SALIENCY DETECTION BY DEEP NETWORK WITH BOUNDARY REFINEMENT AND GLOBAL CONTEXT

Xin Tan[*1], *Hengliang Zhu*[1], *Zhiwen Shao*[1], *Xiaonan Hou*[1], *Yangyang Hao*[1] *and Lizhuang Ma*[*1,2]

[1]Department of Computer Science and Engineering, Shanghai Jiao Tong University, China
[2]School of Computer Science and Software Engineering, East China Normal Univeristy, China
tanxin2017@sjtu.edu.cn, ma-lz@cs.sjtu.edu.cn

ABSTRACT

A novel end-to-end fully convolutional neural network for saliency detection is proposed in this paper, aiming at refining the boundary and covering the global context (GBR-Net). Previous CNN based methods for saliency detection are universally accompanied with blurring edge and ambiguous salient object. To tackle this problem, we propose to embed the boundary enhancement block (BEB) into the network to refine edge. It keeps the details by the mutual-coupling convolutional layers. Besides, we employ a pooling pyramid that utilizes the multi-level feature informations to search global context, and it also contributes as an auxiliary supervision. The final saliency map is obtained by fusing the edge refinement with global context extraction. Experiments on four benchmark datasets prove that the proposed saliency detection model gains an edge over the state-of-the-art approaches.

Index Terms— Saliency detection, Boundary refinement, Global context, Pooling pyramid

1. INTRODUCTION

As a classic task in computer vision, saliency detection refers to identify the most important and conspicuous objects or regions in an image. Working as a pre-processing step, it contributes a lot to researches and applications in computer vision, such as object detection [1], image classification [2], semantic segmentation [3] and person re-identification [4]. Although DCNNs are beneficial to saliency detection and reach accurate solutions [5, 6, 7, 8] compared with the hand-crafted features based methods [9, 10, 11] in recent years, it still remains unsolved problems. Because saliency object is variegated as its color, location, size and category from the raw image, it's hard to give a generalized detection method.

As an important step in image understanding, saliency detection attracts a lot of attention. At present, a number of deep neural networks based methods are applied for saliency detection. MDF [12] utilizes multiscale deep features to train the visual saliency model. MTDS [13] models the semantic properties of salient objects effectively, it takes a data-driven strategy for encoding the underlying saliency prior information. DCL [6] proposes a deep network with two complementary components, one is the pixel-level fully convolutional stream and the other is segment-wise spatial pooling stream. As Fig. 1(d, e, f) show, these methods can approximately model the shape of salient object, but they are not highlighted compared with the ground truth. On the other hand, DHS [14] uses the global structured saliency cues to perform the global prediction, then adopts a hierarchical recurrent convolutional neural network to refine the details. UCF [7] and RFCN [5] are both improve the accuracy with learning abundant high-level information. ELD [8] generates the saliency maps with high and low level features fused. As Fig. 1(g, h, i, j) illustrate, these approaches highlight the salient objects but the boundary is fuzzy. Especially the connection between the multi salient objects is ambiguous.

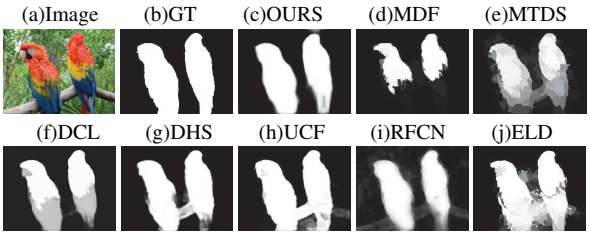


Fig. 1. Examples of the weakness in conventional methods. From top left to bottom right: image, ground truth mask, our saliency maps, and saliency maps of other five latest approaches, including MDF [12], MTDS [13], DCL [6], DHS [14], UCF [7], RFCN [5], ELD [8].

A lot of researches focus on highlighting the salient object and refining edge, but it still needs improvements for realistic requirements. To cope with above drawbacks in conventional methods, we use a single end-to-end network to model the raw image for saliency detection. To refine the boundary, we design a boundary enhancement block (BEB) embedded into

*Corresponding author. Thanks to National Natural Science Foundation of China (No. 61472245 and 61502220), and the Science and Technology Commission of Shanghai Municipality Program (No. 16511101300) for funding.

978-1-5386-1737-3/18/$31.00

the network, which keeps the details by the mutual-coupling convolutional layers. To gather the global information, we utilize the multiple convolution layers by pooling pyramid and combine them as the extra supervision, such that we can acquire semantic information at different levels but very few extra costs.

Based on the above considerations and motivations, a novel end-to-end fully convolutional neural network for saliency detection is proposed, which can highlight the salient objects and preserve the boundary information. In summary, this paper has the following contributions:

- A new saliency detection framework is proposed to refine the boundary and gather the global context simultaneously. Experiments on several benchmark datasets demonstrate that our deep model performs favorably against the state-of-the-art methods.
- We design a boundary enhancement block consisting of mutual-coupling convolutional layers through which we can maintain the completeness of salient object boundary information.
- We also employ the pooling pyramid for global context gathering in our deep network to fuse the semantic information at different levels, such that we can obtain more global information.

2. THE PROPOSED APPROACH

In this paper, we propose an end-to-end fully convolutional network to search the global context and refine the boundary at the same time with very few extra costs. We use the raw image I as the input image without any pre-processing. Section 2.1 describes the network architecture integrally and expounds the global context gathering part. Section 2.2 introduces the BEB used for boundary refinement in detail.

2.1. Network Architecture

As illustrated in Fig. 2, the network architecture consists of a general VGG16 [15], which is pre-trained on ImageNet. To refine the boundary, three BEBs are embedded into the network. We also employ the pooling pyramid to combine the feature maps of conv3-3, conv4-3, conv5-3 for gathering global context information. It is shown as the upper branch with blue mask in Fig.2. Pooling pyramid is helpful to further improve detecting the salient objects with accurate boundary, since it's covering the multi-level context structures which are significant for dense prediction. Although the resolution of feature map is various, global average pooling is to change F_i into 1×1 resolution. It can be formulated as

$$F_{pp} = \mathbf{CAT}(\tilde{F}_{3-3}, \tilde{F}_{4-3}, \tilde{F}_{5-3}), \tag{1}$$

where $\tilde{F}_i$ is the output of F_i after global average pooling. F_{pp} is the result of pooling pyramid by concatenation. Pooling pyramid can maintain the global context structures while it loses details. Then, we change the resolution of combined feature map F_{pp} to 64×64 by upsampling in our case so that it can be fused with conv5-3. It can be formulated as

$$F_{ppout} = \mathbf{SUM}(F_{5-3}, \mathbf{Upsamp}(F_{pp})), \tag{2}$$

where F_{ppout} is the output of the pooling pyramid branch. The upper branch is regarded as extra supervision and its output is fused to the final result.

The loss function of the network is the cross entropy between saliency map S_j and ground truth G. It has five supervisions in the network: one of them is master loss and others are auxiliary. The j-th loss function is defined as:

$$\begin{aligned} Loss_j(\Theta) = &\sum_{S_j} S_j^i log P(S_j^i = 1 | G^i, \Theta) \\ &+ \sum_{S_j} (1 - S_j^i) log P(S_j^i = 0 | G^i, \Theta), \end{aligned} \tag{3}$$

where Θ is the collection of all parameters of our GBRNet. Since we compute the loss function in pixels, we denote the saliency map as $S_j = \{S_j^i, j = 1, ..., 5, i = 1, ..., |S|\}$ and the ground truth as $G = \{G^i, i = 1, ..., |G|\}$. The final supervision for our model is the combined effect of all loss functions. Multiple supervisions jointly train our model and optimize the parameters for highlighting salient objects and refining the contour.

2.2. Boundary Enhancement Block

Inspired by GCNet [16], in which they proposed a BR module to refine the boundary in conjunction with the large kernel at the end of the network, we design a boundary enhancement block(BEB) to optimize the blurring edge in training stage. BEBs are used for modeling the boundary representations from low-level stages to high-level stages. In some early methods, boundary information is not taken into account. Recently, researchers start to employ a special network to refine the boundary information by training the local context. Although it does work in some cases, the extra network branch requires pre-processing training data, such as superpixel segmented images [17]. The BEB is embedded into the main network structure directly which is illustrated in Fig. 3(a).

Given a BEB, the inputs of which are the outputs of conv2-2, conv3-3, conv4-3 layers F_i. Here, we employ three branches to refine the i-th block's input and the corresponding output remains the same resolution of the input. One advantage of BEB is that it maintains the characteristics of the original network in spite of the depth increasing. Thus, there is no need to do any adjustment after embedment. Meanwhile, we can objectively evaluate the contributions of BEB compared with the general VGG16 [15].

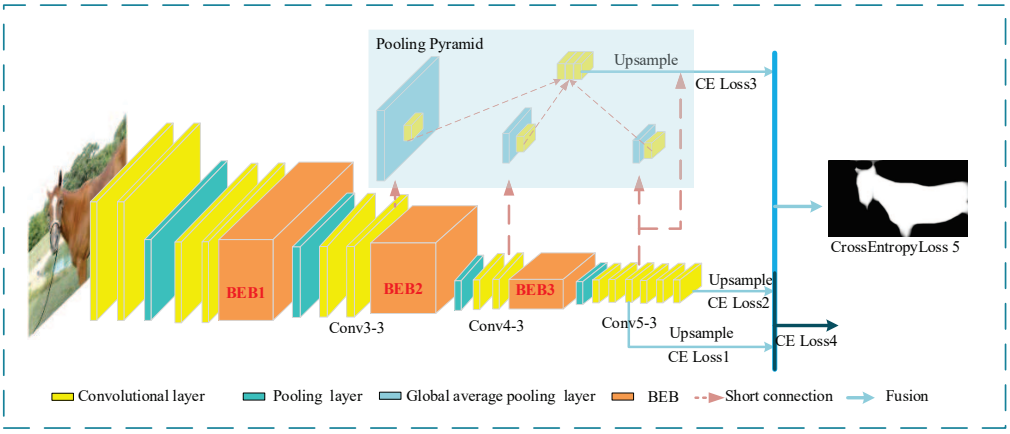
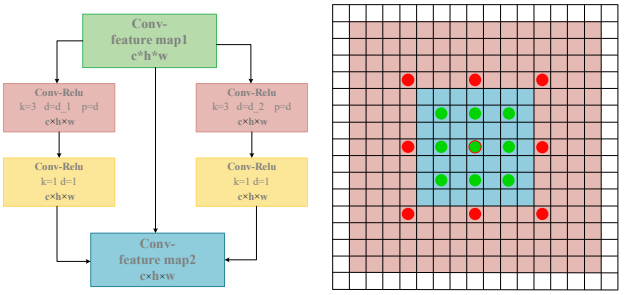


Fig. 2. The structure of the proposed network for boundary refinement and global context gathering(GBRNet). The lower branch is the VGG16 embedded with BEB for boundary refinement. The higher branch is the pooling pyramid for global context gathering.

Structure details. As illustrated in Fig. 3(a), one branch connects the input and output directly without any operation. Another two branches, $F_i{*_{d1}}$ with left dilated convolution operation and $F_i{*_{d2}}$ with right dilated convolution operation, are similar. Both $F_i{*_{d1}}$ and $F_i{*_{d2}}$ branches consist of a dilation convolutional layer following a 1×1 convolution kernel and the output dimensions is the same as the F_i. A minute difference in operation of left and right dilated convolution is dilation size for detecting the information at multi spatial-scale in an image. So BEB can be formulated as

$$F_i = F_i + \sum_{j\epsilon(0,1)} F_i{*_{dj}}. \tag{4}$$



(a) The structure of BEB.

(b) The explanation for the intercoupling kernel.

Fig. 3. The BEB and the intercoupling kernel.

Intercoupling kernel. The $*_d$ is defined as the dilated convolution operation [18]. $F : \mathbb{Z}^2 \to \mathbb{R}$ is a discrete function, and $k : \Omega_r \to \mathbb{R}$ is a 3×3 discrete filter, where $\Omega_r = [-r, r]^2 \cap \mathbb{Z}^2$. The * is the discrete convolution operator. Such that the $*_d$ is defined as

$$(F{*_d}k)(p) = \sum_{s+dt=p} F(s)k(t). \tag{5}$$

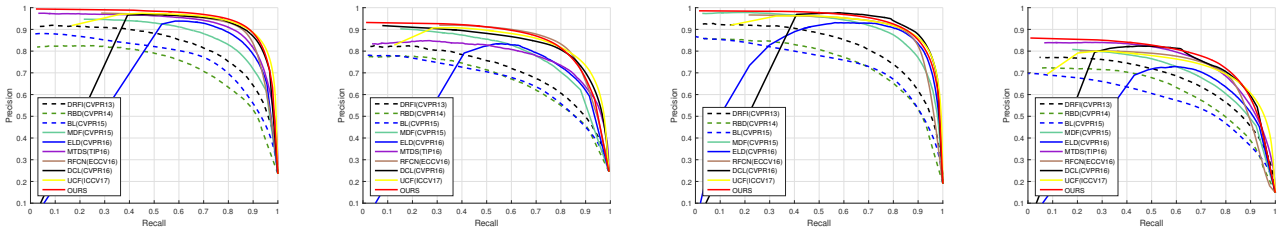
Let $*_{d1}$ and $*_{d2}$ be the dilation factor of the left and right branches, respectively. When they work at the same time, it is easy to obtain that the respective field of each kernel is $(4d-1)^2$. As for $F_i{*_{d1}}$ and $F_i{*_{d2}}$, dilation size is different, since the right one is always twice larger than the left one. Therefore, such two dilation kernels are intercoupling all the time. Considering the dilation size [2,4] illustrated in Fig. 3(b), the red points stand for d_2=4, and red mask is the respective filed of kernel size with 4-dilation equaling to $(4 \times 4 - 1)^2 = 225$. Although the large dilation is helpful to extract regions with wider respective field and keep the resolution, it still ignores the central part since most parameters among the central kernel are zero. As a remedy, BEB takes another 2-dilated kernel in blue mask, with its respective field equaling to $(4 \times 2 - 1)^2 = 49$, to cover the central region again to emphasize the most important central region as an intercoupling kernel.

Fig. 4. The performance of BEB and pooling pyramid (PP). From left to right: image, without BEB or PP, only with BEB, with BEB and PP and ground truth mask.

Fig. 4 shows the performance of BEB and pooling pyramid. BEB can refine the boundary and highlight the object distinctly. Pooling pyramid can also further improve the accuracy by clearing background.

3. EXPERIMENTS

In this section, we demonstrate the effectiveness of proposed network and compare against state-of-the-art methods on four salient detection benchmarks.



(a) ECSSD (b) PASCAL-S (c) HKU-IS (d) DUT-OMRON

Fig. 5. Comparison of PR curves between our proposed algorithm and other state-of-the-art methods.

Table 1. Comparison of F-measure (higher value is better) and MAE (lower value is better) on ECSSD, PASCAL-S, HKU-IS and DUT-OMRON datasets. The best two results are shown in red and blue fonts, respectively.

Method	ECSSD		PASCAL-S		HKU-IS		DUT-OMRON	
	maxFm	MAE	maxFm	MAE	maxFm	MAE	maxFm	MAE
DRFI(CVPR13) [19]	0.786	0.164	0.690	0.281	0.783	0.143	0.664	0.150
RBD(CVPR14) [9]	0.716	0.171	0.655	0.273	0.726	0.141	0.630	0.141
BL(CVPR15) [10]	0.755	0.217	0.659	0.318	0.723	0.206	0.580	0.240
MDF(CVPR15) [12]	0.831	0.105	0.764	0.142	0.861	0.129	0.694	0.092
ELD(CVPR16) [8]	0.868	0.079	0.771	0.121	0.881	0.063	0.705	0.091
MTDS(TIP16) [13]	0.882	0.122	0.760	0.175	-	-	0.745	0.120
RFCN(ECCV16) [5]	0.898	0.095	0.832	0.118	0.888	0.080	0.738	0.095
DCL(CVPR16) [6]	0.901	0.075	0.810	0.115	0.907	0.055	0.756	0.086
UCF(ICCV17) [7]	0.903	0.069	0.818	0.116	0.888	0.061	0.730	0.120
GBR(OURS)	0.909	0.066	0.824	0.107	0.893	0.055	0.758	0.074

3.1. Datasets and Evaluation Criteria

We evaluate our method on four standard benchmark datasets: ECSSD [20], PASCAL-S [21], HKU-IS [17] and DUT-OMRON [22]. We take three metrics: precision-recall curve (PR curve), F-measure and mean absolute error (MAE), to evaluate the competitiveness of all salient detection algorithms. The PR curves are obtained by the generated binary mask from the saliency map compared with the ground truth. And F-measure is formulated as

$$F_w = \frac{1+\omega^2 * precision * recall}{\omega^2 * precision + recall}, \tag{6}$$

where w^2 equals to 0.3 like the most precious works. Then let $\hat{S}$ and $\hat{G}$ remark the saliency map and the ground truth which are normalized to [0,1] so that the MAE is computed by

$$MAE = \frac{1}{H \times W}\sum_{i=1}^{H}\sum_{j=1}^{W}|\hat{S}(i,j) - \hat{G}(i,j)|, \tag{7}$$

To clearly evaluate the performance of these saliency algorithms, we also provide few quantitative results of F-measure and MAE with nine representative algorithms, including six deep learning based methods, on four datasets (Table1). It can be seen that our algorithm significantly outperforms other state-of-the-art algorithms.

3.2. Implementation Details

We train our model on MSRA-B [23] dataset with 5,000 images and MSRA10K [24] dataset with 10,000 images. As for BEB, we adapt the dilation size as [2,4], [4,8], [6,12] on the 3×3 kernel, respectively. The number of channels is set to 128, 256 and 512, respectively. Newly added layers are defaultly initialized with Gaussian distribution with standard deviation 0. It is fine-tuned based on the pre-trained VGG16 [15] for better comparison with previous works. The hyper-parameters used in this work contain base learning rate (1e-8), learning policy (step), stepsize (7,500), momentum (0.90) and weight decay (0.0005). The entire networks are implemented on the publicly available platform Caffe [25].

In the training phase, all five loss functions are taken for training. Loss5 is from the master branch and Loss3 is from the pooling pyramid branch. The deep supervision at the end of the network aims to reuse the feature maps to the maximum extent with least costs. In the testing phase, only the optimized master branch is used to predict the final result and another four auxiliary ones are abandoned.

3.3. Performance Comparison

We compare the proposed saliency detection method with 9 latest state-of-the-art methods, including **a) 6 deep learning based methods:** MDF [12], ELD [8], MTDS [13], RFCN [5],

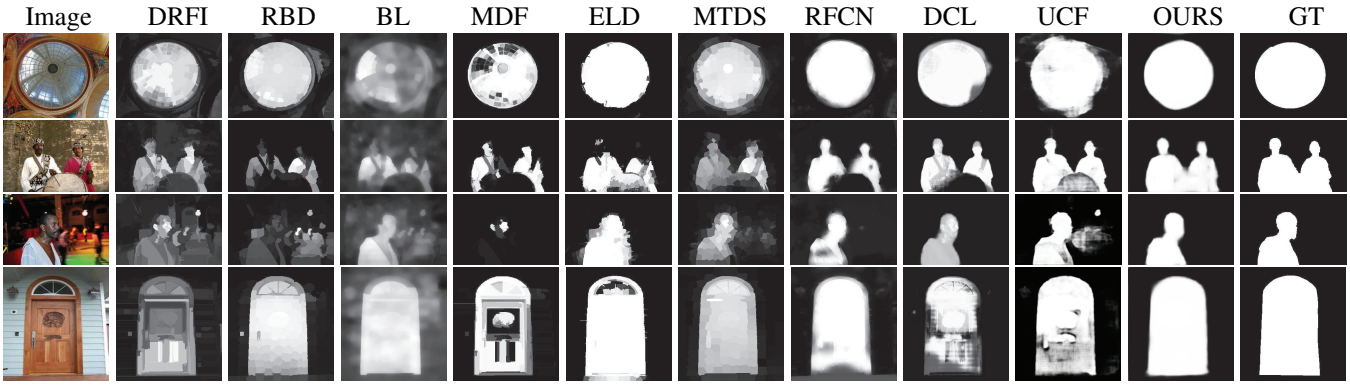


Fig. 6. Visual comparison with other state-of-the-art methods.

DCL [6], UCF [7] and **b) 3 classic methods:** DRFI [19], RBD [9], BL [10]. For fair comparison, we use either the saliency maps or the implementations of these methods provided by the authors.

F-measure and MAE. Table 1 shows that the proposed model performs favorably against almost all the existing state-of-the-art methods according to the evaluation criteria, including deep learning based methods and classic ones. Our method achieves top two on four test datasets over F-measure and MAE, it shows the effectiveness of the proposed model. Especially the GBRNet outperforms other methods over MAE on all four datasets. On ECSSD and DUT-OMRON datasets, which contains 1,000 and 5,618 images with pixel-wise annotation of salient object, respectively, our model also performs best over F-measure. On PASCAL-S and HKU-IS datasets, the GBRNet performs second place over F-measure.

PR curve. As illustrated in Fig. 5, our method performs best at the beginning of the PR curve, since our saliency maps are closely match to the ground truth masks. At the end of the PR curve, our method is still very competitive with other approaches. It can be reasonable believed that our method is very potential and efficient to the multiple salient objects detection task with the augmentation of the training dataset.

Visual comparison. Fig. 6 shows the visual comparison between our proposed approach for saliency detection and other mentioned methods. We can see that our saliency maps obtain the accurate shape of the salient objects, meanwhile, it preserves the smooth boundary which is significant for salient objects.

Running time. It takes us about 12 hours to train our model on a NVIDIA GTX-1080Ti GPU and Intel E5-2630 CPU processor. It only cost 0.045s for each image of average size 400×300 without any pre/post-processing. It's high-efficiency compared with most existed convolutional methods. It's easy to extend the proposed method and apply it to videos.

3.4. Ablation Studies

To strictly evaluate the effect of BEB and pooling pyramid, respectively, we perform a contrast experiment. For fair comparison, we only take out BEBs and pooling pyramid with Loss3 of GBRNet (see Fig. 2) as the baseline model. And we embed BEB/pooling pyramid into baseline model to evaluate them, respectively. Finally we put both BEB and pooling pyramid into the network as the final model to evaluate the performance. As shown in table 2, the baseline model with BEB and pooling pyramid performs best and the baseline model only with BEB/pooling pyramid has the nearly second highest performance. Especially, on ECSSD dataset, BEB can decrease by 12% over MAE (from 0.076 to 0.067) and pooling pyramid increases by 1% over F-measure (from 0.896 to 0.909). On PASCAL-S dataset, BEB and pooling pyramid both can decreases by 6.8% over MAE (from 0.118 to 0.111/0.110). It demonstrates that BEB and pooling pyramid are works a lot for the saliency detection.

Table 2. Comparison of F-measure and MAE on ECSSD and PASCAL-S datasets to evaluate the performance of BEB and pooling pyramid. 'PP' in the third row refers to the pooling pyramid.

Method	ECSSD		PASCAL-S	
	maxFm	MAE	maxFm	MAE
Baseline	0.896	0.076	0.812	0.118
Baseline+BEB	0.899	0.067	0.818	0.110
Baseline+PP	0.902	0.070	0.818	0.111
Baseline+BEB+PP	0.909	0.066	0.824	0.107

4. CONCLUSIONS

In this paper, we propose a novel fully convolutional networks for salient detetcion without any pre/post-processing. It can refine the boundary and cover the global context at the same time. This proposed method solves the problem that salient objects are always ambiguous with blurring edge in previous

networks. BEBs are embedded into the VGG16 to keep the edge details with the mutual-coupling convolutional kernels, which can emphasizes the most important central region. And the pooling pyramid is utilized to search the global context. Experimental results on four standard benchmark datasets, ECSSD, Pascal-S, HKU-IS and DUT-OMRON show our proposed method outperforms other state-of-the-art methods.

5. REFERENCES

[1] Vidhya Navalpakkam and Laurent Itti., "An integrated model of top-down and bottom-up attention for optimizing detection speed," in *CVPR*, 2006, pp. 2049–2056.

[2] Yizhou Yu Ruobing Wu and Wenping Wang, "Scale: Supervised and cascaded laplacian eigenmaps for visual object recognition based on nearest neighbors," in *CVPR*, 2013, pp. 867–874.

[3] Michael Donoser, Martin Urschler, Martin Hirzer, and Horst Bischof, "Saliency driven total variation segmentation," in *ICCV*, 2009, pp. 817–824.

[4] Rui Zhao, Wanli Ouyang, and Xiaogang Wang, "Unsupervised salience learning for person re-identification," in *CVPR*, 2013, pp. 3586–3593.

[5] Linzhao Wang, Lijun Wang, Huchuan Lu, Pingping Zhang, and Ruan Xiang, "Saliency detection with recurrent fully convolutional networks," in *ECCV*, 2016, pp. 825–841.

[6] Guanbin Li and Yizhou Yu, "Deep contrast learning for salient object detection," in *CVPR*, 2016, pp. 478–487.

[7] Pingping Zhang, Dong Wang, Huchuan Lu, Hongyu Wang, and Baocai Yin, "Learning uncertain convolutional features for accurate saliency detection," in *ICCV*, 2017.

[8] Gayoung Lee, Yu Wing Tai, and Junmo Kim, "Deep saliency with encoded low level distance map and high level features," in *CVPR*, 2016, pp. 660–668.

[9] Wangjiang Zhu, Shuang Liang, Yichen Wei, and Jian Sun, "Saliency optimization from robust background detection," in *CVPR*, 2014, pp. 2814–2821.

[10] Na Tong, Huchuan Lu, Ruan Xiang, and Ming Hsuan Yang, "Salient object detection via bootstrap learning," in *CVPR*, 2015, pp. 1884–1892.

[11] Hengliang Zhu, Bin Sheng, Xiao Lin, Yangyang Hao, and Lizhuang Ma, "Foreground object sensing for saliency detection," pp. 111–118, 2016.

[12] Guanbin Li and Yizhou Yu, "Visual saliency based on multiscale deep features," in *CVPR*, 2015, pp. 5455–5463.

[13] Xi Li, Liming Zhao, Lina Wei, Ming-Hsuan Yang, Fei Wu, Yueting Zhuang, Haibin Ling, and Jingdong Wang, "Deepsaliency: Multi-task deep neural network model for salient object detection.," *IEEE Transaction on Image Processing*, vol. 25, no. 8, pp. 3919, 2016.

[14] Nian Liu and Junwei Han, "Dhsnet: Deep hierarchical saliency network for salient object detection," in *CVPR*, 2016, pp. 678–686.

[15] Karen Simonyan and Andrew Zisserman, "Very deep convolutional networks for large-scale image recognition," *Computer Science*, 2014.

[16] Chao Peng, Xiangyu Zhang, Gang Yu, Guiming Luo, and Jian Sun, "Large kernel matters – improve semantic segmentation by global convolutional network," in *CVPR*, 2017.

[17] Zhao Rui, Ouyang Wanli, Li Hongsheng, and Wang Xiaogang, "Saliency detection by multi-context deep learning," in *CVPR*, 2015, pp. 1265–1274.

[18] Fisher Yu and Vladlen Koltun, "Multi-scale context aggregation by dilated convolutions," in *ICLR*, 2016.

[19] Huaizu Jiang, Jingdong Wang, Zejian Yuan, Yang Wu, Nanning Zheng, and Shipeng Li, "Salient object detection: A discriminative regional feature integration approach," in *CVPR*, 2013, pp. 2083–2090.

[20] Qiong Yan, Li Xu, Jianping Shi, and Jiaya Jia, "Hierarchical saliency detection," in *CVPR*, 2013, pp. 1155–1162.

[21] Yin Li, Xiaodi Hou, Christof Koch, James M. Rehg, and Alan L. Yuille, "The secrets of salient object segmentation," in *CVPR*, 2014, pp. 280–287.

[22] Chuan Yang, Lihe Zhang, Huchuan Lu, Ruan Xiang, and Ming Hsuan Yang, "Saliency detection via graph-based manifold ranking," in *CVPR*, 2013, pp. 3166–3173.

[23] Zejian Yuan, Zejian Yuan, Jian Sun, Nanning Zheng, Nanning Zheng, Xiaoou Tang, and Heung Yeung Shum, "Learning to detect a salient object," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 33, no. 2, pp. 353, 2011.

[24] Ming Ming Cheng, Guo Xin Zhang, N. J. Mitra, Xiaolei Huang, and Shi Min Hu, "Global contrast based salient region detection," in *CVPR*, 2011, pp. 409–416.

[25] Yangqing Jia, Evan Shelhamer, Jeff Donahue, Sergey Karayev, Jonathan Long, Ross Girshick, Sergio Guadarrama, and Trevor Darrell, "Caffe: Convolutional architecture for fast feature embedding," in *ACM MM*, 2014, pp. 675–678.