Robust Background Exclusion for Salient Object Detection

Yuming Hu[1], Quan Zhou[1], Guangwei Gao[1], Zhijun Yao[2], Weihua Ou[3], Longin Jan Latecki[4]
[1]Key Lab of Ministry of Education for Broad Band Communication and Sensor Network Technology,
Nanjing University of Posts and Telecommunications, Nanjing, China
[2]The 723 Institute of China Shipbuilding Industry Corporation, Yangzhou, China
[3]School of Big Data and Computer Science, Guizhou Normal University, Guiyang, China
[4]Department of Computer and Information Science, Temple University, Philadelphia, USA

Abstract—**Recently, investigating boundary prior to aid other low-level image cues, have gained great attention in salient object detection. Although the salient regions are mostly located in the image center, the inverse might not necessarily be true. In addition, such kind of center-bias assumption is very simple and fragile, especially when salient regions often touch the image boundary or the images are with complex background (e.g., small scale and high contrast patterns). To this end, this paper presents a new framework to address these two issues. First, we propose a robust background descriptor, called *boundary connectivity*, which is calculated to measure how heavily a region is connected to image boundary. This measure has an intuitive geometrical interpretation that are absent in previous saliency formulation. In order to reduce the effect of imprecise object boundary, we also propose a multiple layer over-segmentation framework to integrate multiple low-level cues, including our background measure, to highlight clean and uniform saliency maps. The experiment results demonstrate that our method achieves state-of-the-art results on MSRA-1000 datasets.**

Index Terms—**salient object detection, boundary connectivity, multiple layer over-segmentation.**

I. Introduction

Our human beings have the powerful capability to quickly locate the most interesting parts, when first glancing at the given scene. Such image parts are considered as *salient object/regions* since they attract greater attention than other parts by the human visual system (HVS). Form the perspective of computer vision, the task of *salient object detection* is defined as the binary segmentation problem of separating the salient objects from the background [1]. In recent years, salient object detection has become an important and active research topic in both neuroscience and computer vision[2], [3], [4], [5], [6], [7], [8], since it has served as a pre-processing procedure for many vision tasks, such as image compression [9], stylized rendering [10], object recognition [11], visual tracking [12], and image retargeting [13], etc.

According to whether the detection procedure requires human interaction or not, existing methods are divided into two categories: top-down (supervised) and bottom-up (unsupervised) approaches. The first category often describes the saliency by the visual knowledge constructed from the training process, and then use such knowledge for saliency detection on the test images [1], [14], [15]. On the contrary, the second one usually determines the saliency of a pixel based on low-level stimuli-driven features without any prior of the salient region or object [2], [3], [4], [6].

Due to the high computational efficiency and scalability to large scale datasets, mainly previous methods are favor to bottom-up approaches, relying on the assumptions about the properties of objects and backgrounds. The mostly utilized assumption is that the visual saliency can be measured from distinctiveness between center patches/regions and surrounding context. This is called *contrast prior* and is widely used in existing salient object detection models [1], [3], [4], [7], [10]. Besides contrast prior, several recent approaches attempt to exploit image *boundary prior* information [2], [5], [6], [16], [17], [18], i.e., image boundary regions are always backgrounds, to inhibit the influence of background to the true salient regions. Although these background prior based models have achieved promising results, they still suffer from following limitations:

- They simply treat all image boundary as background. However, the salient objects may locate in or slightly touch the image boundary, leading to the failure using such heuristic assumption.
- The background itself may be complex and exhibit variety of visual appearance, e.g., containing small scale and high contrast patterns. Without assistance of high-level knowledge, it is often difficult to exclude such noisy background and may highlight wrong pixels.

This work presents a new framework to address above problems. By investigating the connective property between image regions and boundaries, we propose a reliable background descriptor, named *boundary connectivity*, which is our first contribution. Instead of assuming image boundary is always background [8], [19], [18], our boundary connectivity judges a region belongs to background only when it *heavily* connects to image boundary. Since this descriptor characterizes a geometrical interpretation of image regions with respect to image boundaries, it is more stable with respect to the variations of image content. The main advantage of boundary connectivity lies in the fact that it has similar distributions of scores across different images and are directly comparable.

978-1-5090-2860-3/16/$31.00 ©2016 IEEE

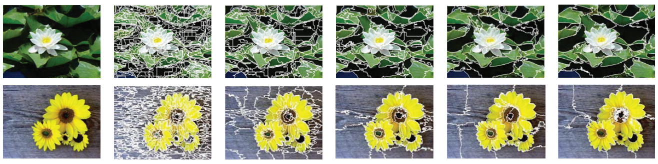


Fig. 1. Two examples of multiple layer over-segmentation. From left to right are the original images and their segmentation results. Different superpixels are separated by white boundaries. (Best viewed in color)

Consequently, it can significantly enhance traditional contrast computation, e.g., color contrast, leading to highlight entire salient object regions and excludes background regions.

It is well known that salient object detection heavily depends on the accuracy of the segmentation. Yet, this is usually explored in one scale over-segmentation framework, always resulting in inaccurate object shape delineation [10], [20]. In the coarse-scale segmentation, some superpixels may contain the salient object, yet ignore the local details of heterogeneous regions. On the other hand, in the fine-scale segmentation, the superpixels tend to have object boundaries with high precision, but neglect the integrity of entire salient object. Our second contribution is a principled framework that estimates pixel saliency using multiple layer over-segmentation. In each segmentation layer, background regions are constrained to have low saliency using our background connectivity measure, while object regions are constrained to take high saliency only using color contrast with respect to background regions. Finally, a smoothness integration is adopted to predict pixel saliency, which ensures that the saliency map is uniform to stress the whole salient objects. Our approach combines contrast feature maps in an intuitive, efficient, and straightforward manner, which is significantly different from complex CRF/MRF-based optimization methods [1], [8], [21].

We evaluated our method on MSRA-1000 dataset [1], and compared with 12 main-stream saliency models [3], [14], [22], [23], [24], [25], [26], [27], [28], [29], [30], [31] as well as with manually produced ground truth annotations. The whole paper is organized as follows: Section II briefly presents our multiple layer over-segmentation using the technique of [32], and the details of our saliency computation is introduced in Section III. Section IV shows the experimental results, and the conclusion remark is given in Section V.

II. Multiple Layer Over-Segmentation

Using over-segmentation technique has been found useful for salient object detection by other researchers [8]. Thus, our first step is to form multiple layer over-segmentation from raw pixel intensities of input image. These segmentations are able to produce a set of superpixels, corresponding to small, homogeneous image regions with different size.

As shown in Fig. 1, an input image $\mathcal{I}$ is first partitioned into a series of super-pixels using mean shift segmentation technique [32] with different parameter settings. The involved parameters are the spatial resolution parameter h_s, the range resolution parameter h_r and the size of smallest segments Min. In practice, we initialize these parameters as $h_s = 10$, $h_r = 2$, and $Min = 100$. Then multiple layers of over-segmentations are produced by incremental increasing these parameters with updated step as 10, 2, and 100, respectively. Ideally, we would like to consider as many as possible of over-segmentation so that we can make full use of region cues. However, this would require a significant computational efforts, and thus only 5 segmentation layers are considered in our method.

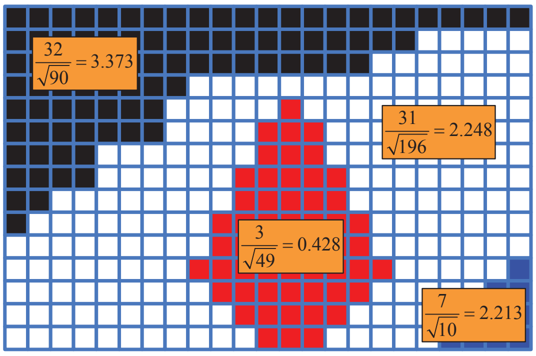

$\frac{32}{\sqrt{90}} = 3.373$
$\frac{31}{\sqrt{196}} = 2.248$
$\frac{3}{\sqrt{49}} = 0.428$
$\frac{7}{\sqrt{10}} = 2.213$

Fig. 2. A toy example of boundary connectivity. The synthetic image consists of four regions, where the boundary connectivity values overlaid correspondingly. (Best viewed in color)

From Fig. 1, although generated superpixels tend to be highly irregular in size and shape, the advantage of using [32] lies in the fact that it can often group large homogeneous regions with similar appearance, while dividing heterogeneous regions into many smaller ones. Alternative over-segmentation approaches include hierarchical segmentation [8], graph-based segmentation [33], and normalized cuts [34]. These methods require much more processing time to produce superpixels or the generated superpixels have imprecise boundaries.

III. Measuring Visual Saliency

In this section, we first introduce the background connectivity descriptor, and then describe our color contrast computation for producing saliency maps.

A. Background Connectivity Measure

Observing nature images, we found that object and background regions usually have different spatial layout, for instance, *object regions tend to be much less connected to image boundaries than background ones.*

Fig. 2 shows a toy example that verifies this observation. The synthetic image contains four regions (white, black, red and blue), where each square represents a pixel. From human perception, the red region is definitely a salient object as it is compact, regular, and only slightly connects image boundary (only three pixels). On the contrary, the white and black regions are clearly backgrounds as they occupy large image area, and significantly touch the image boundary. Finally, only a small amount of the blue region touches the image boundary (6 pixels), yet its size is also too small so that it looks more like a partially cropped object, e.g., small scale and high contrast

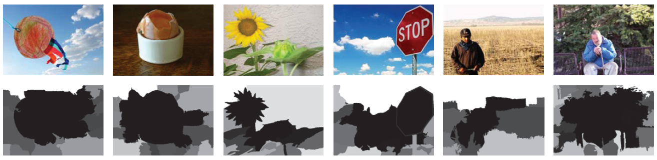


Fig. 3. Illustration of boundary connectivity for some images. White color means high boundary connectivity value, while black color represents low boundary connectivity value. (Best viewed in color)

noisy patterns. As a result, it also should be assigned a low saliency value. According to above observation, we propose a novel measure, named *boundary connectivity*, to quantify how heavily a segmented superpixel r_i is connected to the image boundaries. It is defined as:

$$BC(r_i) = \frac{|\{p(x,y)|p(x,y) \in r_i, p(x,y) \in Bnd\}|}{\sqrt{|A(r_i)|}} \tag{1}$$

where $p(x,y)$ denotes a pixel located with coordinate (x,y), and Bnd is the set of boundary pixels. $|\cdot|$ presents pixel counting operation, and $A(r_i) = |\{p(x,y)|p(x,y) \in r_i\}|$ is the area of superpixel r_i.

In order to achieve scale invariance, we compute the square root of the area of r_i in Eqn. (1), which ensures the stability across image regions with different resolutions. The definition of background connectivity has an intuitive geometrical interpretation: *it is the ratio of a region's boundary perimeter to the square root of its area.* As shown in Fig. 3, the boundary connectivity usually assigns large values for background regions and small values for object regions, whether salient object are in the center of image, or slightly touch, even heavily connected to image boundary.

B. Robust Background Region Abstraction

According to Eqn. (1), boundary connectivity is with similar distributions of values across different images. As a result, it is able to detect the background at a high precision only using a single threshold.

In j^{th} segmentation layer of one input image, the boundary connectivity value b_m^j, corresponding to superpixel r_m^j, is computed based on Eqn. (1). Then we select the average boundary connectivity value across all superpixels as single threshold θ:

$$\theta = \frac{1}{M}\sum_{m=1}^{M} b_m^j \tag{2}$$

where M is the total number of superpixels in j^{th} segmentation layer. Those superpixels, whose boundary connectivity value is higher than θ, are selected as background regions. Using such simple criteria is more robust with respect to the variations of image content, e.g., it naturally handles images with multiple salient objects, or purely background images without objects.

C. Color Contrast Saliency

It often happens that salient object may not perfectly locate in the center of image, while the color of object regions is still quite different with respect to other regions. Instead of computing saliency based on entire image [10], here we calculate the contrast based on background regions.

Due to the human visual perception, we first transfer RGB color space to CIELab color space, and thus superpixel r_i^j in j^{th} layer of over-segmentation is represented by its mean color $\bar{c}(r_i^j)$. Let $\{B_1^j, B_2^j, \cdots, B_N^j\}$ be the selected background regions according to Section III-B, then we calculate the color contrast value for r_i^j as:

$$CS(r_i^j) = \frac{1}{N}\sum_{n=1}^{N}\sum_{*=L,a,b} d_*(r_i^j, B_n^j) \tag{3}$$

where $d_*(r_i^j, B_n^j) = \sqrt{[\bar{c}_*(r_i^j) - \bar{c}_*(B_n^j)]^2}$ is the Euclidean distance quantified the difference in each color channel.

In practice, we found the this measure to be of higher significance and discriminative power. Therefore, we employ an exponential function to enhance color contrast saliency in each layer of segmentation:

$$s(r_i^j) = \exp\{\alpha \cdot CS(r_i^j)\} - 1 \tag{4}$$

where α is a scaling parameter, and set as 6 empirically. From Eqn. (4), if a region has great color difference with respect to background regions, it will get a large value leading to high contrast saliency for that region overall. Note that if r_i^j itself is a background regions, the color contrast saliency will be 0. Finally, saliency value of superpixel r_i^j is assigned to the contained pixel $p(x,y)$, and the corresponding color contrast saliency is presented by $s^j(p)$.

D. Combined Saliency

It is well known that the integration of low level cues from different segmentation layers can yield better results [1], [8], [20]. Instead of using heuristic ways, e.g., weighted summation or multiplication [4], [35], we employ an nonlinear operator to combine color contrast to generate final saliency map.

More preciously, we assume that the color contrast saliency maps in each segmentation layer are independent, and start by normalizing them to the range $[0,1]$ following [4]:

$$s^j(p) = \frac{s^j(p) - s^j_{\min}(p)}{s^j_{\max}(p) - s^j_{\min}(p)} \tag{5}$$

where $s^j_{\max}(p)$ and $s^j_{\min}(p)$ are the maximum and minimum value in j^{th} segmentation layer. Then the final saliency map $s(p)$ is calculated using the minimum value among all segmentation layers:

$$s(p) = \min_j s^j(p) \tag{6}$$

Using this nonlinear operation is quite effective to exclude background, and maintains good object outlines.

IV. Experimental Results

In order to evaluate our proposed method, we carried out experiments on MSRA-1000 benchmark dataset using the Precision-Recall curve (PRC) and F-measure [10], [36]. MSRA-1000 dataset contains 1000 images with resolution of

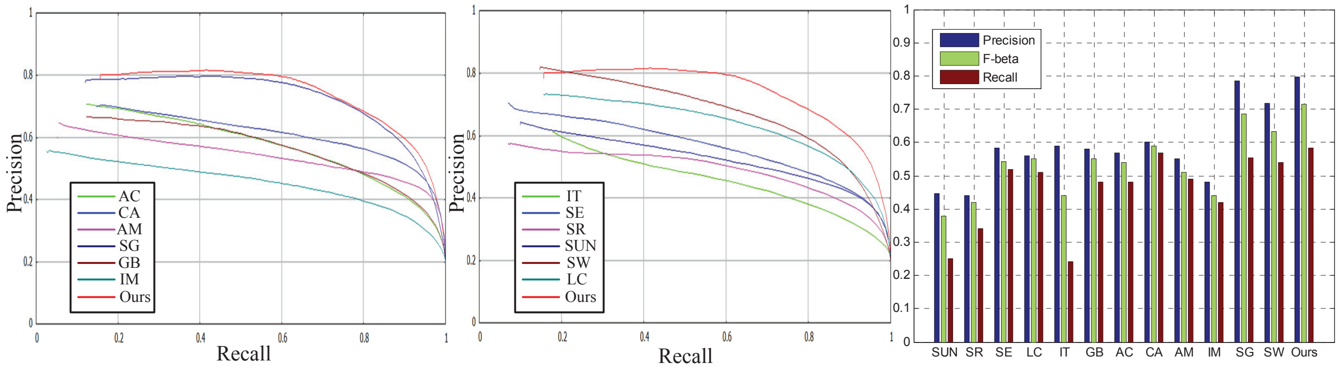
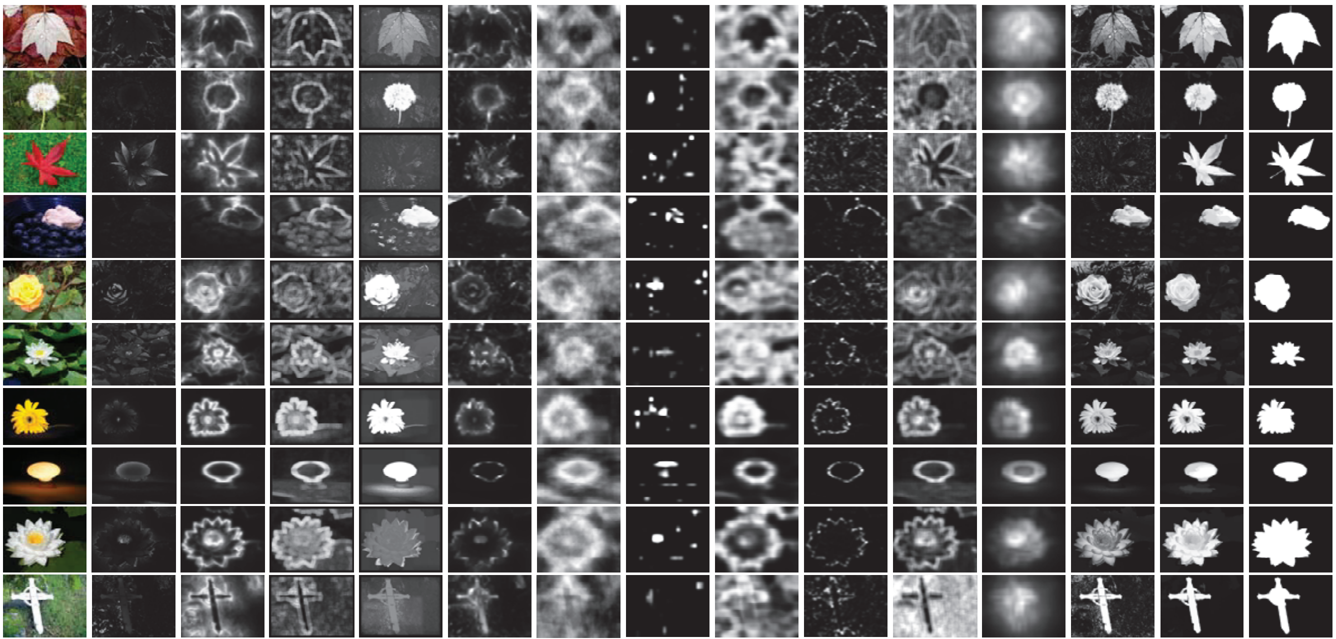


Fig. 4. Quantitative comparison with all state-of-the-art methods. Left and middle: PRC of our method compared with CA [14], AC [26], IT [3], LC [23], SR [22], GB [24], AM [25], SG [27], IM [28], SUN [31], SE [29], and SW [30]. Right: Average precision, recall and F-measure value. (Best viewed in color)



(a) image (b) AC (c) CA (d) AM (e) SG (f) GB (g) IM (h) IT (i) SE (j) SR (k) SUN (l)SW (m) LC (n) Ours (o) GT

Fig. 5. Visual comparison of previous approaches with our method. See the legend of Fig. 4 for the references to all methods. (Best viewed in color)

approximate 400×300 or 300×400 pixels, and provides accurate object-contour-based ground truth.

A. Baselines

We selected 12 state-of-the-art models as baselines for comparison, including spectral residual saliency (SR [22]), spatiotemporal cues (LC [23]), attention measure (IT [3]), graph-based saliency (GB [24]), visual search (AM [25]), saliency segmentation (AC [26]), context-aware saliency (CA [14]), segmenting saliency (SG [27]), saliency estimation (IM [28]), self-resemblance saliency (SE [29]), spatial dissimilarity saliency (SW [30]), and nature statistic saliency (SUN [31]).

B. Evaluation Metrics

In order to quantitatively evaluate the effectiveness of our method, we conducted experiments based on the following widely used criteria. The PRC is used to evaluate the similarity between the predicted saliency maps and the ground truth [10], [20]. Precision corresponds to the percentage of salient pixels correctly assigned, while recall corresponds to the fraction of detected salient pixels in relation to the ground truth number of salient pixels. An alternative criterion to evaluate the overall performance is the F-measure [10], [36], which is utilized to weight harmonic mean measurement of precision and recall. The F-measure is defined as:

$$F_\beta = \frac{(1+\beta^2) \times Precision \times Recall}{\beta^2 \times Precision + Recall} \tag{7}$$

where $\beta^2 = 0.3$ following [10], [36].

C. Overall Results

We implemented our method and the state-of-the-art models using a Dual Core 2.6 GHz machine with 4GB memory.

The precision-recall curve (PRC) is illustrated in Fig. 4. It clearly shows that our method outperforms other approaches. It is interesting to note that the minimum recall value of our methods starts from 0.17, and the corresponding precision is higher than those of the other methods. This probably because the saliency maps computed by our methods contain more pixels with the saliency value 255. We also evaluate our method in terms of F-measure[36] in the right panel of Fig. 4. Our method achieves higher F-measure value (ours = 0.713) than other competitive models (CA = 0.592, SW = 0.634, and SG = 0.685).

Visual comparison with different methods on MSRA-1000 dataset are shown in Fig. 5. Compared with these models, our method is very effective in eliminating the cluttered backgrounds, and uniformly highlighted salient regions with well-defined object shapes, no matter whether salient objects locate in image center, or far away from image center, even on the image boundary.

V. Conclusion and Future Work

In this paper, we present a robust background exclusion model for salient object detection. The key advantages of our method are: (1) background regions can be automatically produced according to the new background measure, called background connectivity; (2) using the contrast against background regions makes our method superior than the methods computing contrast with respect to the entire image; (3) employing multiple layer over-segmentation framework is able to integrate low-level visual cues from different layers for salient object detection, while maintaining well object shape declination. Our method has been tested on MSRA-1000 dataset, and achieves state-of-the-art results. The future work includes incorporating top-down priors to further improve the performance, as well as [21] does.

Acknowledgment

This work was partially supported by NSFC (No. 61401228, 61402122, 61401231, 61403350), Postdoctoral Science Foundation of Jiangsu Province (No. 1501019A), China Postdoctoral Science Foundation (No. 2015M581841), and the 2014 ph.D. Recruitment Program of Guizhou Normal University. Quan Zhou is the corresponding author.

References

[1] T. Liu, Z. Yuan, J. Sun, J. Wang, N. Zheng, X. Tang, and H.Y. Shum, “Learning to detect a salient object,” *TPAMI*, vol. 33, no. 2, pp. 353–367, 2011.

[2] Junling Li, Fang Meng, and Yichun Zhang, “Saliency detection using a background probability model,” in *ICIP*, 2015, pp. 2189–2193.

[3] L. Itti, C. Koch, and E. Niebur, “A model of saliency-based visual attention for rapid scene analysis,” *TPAMI*, vol. 20, no. 11, pp. 1254–1259, 1998.

[4] Z. Quan, C. Ji, R. Shiwei, Z. Yu, C. Jun, and Wenyu L., “On contrast combinations for visual saliency detection,” in *ICIP*, 2013, pp. 2665–2669.

[5] Junwei Han, Dingwen Zhang, Xintao Hu, Lei Guo, Jinchang Ren, and Feng Wu, “Background prior-based salient object detection via deep reconstruction residual,” *TCSVT*, vol. 25, no. 8, pp. 1309–1321, 2015.

[6] Hongyang Li, Huchuan Lu, Zhe Lin, Xiaohui Shen, and Brian Price, “Inner and inter label propagation: salient object detection in the wild,” *TIP*, vol. 24, no. 10, pp. 3176–3186, 2015.

[7] B. Alexe, T. Deselaers, and V. Ferrari, “What is an object?,” in *CVPR*, 2010, pp. 73–80.

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

[9] C. Guo and L. Zhang, “A novel multiresolution spatiotemporal saliency detection model and its applications in image and video compression,” *TIP*, vol. 19, no. 1, pp. 185–198, 2010.

[10] M.M. Cheng, G.X. Zhang, N.J. Mitra, X. Huang, and S.M. Hu, “Global contrast based salient region detection,” in *CVPR*, 2011, pp. 409–416.

[11] Ueli Rutishauser, Dirk Walther, Christof Koch, and Pietro Perona, “Is bottom-up attention useful for object recognition?,” in *CVPR*, 2004, pp. 37–44.

[12] V. Mahadevan and N. Vasconcelos, “Saliency-based discriminant tracking,” in *CVPR*, 2009, pp. 1007–1013.

[13] Xi Li, Yao Li, Chunhua Shen, Anthony Dick, and Anton Van Den Hengel, “Contextual hypergraph modeling for salient object detection,” in *ICCV*, 2013, pp. 3328–3335.

[14] S. Goferman, L. Zelnik-Manor, and A. Tal, “Context-aware saliency detection,” in *CVPR*, 2010, pp. 2376–2383.

[15] T. Judd, K. Ehinger, F. Durand, and A. Torralba, “Learning to predict where humans look,” in *ICCV*, 2009, pp. 2106–2113.

[16] Tong Zhao, Lin Li, Xinghao Ding, Yue Huang, and Delu Zeng, “Saliency detection with spaces of background-based distribution,” *IEEE SPL*, vol. 23, no. 5, pp. 683–687, 2016.

[17] Chintak Sheth and R Venkatesh Babu, “Object saliency using a background prior,” in *ICASSP*, 2016, pp. 1931–1935.

[18] Y. Wei, F. Wen, W. Zhu, and J. Sun, “Geodesic saliency using background priors,” in *ECCV*, 2012.

[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] F. Perazzi, P. Krahenbuhl, Y. Pritch, and A. Hornung, “Saliency filters: Contrast based filtering for salient region detection,” in *CVPR*, 2012, pp. 733–740.

[21] X. Shen and Y. Wu, “A unified approach to salient object detection via low rank matrix recovery,” in *CVPR*, 2012, pp. 853–860.

[22] X. Hou and L. Zhang, “Saliency detection: A spectral residual approach,” in *CVPR*, 2007, pp. 1–8.

[23] Y. Zhai and M. Shah, “Visual attention detection in video sequences using spatiotemporal cues,” in *ACMMM*, 2006, pp. 815–824.

[24] Harel J., Koch C., and Pernoa P., “Graph-based visual saliency,” in *NIPS*, 2006, pp. 545–552.

[25] Neil DB Bruce and John K Tsotsos, “Saliency, attention, and visual search: An information theoretic approach,” *Journal of vision*, vol. 9, no. 3, pp. 1–24, 2009.

[26] R. Achanta, F. Estrada, P. Wils, and S. Susstrunk, “Salient region detection and segmentation,” *Computer Vision Systems*, pp. 66–75, 2008.

[27] Esa Rahtu, Juho Kannala, Mikko Salo, and Janne Heikkilä, “Segmenting salient objects from images and videos,” in *ECCV*, 2010, pp. 366–379.

[28] Naila Murray, Maria Vanrell, Xavier Otazu, and C Alejandro Parraga, “Saliency estimation using a non-parametric low-level vision model,” in *CVPR*, 2011, pp. 433–440.

[29] Hae Jong Seo and Peyman Milanfar, “Static and space-time visual saliency detection by self-resemblance,” *Journal of vision*, vol. 9, no. 12, pp. 15–15, 2009.

[30] L. Duan, C. Wu, J. Miao, L. Qing, and Y. Fu, “Visual saliency detection by spatially weighted dissimilarity,” in *CVPR*, 2011, pp. 473–480.

[31] L. Zhang, M.H. Tong, T.K. Marks, H. Shan, and G.W. Cottrell, “Sun: A bayesian framework for saliency using natural statistics,” *Journal of Vision*, vol. 8, no. 7, pp. 1–20, 2008.

[32] Dorin Comaniciu and Peter Meer, “Mean shift: A robust approach toward feature space analysis,” *TPAMI*, vol. 24, no. 5, pp. 603–619, 2002.

[33] Pedro F Felzenszwalb and Daniel P Huttenlocher, “Efficient graph-based image segmentation,” *IJCV*, vol. 59, no. 2, pp. 167–181, 2004.

[34] J. Shi and J. Malik, “Normalized cuts and image segmentation,” *TPAMI*, vol. 22, no. 8, pp. 888–905, 2000.

[35] Ali Borji, Dicky N Sihite, and Laurent Itti, “Salient object detection: A benchmark,” in *ECCV*, pp. 414–429. 2012.

[36] R. Achanta, S. Hemami, F. Estrada, and S. Susstrunk, “Frequency-tuned salient region detection,” in *CVPR*, 2009, pp. 1597–1604.