Multi-Scale Salient Object Detection with Pyramid Spatial Pooling

Jing Zhang*, Yuchao Dai†, Fatih Porikli† and Mingyi He*
* School of Electronics and Information, Northwestern Polytechnical University, China.
E-mail: zjnwpu@gmail.com; myhe@nwpu.edu.cn
† Research School of Engineering, Australian National University, Australia.
E-mail: yuchao.dai@anu.edu.au; fatih.porikli@anu.edu.au

***Abstract*—Salient object detection is a challenging task in complex compositions depicting multiple objects of different scales. Albeit the recent progress thanks to the convolutional neural networks, the state-of-the-art methods still fall short to handle such real-life scenarios.**

In this paper, we propose a new method that exploits both multi-scale feature fusion and pyramid spatial pooling to detect salient object regions in varying sizes. Our framework consists of a front-end network and two multi-scale fusion modules. The front-end network learns an end-to-end mapping from the input image to a saliency map, where a pyramid spatial pooling is incorporated to aggregate rich context information from different spatial receptive fields. The multi-scale fusion module integrates saliency cues across different layers, that is from low-level detail patterns to high-level semantic information by concatenating feature maps, to segment out salient objects with multiple scales. Extensive experimental results on eight benchmark datasets demonstrate the superior performance of our method compared with existing methods.

I. Introduction

Salient object detection aims at assigning each pixel in the image a saliency label, thus predicting prominent and important regions of the scene. It is intrinsic to many computer vision tasks, such as context-aware image editing [1] and semantic image labeling [2]. In general, traditional methods either employ handcrafted features such as color, contrast, and texture based descriptors, or compute variants of appearance uniqueness and region compactness based on statistical priors, e.g. center prior [3] and boundary prior [4]. These methods achieve acceptable results on relatively simple datasets, but their saliency maps deteriorate when the input images are cluttered and complicated.

Data-driven approaches, in particular, deep learning has recently achieved significant success in high-level computer vision tasks such as image classification [5], and semantic segmentation [6]. These approaches have also been extended to salient object detection, e.g. [7] [8] [9] [10] [11] [12] [13] that learn high-level feature representations of salient objects, outperforming traditional handcrafted methods with a large margin. Even though the success provided mainly by fully convolutional neural networks, this task is still very challenging for complex scenarios with multiple objects in different scales. The state-of-the-art deep saliency methods fail to address these difficult and practical scenarios.

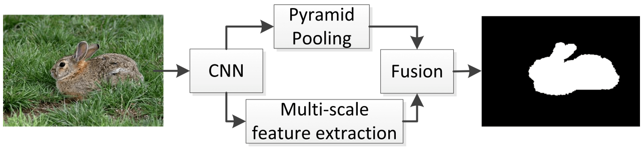

Fig. 1. Given an RGB image as input, our framework employs both multi-scale features and pyramid spatial pooling to learn the saliency map with rich semantic information. Multi-scale feature extraction is performed to aggregate different level side-outputs from the front-end model. Pyramid spatial pooling aims at extracting rich context information for saliency from different regions.

To provide an efficient and robust solution for detection of multiple varying size salient objects, we introduce multi-scale feature fusion and incorporate them in pyramid spatial pooling as demonstrated in Fig. 1. First, inspired by context-based semantic segmentation networks [14] and skip-layer edge detection networks [15], we learn a front-end deep model that is built on a very deep classification network, ResNet [5]. This front-end deep model captures the mapping from input color image to output saliency map, where pyramid spatial pooling is used to integrate rich context cues for saliency from different regions. Second, to fuse features at varying scales from the front-end model, we propose two multi-scale feature fusion modules that aggregate low-level patterns and high-level semantic cues for saliency. In this way, our model effectively partitions out multiple salient objects and resolves multi-scale issues. Our framework is illustrated in Fig. 2.

The main contributions of the paper can be summarized as:

1) We introduce a front-end network to learn an end-to-end mapping from the input image to saliency map by incorporating pyramid spatial pooling.
2) We propose a multi-scale fusion module to integrate saliency cues across different layers from low-level details to high-level semantic information.
3) Our method achieves the state-of-the-art performance on eight large benchmark datasets.

II. Related Work

Prior to the deep learning revolution, conventional salient object detection methods were mainly based on low-level handcrafted features. Here, we provide an brief overview of deep learning based approaches. We refer readers to [16]

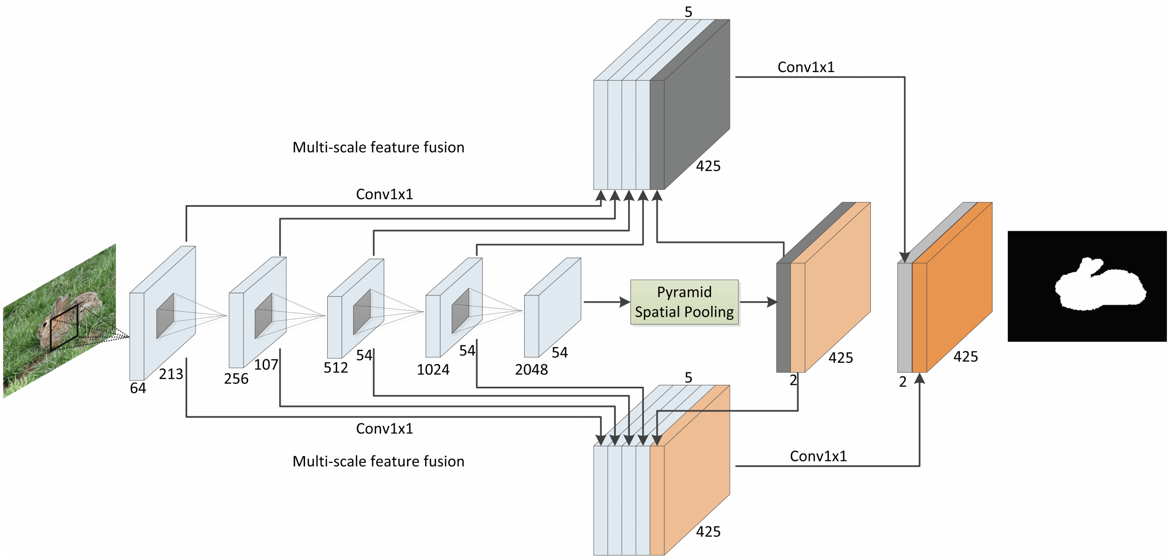

Fig. 2. For a given input image, the repurposed ResNet-101 [5] and a pyramid spatial pooling module generate the front-end two-channel saliency map, with the first channel being the background map and the second channel the saliency. Then, four mid-layer outputs of the ResNet-101 network are concatenated together with the front-end saliency map, fusing multi-scale features. Each of these two concatenated feature maps is fed to a 1×1 convolutional layer to obtain the final two-channel saliency map.

and [17] for in-depth surveys and benchmark comparisons of handcrafted feature based methods.

Deep learning approaches derive a mapping from the input image to the saliency map by employing convolutional and fully connected neural networks. By jointly modeling global and local context, Li and Yu [9] proposed a multi-context deep convolutional neural network for saliency detection. Along the same pipeline, [18] integrated deep features and handcrafted features. The work in [10] formulated saliency detection as a two-stage estimation problem, where a local estimation stage and a global search stage are performed to predict saliency scores for salient regions. An end-to-end contrast network was applied in [11] to produce pixel-level saliency maps. Similarly, Deep Image Saliency Computing (DISC) [13] aimed at computing fine-grained image saliency. A recurrent attentional convolution-deconvolution network was incorporated in [19]. In [20], both CNN features and low-level features were integrated. Li et al. [8] presented a multi-task learning framework where saliency detection and semantic segmentation are jointly learned. Liu and Han [21] utilized an end-to-end hierarchical network for the same purpose.

Multi-scale feature fusion has been shown as a key element in achieving the state-of-the-art performance on semantic segmentation [22] [14] [23] and edge detection [24]. There are mainly three types of network structures that exploit multi-scale features. The first type is the skip-net based methods, where features from intermediate layers are combined to achieve a specific task. By introducing supervision to side-outputs of the network, the Holistically-Nested Edge Detector (HED)[24] model led to considerable improvements over generic fully connected network models for edge detection. Within this structure, Cheng et al. [7] proposed a saliency detection method by introducing short connections to the skip-layer structures of [24]. The second type of structure is to fuse features of different context. For example, [23] and [6] introduced the dilated convolutions that support an exponential expansion of the receptive field without loss of resolution. The work in [14] introduced a pyramid pooling module to exploit the capability of context information by using a multi-scale pooling layer. The third type of network structure is the share-net based methods where the input image is resized to several scales with each passes through a shared-weight deep network. Along this line, [22] introduced an attention mechanism that learns to softly weight the multi-scale features at each pixel location.

Our framework is based on the fully convolutional neural networks yet it consists of both pyramid spatial pooling and multi-scale feature fusion. In contrast to existing methods, in particular to multi-scale feature based saliency [7], [25] and [11], our framework allows efficient aggregation of saliency cues across different layers and also incorporation of rich context information from different spatial regions.

III. Our Framework

Targeting at segmenting out salient objects with different scales, we propose an end-to-end fully convolutional network based framework to leverage on both high-level semantic cues and low-level information. Our framework exploits the rich context information from different spatial regions by using pyramid spatial pooling.

Firstly, we repurpose a deep convolutional neural network, PSPNet [14], adapting it from semantic segmentation to saliency detection. By using four scale pyramid pooling, PSPNet is able to capture both global context (with receptive field of the whole image) and local context (with receptive field of $1/4$, $1/9$ and $1/36$ respectively). This network constitutes our front-end saliency detection model.

Secondly, we fuse high-level semantic information and low-level feature details from different layers of the front-end network to learn the multi-scale saliency cues. These together constitute our multi-scale fusion model. Given an input color image I, the front-end model produces a two-channel feature map S_f which is $1/8$ size of I. Furthermore, we add one 1×1 convolutional layer to the last convolutional layer of each lower block of our front-end model (conv1_3_3×3/relu, conv2_3/relu, conv3_4/relu and conv4_23/relu layers respectively of [14]) to map the lower-level features to one-channel feature map.

Lastly, we concatenate the above four feature maps from the front-end model with each of the front-end (two channel) feature maps, and apply a 1×1 convolutional layer to map the concatenated feature map to a one-channel feature map. This procedure is illustrated in Fig. 2. We train our model in a jointly supervised manner, where the loss is evaluated at both the front-end module and the multi-scale fusion module.

A. Front-end Model

Our front-end saliency detection network is built upon a semantic segmentation net, i.e., PSPNet [14] and DeepLab [6], where we repurpose a deep convolutional neural network (ResNet-101 [5]) originally designed for image classification to the task of semantic segmentation by 1) transforming all fully connected layers to convolutional layers, 2) increasing feature resolution through dilated convolutional layers [6] [23], and 3) introducing a pyramid spatial pooling module to explore the potential of different region-based contextual information. Under our framework, the spatial resolution of the output feature map is increased four times, which is superior to [10] and [9].

With the increasing size of the receptive fields, more contextual information is extracted through continuous convolutions. This motivates building deeper networks to achieve larger receptive fields. Even though the theoretical receptive field of ResNet [5] is already larger than its input image, it has been shown in [26] that the empirical receptive field of CNN is much smaller than the theoretical one. To attain and take advantage of contextual information from different regions, we propose a pyramid spatial pooling module, similar to the one used in semantic segmentation [14]. Specifically, we apply a four-scale spatial pooling (average pooling) to the feature maps. After spatial pooling, each feature map is fed through a convolutional layer and upsampled to the same dimension. The concatenated feature maps are then supplied to a two-layer convolutional network in order to generate a two-channel feature map, which corresponds to our saliency map. This pyramid spatial pooling as illustrated in Fig. 3 produces a useful contextual prior that is essential for saliency detection.

In our experiments, we randomly cropped a 425×425 regions of the input image. For the pyramid spatial pooling module, four scale average pooling is utilized, where kernel sizes are set as $k = 54, 27, 18, 9$, respectively. We initialized the front-end model using PSPNet for scene parsing [14] on the PASCAL VOC 2012 dataset. We used the "Softmaxwith-loss" as our loss function at the last layer.

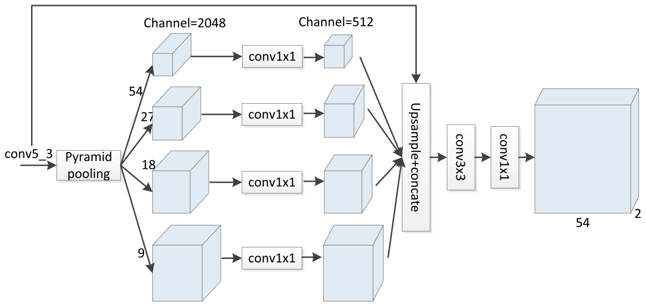

Fig. 3. Pyramid spatial pooling. Given feature map from layer "conv5_3", the pooling module uses four different scales. The concatenated feature maps are then fed into the convolutional layers to map the pooling features to a two-channel feature map.

B. Fusion of Multi-scale Features

We have two insights to leverage on cues that are available in multiple scales. Firstly, the receptive fields of lower-level features especially the bottom-level is quite small, which makes it unsuitable to perform classification in an earlier stage. Secondly, low-level features can provide favorable details while high-level features own rich semantic information. Based on these, we use side-outputs of different layers of the network such that we employ feature extraction modules instead of applying classification module after each lower block. This allows us to extract feature maps from lower level of network to a one-channel feature map instead of two-channel saliency map.

We define the four side-outputs from our front-end model as A_1, A_2, A_3, A_4 (conv1_3_3×3/relu, conv2_3/relu, conv3_4/relu and conv4_23/relu layers respectively). Then a conv1×1 layer is added in the end of A_i $(i = 1...4)$ to map A_i to a one-channel feature map F_1, F_2, F_3, F_4. Further, we define the two-channel feature map from our front-end model as $S = \{S_b, S_f\}$, where S_b is the background map and S_f is the saliency map with each pixel represents the possibility for it to be background or salient. Finally, we fuse the above features as:

1) Concatenating multi-level features with each channel of the front-end feature map, with the background channel: $F_b = \{F_1, F_2, F_3, F_4, S_b\}$ and the salient foreground channel: $F_f = \{F_1, F_2, F_3, F_4, S_f\}$.
2) Extracting new features from both low-level and high-level features. One $conv1 \times 1$ layer is used to map F_b and F_f to a one-channel feature map S_b and S_f respectively.
3) Our final output is a two-channel feature map: $S_{fuse} = \{S_{nb}, S_{nf}\}$.

In both the feature extraction stage and the feature fusion stage, weights are initialized with the "MSRA" policy, and bias is initialized with a constant. We upsample the one-channel feature map from each side-output to the original image size before the last fusion stage.

C. Training Details

We trained our model using Caffe [27] where we stopped the training when the accuracy score on the training data remained unchanged for at least 200 iterations with a maximum iteration number of 10K. We used the stochastic gradient descent method with the momentum value of 0.9 and decreased the learning rate 90% when the training loss did not decrease. The learning rate is initialized as 1e-3 with the "poly" decay policy. For validation, we set "test_iter" as 500 (test batch size 1) to cover the full 500 validation images. The whole training process takes 30 hours with training batch size 1 and "iter_size" 20 on a PC with an NVIDIA Quadro M4000 GPU. At the inference stage, it takes 0.2 seconds on average to predict the saliency map for a 425×425 image.

IV. Experimental Results

A. Setup

Dataset: We have evaluated our method on eight saliency benchmark datasets. 2,500 images from the MSRA-B dataset [28] are used for training and 500 images are used for validation, with the remaining 2,000 for testing. The other seven datasets are ECSSD [29], DUT [30], SED1 and SED2 [31], PASCAL-S [32], HKU-IS [9] and THUR [33] dataset.

Compared methods: We evaluated eight state-of-the-art deep learning based saliency detection methods: DSS [7], DC [11], MDF [9], DeepMC [10], DMT [8], DISC [13], RFCN [12] and LEGS [34], and four traditional saliency detection methods: DRFI [35] and RBD [4], DSR [36] and MC [37] which were proven in [17] as the state-of-the-art before the era of deep learning.

Evaluation metrics: We use two evaluation metrics, namely mean absolute error (MAE), and Precision-Recall (PR) curve. MAE can provide a better estimate of the dissimilarity between the estimated saliency map and the ground truth saliency map. It is the average per-pixel difference between the ground truth and the estimated binary saliency map, normalized to [0, 1], which is defined as:

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

where W and H are the width and height of the respective saliency map S, GT is the ground truth saliency map.

The PR curves are obtained by binarizing the saliency map in the range of [0 255], where $Precision$ corresponds to the percentage of salient pixels being correctly detected, and $Recall$ corresponds to the fraction of detected salient pixels in relation to the ground truth number of salient pixels.

B. Comparisons with the State-of-the-art

Quantitative Comparisons: We compared our method with eight deep learning based methods and four traditional methods. Results are shown in Table I and Fig.4, where "OUR" represents results of our model. Table I shows that for those eight datasets, the state-of-the-art deep learning based methods outperform the best traditional methods with 3%-10% decrease in MAE, which further proves the superiority of deep features for saliency detection. Our method ranks 1st in seven out of eight benchmark datasets except for the ECSSD dataset, where our method ranks the 2nd. Especially for the DUT, SED1 and THUR dataset, our method achieves about 2% decrease in MAE. We further compare the PR curve on the above eight datasets and the results are shown in Fig. 4. The results show that for the DUT dataset and PASCAL-S dataset, our method outperform existing deep learning based methods with a big margin, and for the remaining six datasets, our method achieves consistent better performance.

Qualitative Comparisons: Fig. 5 demonstrates several visual comparisons; as visible our method outperforms the competing methods. The tested samples in the first two rows of Fig. 5 are in very low contrast where most of the existing methods failed to capture the whole salient object, especially for DeepMC [10] and MDF [9]. In contrast, our method successfully captures the salient objects with much sharper edges preserved. The sample in the fourth row of Fig. 5 depicts a simple scenario where most of the competing methods can generate satisfactory results except for DSS [7] and MDF [9] where the green capsicum is missed. For this image, our method achieves the best result with most of the salient regions highlight equally. The salient object in the fourth row has strong internal contrast, which lead to much false detection especially for MDF. Still, our method achieves consistently better results inside the salient object and the most of the background is accurately suppressed.

C. Ablation Analysis

As our multi-scale saliency detection model consists of both front-end saliency detection and multi-scale feature fusion based structure, it is interested to analyze the contribution of each component. To this end, we define the front-end model as our basic model, where we change "num_output" of the last classification layer to 2 to match our case. We trained the front-end module for saliency detection individually, and the performance is shown in Table II, where "Basic" represents our front-end model. On eight benchmark datasets, our model of using multi-scale fusion strategy achieves consistent better performance compared with the basic model with front-end model only, which further proves the effectiveness of our method.

V. Conclusions

With the use of fully convolutional neural network, salient object detection has witnessed great progress and performance leap. However, salient object detection is still challenging for complex scenarios with objects of multiple scales and the current state-of-the-art methods fail to handle these difficult and practical scenarios.

In this paper, by integrating low-level feature with high-level features through a skip-architecture in a pyramid spatial pooling structure, we propose a multi-scale saliency detection

TABLE I
MAE FOR DIFFERENT METHODS INCLUDING OURS ON EIGHT BENCHMARK DATASETS.(BEST ONES IN BOLD)

	DSS	DC	MDF	DeepMC	DMT	DISC	RFCN	DRFI	RBD	DSR	MC	OUR
ECSSD	**0.0628**	0.0906	0.1081	0.1019	0.1601	0.0699	0.0973	0.1719	0.1739	0.1742	0.2037	0.0654
DUT	0.0760	0.0971	0.0916	0.0885	0.0758	0.1182	0.0945	0.1496	0.1467	0.1374	0.1863	**0.0579**
SED1	0.0887	0.0886	0.1198	0.0881	-	0.0772	0.1020	0.1454	0.1407	0.1614	0.1620	**0.0676**
SED2	0.1014	0.1014	0.1171	0.1162	0.1074	0.1203	0.1140	0.1373	0.1316	0.1457	0.1848	**0.0863**
PASCAL-S	0.1546	0.1614	0.2069	0.1928	0.2103	-	0.1662	0.2556	0.2418	0.2600	0.2719	**0.1502**
MSRA-B	0.0474	0.0467	0.1040	0.0491	0.0658	0.0536	0.0620	0.1229	0.1171	0.1207	0.1441	**0.0379**
HKU-IS	0.0520	0.0730	-	0.0913	-	0.1023	0.0798	0.1445	0.1432	0.1404	0.1840	**0.0443**
THUR	0.1142	0.0959	0.1029	0.1025	0.0854	-	0.1003	0.1471	0.1507	0.1408	0.1838	**0.0670**

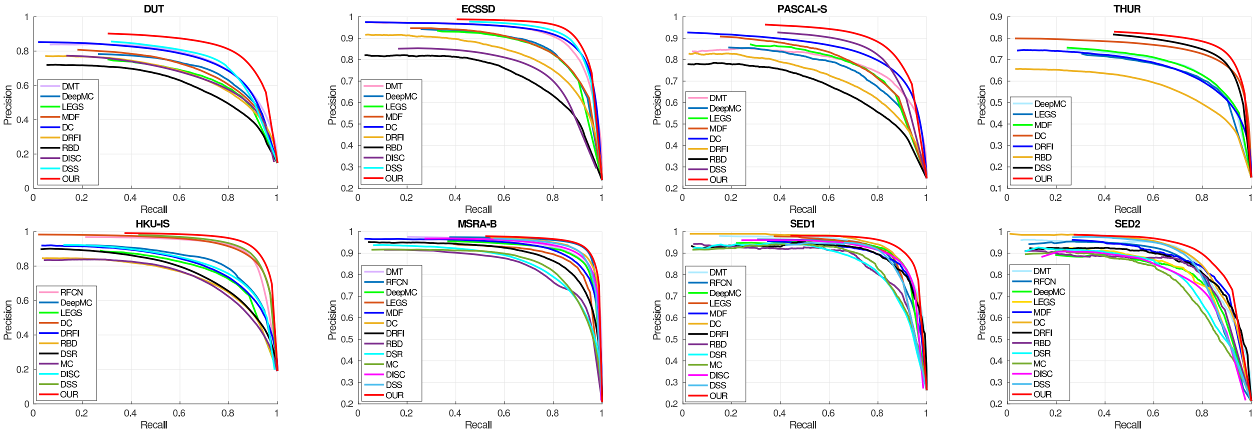

Fig. 4. PR curve of our method and competing methods on eight benchmark datasets.

TABLE II
MAE FOR OUR METHOD AND THE BASIC MODEL ON EIGHT BENCHMARK DATASETS.(BEST ONES IN BOLD)

	ECSSD	DUT	SED1	SED2	PASCAL-S	MSRA-B	HKU-IS	THUR
Basic	0.0759	0.0658	0.0773	0.0956	0.1593	0.0425	0.0526	0.0746
OUR	**0.0654**	**0.0579**	**0.0676**	**0.0863**	**0.1502**	**0.0379**	**0.0443**	**0.0670**

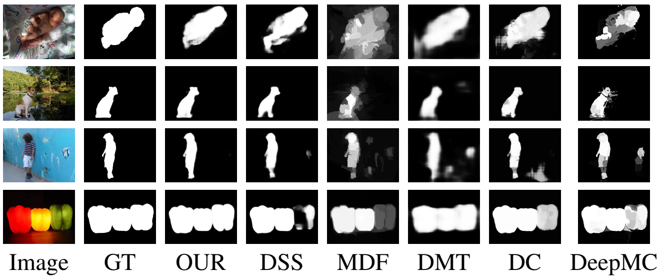

Fig. 5. Qualitative comparisons of our method with the state-of-the-art.

model that can handle salient objects of multi-scale in resolution as well as those exist in complex scenarios. Extensive evaluations on eight benchmark datasets demonstrate that the proposed method outperforms the state-of-the-art approaches with a large margin.

ACKNOWLEDGEMENT

This work was done when Jing Zhang was a joint PhD student between Northwestern Polytechnical University and the Australian National University. This work was supported in part by Natural Science Foundation of China grants (61420106007, 61671387) and the Australian Research Council grants (DE140100180, DP150104645).

REFERENCES

[1] G. Zhang, M. Cheng, S. Hu, and R. R. Martin, "A shape-preserving approach to image resizing," *Computer Graphics Forum*, vol. 28, no. 7, pp. 1897–1906, 2009. 1

[2] S.J. Oh, R. Benenson, A. Khoreva, Z. Akata, M. Fritz, and B. Schiele, "Exploiting saliency for object segmentation from image level labels," *CoRR*, vol. abs/1701.08261, 2017. 1

[3] S. Goferman, L. Zelnik-Manor, and A. Tal, "Context-aware saliency detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 10, pp. 1915–1926, Oct 2012. 1

[4] W. Zhu, S. Liang, Y. Wei, and J. Sun, "Saliency optimization from robust background detection," in *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, 2014, pp. 2814–2821. 1, 4

[5] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, June 2016, pp. 770–778. 1, 2, 3

[6] L. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "Semantic image segmentation with deep convolutional nets and fully connected crfs," *CoRR*, vol. abs/1412.7062, 2014. 1, 2, 3

[7] Q. Hou, M. Cheng, X. Hu, A. Borji, Z. Tu, and P. H. S. Torr, "Deeply supervised salient object detection with short connections," *CoRR*, vol. abs/1611.04849, 2016. 1, 2, 4

[8] 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 Trans. Image Proc.*, vol. 25, no. 8, pp. 3919–3930, Aug 2016. 1, 2, 4

[9] G. Li and Y. Yu, "Visual saliency based on multiscale deep features," in *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, June 2015, pp. 5455–5463. 1, 2, 3, 4

[10] R. Zhao, W. Ouyang, H. Li, and X. Wang, "Saliency detection by multi-context deep learning," in *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, 2015, pp. 1265–1274. 1, 2, 3, 4

[11] G. Li and Y. Yu, "Deep contrast learning for salient object detection," in *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, June 2016, pp. 478–487. 1, 2, 4

[12] L. Wang, L. Wang, H. Lu, P. Zhang, and X. Ruan, "Saliency detection with recurrent fully convolutional networks," in *Proc. Eur. Conf. Comp. Vis.*, 2016, pp. 825–841. 1, 4

[13] T. Chen, L. Lin, L. Liu, X. Luo, and X. Li, "Disc: Deep image saliency computing via progressive representation learning," *IEEE Trans. Neural Networks Learning Syst.*, vol. 27, no. 6, pp. 1135–1149, June 2016. 1, 2, 4

[14] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," *CoRR*, vol. abs/1612.01105, 2016. 1, 2, 3

[15] M. Liu S. Ramalingam Z. Yu, C. Feng, "Casenet: Deep category-aware semantic edge detection," *CoRR*, vol. abs/1705.09759, 2017. 1

[16] Ali Borji, Ming-Ming Cheng, Huaizu Jiang, and Jia Li, "Salient object detection: A survey," *CoRR*, vol. abs/1411.5878, 2014. 1

[17] A. Borji, M. Cheng, H. Jiang, and J. Li, "Salient object detection: A benchmark," *IEEE Trans. Image Proc.*, vol. 24, no. 12, pp. 5706–5722, 2015. 2, 4

[18] G. Li and Y. Yu, "Visual saliency detection based on multiscale deep CNN features," *CoRR*, vol. abs/1609.02077, 2016. 2

[19] J. Kuen, Z. Wang, and G. Wang, "Recurrent Attentional Networks for Saliency Detection," *ArXiv e-prints*, Apr. 2016. 2

[20] G. Lee, Y. Tai, and J. Kim, "Deep saliency with encoded low level distance map and high level features," *CoRR*, vol. abs/1604.05495, 2016. 2

[21] N. Liu and J. Han, "Dhsnet: Deep hierarchical saliency network for salient object detection," in *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, June 2016, pp. 678–686. 2

[22] L. C. Chen, Y. Yang, J. Wang, W. Xu, and A. L. Yuille, "Attention to scale: Scale-aware semantic image segmentation," in *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, June 2016, pp. 3640–3649. 2

[23] F. Yu and V. Koltun, "Multi-Scale Context Aggregation by Dilated Convolutions," *ArXiv e-prints*, Nov. 2015. 2, 3

[24] S. Xie and Z. Tu, "Holistically-nested edge detection," in *Proc. IEEE Int. Conf. Comp. Vis.*, 2015, pp. 1395–1403. 2

[25] G. Li, Y. Xie, L. Lin, and Y. Yu, "Instance-Level Salient Object Segmentation," *ArXiv e-prints*, Apr. 2017. 2

[26] B. Zhou, A. Khosla, À. Lapedriza, A. Oliva, and A. Torralba, "Object detectors emerge in deep scene cnns," *CoRR*, vol. abs/1412.6856, 2014. 3

[27] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell, "Caffe: Convolutional architecture for fast feature embedding," *arXiv preprint arXiv:1408.5093*, 2014. 4

[28] T. Liu, J. Sun, N. Zheng, X. Tang, and H. Shum, "Learning to detect a salient object," in *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, 2007, pp. 1–8. 4

[29] Q. Yan, L. Xu, J. Shi, and J. Jia, "Hierarchical saliency detection," in *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, 2013, pp. 1155–1162. 4

[30] C. Yang, L. Zhang, H. Lu, X. Ruan, and M. Yang, "Saliency detection via graph-based manifold ranking," in *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, 2013, pp. 3166–3173. 4

[31] S. Alpert, M. Galun, A. Brandt, and R. Basri, "Image segmentation by probabilistic bottom-up aggregation and cue integration," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 2, pp. 315–327, Feb 2012. 4

[32] Y. Li, X. Hou, C. Koch, J. M. Rehg, and A. L. Yuille, "The secrets of salient object segmentation," in *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, 2014, pp. 280–287. 4

[33] M. Cheng, N. J. Mitra, X. Huang, and S. Hu, "Salientshape: group saliency in image collections," *The Visual Computer*, vol. 30, no. 4, pp. 443–453, 2014. 4

[34] L. Wang, H. Lu, X. Ruan, and M. H. Yang, "Deep networks for saliency detection via local estimation and global search," in *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, June 2015, pp. 3183–3192. 4

[35] H. Jiang, J. Wang, Z. Yuan, Y. Wu, N. Zheng, and S. Li, "Salient object detection: A discriminative regional feature integration approach," in *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, 2013, pp. 2083–2090. 4

[36] X. Li, H. Lu, L. Zhang, X. Ruan, and M. Yang, "Saliency detection via dense and sparse reconstruction," in *Proc. IEEE Int. Conf. Comp. Vis.*, Dec 2013, pp. 2976–2983. 4

[37] B. Jiang, L. Zhang, H. Lu, C. Yang, and M. Yang, "Saliency detection via absorbing markov chain," in *Proc. IEEE Int. Conf. Comp. Vis.*, 2013, pp. 1665–1672. 4