	0.059
DSS [29]	447.3	MSRA-B	2,500	0.826	0.791	0.057	0.910	**0.895**	0.041	0.841	0.793	0.121	0.772	0.729	0.066	0.831	-	0.093	0.916	**0.901**	0.053
PAGRN [30]	-	DUTS-TR	10,553	0.855	0.788	0.056	0.918	886	0.048	-	-	-	0.771	0.711	0.071	0.856	**0.807**	0.093	0.927	**0.894**	0.061
BMPM [15]	-	DUTS-TR	10,553	0.851	0.751	0.049	0.921	0.871	0.039	**0.855**	0.763	0.107	0.774	0.692	0.064	0.862	0.769	0.074	0.929	0.869	0.045
AFNet [31]	143.9	DUTS-TR	10,553	**0.862**	**0.797**	**0.046**	**0.923**	0.888	**0.036**	**0.856**	**0.809**	0.109	0.797	**0.738**	0.057	**0.868**	**0.826**	0.071	0.935	**0.908**	0.042
RAS [32]	**81.0**	MSRA-B	2,500	0.831	0.755	0.060	0.913	0.871	0.045	0.850	0.799	0.124	0.786	0.713	0.062	0.837	0.785	0.104	0.921	0.889	0.056
PiCANet [33]	153.3	DUTS-TR	10,553	0.851	0.755	0.054	0.921	0.870	0.042	0.853	0.791	0.102	0.794	0.710	0.068	**0.868**	0.801	0.077	0.931	0.884	0.047
DAANet-A	89.8	DUTS-TR	10,553	0.867	0.795	**0.043**	0.925	0.890	0.035	0.853	0.751	0.108	0.796	0.737	**0.057**	0.862	0.776	**0.073**	**0.935**	0.882	0.043
ResNet backbone																					
DGRL [34]	648.0	DUTS-TR	10,553	0.829	0.798	0.050	0.921	0.890	0.036	0.845	**0.799**	**0.104**	0.774	0.733	0.062	0.854	**0.825**	**0.072**	0.922	0.906	**0.041**
SRM [35]	213.1	DUTS-TR	10,553	0.827	0.757	0.059	0.906	0.874	0.046	0.843	**0.800**	0.127	0.769	0.707	0.069	0.847	0.801	0.085	0.917	0.892	0.054
PiCANet-R [33]	197.2	DUTS-TR	10,553	**0.860**	0.764	0.051	0.919	0.870	0.043	0.853	0.785	**0.103**	**0.803**	0.717	0.065	0.857	0.792	0.076	0.935	0.886	0.046
BASNet [36]	348.5	DUTS-TR	10,553	0.859	**0.796**	0.048	**0.928**	**0.896**	**0.032**	0.851	0.745	0.113	**0.805**	**0.755**	**0.057**	**0.862**	0.779	0.077	**0.942**	0.879	**0.037**
DAANet-B	229.0	DUTS-TR	10,553	**0.870**	**0.801**	**0.042**	**0.928**	**0.892**	**0.034**	**0.855**	0.757	**0.105**	**0.802**	**0.748**	**0.054**	0.860	0.781	**0.067**	**0.940**	0.886	**0.040**
MobileNet backbone																					
DAANet-C	**15.8**	DUTS-TR	10,553	0.836	0.767	0.051	0.908	0.876	0.043	0.828	0.745	0.123	0.785	0.722	0.061	0.840	0.772	0.082	0.922	0.875	0.051

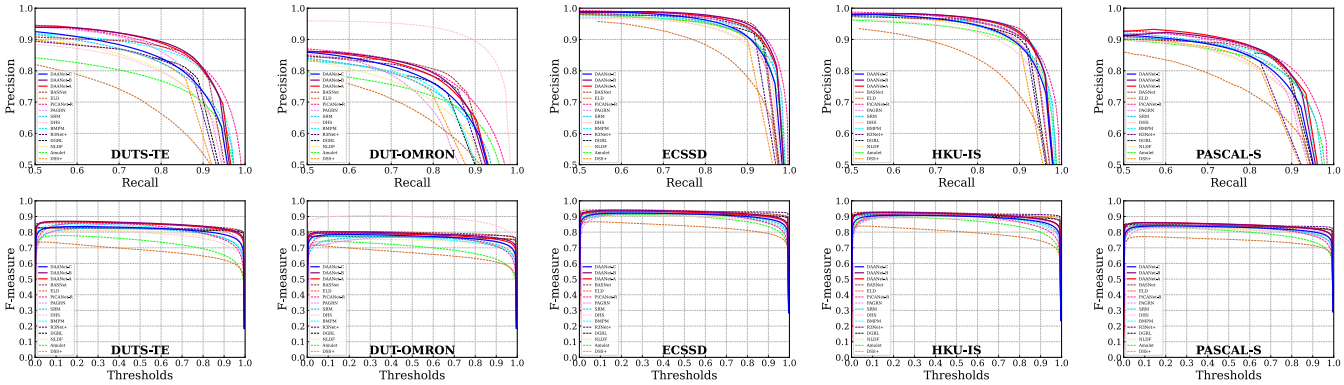

Fig. 4. Illustration of PR curves (the first column) and F-measure curves (the second column) on five benchmark datasets

[34]. As shown in Figure 3, the results show that DAANet significantly improves the quality and accuracy of salient prediction, while DAANet can capture more details and have fewer wrong predictions than others. In the 2nd row, 4th row, and 6th row, many previous approaches generated salient masks have many false-positive cases, such as the area under the desk, the wall is predicted as a part of the chair, but DAANet-B can capture the salient object accurately without these false-positive pixels. In the 5th row, DAANet can accurately detect the small object. These results illustrate that DAANet can accurately detect the salient object from different scenes and it also has the ability to detect small objects.

E. Quantitative Evaluation

To further analyze the performance DAANet, we perform the quantitative evaluation on three configurations of DAANet with 12 state-of-the-art approaches: UCF [27], Amulet [28], DSS [29], PAGRN [30], BMPM [15], AFNet [31], RAS [32], PiCANet [33], DGRL [34], SRM [35], PiCANet-R [33], and BASNet [36]. The three configurations of DAANet are represented as follows:

- **DAANet-A**: VGG16+DAAM+DRB
- **DAANet-B**: ResNet50+DAAM
- **DAANet-C**: MobileNetV2+DAAM

The evaluation metrics and datasets have been mentioned in Section IV-B and IV-A. As shown in Table II, **DAANet-B** achieve the best performance on DUTS-TE dataset with $maxF_\beta$ by 0.870, mF_β by 0.801. and MAE by 0.042, which also be in the lead of other models on the other five benchmark datasets. We also build lightweight approaches with MobileNetV2 backbone, **DAANet-C**, which only has 15.8 MB of parameters in total, but it can achieve the MAE by 0.051, which proves the effectiveness of our proposed DAAM module. Besides, we evaluate DAANet quantitatively by using the PR curves and F-measure curves, as shown in Figure 4, the area under the curve is larger, and the performance of the model is better. We can see from those figures that our proposed ResNet50+DAAM (**DAANet-B**) has the best performance on most datasets.

V. CONCLUSIONS AND FUTURE WORK

In this paper, we propose a pipeline with a dual attention aggregation module (DAAM) and dilated refinement block (DRB) for salient object detection. Our proposed DAANet is an encoder-decoder structure that uses an ImageNet pre-trained backbone as an encoder and adopts the FPN architecture with DAAM modules. The DAAM module can help the basic FPN structure capture more features and details to produce salient masks with higher accuracy. The DRB takes

advantage of dilated depth-wise convolution that achieves the goal of expanding the receptive field and further improving the performance of DAANet. We evaluate DAANet qualitatively and quantitatively, the results validate the effectiveness of DAANet and its core components.

For future work, we plan to focus on further enhancing the capabilities of our proposed DAANet and exploring additional avenues for improvement in salient object detection. In addition, we will also focus on domain-specific adaptations of the model and efficient deployment of DAANet on resource-constrained platforms, such as mobile devices and embedded systems.

REFERENCES

[1] H. Wang, M. S. Pathan, S. Dev, Stereo matching based on visual sensitive information, in: 2021 6th International Conference on Image, Vision and Computing (ICIVC), IEEE, 2021, pp. 312–316.

[2] H. Wang, Y. Li, S. Xi, S. Wang, M. S. Pathan, S. Dev, AMDCNet: An attentional multi-directional convolutional network for stereo matching, Displays 74 (2022) 102243.

[3] P. Dey, B. P. Das, Y. H. Lee, S. Dev, NeSNet: A deep network for estimating near-surface pollutant concentrations, IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing (2023).

[4] J. Walsh, A. Othmani, M. Jain, S. Dev, Using U-Net network for efficient brain tumor segmentation in MRI images, Healthcare Analytics 2 (2022) 100098.

[5] S. Dev, A. Nautiyal, Y. H. Lee, S. Winkler, CloudSegNet: A deep network for nychthemeron cloud image segmentation, IEEE Geoscience and Remote Sensing Letters 16 (12) (2019) 1814–1818.

[6] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, International Conference on Learning Representations (ICLR) (2014).

[7] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 770–778.

[8] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, S. Belongie, Feature pyramid networks for object detection, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 2117–2125.

[9] O. Ronneberger, P. Fischer, T. Brox, U-net: Convolutional networks for biomedical image segmentation, in: International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI), 2015, pp. 234–241.

[10] S. Woo, J. Park, J.-Y. Lee, I. S. Kweon, Cbam: Convolutional block attention module, in: Proceedings of the European Conference on Computer Vision (ECCV), 2018, pp. 3–19.

[11] R. S. Srivatsa, R. V. Babu, Salient object detection via objectness measure, in: IEEE International Conference on Image Processing (ICIP), 2015, pp. 4481–4485.

[12] Y. Wei, F. Wen, W. Zhu, J. Sun, Geodesic saliency using background priors, in: Proceedings of the European Conference on Computer Vision (ECCV), 2012, pp. 29–42.

[13] 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 (CVPR), 2013, pp. 3166–3173.

[14] M.-M. Cheng, N. J. Mitra, X. Huang, P. H. Torr, S.-M. Hu, Global contrast based salient region detection, IEEE Transactions on Pattern Analysis and Machine Intelligence (IEEE T-PAMI) 37 (3) (2014) 569–582.

[15] 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 (CVPR), 2018, pp. 1741–1750.

[16] X. Hu, L. Zhu, J. Qin, C.-W. Fu, P.-A. Heng, Recurrently aggregating deep features for salient object detection, in: Proceedings of the AAAI Conference on Artificial Intelligence (AAAI), Vol. 32, 2018.

[17] 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.

[18] P. Zhang, W. Liu, H. Lu, C. Shen, Salient object detection by lossless feature reflection, IEEE Transactions on Image Processing (IEEE TIP) (2018).

[19] J. Long, E. Shelhamer, T. Darrell, Fully convolutional networks for semantic segmentation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 3431–3440.

[20] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, L.-C. Chen, Mobilenetv2: Inverted residuals and linear bottlenecks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 4510–4520.

[21] 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 (CVPR), 2017, pp. 136–145.

[22] V. Movahedi, J. H. Elder, Design and perceptual validation of performance measures for salient object segmentation, in: IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops (CVPR Workshops), 2010, pp. 49–56.

[23] G. Li, Y. Yu, Visual saliency based on multiscale deep features, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 5455–5463.

[24] Q. Yan, L. Xu, J. Shi, J. Jia, Hierarchical saliency detection, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 1155–1162.

[25] 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 (CVPR), 2014, pp. 280–287.

[26] F. Perazzi, P. Krähenbühl, Y. Pritch, A. Hornung, Saliency filters: Contrast based filtering for salient region detection, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2012, pp. 733–740.

[27] 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 (ICCV), 2017, pp. 212–221.

[28] 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 (ICCV), 2017, pp. 202–211.

[29] Q. Hou, M.-M. Cheng, X. Hu, A. Borji, Z. Tu, P. Torr, Deeply supervised salient object detection with short connections, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 5300–5309.

[30] 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 (CVPR), 2018, pp. 714–722.

[31] 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 (CVPR), 2019.

[32] 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.

[33] 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 (CVPR), 2018, pp. 3089–3098.

[34] 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 (CVPR), 2018, pp. 3127–3135.

[35] T. Wang, A. Borji, L. Zhang, P. Zhang, H. Lu, A stagewise refinement model for detecting salient objects in images, in: Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 4019–4028.

[36] X. Qin, Z. Zhang, C. Huang, C. Gao, M. Dehghan, M. Jagersand, Basnet: Boundary-aware salient object detection, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 7479–7489.

[37] Z. Deng, X. Hu, L. Zhu, X. Xu, J. Qin, G. Han, P.-A. Heng, R^3Net: Recurrent residual refinement network for saliency detection, in: International Joint Conference on Artificial Intelligence (IJCAI), 2018.