

# Boundary-guided Feature Aggregation Network for Salient Object Detection

Yunzhi Zhuge, Pingping Zhang, Huchuan Lu

**Abstract**—Fully convolutional networks (FCN) has significantly improved the performance of many pixel-labeling tasks, such as semantic segmentation and depth estimation. However, it still remains non-trivial to thoroughly utilize the multi-level convolutional feature maps and boundary information for salient object detection. In this paper, we propose a novel FCN framework to integrate multi-level convolutional features recurrently with the guidance of object boundary information. First, a deep convolutional network is used to extract multi-level feature maps and separately aggregate them into multiple resolutions, which can be used to generate coarse saliency maps. Meanwhile, another boundary information extraction branch is proposed to generate boundary features. Finally, an attention-based feature fusion module is designed to fuse boundary information into salient regions to achieve accurate boundary inference and semantic enhancement. The final saliency maps are the combination of the predicted boundary maps and integrated saliency maps, which are more closer to the ground truths. Experiments and analysis on four large-scale benchmarks verify that our framework achieves new state-of-the-art results.

**Index Terms**—Salient object detection, Boundary information extraction, Attention, feature fusion.

## I. INTRODUCTION

Saliency object detection is a fundamental computer vision task which aims to identify the most eye-catching objects and areas in an image [1] [2] [3] [4]. In the past two decades, great success has been made in this pixel-labeling task. However, due to several inevitable factors such as cluttered backgrounds or blurred boundaries, it still remains a difficult task to combine all hand-tuned cues in an appropriate way.

Recently, deep convolutional neural networks (CNNs) have greatly improved the performances of many computer vision tasks, such as image classification [5], semantic segmentation [6] and visual tracking [7], [8]. With the advantages of fully convolutional networks (FCNs) [6], several FCN-based attempts have been performed and delivered state-of-the-art performance in predicting saliency maps [9]–[11]. Nonetheless, existing models mainly focus on utilizing high-level features extracted from last convolutional layers. As a result, they are lack of low-level visual information such as object boundary. Thus, these models tend to predict imperfect results with poorly localized object boundaries.

In this paper, we propose a novel saliency detection method based on multi-level features and boundary cues. To make full use of the multi-level convolutional features and boundary information, we present a boundary-guided feature aggregating architecture, which simultaneously generates and merges multi-level saliency maps and boundary prediction maps to obtain accurate saliency maps.

Our framework has two streams for saliency prediction. In the main stream, we predict saliency maps with the incorporated features maps at different resolutions, and these predicted saliency maps are recursively sent to the refinement stage as the inputs. In another stream, boundary features are extracted through a boundary extraction structure. To utilize the boundary information, we propose an attention-based feature fusion module to integrate two-stream features.

Our main contributions are as follows:

- We propose a boundary-guided feature aggregation network, termed as BFANet, to utilize multi-level convolutional features and boundary cues for salient object detection. The BFANet first extracts multi-level features, then integrates them into multiple resolutions. The boundary information fusion is performed to enhance these features to generate finer saliency maps.
- We propose a boundary extraction network as a branch of the BFANet, which generates boundary feature maps under the supervision of boundary ground truth. Besides, we also introduce an attention-based feature fusion module to produce saliency maps with robust boundary.
- Extensive experiments on four large-scale benchmarks have shown that our approach performs favorably against other state-of-the-art methods.

## II. OUR PROPOSED METHOD

Our method is mainly motivated by the following facts. First, previous methods usually focus on precisely localizing salient objects, which more or less neglect the sharpness of boundary areas. To enhance the boundary, we propose a two-stream structure to generate saliency maps with accurate object boundaries. Secondly, features of variant scales contribute differently to detection. However, there still exist questions in how to effectively utilize these features. Therefore, we make a feasible attempt and propose an effective attention-based structure to perform the multi-level feature fusion.

As shown in Fig. 1, our proposed framework is composed of three components: Aggregating Feature Extraction Network (AFEN), Boundary Prediction Network (BPN) and Attention-based Feature Fusion Module (AFFM). In the following subsections, we will elaborate these components in detail.

### A. Aggregating Feature Extraction Network

The residual networks (ResNet) [14] have shown excellent performances in many computer vision tasks. Our feature extraction network is based on the ResNet-101, which extracts multi-level feature maps from raw RGB images for feature

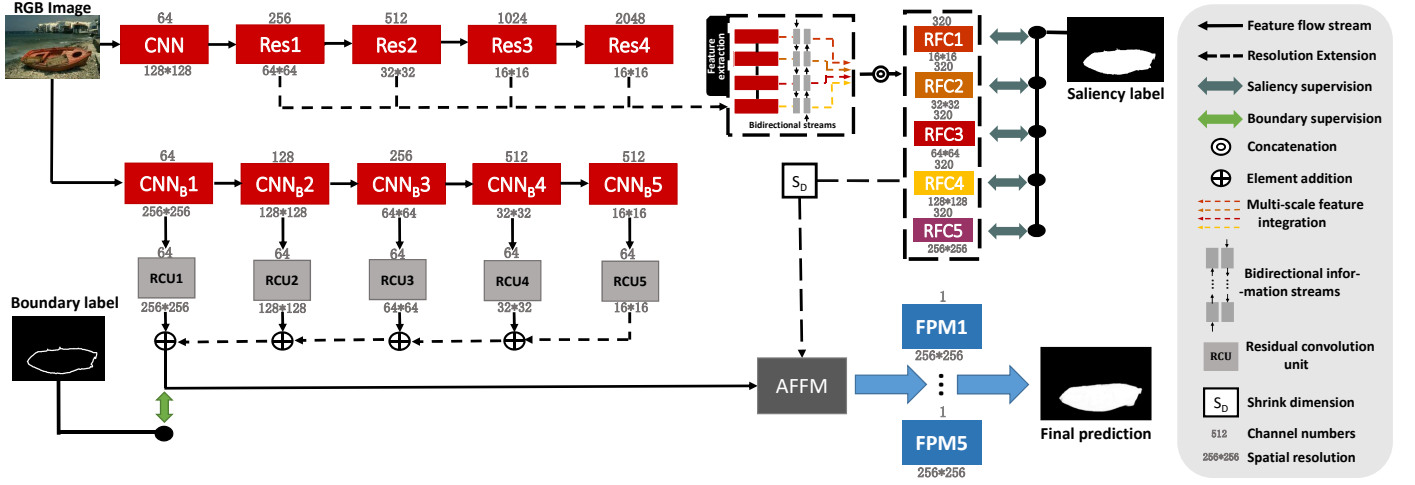


Fig. 1. The overall architecture of our proposed BFANet. The upper stream represents the reproduced Amulet [13], which is used to extract multi-level convolutional features. The details of Amulet are described in Section II-A. In another stream, Boundary Prediction Network (BPN) can progressively enlarge the resolution and refine the details of boundary predictions through a cascaded of residual convolution unit (RCU). Attention-based feature fusion module (AFFM) is employed to integrate multi-level features and boundary cues. After that, we obtain multi-level fused feature maps by parallel fused prediction modules (FPM). The final saliency map is produced with the fused feature maps.

integration and saliency prediction. We modify the ResNet-101 and reduce the resolution of features by a factor of 16 comparing to the input image, as shown in the Fig. 1. The feature maps are extracted from specified convolutional layers. For the balance of resolution, we adopt the resolution-based feature combination structure (RFC) [13] to integrate multi-level convolutional features. The RFC structure unifies convolutional features through shrink and extend operations. More specifically, given an input image  $I$ , the integrated feature maps of scale  $\tau \in [1, \dots, 5]$  are computed by

$$\mathbf{F}^\tau = C_{m=1}^4 (R_m^\tau(\mathbf{F}_m(I); \psi_m)), \quad (1)$$

where  $R_m^\tau(\cdot; \psi_m)$  represents the reshape operator that expands or shrinks the feature maps by a factor of  $\psi_m$ .  $\mathbf{F}_m$  denotes the  $m$ -level feature maps.  $C$  is the concatenation operation in channel-wise. The resolution of generated feature maps  $\mathbf{F}^\tau$  is  $\lfloor \frac{W}{2^\tau}, \frac{H}{2^\tau} \rfloor$ . Besides, to enhance the feature interaction, we adopt the bidirectional information streams [13], which integrate multi-level features in both bottom-up and top-down directions. Although multi-level features have been extracted and fused in this effective way, there still exists a large gap between predicted saliency maps and ground truth. Due to the defects of down-sampling operations, the predicted saliency maps are blurred or occluded on boundary areas.

### B. Boundary Prediction Network

To resolve the blurry boundary problem, we introduce the BPN, which generates boundary predictions to guide saliency prediction. To generate the boundary labels, we apply the open-source ‘‘Canny’’ algorithm [15] to the binary saliency labels, which usually provide in public saliency benchmarks. An example is shown in Fig. 2. The detailed structure of our BPN is shown at the bottom of Fig. 1. We adopt five convolutional blocks of the VGG-16 model [16] to extract multi-scale boundary features. Given an input image, our BPN first extracts five scale feature maps  $\mathbf{B}_f^\tau$ . To progressively merge multi-scale features and enlarge the boundary prediction map, we cascade several Residual Convolution Units (RCU) [17] on

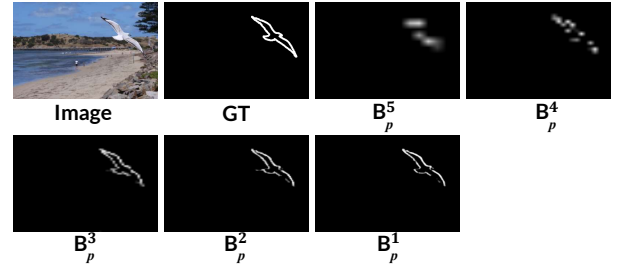


Fig. 2. Illustration of boundary ground truth and boundary prediction maps.  $\mathbf{B}_p^\tau$  ( $\tau = 1, 2, \dots, 5$ ) represents the prediction results of level  $\tau$ . Note that we rescale the predictions to the same size for better visualization.

the side-output feature maps. In the scale  $\tau$ , BPN generates boundary feature maps  $\mathbf{B}^\tau$  by

$$\mathbf{B}^\tau = \begin{cases} ((\mathbf{W}^\tau \star_s \mathbf{B}^{\tau+1}) \oplus \text{RCU}(\mathbf{B}_f^\tau)), & 1 \leq \tau < 5 \\ \text{RCU}(\mathbf{B}_f^\tau), & \tau = 5 \end{cases} \quad (2)$$

where  $\mathbf{B}^\tau$  and  $\mathbf{B}_f^\tau$  represent boundary feature maps and corresponding convolutional feature maps respectively.  $\star_s$  denotes the deconvolution with a stride  $s$  to ensure the same resolution.  $\oplus$  is the element-wise addition. Then we apply a  $1 \times 1$  convolution operation on the boundary feature maps of each scale to generate the boundary prediction map  $\mathbf{B}_p^\tau$ . The comparison of boundary prediction maps is shown in Fig. 2.

### C. Attention-based Feature Fusion Module

To efficiently utilize the boundary information and refine saliency maps, we introduce the attention-based feature fusion module (AFFM), which exploits multiple attention cues [18]. More specifically, we first reduce the channels of saliency features to the same size of boundary features by shrink dimension. Then we perform a global average pooling on the saliency and boundary features to obtain a feature vector  $\mathbf{v} = [v_1, v_2, \dots, v_n]$  ( $n$  is the channel dimension of layers). A spatial softmax operator is applied on the feature vector to generate a normalized weight vector of the feature maps:

$$w_i = \frac{e^{v_i}}{\sum_{i=1}^n e^{v_i}}, \quad (3)$$

TABLE I  
QUANTITATIVE COMPARISON ON FOUR LARGE-SCALE DATASETS. THE BEST TWO RESULTS ARE SHOWN IN RED AND BLUE, RESPECTIVELY.

Dataset	Metric	Ours	Amulet	UCF	DHS	NLDF	RFCN	DS	DCL	ELD	LEGS	MDF	DRFI	BSCA
ECSSD	$F_\beta \uparrow$	<b>0.882</b>	0.867	0.839	0.872	<b>0.878</b>	0.834	0.825	0.829	0.810	0.785	0.807	0.733	0.705
	MAE $\downarrow$	<b>0.051</b>	<b>0.059</b>	0.078	0.059	0.063	0.107	0.122	0.088	0.079	0.118	0.105	0.164	0.182
DUT-OMRON	$F_\beta \uparrow$	<b>0.721</b>	0.669	0.613	/	<b>0.683</b>	0.626	0.603	0.684	0.611	0.591	0.644	0.550	0.509
	MAE $\downarrow$	<b>0.060</b>	0.090	0.132	/	<b>0.079</b>	0.111	0.120	0.097	0.092	0.133	0.092	0.139	0.190
DUTS-TE	$F_\beta \uparrow$	<b>0.763</b>	0.709	0.629	0.724	<b>0.743</b>	0.712	0.632	0.714	0.628	0.585	0.673	0.541	0.499
	MAE $\downarrow$	<b>0.050</b>	0.080	0.117	0.067	<b>0.065</b>	0.091	0.090	0.088	0.093	0.138	0.094	0.175	0.197
HKU-IS	$F_\beta \uparrow$	<b>0.887</b>	0.861	0.808	0.855	<b>0.873</b>	0.835	0.785	0.853	0.769	0.723	0.801	0.722	0.654
	MAE $\downarrow$	<b>0.043</b>	0.053	0.074	0.053	<b>0.048</b>	0.089	0.078	0.072	0.074	0.119	0.089	0.144	0.175

where  $w_i$  is the weight of channel  $i$  and  $\sum_{i=1}^n w_i = 1$ . Subsequently, the fused feature maps  $\mathbf{F}_{\text{fused}}^\tau$  is generated by

$$\mathbf{F}_{\text{fused}}^\tau = (\mathbf{w}_F^\tau \otimes \mathbf{F}^\tau) \oplus (\mathbf{w}_B \otimes \mathbf{B}), \quad (4)$$

where  $\otimes$  is channel-wise product.  $\mathbf{B}$  denotes the boundary feature maps with the resolution of  $256 \times 256$ , which are in accordance with the aggregated saliency feature maps.  $\mathbf{w}_F^\tau$  and  $\mathbf{w}_B$  represents the attention weights for saliency features and boundary features respectively. With the fused feature, five paralleled fused prediction modules (FPM) (each of which is composed of a  $3 \times 3$  convolutional layer and an upsampling layer) are used to predict stage-wise prediction maps. With the stage-wise prediction maps, we add another convolutional layer with a  $1 \times 1$  kernel to predict the final prediction map.

### III. EXPERIMENTS

#### A. Training and Testing Datasets

In the training process, the **DUTS-TR** [12] dataset is chosen as our training dataset, which includes 10, 553 images with accurate pixel-wise annotations. We implement our proposed model based on the Caffe toolbox [19]. We train and test our method with an NVIDIA 1080 GPU (with 8G memory). The input image is uniformly resized into  $256 \times 256 \times 3$  pixels and subtracted the ImageNet mean [20]. We find our model with this resolution achieves both effectiveness and efficiency. We adopt the sigmoid cross entropy as the loss function for both saliency and boundary prediction. Following previous works [13], [21], [22], we train the model until its training loss converges. The weights of FCN backbones are initialized from the VGG-16 [16] and ResNet-101 [14] models. For other layers, we initialize the weights by the “msra” method. We follow the parameters in [13] and use the standard SGD method with a batch size 8, momentum 0.9 and weight decay 0.0005. We set the learning rate to  $1 \times 10^{-8}$  and decrease it by 10% after every 10 epoch. Our model has a size of 410 MB and runs around 10 *fps* for saliency inference, which is comparable even faster than most of methods.

We evaluate the performance of our method on four large-scale datasets described as follows. **ECSSD** [23] is composed of 1000 images with random objects of different scales. **HKU-IS** [24] includes 4447 images with fine pixel-wise annotations. Images of this dataset are well chosen to include multiple disconnected salient objects or objects touching the image boundary. **DUTS-TE** [12] has 5019 images with accurate pixel-wise annotations. All images are picked from the ImageNet DET test set and the SUN dataset [25]. **DUT-OMRON** [23] has a total of 5168 high-quality images.

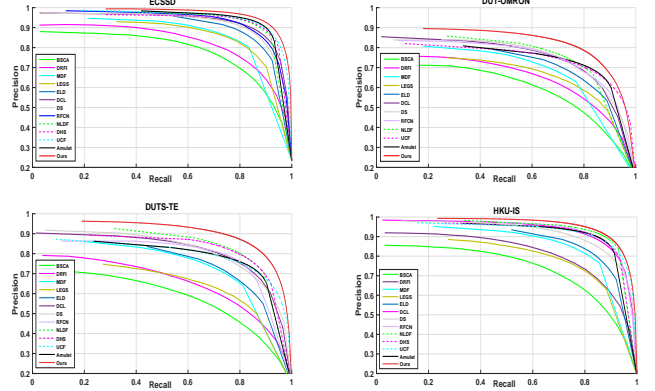


Fig. 3. The PR curves of different state-of-the-art methods.

#### B. Evaluation Metrics.

To evaluate the performance, we adopt three main metrics, *i.e.*, the PR curves, mean F-measure score and mean absolute error (MAE) [26]. Precision-recall (PR) curves can be computed by binarizing the saliency map with a threshold in  $[0, 255]$  and then comparing the binary maps with the ground truth. In many occasions, both precision and recall are important to measure methods. Therefore, F-measure, which is averaged with precision and recall, is proposed to achieve the overall performance evaluation,

$$F_\beta = \frac{(1 + \beta^2) \times \text{Precision} \times \text{Recall}}{\beta^2 \times \text{Precision} + \text{Recall}}. \quad (5)$$

We set  $\beta^2$  to 0.3 to weigh precision more than recall as suggested in [2], [4], [27].

The above evaluations usually assign high saliency scores to salient pixels, which can be unfair especially for the methods which successfully detect non-salient regions, but miss the detection of salient regions. Therefore, we also adopt the MAE metric [26] to measure the average difference between the saliency prediction and the ground truth.

$$MAE = \frac{1}{W \times H} \sum_{x=1}^W \sum_{y=1}^H |S(x, y) - G(x, y)|, \quad (6)$$

where  $S$  is the predicted saliency map and  $G$  is the binary ground truth mask. It indicates how similar a saliency map is compared to the ground truth.

#### C. Comparison with Other Methods

We compared our method with other 12 algorithms, including 10 deep learning based algorithms (Amulet [13], UCF [21],



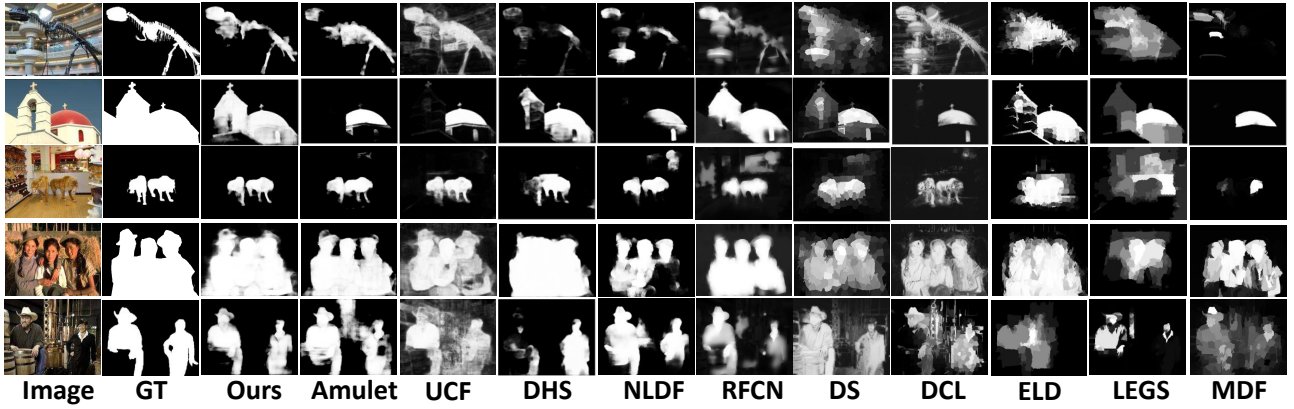


Fig. 4. Comparisons of saliency maps with state-of-the-art methods. Due to the limitation of space, we do not show results of DRFI and BSCA methods.

NLDF [28], DHS [11], DS [29], DCL [30], ELD [10], RFCN [9], LEGS [4], MDF [24]) and 2 conventional algorithms (DRFI [31], BSCA [32]). We compute saliency maps with the original implementations or use them provided by the authors.

**Quantitative Results** As shown in Tab. I and Fig. 3, our model consistently outperforms other methods across all the datasets in terms of all evaluation metrics, which convincingly demonstrates the effectiveness of the proposed method.

**Qualitative Results.** Fig. 4 shows visual comparison between our method and other algorithms. As shown in the 1st row, the foreground is very complex, while our method successfully captures the main components against the competing algorithms. The object in 2nd row is an unusual roof. Many algorithms fail to capture the semantic structure, while our method successfully highlights it. Our proposed method also perform better on images with similar color distribution between foreground and background (3rd row). For disconnected objects (last two rows), our method still performs well, while other algorithms misjudge the interferences.

#### D. Ablation Studies

**Effects of the boundary branch.** We verify the effectiveness of boundary branch in our framework. The compared models include: (1) With the same ResNet-101, we construct a baseline network, which is composed of encoder-decoder and feature combination structure [13]. The multi-level features are used to generate stage-wise prediction maps. The prediction map is generated by merging stage-wise prediction maps. (2) We add boundary-stream to the baseline network and use concatenation operations to combine features of both branches for saliency detection. The prediction scheme is similar with baseline network. Different from (1), features of each level incorporate both saliency and boundary information. We name this setting as *Boundary*<sup>+</sup>. (3) To verify the effectiveness of cascaded RCUs, we implement the boundary-stream approach without RCUs, named *Boundary*<sup>-</sup>. (4) Finally, AFFM is added to fuse saliency and boundary features to generate attentive features, resulting in our final model *AFFM*<sup>+</sup>.

We perform detailed experiments on two datasets, *i.e.*, ECSSD and DUT-OMRON. The results are shown in Tab. II. From quantitative evaluations and qualitative comparisons, one can observe that our proposed components effectively enhance

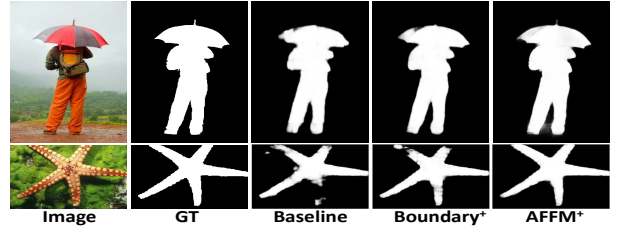


Fig. 5. Qualitative comparisons with different model settings.

saliency detection performance. Especially, the boundary prediction module improves the MAE with 4%. Qualitative comparisons with different model settings are shown in Fig. 5. Compared with the baseline, our boundary-guided prediction is much better in predicting the boundary details. The proposed method indeed produces saliency map with sharp boundaries. The low MAE metric also verifies this fact.

TABLE II  
QUANTITATIVE COMPARISON OF DIFFERENT SETTINGS. THE BEST RESULTS ARE IN BOLD. THE VGG-16 [16] VERSION IS *AFFM*<sup>+</sup>•.

*	ECSSD		DUT-OMRON	
	$F_\beta \uparrow$	MAE $\downarrow$	$F_\beta \uparrow$	MAE $\downarrow$
<i>Baseline</i>	0.8753	0.0568	0.7008	0.0761
<i>Boundary</i> <sup>-</sup>	0.8712	0.0576	0.6944	0.0779
<i>Boundary</i> <sup>+</sup>	0.8797	0.0521	0.7083	0.0712
<b><i>AFFM</i><sup>+</sup></b>	<b>0.8821</b>	<b>0.0509</b>	<b>0.7213</b>	<b>0.0601</b>
<i>AFFM</i> <sup>+</sup> •	0.8756	0.0558	0.7141	0.0669

**Effects of merging multi-scale predictions.** To verify the effects of the multi-scale FPM, we also perform a series experiments on ECSSD dataset. Firstly, we evaluate the results of a single FPM. Then we progressively add more FPMs to obtain the merged multi-scale results. The quantitative results are listed in Tab. III. The “12345” is our final model. From the results, we can observe that adding more FPMs can integrate more information, thus improving the performances.

TABLE III  
QUANTITATIVE RESULTS OF MERGING MULTI-SCALE PREDICTIONS.

Metric	1	2	3	4	5	45	345	2345	12345
$F_\beta \uparrow$	0.839	0.863	0.872	0.876	0.879	0.877	0.879	0.881	0.882
MAE $\downarrow$	0.060	0.057	0.054	0.053	0.052	0.053	0.052	0.051	0.051

## IV. CONCLUSION

In this paper, we propose a boundary-guided aggregating feature fusion network for salient object detection. Different from the methods directly introduce high-level features into shallow layers, our method integrates feature maps into multiple resolutions. The proposed attention-based feature fusion module can effectively refine the results with clear boundaries.

## REFERENCES

- [1] L. Itti, C. Koch, and E. Niebur, "A model of saliency-based visual attention for rapid scene analysis," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 1254–1259, 1998.
- [2] R. Achanta, S. Hemami, F. Estrada, and S. Susstrunk, "Frequency-tuned salient region detection," in *Proc. of IEEE Conference on Computer Vision and Pattern Recognition*, 2009, pp. 1597–1604.
- [3] N. Tong, H. Lu, R. Xiang, and M. H. Yang, "Salient object detection via bootstrap learning," in *Proc. of IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 1884–1892.
- [4] L. Wang, H. Lu, R. Xiang, and M. H. Yang, "Deep networks for saliency detection via local estimation and global search," in *Proc. of IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 3183–3192.
- [5] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems*, 2012, pp. 1097–1105.
- [6] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proc. of IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 3431–3440.
- [7] L. Wang, W. Ouyang, X. Wang, and H. Lu, "Stct: Sequentially training convolutional networks for visual tracking," in *Proc. of IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 1373–1381.
- [8] P. Zhang, D. Wang, H. Lu, and H. Wang, "Non-rigid object tracking via deep multi-scale spatial-temporal discriminative saliency maps," *arXiv:1802.07957*, 2018.
- [9] L. Wang, L. Wang, H. Lu, P. Zhang, and R. Xiang, "Saliency detection with recurrent fully convolutional networks," in *Proc. of IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 825–841.
- [10] G. Lee, Y. W. Tai, and J. Kim, "Deep saliency with encoded low level distance map and high level features," in *Proc. of IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 660–668.
- [11] N. Liu and J. Han, "Dhsnet: Deep hierarchical saliency network for salient object detection," in *Proc. of IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 678–686.
- [12] L. Wang, H. Lu, Y. Wang, M. Feng, D. Wang, B. Yin, and X. Ruan, "Learning to detect salient objects with image-level supervision," in *Proc. of IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 3796–3805.
- [13] P. Zhang, D. Wang, H. Lu, H. Wang, and X. Ruan, "Amulet: Aggregating multi-level convolutional features for salient object detection," in *Proc. of IEEE International Conference on Computer Vision*, 2017, pp. 202–211.
- [14] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. of IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 770–778.
- [15] J. Canny, "A computational approach to edge detection," *IEEE Transactions on Pattern Analysis & Machine Intelligence*, vol. 8, no. 6, pp. 679–698, 1986.
- [16] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *Computer Science*, 2014.
- [17] G. Lin, A. Milan, C. Shen, and I. Reid, "Refinenet: Multi-path refinement networks for high-resolution semantic segmentation," in *Proc. of IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 1925–1934.
- [18] L. Chen, H. Zhang, J. Xiao, L. Nie, J. Shao, W. Liu, and T. S. Chua, "Sca-cnn: Spatial and channel-wise attention in convolutional networks for image captioning," in *Proc. of IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 6298–6306.
- [19] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell, "Caffe: Convolutional architecture for fast feature embedding," in *ACM international conference on Multimedia*, 2014, pp. 675–678.
- [20] J. Deng, W. Dong, R. Socher, L. J. Li, K. Li, and F. F. Li, "Imagenet: A large-scale hierarchical image database," in *Proc. of IEEE Conference on Computer Vision and Pattern Recognition*, 2009, pp. 248–255.
- [21] P. Zhang, D. Wang, H. Lu, H. Wang, and B. Yin, "Learning uncertain convolutional features for accurate saliency detection," in *Proc. of IEEE International Conference on Computer Vision*, 2017, pp. 212–221.
- [22] P. Zhang, W. Liu, H. Lu, and C. Shen, "Salient object detection by lossless feature reflection," in *Proc. of International Joint Conference on Artificial Intelligence (IJCAI)*, 2018, pp. 1149–1155.
- [23] C. Yang, L. Zhang, H. Lu, R. Xiang, and M. H. Yang, "Saliency detection via graph-based manifold ranking," in *Proc. of IEEE Conference on Computer Vision and Pattern Recognition*, 2013, pp. 3166–3173.
- [24] G. Li and Y. Yu, "Visual saliency based on multiscale deep features," in *Proc. of IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 5455–5463.
- [25] J. Xiao, J. Hays, K. A. Ehinger, A. Oliva, and A. Torralba, "Sun database: Large-scale scene recognition from abbey to zoo," in *Proc. of IEEE Conference on Computer Vision and Pattern Recognition*, 2010, pp. 3485–3492.
- [26] A. Borji, M.-M. Cheng, H. Jiang, and J. Li, "Salient object detection: A benchmark," *IEEE Transactions on Image Processing*, vol. 24, no. 12, pp. 5706–5722, 2015.
- [27] A. Borji, "What is a salient object? a dataset and a baseline model for salient object detection," *IEEE Transactions on Image Processing*, vol. 24, no. 2, pp. 742–56, 2015.
- [28] Z. Luo, A. Mishra, A. Achkar, J. Eichel, S. Li, and P. M. Jodoin, "Non-local deep features for salient object detection," in *Proc. of IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 6593–6601.
- [29] X. Li, L. Zhao, L. Wei, M. H. Yang, F. Wu, Y. Zhuang, H. Ling, and J. Wang, "Deepsaliency: Multi-task deep neural network model for salient object detection," *IEEE Transactions on Image Processing*, vol. 25, no. 8, p. 3919, 2015.
- [30] G. Li and Y. Yu, "Deep contrast learning for salient object detection," in *Proc. of IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 478–487.
- [31] H. Jiang, J. Wang, Z. Yuan, Y. Wu, N. Zheng, and S. Li, "Salient object detection: A discriminative regional feature integration approach," in *Proc. of IEEE Conference on Computer Vision and Pattern Recognition*, 2013, pp. 2083–2090.
- [32] Y. Qin, H. Lu, Y. Xu, and H. Wang, "Saliency detection via cellular automata," in *Proc. of IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 110–119.