Gaze-Based Object Segmentation

Ran Shi, Ngi King Ngan, *Fellow, IEEE*, and Hongliang Li, *Senior Member, IEEE*

***Abstract*—This letter addresses the problem of object segmentation with a gaze map and a group of candidate regions. First, we analyze distribution characteristics of gazes when people look at an image. Then, we summarize different cases of the group of candidate regions. Based on our analysis and summary, we develop three measures to evaluate the likelihood of a candidate region belonging to the target object and a pooling method to create a likelihood map of this object. Finally, the measures and pooling method are integrated with a proposed iterative strategy for generating the segmentation result. Experimental results demonstrate that our method can handle different types of gaze maps and different groups of candidate regions, and the overall performance of our method is better than that of the state-of-the-art method.**

***Index Terms*—Gaze distribution, interactive segmentation, iterative screening.**

I. Introduction

OBJECT segmentation is an important technique in image processing and computer vision. According to the requirement of users' input, it can be classified into interactive segmentation and automatic segmentation. The common interactive way requests users to draw some scribbles in an image [1] with a mouse. These scribbles label parts of the object and background as "seeds" to construct the object and background models. With the development of an eye tracking technique, many portable eye trackers even a web camera can record locations of gazes when people look at an image. It inspires researchers to employ gazes as an alternative interactive way instead of the scribbles. Compared with the scribble which is an explicit input, the gaze as an implicit input is more intuitive and convenient. It makes the interaction more direct and frees our hands. One typical example is that we can perform content-based image cropping by just looking at this image [2]. The gaze-based object segmentation is not equivalent to the salient object segmentation. Its scope is wider since any objects in an image can be segmented by the gaze, not only the salient one. But if we use the predicted gaze generated by the saliency-based method instead of the real gaze, the process becomes automatic saliency-aware object segmentation [3], [4]. In this letter, we mainly focus on the interactive object segmentation based on the real gaze.

Some researchers [2], [5]–[7] proposed their own interactive object segmentation methods based on the gaze information. In [5], Sadeghi *et al.* developed a new graphical user interface controlled by the gazes. It requests the users to look at an object which they want to segment and the background, respectively. This method simply treats the gaze as another form of the scribble. Although it can obtain the exact object and background seeds, the work load of the user is doubled. A better way is to let the user observe the image and then look at the target object. However, this way cannot provide exact information about the background. In [2], an image is presegmented into several superpixels [8]. According to the distribution of the gazes on the superpixels, these superpixels are labeled as "object seeds," "background seeds," and "unknown." This method fixes the number of object and background seeds, which cannot guarantee the quality of the models. In [6], two saliency maps are introduced to assist the estimation of background seeds. However, this method can only be applied to a salient object segmentation task in a simple scene. In [7], Li *et al.* proposed a method combining the object proposal with gaze information. The object proposal [9] is generally used for object detection and recognition tasks. It can provide thousands of candidate regions as potential target objects for detection and recognition. In this method, the gaze information is obtained from either a blurred real gaze map (RGM) or a gaze map predicted by a bottom-up method [10]. This differs from some existing saliency models aim at detecting the salient region by measuring the global contrast [3], utilizing the consistency properties [11], employing the clustering [12], or transferring the saliency from the annotated images [13]. The gaze prediction method simulates human eye movement behaviors to automatically predict the gaze location. By using the gaze map, Li's method avoids the selection of background seeds and converts the selection problem into evaluating the probability of each candidate region being looked at by the viewers. It extracts the shape features of the candidate region and the gaze distribution features within the candidate region. Using these two types of features, a learning-based method is developed, which is achieved by learning a scoring function for each candidate region. One challenging problem for Li's method is that the candidate region itself may not be precise. Additionally, because the gaze information actually reflects how an image is being observed in terms of viewer's perception, there should still be some gazes located on the background. All methods mentioned above do not screen those gazes on the background, so the accuracy of their segmentation results may be reduced.

In this letter, we propose a gaze-based interactive segmentation method using an iterative strategy. Similar to Li's method,

Manuscript received May 31, 2017; revised July 19, 2017; accepted August 5, 2017. Date of publication August 14, 2017; date of current version August 30, 2017. This work was supported in part by a grant from the Research Grants Council of the Hong Kong SAR, China, under Project CUHK 14201115. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Jun Fang. *(Corresponding authors: Ran Shi.)*

R. Shi is with the Department of Electronic Engineering, The Chinese University of Hong Kong, Shatin, Hong Kong (e-mail: rshi@ee.cuhk.edu.hk).

K. N. Ngan is with the Department of Electronic Engineering, The Chinese University of Hong Kong, Shatin, Hong Kong, and also with the School of Electronic Engineering, University of Electronic Science and Technology of China, Chengdu 610051, China (e-mail: knngan@ee.cuhk.edu.hk).

H. Li is with the School of Electronic Engineering, University of Electronic Science and Technology of China, Chengdu 610051, China (e-mail: hlli@uestc.edu.cn).

Color versions of one or more of the figures in this letter are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/LSP.2017.2739200

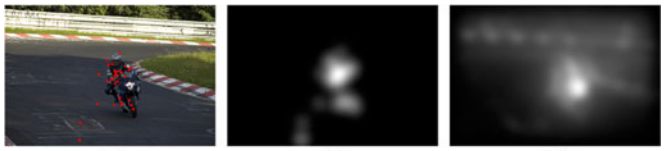

Fig. 1. Examples of different gaze maps. (a) RGM, where the original image is with red real gazes [14]; (b) Blurred RGM of (a); (c) PGM by [10].

Fig. 2. One example of a group of 20 candidate regions for the original image of Fig. 1(a) [9].

we also use one gaze map and a group of candidate regions as inputs. The contributions of our letter are as follows:

1) We analyze the distribution characteristics of the real gazes located on the target object, which correspond to human visual processing mechanisms.
2) We propose a new pooling method to fuse different candidate regions based on their qualities and appearance frequencies.
3) We utilize an iterative strategy to screen the gazes on the background region.

This letter is organized as follows. Section II describes our interactive object segmentation method in detail. Experimental results are presented in Section III. We conclude our letter in Section IV.

II. Proposed Method

Our method is based on a gaze map and a group of K candidate regions for one image I and aims at segmenting the target object which the viewer is looking at. We can obtain a set of gaze locations $\mathbf{g}_i$ and a corresponding set of weights of reliability $\mathbf{w}_i$ for $i \in [1...N]$, where N is the total number of gazes and the weight depends on the pixel value. In order to utilize distribution characteristics of gazes to segment the object, we analyze Pascal-S database [14] and Judd database [15], which collect the real gazes of viewers when they look at an image. One example is shown in Fig. 1(a), where we indicate the gazes with red dots on the image. The distribution characteristics of the gazes are enumerated as below:

1) a considerable number of gazes are located within the target objects and the rest of the gazes are scattered on the background.
2) the minimum bounding box of all gazes located within one target object can roughly cover most part of the object.
3) the gazes located within one target object are mostly concentrated on a certain part of the object.

Especially, characteristics 2) and 3) correspond to two human visual processing mechanisms: the ventral stream and the dorsal stream, which are responsible for the localization and identification of one object, respectively [16].

The group of K candidate regions can be provided by object proposal algorithms [9]. One example is shown in Fig. 2. However, the performance of the same object proposal algorithm is quite different for different images. Generally speaking, there are four cases of a group of candidate regions, though they are not all mutually exclusive:

1) many candidate regions are similar to the target object;
2) a few candidate regions are similar to the target object;
3) a merger of several candidate regions is similar to the target object; and
4) an intersection of several candidate regions is similar to the target object.

By considering the distribution characteristics of the gazes and different cases of the candidate regions, we propose an iterative strategy to screen the gazes at the background as much as possible and exclude poor candidate regions.

In the initial iteration, we first screen gazes that are not located at any candidate regions. Meanwhile, the candidate regions, which do not include any gazes, are excluded too. Supposed there are total T candidate regions remaining. Then, we assign a score to each candidate region according to the distribution of the gazes located on it. This score indicates the probability of viewers looking at this region. For one candidate region cr_t with M gazes, its score $S(cr_t)$ is evaluated in terms of three measures:

$$S(cr_t) = \text{Ratio}(cr_t) \cdot \text{Range}(cr_t) \cdot \text{Con}(cr_t). \tag{1}$$

For $\text{Ratio}(cr_t)$, it is not only used to calculate the portion of gazes occupied by cr_t but considers its relative area:

$$\text{Ratio}(cr_t) = \frac{\sum_{\mathbf{g}_j \in cr_t} \mathbf{w}_j}{W} \cdot \left(1 - \frac{A(cr_t)}{A(I)}\right) \tag{2}$$

where W is the total weight of all remaining gazes. $A(cr_t)$ and $A(I)$ represent the area of cr_t and I, respectively. A high $\text{Ratio}(cr_t)$ indicates that cr_t can attract many viewers' gazes with a proper size.

We use $\text{Range}(cr_t)$ to evaluate the range of the gazes in cr_t:

$$\text{Range}(cr_t) = \frac{A(B_M \cap cr_t)}{A(cr_t)} \cdot \left(1 + \frac{\sum_{\mathbf{g}_j \in b_M} \mathbf{w}_j}{255 \cdot BN}\right) \tag{3}$$

where B_M is the minimum possible bounding box enclosing all M gazes in cr_t, b_M is a set of gazes, which are located on the sides of B_M and BN is the number of such gazes belonging to b_M. Two hundred fifty five as the upper bound of w_i is used to rescale the latter term. Since these gazes of b_M directly decide the size of B_M, the latter term of $\text{Range}(cr_t)$ is used to enhance the reliability of the range according to their weights. A high $\text{Range}(cr_t)$ indicates the gazes cover most part of cr_t. Corresponding to a high response of the ventral stream, this means viewers have generally completed the localization of cr_t.

For $\text{Con}(cr_t)$, it is used to evaluate the concentration degree of these M gazes:

$$\text{Con}(cr_t) = \left(\frac{1}{M} \sum_{\mathbf{g}_j \in cr_t} \frac{1}{1 + D(\mathbf{g}_j, C_M)}\right) \cdot \left(\frac{1}{255 \cdot M} \sum_{\mathbf{g}_j \in cr_t} \mathbf{w}_j\right) \tag{4}$$

where C_M is the weighted center of M gazes and $D(\mathbf{g}_j, C_M)$ is the distance between $\mathbf{g}_j$ and C_M. A high $\text{Con}(cr_t)$ indicates that most gazes with high weights are close to C_M. Corresponding

Fig. 3. Examples of LM and R in the initial iteration and one refined gaze map after the initial iteration. (a) LM when the input is Fig. 1(a); (b) R based on (a); (c) Refined RGM based on (b).

to a high response of the dorsal stream, it means viewers have identified cr_t.

Having obtained the score of each candidate region, we propose a pooling method for creating a likelihood map LM. This map estimates the likelihood of each pixel belonging to the target object. For a certain candidate region cr_t, if pixel p belongs to it, its score $S(cr_t)$ is transferred to p as $S_p(cr_t)$. Otherwise, $S_p(cr_t)$ is set to 0. Since we cannot exactly decide which case a group of candidate regions belongs to without the ground truth, two types of average values: $\text{Avg}_{\text{all}}(p)$ for all $S_p(cr_t)$ s and $\text{Avg}_{nz}(p)$ for nonzero $S_p(cr_t)$ s are considered in the likelihood function $L(p)$, in order to ensure $L(p)$ is properly estimated for different probable cases of the candidate regions. On the one hand, if p appears many times in the candidate regions, $\text{Avg}_{\text{all}}(p)$ should be high. On the other hand, even though p appears a few times, $\text{Avg}_{nz}(p)$ will still be high if its score is high every time. Therefore, $L(p)$ can achieve a balance between p s frequencies of appearance and scores. Thus, $L(p)$ is estimated as below:

$$L(p) = (\widetilde{\text{Avg}_{\text{all}}}(p) \cdot \widetilde{\text{Avg}_{nz}}(p))^{0.5} \tag{5}$$

$$\text{Avg}_{\text{all}}(p) = \frac{1}{T}\sum_{t=1}^{t=T} S_p(cr_t) \tag{6}$$

$$\text{Avg}_{nz}(p) = \frac{1}{\text{NZ}_p}\sum_{t=1}^{t=T} S_p(cr_t) \tag{7}$$

$$S_p(cr_t) = \begin{cases} S(cr_t) & p \in cr_t \\ 0 & o.w. \end{cases} \tag{8}$$

where $\widetilde{\text{Avg}_{\text{all}}}(p)$ and $\widetilde{\text{Avg}_{nz}}(p)$ are the normalization forms and NZ_p is the number of times of nonzero $S_p(cr_t)$s. Then, LM is created by linearly mapping $L(p)$ to the range of $[0, 255]$. Finally, we use OTSU [17] to generate a segmentation result R of the initial iteration. Examples of LM and R are shown in Fig. 3. The lighter pink region in Fig. 3(a) indicates high likelihood values. We can see that our method can generate a good initial segmentation result.

In the subsequent iteration, we only retain the gazes located in R as shown in Fig. 3(c), and then exclude some candidate regions according to these new remaining gazes. Then, we perform the estimation and segmentation the same as we have done in the initial iteration. This iterative processing continues until the segmentation result does not change any more.

III. Experimental Results

In order to assess the performance of our proposed method, we tested our method on Pascal-S [14] and IS database [18] consisting of 510 and 235 test images, respectively. For each segmentation result, we adopt the F_α measure [19] and IoU score to evaluate its accuracy compared with its corresponding ground truth. F_α measure is the weighted harmonic mean of precision and recall where the weight α is set to 0.3 as in [14] and IoU score is a ratio of the area of the intersection region over that of union region between the segmentation result and its ground truth. The average F_α measure and IoU score indicate the overall performance of our method on the database. We compared our method with Li's method. For fair comparisons, the inputs of our method are the same as those of Li's method [7]. For the group of candidate regions, it is provided by MCG [9] and K is set to 20 [14]. For the gaze map, there are two types of gaze maps adopted by Li's method: the first type is a RGM [14] blurred by a Gaussian kernel with $\sigma = 0.03$ of the image width (BRGM) [7], as shown in Fig. 1(b); the second type is a predicted gaze map (PGM) by GBVS [10], as shown in Fig. 1(c). Each type has its own corresponding trained model. If RGM and BRGM are used as the inputs, both methods are the gaze-based interactive segmentation method. However, if PGM is used as the input, both methods become gaze-based automatic salient object segmentation method. Therefore, we also compare our method with two other salient object segmentation methods PCAS [20] and SF [21]. In order to make the PGM and BRGM more similar to the RGM, we threshold them by their means and rescale them into [0, 255] as the weight of the gaze. Moreover, we directly use the RGM as the input, which is not considered in Li's experiments while the weights of all gazes are set to 255. Therefore, we only evaluate our own performance when the input gaze map is RGM. The performances of different methods on two databases and the iteration times of our method are shown in Table I.

Comparing the interactive segmentation methods, the overall performance of our method with BRGM outperforms that of Li's method in terms of the above two measures. However, compared with Li's method, our method achieves higher F_α measure and lower IoU score on Pascal-S database. The reason is that our method is designed to discard the unreliable gazes according to the iterative segmentation result. However, when the target object is very large, the distribution of gazes becomes much more scattered. Our method may mistakenly treat some object gazes as background gazes and discard them so that the final segmentation result is incomplete. It can also be illustrated by the precision-recall curves of the likelihood maps on Pascal-S database, as shown in Fig. 4. We can see that our methods tend to generate results with higher precision and lower recall against Li's method. Since F_α measure encourages high precision while IoU score encourages high recall, our method achieves higher F_α measure and lower IoU score. For IS database, it includes many images with small target objects. So, the number of gazes in the background increases and the poor candidate regions may appear more often. Therefore, our performance on IS database is worse than that on Pascal-S database. However, our performance is still better than that of Li's method. It demonstrates better robustness of the proposed distribution characteristics of the gazes and the pooling method, and effectiveness of our iterative strategy. If we use BRGM as the input, the performance is even better. The main reason is that the blurring operation decreases the weights of the scattered gazes so that the potential risk of these gazes located at the background is reduced. A comparison of the automatic segmentation methods also shows that the performances of our method are competitive. It demonstrates that our method can be treated

TABLE I
F_α MEASURES AND IOU SCORES OF DIFFERENT SEGMENTATION METHODS ON PASCAL-S AND IS DATABASES AND THE ITERATION TIMES OF OUR METHODS

Method	F_α measure								
	Pascal-S			IS			Average		
PCAS	**0.596**			**0.612**			**0.601**		
SF	**0.510**			**0.494**			**0.505**		
	PGM	RGM	BRGM	PGM	RGM	BRGM	PGM	RGM	BRGM
Li	**0.683**	\	0.698	**0.485**	\	0.531	**0.621**	\	0.645
OUR	**0.686**	0.685	0.711	**0.591**	0.536	0.681	**0.656**	0.638	0.702
Method	IoU score								
	Pascal-S			IS			Average		
PCAS	**0.407**			**0.382**			**0.399**		
SF	**0.365**			**0.298**			**0.344**		
	PGM	RGM	BRGM	PGM	RGM	BRGM	PGM	RGM	BRGM
Li	**0.470**	\	0.518	**0.276**	\	0.308	**0.409**	\	0.452
OUR	**0.494**	0.490	0.487	**0.388**	0.346	0.426	**0.461**	0.445	0.468
Method	Iteration times								
	Pascal-S			IS			Average		
Our	PGM	RGM	BRGM	PGM	RGM	BRGM	PGM	RGM	BRGM
	2.61	2.44	2.52	**2.57**	2.57	2.50	**2.60**	2.48	2.51

The automatic segmentation results are shown in bold.

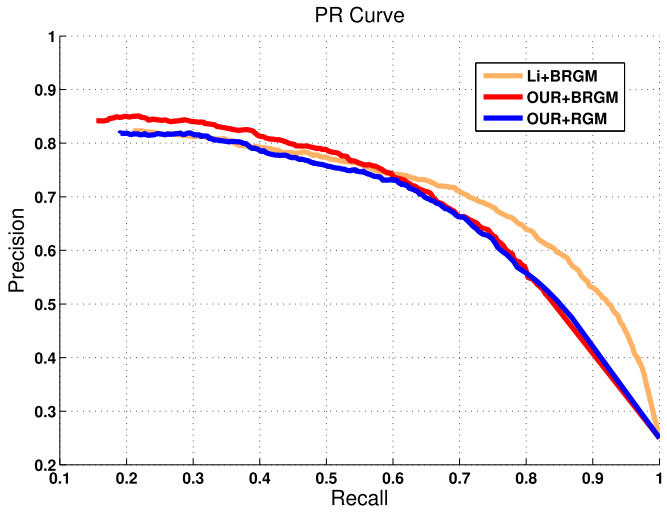

Fig. 4. Precision-recall curves of different methods on Pascal-S database.

TABLE II
ABLATION STUDY OF THREE PROPOSED MEASURES

	Ratio	Ratio+Range	Ratio+Range+Con
F_α measure	0.622	0.660	0.685

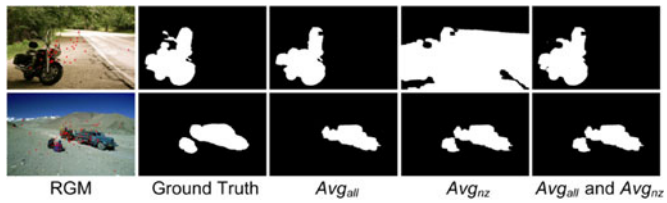

Fig. 5. Two groups of segmentation examples generated using different types of average values.

as an alternative salient object segmentation method using the PGM. Furthermore, our method can quickly converge after two or three iterations for all three types of inputs on these two databases.

In order to further analyze our method, we perform an ablation study on the three proposed measures. We used RGM as the input and tested them on Pascal-S database. The results are shown in Table II. We can see that the Ratio measure can be treated as a base and the other two measures play significant roles in improving the performance. Moreover, two groups of segmentation examples are shown in Fig. 5. They illustrate that the combination of Avg_{all} and Avg_{nz} can balance their effects to adapt to different images.

IV. CONCLUSION

The gaze is an intuitive input for object segmentation. In order to avoid the selection of the background seeds, our method utilizes the gaze information and the object proposal to compose the basic framework. In our method, the iterative strategy is adopted to select more reliable the gazes and candidate regions. Three measures based on the gaze information, i.e., ratio, range, and concentration, are designed to evaluate the probability of one candidate region belonging to the target object. We also develop a new pooling method for handling different probable cases of the candidate regions. According to the experimental results, the most suitable type of gaze map is the blurred RGM. Since it is more accurate than the PGM and has higher tolerance for scattered gazes compared with the RGM, our method can achieve the best performance with this gaze map.

REFERENCES

[1] Y. Y. Boykov and M.-P. Jolly, "Interactive graph cuts for optimal boundary & region segmentation of objects in nd images," in *Proc. IEEE Int. Conf. Comput. Vision*, 2001, vol. 1, pp. 105–112.

[2] A. Santella, M. Agrawala, D. DeCarlo, D. Salesin, and M. Cohen, "Gaze-based interaction for semi-automatic photo cropping," in *Proc. SIGCHI Conf. Human Factors Comput. Syst.*, 2006, pp. 771–780.

[3] M.-M. Cheng, N. J. Mitra, X. Huang, P. H. S. Torr, and S.-M. Hu, "Global contrast based salient region detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 3, pp. 569–582, Mar. 2015.

[4] W. Wang, J. Shen, and F. Porikli, "Saliency-aware geodesic video object segmentation," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, 2015, pp. 3395–3402.

[5] M. Sadeghi, G. Tien, G. Hamarneh, and M. S. Atkins, "Hands-free interactive image segmentation using eyegaze," in *Proc. SPIE Med. Imag.*, 2009, pp. 72601H–72601H.

[6] X. Tian and C. Jung, "Point-cut: Fixation point-based image segmentation using random walk model," in *Proc. IEEE Int. Conf. Image Process.*, 2015, pp. 2125–2129.

[7] Y. Li, X. Hou, C. Koch, J. M. Rehg, and A. L. Yuille, "The secrets of salient object segmentation," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, 2014, pp. 280–287.

[8] C. M. Christoudias, B. Georgescu, and P. Meer, "Synergism in low level vision," in *Proc. IEEE Int. Conf. Pattern Recognit.*, 2002, vol. 4, pp. 150–155.

[9] P. Arbeláez, J. Pont-Tuset, J. T. Barron, F. Marques, and J. Malik, "Multiscale combinatorial grouping," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, 2014, pp. 328–335.

[10] J. Harel, C. Koch, and P. Perona, "Graph-based visual saliency," in *Proc. 19th Int. Conf. Neural Inf. Process. Syst.*, 2006, vol. 1, pp. 545–552.

[11] X. Cao, Z. Tao, B. Zhang, H. Fu, and W. Feng, "Self-adaptively weighted co-saliency detection via rank constraint," *IEEE Trans. Image Process.*, vol. 23, no. 9, pp. 4175–4186, Sep. 2014.

[12] H. Fu, X. Cao, and Z. Tu, "Cluster-based co-saliency detection," *IEEE Trans. Image Process.*, vol. 22, no. 10, pp. 3766–3778, Oct. 2013.

[13] W. Wang, J. Shen, L. Shao, and F. Porikli, "Correspondence driven saliency transfer," *IEEE Trans. Image Process.*, vol. 25, no. 11, pp. 5025–5034, Nov. 2016.

[14] Y. Li, X. Hou, C. Koch, J. M. Rehg, and A. L. Yuille, "Pascal-s salient object dataset," 2014. [Online]. Available: http://cbi.gatech.edu/salobj/

[15] T. Judd, K. Ehinger, F. Durand, and A. Torralba, "Learning to predict where humans look," in *Proc. IEEE Int. Conf. Comput. Vision*, 2009, pp. 2106–2113.

[16] B. B. Velichkovsky, M. A. Rumyantsev, and M. A. Morozov, "New solution to the midas touch problem: Identification of visual commands via extraction of focal fixations," *Procedia Comput. Sci.*, vol. 39, pp. 75–82, 2014.

[17] N. Otsu, "A threshold selection method from gray-level histograms," *Automatica*, vol. 11, no. 285–296, pp. 23–27, 1975.

[18] J. Li, M. D. Levine, X. An, X. Xu, and H. He, "Visual saliency based on scale-space analysis in the frequency domain," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 4, pp. 996–1010, Apr. 2013.

[19] A. Borji, "What is a salient object? A dataset and a baseline model for salient object detection," *IEEE Trans. Image Process.*, vol. 24, no. 2, pp. 742–756, Feb. 2015.

[20] R. Margolin, A. Tal, and L. Zelnik-Manor, "What makes a patch distinct?" in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, 2013, pp. 1139–1146.

[21] F. Perazzi, P. Krähenbühl, Y. Pritch, and A. Hornung, "Saliency filters: Contrast based filtering for salient region detection," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, 2012, pp. 733–740.