SALIENT OBJECT DETECTION VIA FAST R-CNN AND LOW-LEVEL CUES

Xiang Wang   *Huimin Ma*   *Xiaozhi Chen*

Tsinghua National Laboratory for Information Science and Technology(TNList)
Department of Electronic Engineering, Tsinghua University, Beijing 100084, China
{wangxiang14, chenxz12}@mails.tsinghua.edu.cn, mhmpub@tsinghua.edu.cn

ABSTRACT

Recent advances in salient object detection have exploited the deep Convolutional Neural Network (CNN) to represent high-level semantic, however, due to the presence of convolutional and pooling layers, it is difficult for CNN to generate saliency map with sharp boundaries. In this paper, we propose multi-scale mask-based Fast R-CNN framework which generate saliency score of each region. Since the regions are segmented using edge-preserved methods, the results are naturally with sharp boundaries. To consider context information, we also propose low-level contrast and backgroundness prior which are complementary with high-level semantic. Finally, an edge-based propagation method which takes advantages of edge information is proposed to refine the saliency map. Experiments on three benchmark datasets demonstrate that the proposed method outperforms previous methods and achieves state-of-the-art performance.

Index Terms— Salient object detection, Fast R-CNN, backgroundness prior, edge-based propagation

1. INTRODUCTION

Salient object detection which aims to detect object that attracts people's attention has been widely researched in recent years. Traditional bottom-up methods mostly rely on some priors or assumptions, such as center-surround difference [1, 2], uniqueness prior [3, 4] and backgroundness prior [5, 6]. Those methods mainly focus on finding the difference between salient objects and background. However, bottom-up methods may fail in representing high-level semantic feature, thus making it difficult to achieve high performance for images with low-contrast or complex scenes.

Recently, the deep Convolutional Neural Network (CNN) has attracted a lot of attention for its strong ability in representing high-level semantic feature. Several methods [7, 8, 9] based on CNN have been proposed to detect salient object and have achieved state-of-the-art results. However, due to the presence of convolutional layers with large receptive fields and pooling layers, the output of CNN is coarse and has non-sharp boundaries [10]. In this paper, we propose a mask-based Fast R-CNN [11] framework to address this issue. As shown in Fig. 1, by adding region mask at the RoI pooling layer, the Fast R-CNN framework can extract high-level semantic feature and output saliency score of each region. Then the saliency map of the entire image is formed by saliency score of each region. Since the regions are segmented by edge-preserved method, the saliency map generated by our method is naturally with sharp boundaries.

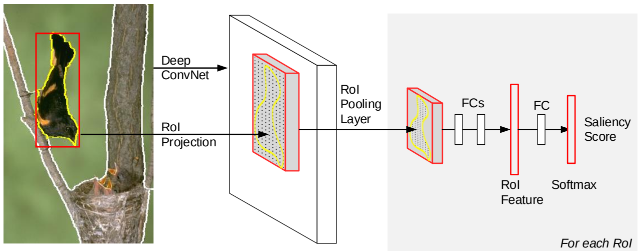


Fig. 1. Mask-based Fast R-CNN architecture.

On the other hand, most previous works are based on superpixel considering computational efficiency and that superpixel can preserve object boundaries. However, superpixel segmentation may cause object being segmented into many parts, making it difficult to highlight the whole salient object uniformly. Recent advances in edge detection [12, 13] have achieved highly satisfactory performance which makes it practical to use edge information to help better detect salient objects. In this paper, we also propose to use regions segmented by edges that is complementary with superpixel regions.

The main contributions of this paper are threefold. First, a multi-scale mask-based Fast R-CNN framework is proposed to generate saliency score of each region. Second, to consider context information, low-level contrast and backgroundness are proposed to be complementary with high-level semantic. Third, an edge-based propagation method which takes advantages of edge information is proposed to refine saliency map. With this propagation, salient objects are highlighted uniformly and the background are suppressed. Fig. 2 shows the pipeline of our method.

The rest of this paper is organized as follows. Sec. 2 introduces the details of the proposed method, Sec. 3 presents the experiments and analysis, conclusion is made in Sec. 4.

This work was supported by National Natural Science Foundation of China (No. 61171113).

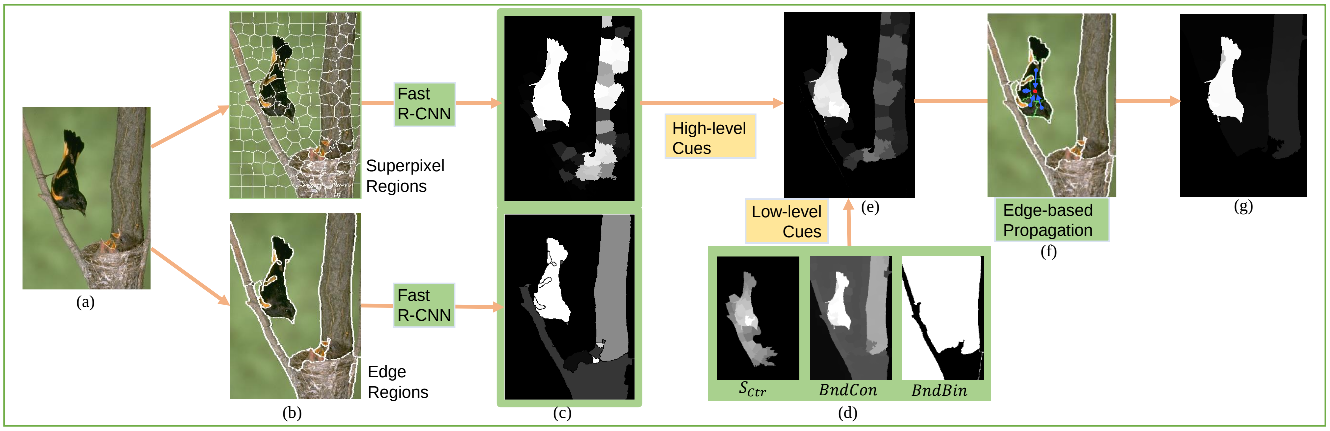


Fig. 2. Pipeline of our method. Given an image (a), we segment it into multi-scale regions (b), and generate saliency score (c) via mask-based Fast R-CNN. To consider context information, low-level cues (d) including contrast and backgroundness are proposed. We combine the high-level and low-level cues to form the coarse map (e) and refine it via an edge-based propagation (f) to get the final saliency map (g).

2. THE PROPOSED METHOD

2.1. Mask-based Fast R-CNN for Saliency Estimation

Fast R-CNN [11] is an efficient and general framework for object classification in which the convolutional layers are shared on the entire image and the feature of each region is extracted by the RoI pooling layer. Recently, this framework has also been applied to pixel-wise applications, such as semantic segmentation [14]. However, these methods require computing masks as input of the network. In this paper, we apply Fast R-CNN framework to salient object detection by considering it as a binary classification problem. The architecture of our proposed mask-based Fast R-CNN is shown as Fig. 1. Given an image, we first segment it into regions and use them as masks. Then for each region, we use its external rectangle as proposal and put it into Fast R-CNN. At the RoI pooling layer, we only pool the data inside the mask while leave the outside pooling data as zero. By using a softmax at the last layer, the network generates a score of region being salient. At the training stage, a region is considered as salient/background if more than 80% of its pixels are located inside/outside ground truth. With this network, we infer saliency score of each region to form the saliency map of the entire image. We also take the data of the last convolutional layer as the feature of each region.

The purpose of salient object detection is to uniformly highlight salient objects and suppress background regions. Most previous works are based on superpixels, however, when an object is segmented into dozens of superpixels, it will be difficult to uniformly highlight the whole object. Recent advances in edge detection have achieved highly satisfactory performance which makes it practical to use edge information to help better detect salient objects. So in our work, we also consider larger scale regions which are segmented by edges (denoted as edge regions). We use method in [13] to get object edges and thinning them use method in [12]. These two scales regions are complementary since superpixels can generate results with high resolution and edge regions can preserve compactness of objects. We denote the saliency map generated by Fast R-CNN with superpixel regions and edge regions as S_S and S_E , respectively.

While the Fast R-CNN framework can represent high-level semantic feature, it ignores the context information. So we also consider low-level backgroundness prior and contrast to address this issue.

2.2. Backgroundness Prior

We propose two kinds of backgroundness priors based on edge regions, with one is real-valued which measures the probability of regions being background, and another one is binary which directly removes background regions.

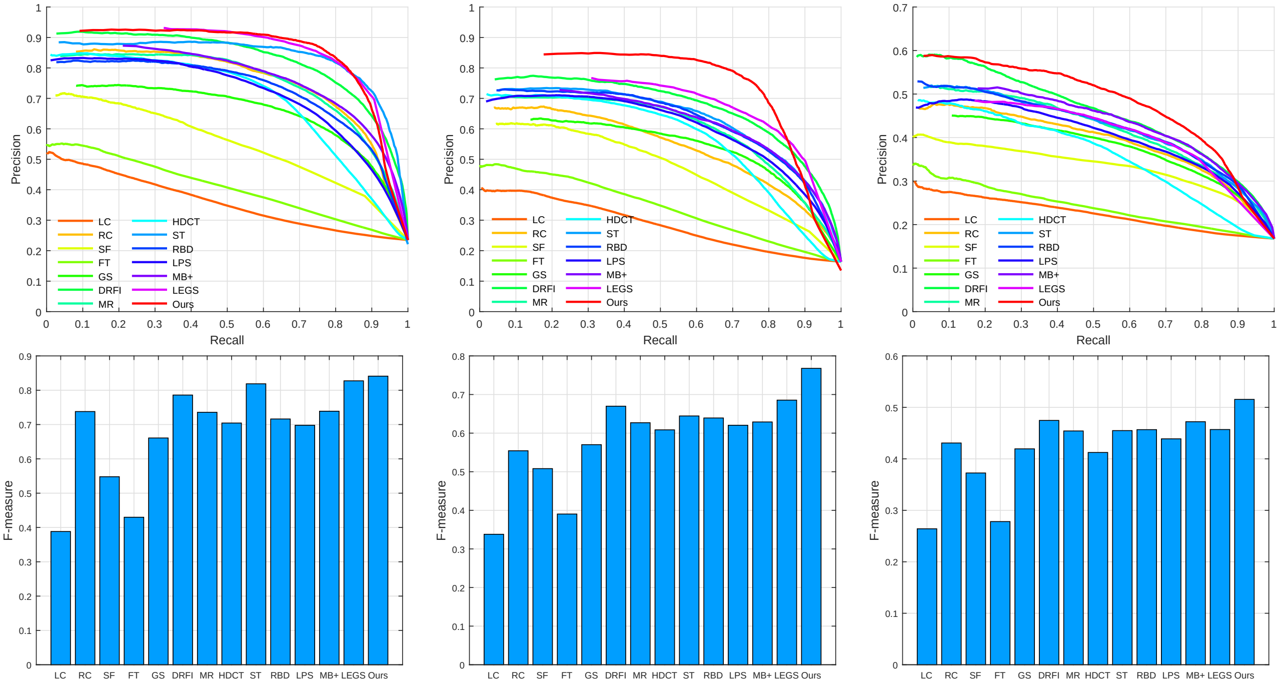
The first kind of backgroundness is boundary connectivity [6] which measures regions' backgroundness as:

$$BndCon(r) = \frac{Len_{bnd}(r)}{\sqrt{Area(r)}}, \tag{1}$$

in which $Len_{bnd}(r)$ denotes the length that region r connected with image boundaries and $Area(r)$ denotes the area of region r.

While previous backgroundness methods usually take image boundaries as background seeds, they will be fail when the objects touch the boundaries. Based on our observation, salient objects may touch one or two sides of image boundaries, they hardly touch two opposite sides of image boundaries. So if a region touches at least two opposite sides of image boundaries, we can consider it as background and directly remove it. The second backgroundness is defined as:

$$BndBin(r) = \begin{cases} 1, & \text{if } r \text{ touches opposite boundaries,} \\ 0, & \text{others.} \end{cases} \tag{2}$$



ECSSD DUT-OMRON JuddDB

Fig. 3. Comparison with state-of-the-art methods on three benchmark datasets. The first row shows the PR curves and the second row shows the F-measure. Best viewed in color.

With these two kinds of backgroundness, we can define the backgroundness prior as:

$$S_B = exp(-\frac{BndCon}{2\sigma_{BC}^2}) \times (1 - BndBin). \tag{3}$$

To consider context information and be complementary with saliency maps generated by Fast R-CNN, we propose a region contrast weighted by spatial distance and boundary connectivity as:

$$S_{Ctr}(p) = \sum_{i=1}^{N} d_f(p, p_i) w_{pos}(p, p_i) BndConSP(p_i), \tag{4}$$

in which $d_f(p, p_i)$ denotes the feature distance between superpixel p and p_i and the feature is extracted from the last fully convolutional layer. $w_{pos} = exp(-\frac{d_{pos}(p,p_i)}{2\sigma_{pos}})$ represents spatial weight and $d_{pos}(p, p_i)$ is the spatial distance of two superpixels' centers. We set $\sigma_{pos} = 0.25$ as in [15]. $BndConSP$ is computed using $BndCon$ by averaging it in each superpixel region.

Up to now, we get a coarse saliency map which considers high-level semantic feature, low-level contrast and backgroundness prior:

$$S_{coarse} = (S_E + S_S + S_{Ctr}) \times S_B. \tag{5}$$

2.3. Edge-based Propagation

Note that S_S and S_{Ctr} are based on superpixel regions, to take advantages of edge information and further refine the saliency map, we propose an edge-based propagation method. First, we convert S_{coarse} into superpixel level, namely, averaging it in each superpixel region. Then in the propagation, each superpixel region i is refined using the weighted sum of regions $PropR(i)$ which are not only adjacent but also within the same edge region (A superpixel is considered inside an edge region if more than 80% of its pixels are located in the edge region). The final saliency map (named FL) is formulated as:

$$S_{FL}(i) = \frac{1}{\sum\limits_{j \in PropR(i)} W(i,j)} \sum_{j \in PropR(i)} W(i,j) S_{coarse}(j), \tag{6}$$

with $W(i, j)$ denotes the weight which is formulated by feature similarity and spatial distance similar with (4).

$$W(i,j) = d_f(i,j) w_{pos}(i,j) \tag{7}$$

Fig. 2 (f) shows an example. With the edge-based propagation, the saliency maps are more uniform and background regions are also suppressed.

3. EXPERIMENTS

3.1. Setup

We evaluate our method on three challenging benchmark datasets: ECSSD [16] , DUT-OMRON [17] and JuddDB [18]. The ECSSD dataset contains 1000 images with complex scenes. The DUT-OMRON dataset contains 5168 images with more complex and natural scenes. The JuddDB dataset contains 900 images with complex natural scenes and high resolution (most are 1024×768). It is the most challenging dataset according to the benchmark in [19].

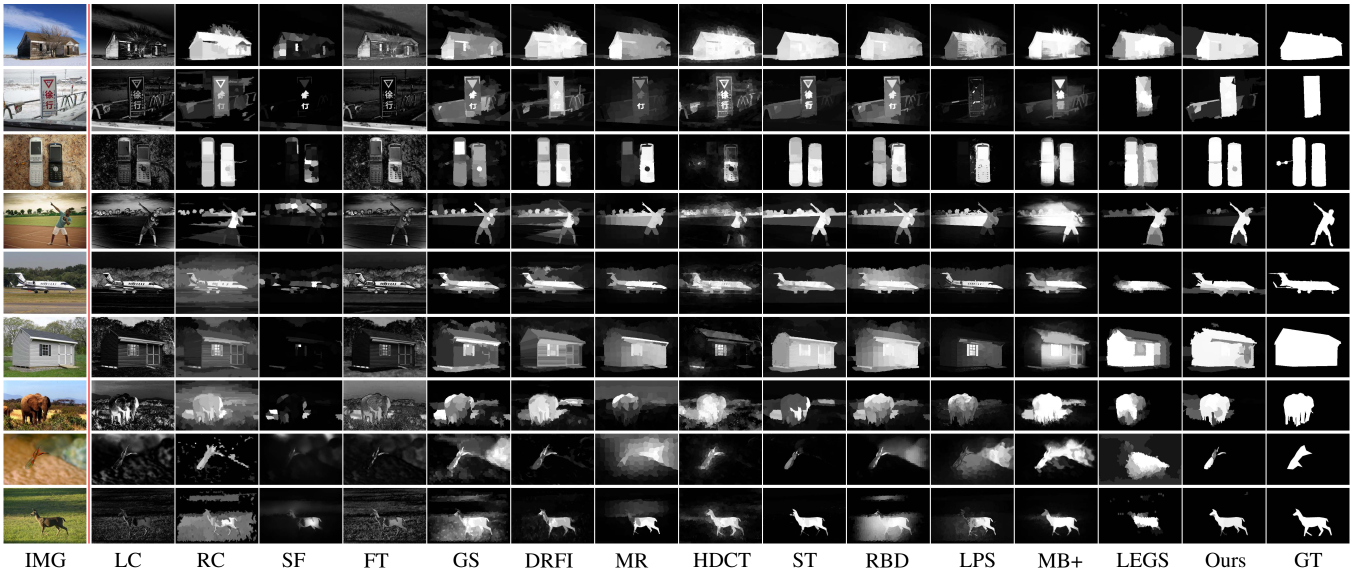


Fig. 4. Qualitative comparison with state-of-the-art methods on three benchmark datasets.

In our work, we randomly sample 4000 images from the DUT-OMRON dataset to train the proposed mask-based Fast R-CNN. The remaining images are used for evaluations. The network we used is VGG16 [20] and we implement it using Caffe framework [21]. The code will be available online.

We evaluate the performance using precision-recall (PR) curves and F-measure. The saliency maps are first normalized to $[0, 255]$, and then the precision and recall are computed by binarizing them with 256 thresholds and comparing them with ground truth. The PR curves are computed by averaging them on each dataset. The F-measure considers both precision and recall which is computed as: $F_\beta = \frac{(1+\beta^2)Precision \times Recall}{\beta^2 Precision + Recall}$, we set $\beta^2 = 0.3$ as most previous works [22, 23] to emphasize the precision. The final F-measure is the maximal F_β computed by 256 precision-recall pairs in the PR curves [19].

3.2. Evaluation of the Effectiveness of Our Method

To verify the effectiveness of the proposed method, we output all the intermediate results and compare them with the final result. Table 1 shows the results with F-measure compared. We can see that $S_{Ctr} + S_S + S_E$ outperforms their individual result, which demonstrates that they are complementary. After considering the backgroundness prior and the edge-based propagation, the performance increases gradually, which shows that the backgroundness prior and edge-based propagation are also effective.

3.3. Comparison with State-of-the-Art Methods

We compare our method with 13 state-of-the-art methods: LC [24], RC [23], SF [15], FT [22], GS [5], DRFI [25] MR [17], HDCT [26], ST [27], RBD [6], LPS [28], MB+ [29], LEGS [9]. Fig. 3 shows the PR curves and F-measure on three benchmark datasets. We can see that our method outperforms all other methods. Note that we only train the Fast R-CNN on parts of the DUT-OMRON dataset, our method performs well on other datasets, which shows that our method has strong generalization ability. In addition, from Table 1 and Fig. 3, we can see that our intermediate result S_E already outperforms other methods on the DUT-OMRON and JuddDB datasets, which demonstrates that the edge information is an important cue for images with complex scenes.

Table 1. Evaluation of the Effectiveness of Our Method

	ECSSD	DUT-OMRON	JuddDB
S_{Ctr}	0.581	0.451	0.326
S_S	0.801	0.712	0.440
S_E	0.815	0.743	0.482
$S_{Ctr} + S_S + S_E$	0.831	0.752	0.488
S_{coarse}	0.837	0.754	0.500
S_{FL}	**0.841**	**0.768**	**0.516**

Fig 4 shows the qualitative comparison. We can see that our results are much better and with clear boundaries.

4. CONCLUSION

In this paper, we propose a multi-scale mask-based Fast R-CNN framework to generate saliency map and to extract high-level semantic feature. While saliency map via Fast R-CNN may lose the context information, we also consider low-level contrast and backgroundness prior which are complementary with high-level semantic. An edge-based propagation method is proposed to further refine the saliency map by taking advantages of edge information. Experimental results on three benchmark datasets show that the proposed method outperforms previous methods and achieves state-of-the-art performance, we can also conclude that the edge information is important especially for images with complex scenes.

5. REFERENCES

[1] Laurent Itti, Christof Koch, and Ernst Niebur, “A model of saliency-based visual attention for rapid scene analysis,” *IEEE TPAMI*, 1998.

[2] Tie Liu, Jian Sun, Nan-Ning Zheng, Xiaoou Tang, and Heung-Yeung Shum, “Learning to detect a salient object,” in *CVPR*, 2007.

[3] Keyang Shi, Keze Wang, Jiangbo Lu, and Liang Lin, “Pisa: pixelwise image saliency by aggregating complementary appearance contrast measures with spatial priors,” in *CVPR*, 2013.

[4] Peng Jiang, Haibin Ling, Jingyi Yu, and Jingliang Peng, “Salient region detection by ufo: Uniqueness, focusness and objectness,” in *ICCV*, 2013.

[5] Yichen Wei, Fang Wen, Wangjiang Zhu, and Jian Sun, “Geodesic saliency using background priors,” in *ECCV*. 2012.

[6] Wangjiang Zhu, Shuang Liang, Yichen Wei, and Jian Sun, “Saliency optimization from robust background detection,” in *CVPR*, 2014.

[7] Tianshui Chen, Liang Lin, Lingbo Liu, Xiaonan Luo, and Xuelong Li, “DISC: Deep image saliency computing via progressive representation learning,” *IEEE TNNLS*, 2016.

[8] Xi Li, Liming Zhao, Lina Wei, MingHsuan Yang, Fei Wu, Yueting Zhuang, Haibin Ling, and Jingdong Wang, “DeepSaliency: Multi-task deep neural network model for salient object detection,” *arXiv preprint arXiv:1510.05484*, 2015.

[9] Lijun Wang, Huchuan Lu, Xiang Ruan, and Ming-Hsuan Yang, “Deep networks for saliency detection via local estimation and global search,” in *CVPR*, 2015.

[10] Shuai Zheng, Sadeep Jayasumana, Bernardino Romera-Paredes, Vibhav Vineet, Zhizhong Su, Dalong Du, Chang Huang, and Philip Torr, “Conditional random fields as recurrent neural networks,” in *ICCV*, 2015.

[11] Ross Girshick, “Fast R-CNN,” in *ICCV*, 2015.

[12] Piotr Dollár and C Lawrence Zitnick, “Structured forests for fast edge detection,” in *ICCV*, 2013.

[13] Saining Xie and Zhuowen Tu, “Holistically-nested edge detection,” in *ICCV*, 2015.

[14] Jifeng Dai, Kaiming He, and Jian Sun, “Instance-aware semantic segmentation via multi-task network cascades,” *arXiv preprint arXiv:1512.04412*, 2015.

[15] Federico Perazzi, Philipp Krähenbühl, Yael Pritch, and Alexander Hornung, “Saliency filters: Contrast based filtering for salient region detection,” in *CVPR*, 2012.

[16] Qiong Yan, Li Xu, Jianping Shi, and Jiaya Jia, “Hierarchical saliency detection,” in *CVPR*, 2013.

[17] Chuan Yang, Lihe Zhang, Huchuan Lu, Xiang Ruan, and Ming-Hsuan Yang, “Saliency detection via graph-based manifold ranking,” in *CVPR*, 2013.

[18] Ali Borji, “What is a salient object? a dataset and a baseline model for salient object detection,” *IEEE TIP*, 2015.

[19] A. Borji, Ming-Ming Cheng, Huaizu Jiang, and Jia Li, “Salient object detection: A benchmark,” *IEEE TIP*, 2015.

[20] Karen Simonyan and Andrew Zisserman, “Very deep convolutional networks for large-scale image recognition,” *ICLR*, 2015.

[21] 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.

[22] Ravi Achanta, Sheila Hemami, Francisco Estrada, and Sabine Susstrunk, “Frequency-tuned salient region detection,” in *CVPR*, 2009.

[23] Ming-Ming Cheng, Guo-Xin Zhang, Niloy J Mitra, Xiaolei Huang, and Shi-Min Hu, “Global contrast based salient region detection,” in *CVPR*, 2011.

[24] Yun Zhai and Mubarak Shah, “Visual attention detection in video sequences using spatiotemporal cues,” in *ACM MM*, 2006.

[25] 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.

[26] Jiwhan Kim, Dongyoon Han, Yu-Wing Tai, and Junmo Kim, “Salient region detection via high-dimensional color transform,” in *CVPR*, 2014.

[27] Zhi Liu, Wenbin Zou, and Olivier Le Meur, “Saliency tree: A novel saliency detection framework,” *IEEE TIP*, 2014.

[28] Hongyang Li, Huchuan Lu, Zhe Lin, Xiaohui Shen, and Brian Price, “Inner and inter label propagation: Salient object detection in the wild,” *IEEE TIP*, 2015.

[29] Jianming Zhang, Stan Sclaroff, Zhe Lin, Xiaohui Shen, Brian Price, and Radomir Mech, “Minimum barrier salient object detection at 80 fps,” in *ICCV*, 2015.