SALIENT OBJECT DETECTION FOR RGB-D IMAGE VIA SALIENCY EVOLUTION

Jingfan Guo[1,2], *Tongwei Ren*[1,2,*], *Jia Bei*[1,2]

[1] State Key Laboratory for Novel Software Technology, Nanjing University, China
[2] Software Institute, Nanjing University, China
guojf@smail.nju.edu.cn, rentw@nju.edu.cn, beijia@nju.edu.cn

ABSTRACT

Salient object detection aims to detect the most attractive objects in images, which has been widely used as a fundamental of various multimedia applications. In this paper, we propose a novel salient object detection method for RGB-D images based on evolution strategy. Firstly, we independently generate two saliency maps on color channel and depth channel of a given RGB-D image based on its super-pixels representation. Then, we fuse the two saliency maps with refinement to provide an initial saliency map with high precision. Finally, we utilize cellular automata to iteratively propagate saliency on the initial saliency map and generate the final detection result with complete salient objects. The proposed method is evaluated on two public RGB-D datasets, and the experimental results show that our method outperforms the state-of-the-art methods.

Index Terms— Salient object detection, saliency evolution, RGB-D image

1. INTRODUCTION

Salient object detection aims to detect the most attractive objects for human beings in a given image [1], which has been widely used as an important fundamental of various multimedia applications, including image representation [2, 3], object classification [4, 5], social media mining [6], and video analysis [7, 8].

In order to accurately detect salient objects, a large number of strategies have been proposed, among which evolution strategy is demonstrated to be effective in handling the images with complex structures [9–12]. The basic idea of evolution strategy is to formulate salient object detection as a two-step procedure. Some parts of salient objects are firstly detected as the initial detection result, and then the rest of salient objects are further complemented.

The effectiveness of evolution strategy derives from its consistency to the mechanism of human vision system. When human view an image, some fixation points are firstly generated in seeking salient objects and then the salient objects are completely observed after they are found [13]. Moreover, evolution strategy can effectively reduce the difficulty of realizing automatic salient object detection by computer system, for it decomposes the task of detecting complete salient objects into two steps with different partial objectives, which can be solved separately.

Obviously, two key issues should be considered in salient object detection with evolution strategy. One issue is how to improve the precision of the initial saliency map. Any false detection may be enlarged in the subsequent processing and cause serious error. Current methods usually start from the analysis of boundary bias, i.e., treating the regions dissimilar to image boundary parts as salient objects [14, 15]. However, the effect of this solution is influenced by the layout of salient objects, and it may fail when the background has complex color composition. The other issue is how to generate nearly complete salient objects based on initial saliency maps. Current methods usually formulate it as the problem of saliency propagation. But it is still challenging to improve recall while maintaining high precision.

In this paper, we propose a novel salient object detection method for RGB-D images using saliency evolution strategy. In our method, depth cue is fully explored as well as color cue in each step of saliency evolution. Fig. 1 shows an overview of the proposed method. We firstly over-segment an RGB-D image into super-pixels with the extended SLIC algorithm [16] based on both color cue and depth cue. Then, we estimate two saliency maps based on color cue and depth cue independently, and fuse them with refinement to obtain the initial saliency map with high precision. Finally, we iteratively propagate saliency over the whole image on a graph-based model and generate the final saliency map. With the synergism of color cue and depth cue, the proposed method outperforms the state-of-the-art methods in experiments.

The rest of the paper is organized as follows. In Section 2, we briefly review the existing salient object methods. Then, we present the details of the proposed method in Section 3, and validate its performance on two public datasets in Section 4. Finally, the paper is concluded in Section 5.

This work is supported by National Science Foundation of China (61321491, 61202320), Research Project of Excellent State Key Laboratory (61223003), Research Fund of the State Key Laboratory for Novel Software Technology at Nanjing University (ZZKT2016B09), and Collaborative Innovation Center of Novel Software Technology and Industrialization.

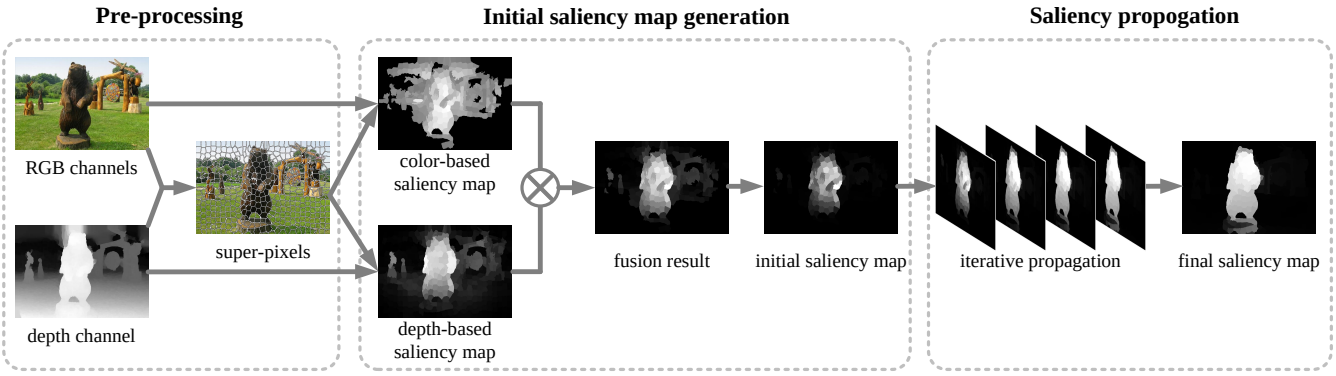


Fig. 1. An overview of the proposed method.

2. RELATED WORK

Salient object detection methods for RGB images are mainly based on global or local color contrast. Achanta *et al.* [17] introduce a straightforward way by simply measure the color distance between each pixel and the average color of the image. Cheng *et al.* [18] propose spatially weighted global color contrast to compute saliency for regions. Recently, graph-based models are widely used in salient object detection for their significant performance. Specifically, Jiang *et al.* [9] employ random walks to formulate the saliency propagation process, in which the duplicated boundary super-pixels are used as the absorbing nodes. Yang *et al.* [10] conduct propagation by manifold ranking. They divide the saliency detection task into two steps: inferring a coarse saliency map by using boundary super-pixels as background seeds, and generating the final result by using foreground queries segmented from the coarse saliency map. Gong *et al.* [11] propose teaching-to-learn and learning-to-teach strategy to improve the propagation quality. They first propagate to obtain initial saliency map by using super-pixels on the boundary and super-pixels out of the convex-hull as background seeds, and further combine it with the convex-hull mask. Then they extract foreground seeds from the initial saliency map and propagate to generate the final saliency map. Qin *et al.* [12] utilize a dynamic evolution model called Cellular Automata for saliency optimization. The initial saliency map is generated by integrating color distinction and spatial distance against boundary super-pixels.

Meanwhile, besides color cue, depth cue is also explored for salient object detection in RGB-D images. Lang *et al.* [19] detect salient objects by integrating global-context depth priors into 2D models. Niu *et al.* [20] propose disparity contrast method, which is extended from [18], for saliency analysis in stereo images. Peng *et al.* [21] detect salient object by combining low-level feature contrast, mid-level region grouping and high-level prior enhancement together. Ju *et al.* [22] present the anisotropic center-surround difference model to measure object-to-surrounding contrast in 3D space.

3. PROPOSED METHOD

3.1. Super-pixel generation based on color and depth

Given an input RGB-D image, we first over-segment it into super-pixels for improving computational efficiency with the intrinsic structure retained. Simple linear iterative clustering (SLIC) algorithm [16] is widely used in super-pixel generation. But the primary SLIC algorithm is only designed for color images, which completely ignores depth cue in super-pixel generation.

In the proposed method, we extend SLIC algorithm by combining depth cue in super-pixel generation. Similar to the primary SLIC algorithm, we cluster the pixels by k-means approach, in which the distance between pixels i and j is calculated as:

$$D_{i,j} = \sqrt{{D_{i,j}^c}^2 + \left(\frac{D_{i,j}^s}{S}\right)^2 m^2}, \quad (1)$$

where $D_{i,j}^c$ and $D_{i,j}^s$ are the color distance and the spatial distance between pixels i and j, respectively; $S = \sqrt{N/k}$ is the grid interval; m weighs the importance between color similarity and spatial proximity, which equals 20 in our implementation as [16].

$D_{i,j}^c$ is calculated as the Euclidean distance in *L*a*b** color space:

$$D_{i,j}^c = \sqrt{(l_i - l_j)^2 + (a_i - a_j)^2 + (b_i - b_j)^2}, \quad (2)$$

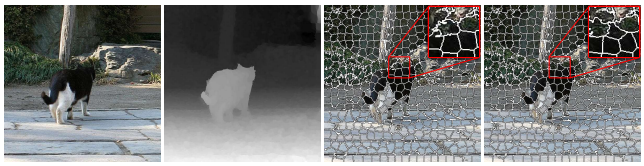
where l_i, a_i and b_i are the color values of pixel i in L, a and b channels, respectively.

And $D_{i,j}^s$ is calculated as the spatial distance in 3D layout:

$$D_{i,j}^s = \sqrt{(x_i - x_j)^2 + (y_i - y_j)^2 + (d_i - d_j)^2}, \quad (3)$$

where x_i and y_i are the horizontal and vertical coordinates of pixel i, respectively; d_i is the depth value of pixel i.

Fig. 2 shows an example of the comparison between the primary SLIC and the extended one for RGB-D images. The



(a) (b) (c) (d)

Fig. 2. Comparison of the primary SLIC and the extended SLIC for RGB-D image. (a) and (b) Color channel and depth channel of RGB-D image. (c) Super-pixels generated by the primary SLIC. (d) Super-pixels generated by the extended SLIC for RGB-D image.

ear of the cat is mixed with the grass by the result generated by the primary SLIC algorithm, but it correctly extracted from background by taking advantage of depth cue.

3.2. Initial saliency map generation

The initial saliency map is generated with two steps. We first generated the saliency maps based on color cue and depth cue, respectively. Then, we fuse these saliency maps and refine it to generate the initial saliency map.

Color-based saliency. We detect the saliency map based on color cue by region contrast and boundary connectivity, in which the saliency value of each super-pixel is calculated as the summation of its color distance to all the other super-pixels weighted by spatial relationships. Inspired by [14, 15, 18], the color-based saliency value of each super-pixel sp_i is calculated as:

$$S_i^c = \sum_{j=1}^{N} \omega_j^b \omega_{i,j}^s \widetilde{D}_{i,j}^c, \tag{4}$$

where N is the number of super-pixels; ω_j^b is the background weight; $\omega_{i,j}^s$ is the spatial weight based on the distance between sp_i and sp_j; $\widetilde{D}_{i,j}^c$ is the color distance between the mean colors of sp_i and sp_j on $L^*a^*b^*$ space, which can be calculated similar to Eq. (2).

As shown in [15], ω_j^b is defined as:

$$\omega_j^b = 1 - exp\Big(-\frac{{B_j}^2}{2{\sigma_b}^2}\Big), \tag{5}$$

where B_j is the boundary connectivity strength of super-pixel sp_j, which denotes the ratio of sp_j's edge length on image boundary to the total length of sp_j's edge; σ_b is a parameter, which equals 1 in our experiments.

To emphasize the color contrast between closer super-pixels, $\omega_{i,j}^s$ is defined as:

$$\omega_{i,j}^s = exp\Big(-\frac{(\widetilde{D}_{i,j}^s)^2}{2{\sigma_s}^2}\Big), \tag{6}$$

where $\widetilde{D}_{i,j}^s$ is the spatial distance between sp_i and sp_j, which is calculated as the Euclidean distance between center locations of sp_i and sp_j without considering depth; σ_s is a parameter, which equals 0.25 in our experiments.

Depth-based saliency. We detect the saliency map based on depth cue by anisotropic center-surround difference [22]. The basic idea is that the salient object usually stands out from its surrounding region and relatively closer to the observer. The depth-based saliency value of each super-pixel sp_i is calculated as:

$$S_i^d = \sum_{\theta} \widetilde{D}^d(\hat{p}_i, P_i(\theta)) \tag{7}$$

where $\hat{p}_i$ is the pixel in the center of sp_i; $P_i(\theta)$ is the set of pixels on the scanning radius emitting from $\hat{p}_i$ with angle θ, in which the lengths of radius depend on the location of $\hat{p}_i$ and it is not more than half length of the image diagonal; $\widetilde{D}^d(\hat{p}_i, P_i(\theta))$ is the Manhattan distance between $\hat{p}_i$ and $P_i(\theta)$, which is defined as:

$$\widetilde{D}^d(\hat{p}_i, P_i(\theta)) = \varphi(d_i, \min_{p_j \in P_i(\theta)} d_j), \tag{8}$$

where d_i is the depth value of pixel $\hat{p}_i$, and $\varphi(,)$ is defined as:

$$\varphi(d_i, d_j) = \begin{cases} d_i - d_j, & d_i > d_j \\ 0, & otherwise \end{cases} \tag{9}$$

Same as [22], we use eight scan radiuses for each super-pixel, i.e., $\theta = \{0, \pi/4, \ldots, 7\pi/4\}$ in our experiments.

Saliency fusion and refinement. Based on Eq.(4) and (7), we obtain the initial saliency value for each super-pixel by fusing the color-based saliency and depth-based saliency, and refine the fusion result by depth-biased weighting:

$$S_i^{ini} = \omega_i^r (S_i^c \otimes S_i^d), \tag{10}$$

where $\otimes$ denotes the fusion strategy of color-based saliency and depth-based saliency, which is element-wise product in our experiments; ω_i^r is the refinement weight, which is defined as:

$$\omega_i^r = \begin{cases} 1, & d_i \geq t \\ \frac{d_i}{t}, & d_i < t \end{cases} \tag{11}$$

where d_i is the depth value of pixel $\hat{p}_i$ in the center of super-pixel sp_i; t is a threshold to decrease the saliency values of the pixels with small depth values, which equals the median value of the depth values of all pixels in each image.

3.3. Saliency propagation

As shown in Fig. 1, the initial saliency map has high precision but fails in providing complete salient objects. It requires further processing to enhance the saliency values of the undetected parts of salient objects in initial saliency maps.

Inspired by [12], we utilize cellular automata [23] to iteratively propagate saliency maps and increase the completeness of salient objects. Each super-pixel is

represented as a cell in the automata, and its saliency value is treated as the state of the cell. In each iteration, the propagation of a super-pixel is simultaneously determined by its current saliency value as well as the saliency values of its neighbors weighted by their feature similarities. Here, we utilize both color cue and depth in feature similarity measurement of two adjacent super-pixels sp_i and sp_j:

$$F_{i,j} = exp\Big(- \frac{\widetilde{D}^c_{i,j} + \widehat{D}^d_{i,j}}{2\sigma_f^2} \Big), \tag{12}$$

where $\widetilde{D}^c_{i,j}$ is the color distance between super-pixels sp_i and sp_j with the same definition as Eq. (4); $\widehat{D}^d_{i,j}$ is the Manhattan distance between the average depth of super-pixel sp_i and sp_j, which is different from $\widetilde{D}^d_{i,j}$ in Eq. (8). While two super-pixels are not adjacent, their feature similarity equals zero. Similar to [12], all the super-pixels around image boundary are regarded as adjacent in our implementation.

Based on feature similarity, the saliency value of each super-pixel is iteratively propagated according to its saliency value and the saliency values of its neighbors:

$$S^*_i = \alpha_i S_i + (1 - \alpha_i) \sum_{j=1}^{N} \omega^F_{i,j} S_j, \tag{13}$$

where S_i and S^*_i are the saliency values of super-pixel sp_i before and after one propagation; N is the number of super-pixels; α_i is a parameter to balance the influence of a super-pixel's current saliency value and the saliency values of its neighbors, which is defined as:

$$\alpha_i = m \left(\max_{j=1,\ldots,N} F_{i,j} \right)^{-1} + n, \tag{14}$$

where m and n are parameters to retain propagation stability, which equals 0.6 and 0.2 in our experiments; $\omega^F_{i,j}$ is used to weight the influences of the neighboring super-pixels:

$$\omega^F_{i,j} = \frac{F_{i,j}}{\sum_{k=1}^{N} F_{i,k}}. \tag{15}$$

Saliency propagation is initialized with the saliency map $\mathbf{S}^{ini}$ generated by Eq. (10), and the number of propagation iteration is set to 20.

4. EXPERIMENTS

4.1. Datasets and evaluation metric

The proposed method is validated on two public RGB-D image datasets for salient object detection, RGBD1000 [21] and NJU2000 [22]. NJU2000 dataset consists of 2,000 pairs of stereo images collected from 3D movies and taken by stereo cameras, in which the depth maps of left views are generated by depth estimation. And RGBD1000 dataset consists of 1,000 RGB-D images captured by Microsoft Kinect. Both these two datasets provide manually labeled ground truths of salient objects for evaluation.

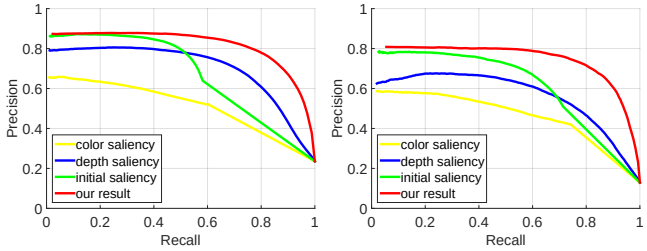


Fig. 3. Component evaluation of the proposed method. (a) NJU2000 dataset. (b) RGBD1000 dataset.

In performance evaluation, precision-recall (PR) curve, weighted F_β-measure F^ω_β ($\beta^2 = 0.3$ to emphasize precision) [24] and mean absolute error (MAE) are used to provide comprehensive evaluation. The average running time with Matlab implementation is about 2.1 seconds per image.

4.2. Experimental results

4.2.1. Component evaluation

We first evaluate the effectiveness of each component in the proposed method, including saliency fusion and saliency propagation. As shown in Fig. 3, the initial saliency map generated by fusion obviously outperforms the saliency maps based on color cue and depth cue in precision, but it suffers a serious problem in recall. Based on the initial saliency map, the following saliency propagation can effectively improve recall performance and generate complete salient objects.

4.2.2. Comparison with the state-of-the-art methods

To illustrate the performance of the proposed method, we further compare it with the state-of-the-art salient object detection methods. Most of the saliency maps used in comparison are generated by the source codes provided by the authors except for DP and SS. Fig. 4 shows some examples of the saliency maps generated by different methods.

We compare the proposed method with six salient object detection for RGB images, including FT [17], RC [18], MC [9], GMR [10], RBD [15] and BSCA [12]. All these methods do not consider depth cue in salient object detection. As shown in Fig. 5 and Table 1, the proposed method provides better PR curves, the highest F^ω_β values and the lowest MAE values on both the two datasets.

We also compare the proposed method with four salient object detection methods for RGB-D images, including DP [19], SS [20], SD [21] and ACSD [22]. As shown in Fig. 6 and Table 1, the proposed method also outperforms these existing methods on all the evaluation metrics, for we fully

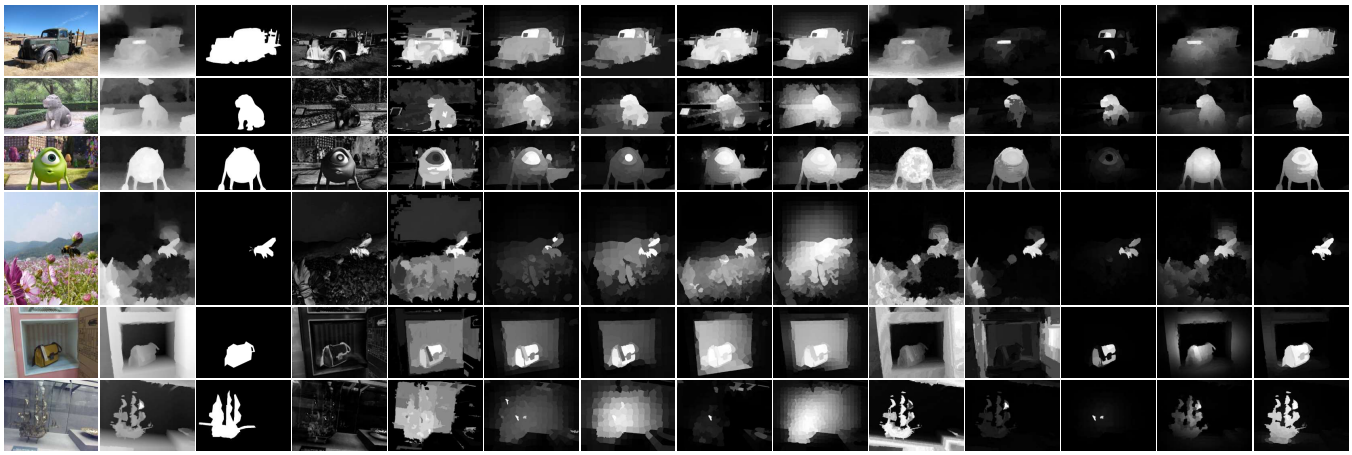


Fig. 4. Examples of comparison with the state-of-the-art methods. (a) and (b) Color channels and depth channels of RGB-D images. (c) Manually labeled ground truths. (d)-(m) Saliency maps generated by FT [17], RC [18], MC [9], GMR [10], RBD [15], BSCA [12], DP [19], SS [20], SD [21] and ACSD [22], respectively. (n) Our results.

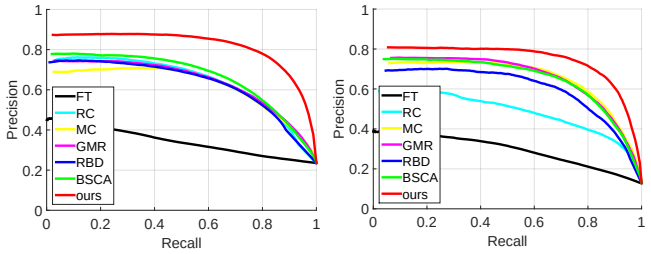


Fig. 5. Comparison with the state-of-the-art salient detection methods for RGB images. (a) NJU2000 dataset. (b) RGBD1000 dataset.

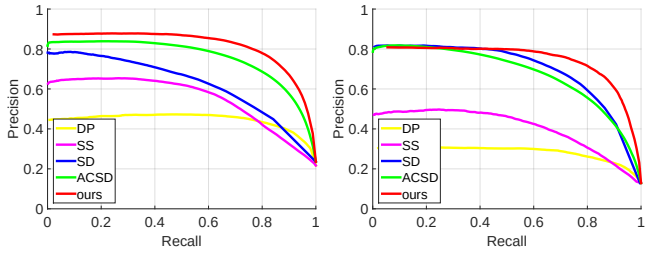


Fig. 6. Comparison with the state-of-the-art salient object detection methods for RGB-D images. (a) NJU2000 dataset. (b) RGBD1000 dataset.

integrate both global depth cue and local depth cue in the proposed framework.

4.3. Discussion

In the experiments, we also find some limitations of the proposed method. As shown in the top row of Fig. 7, if the initial saliency map is completely wrong caused by the invalidation of color cue or depth cue, the proposed method cannot correctly detect the salient objects. Moreover, as shown in the bottom row of Fig. 7, if the salient object consists of completely different parts and the initial saliency map only contains one of them, it is difficult for the proposed method to detect the complete salient object.

5. CONCLUSION

In this paper, we propose a salient object detection for RGB-D images based on evolution strategy. It fully explores the potential of color cue and depth cue in the whole procedure of salient object detection, including super-pixel generation, initial saliency map generation and saliency propagation. And the two-step saliency evolution strategy ensures the high precision and completeness of the detected salient objects. The experimental results show that the proposed method outperforms the state-of-the-art methods for both RGB images and RGB-D images.

6. REFERENCES

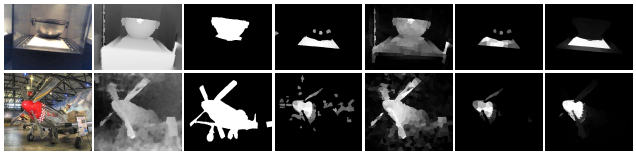
[1] A. Borji, M.-M. Cheng, H. Jiang, and J. Li, "Salient object detection: A benchmark," *TIP*, vol. 24, no. 12, pp. 5706–5722, 2015.

[2] J. Wang, L. Quan, J. Sun, X. Tang, and H.-Y. Shum, "Picture collage," in *CVPR*, 2006, pp. 347–354.

[3] T. Ren, Y. Liu, and G. Wu, "Image retargeting based on

Table 1. Comparison with the state-of-the-art methods for RGB images and RGB-D images on F_β^ω and MAE.

	NJU2000		RGBD1000	
	F_β^ω	MAE	F_β^ω	MAE
FT [17]	0.2009	0.2973	0.1583	0.2175
RC [18]	0.4025	0.2306	0.1689	0.2856
MC [9]	0.3749	0.2278	0.3018	0.1735
GMR [10]	0.4265	0.2174	0.3838	0.1593
RBD [15]	0.4678	0.1939	0.4300	0.1222
BSCA [12]	0.4040	0.2270	0.2886	0.1961
DP [19]	0.3062	0.2896	0.1654	0.3305
SS [20]	0.3507	0.2102	0.2323	0.1750
SD [21]	0.3430	0.2144	0.4647	0.1091
ACSD [22]	0.4318	0.2031	0.3310	0.1452
ours	**0.6009**	**0.1634**	**0.5487**	**0.0974**



(a) (b) (c) (d) (e) (f) (g)

Fig. 7. Examples of failure results. (a) and (b) Color channels and depth channel of RGB-D images. (c) Ground truths. (d) and (e) Saliency maps generated based on color and depth, respectively. (f) Initial saliency maps. (g) Our results.

global energy optimization," in *ICME*, 2009, pp. 406–409.

[4] C. Kanan and G. Cottrell, "Robust classification of objects, faces, and flowers using natural image statistics," in *CVPR*, 2010, pp. 2472–2479.

[5] B.-K. Bao, G. Zhu, J. Shen, and S. Yan, "Robust image analysis with sparse representation on quantized visual features," *TIP*, vol. 22, no. 3, pp. 860–871, 2013.

[6] J. Sang, C. Xu, and J. Liu, "User-aware image tag refinement via ternary semantic analysis," *TMM*, vol. 14, no. 3, pp. 883–895, 2012.

[7] Y.-F. Ma, X.-S. Hua, L. Lu, and H.-J. Zhang, "A generic framework of user attention model and its application in video summarization," *TMM*, vol. 7, no. 5, pp. 907–919, 2005.

[8] S.-H. Zhong, Z. Ma, C. Wilson, Y. Liu, and J. I. Flombaum, "Why do people appear not to extrapolate trajectories during multiple object tracking? a computational investigation," *JOV*, vol. 14, no. 12, pp. 12–12, 2014.

[9] B. Jiang, L. Zhang, H. Lu, C. Yang, and M.-H. Yang, "Saliency detection via absorbing markov chain," in *ICCV*, 2013, pp. 1665–1672.

[10] C. Yang, L. Zhang, H. Lu, X. Ruan, and M.-H. Yang, "Saliency detection via graph-based manifold ranking," in *CVPR*, 2013, pp. 3166–3173.

[11] C. Gong, D. Tao, W. Liu, S. J. Maybank, M. Fang, K. Fu, and J. Yang, "Saliency propagation from simple to difficult," in *CVPR*, 2015, pp. 2531–2539.

[12] Y. Qin, H. Lu, Y. Xu, and H. Wang, "Saliency detection via cellular automata," in *CVPR*, 2015, pp. 110–119.

[13] R. Desimone and J. Duncan, "Neural mechanisms of selective visual attention," *Annu. Rev. Neurosci.*, vol. 18, no. 1, pp. 193–222, 1995.

[14] Y. Wei, F. Wen, W. Zhu, and J. Sun, "Geodesic saliency using background priors," in *ECCV*, 2012, pp. 29–42.

[15] W. Zhu, S. Liang, Y. Wei, and J. Sun, "Saliency optimization from robust background detection," in *CVPR*, 2014, pp. 2814–2821.

[16] R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, and S. Susstrunk, "Slic superpixels compared to state-of-the-art superpixel methods," *TPAMI*, vol. 34, no. 11, pp. 2274–2282, 2012.

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

[18] M.-M. Cheng, N. J. Mitra, X. Huang, Philip H. S. Torr, and S.-M. Hu, "Global contrast based salient region detection," *TPAMI*, vol. 37, no. 3, pp. 569–582, 2015.

[19] C. Lang, T. V. Nguyen, H. Katti, K. Yadati, M. Kankanhalli, and S. Yan, "Depth matters: Influence of depth cues on visual saliency," in *ECCV*, pp. 101–115. 2012.

[20] Y. Niu, Y. Geng, X. Li, and F. Liu, "Leveraging stereopsis for saliency analysis," in *CVPR*, 2012, pp. 454–461.

[21] H. Peng, B. Li, W. Xiong, W. Hu, and R. Ji, "Rgbd salient object detection: a benchmark and algorithms," in *ECCV*, 2014, pp. 92–109.

[22] R. Ju, Y. Liu, T. Ren, L. Ge, and G. Wu, "Depth-aware salient object detection using anisotropic center-surround difference," *SPIC*, vol. 38, pp. 115–126, 2015.

[23] J. Von Neumann, "The general and logical theory of automata," *Cereb. Mech. Behav.*, pp. 1–41, 1951.

[24] R. Margolin, L. Zelnik-Manor, and A. Tal, "How to evaluate foreground maps," in *CVPR*, 2014, pp. 248–255.