Ranking Saliency

Lihe Zhang, Chuan Yang, Huchuan Lu, *Senior Member, IEEE*, Xiang Ruan, and Ming-Hsuan Yang, *Senior Member, IEEE*

Abstract—Most existing bottom-up algorithms measure the foreground saliency of a pixel or region based on its contrast within a local context or the entire image, whereas a few methods focus on segmenting out background regions and thereby salient objects. Instead of only considering the contrast between salient objects and their surrounding regions, we consider both foreground and background cues in this work. We rank the similarity of image elements with foreground or background cues via graph-based manifold ranking. The saliency of image elements is defined based on their relevances to the given seeds or queries. We represent an image as a multi-scale graph with fine superpixels and coarse regions as nodes. These nodes are ranked based on the similarity to background and foreground queries using affinity matrices. Saliency detection is carried out in a cascade scheme to extract background regions and foreground salient objects efficiently. Experimental results demonstrate the proposed method performs well against the state-of-the-art methods in terms of accuracy and speed. We also propose a new benchmark dataset containing 5,168 images for large-scale performance evaluation of saliency detection methods.

Index Terms—Saliency detection, manifold ranking, multi-scale graph

1 Introduction

The human visual system can spot salient objects in a cluttered visual scene with selective visual attention. Such capability is also of great importance to computational visual systems to tackle the information overload problem. Saliency detection aims to simulate selective visual attention of humans for identifying the most important and informative parts of a scene. It has been widely applied to numerous vision problems including image segmentation [1], object recognition [2], image compression [3], content based image retrieval [4], to name a few.

Saliency detection in general can be categorized by bottom-up or top-down models. Bottom-up methods [1], [5], [6], [7], [8], [9], [10], [11], [12], [13], [14], [15], [16], [17], [18], [19], [20], [21] are fast, data-driven and pre-attentive. These methods model saliency by visual distinctness or rarity using low-level image information such as contrast, color, texture and boundary. Top-down models [22], [23] analyze task-driven visual attention, which often entail supervised learning with class labels from a large set of training examples. We note that saliency models have been developed to predict visual attention with eye fixation in human vision [5], [6], [7], [8], [9], [10], [24], and salient object detection in computer vision [13], [14], [16], [18], [19], [20], [23]. In this work, we propose a bottom-up model to detect salient objects in images.

- *L. Zhang and H. Lu are with the School of Information and Communication Engineering, Dalian University of Technology, Dalian, Liaoning 116024, China. E-mail: {zhanglihe, lhchuan}@dlut.edu.cn.*
- *C. Yang is with Alibaba Group, Beijing, Hangzhou 311121, China. E-mail: ycscience86@gmail.com.*
- *X. Ruan is with IWAKI Co., Ltd., Chiyoda-ku, Tokyo 101-8558, Japan. E-mail: gen@omm.ncl.omron.co.jp.*
- *M.-H. Yang is with School of Engineering, University of California, Merced, CA 95344. E-mail: mhyang@ucmerced.edu.*

Manuscript received 6 Mar. 2015; revised 23 July 2016; accepted 7 Sept. 2016. Date of publication 13 Sept. 2016; date of current version 11 Aug. 2017. Recommended for acceptance by L. Zelnik-Manor.
For information on obtaining reprints of this article, please send e-mail to: reprints@ieee.org, and reference the Digital Object Identifier below.
Digital Object Identifier no. 10.1109/TPAMI.2016.2609426

Existing algorithms mainly exploit visual cues of foreground objects for saliency detection, e.g., color [16], [25], distinct patterns [26], spatial compactness [17], [27], smooth appearance [28], focusness [29] and objectness measure [19], [30]. Recently, methods that use the background cues to detect salient objects have been developed [31], [32], [33]. The main observation is that a pair of regions from the background are more similar than a region from one foreground object and another region from the background. For saliency detection, an image is represented by a set of nodes to be labeled, and the labeling task (either salient object or background) is formulated as an energy minimization problem [31], [33] or a random walk problem [32] based on this principle.

We observe that background regions are usually similar to one of four image boundaries in terms of local or global appearance. In contrast, foreground regions are similar in terms of coherent and consistent visual appearance. In this work, we exploit these principles to compute pixel saliency based on ranking of image elements. To detect salient objects at different scale, we construct a multi-scale graph to simultaneously capture local and global structure information of an image, where graph nodes are superpixels (at the bottom level) or coarse regions (at the top level). With a multi-scale graph, the proposed model is able to incorporate long-range spatial connections between pairwise pixels, such that pixel-level saliency of detected objects is labeled more uniformly.

Specifically, we model saliency detection as a manifold ranking problem and propose a cascade scheme for graph labeling. Fig. 1 shows the main steps of the proposed algorithm. In the first stage, we exploit the boundary prior [34], [35] by using the nodes on each image side as labeled background queries. From each labeled result, we compute the saliency of nodes based on their relevances (i.e., ranking) to those queries as background labels. The four labeled maps are then integrated to generate a saliency map, which assigns each node a probabilistic measure of belonging to the queried background class. In the second stage, we take

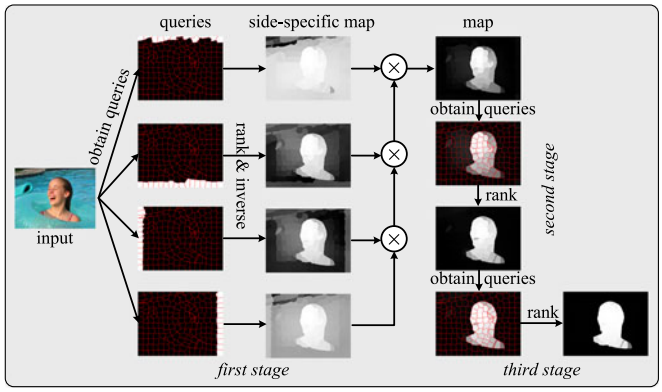


Fig. 1. Main steps of the proposed algorithm. We rank the similarity of image elements with foreground or background cues via graph-based manifold ranking. In this first stage, the boundary priors are used as query where the relevances of regions to each image side are ranked, and then merged to generate a saliency map. In the second stage, all regions with the soft labels are used as queries for the foreground salient objects. The saliency of each node is computed based on its relevance to foreground queries. In the third stage, the saliency probabilistic measures computed in the second stage are used as mid-level features to construct a new image manifold, based on which the regions are ranked to generate the final saliency map.

all nodes with the soft labels as salient queries. The saliency of each node is computed based on its relevance to foreground queries. The ranking procedures are carried out on the image manifold constructed from low-level image features. In the third stage, the saliency probabilistic measures computed in the second stage are used as mid-level features to construct a new image manifold, based on which the nodes are ranked for the final saliency map.

Graph-based manifold ranking is essential for learning an optimal affinity matrix [36] which fully captures intrinsic structure information and describes local and global grouping cues for the graph labeling task. We integrate multi-scale grouping cues across low-level and mid-level image features with graph-based manifold ranking to uniformly highlight the whole salient regions and generate well-defined boundaries of salient objects. The main contributions of this work are summarized as follows:

- We propose a saliency detection algorithm based on graph-based manifold ranking, and analyze the learned affinities of graph nodes, which describe the relevance of foreground and background regions in an image.
- We construct two image manifolds using low-level and mid-level features, and present a cascade saliency detection scheme on a multi-layer representation of an image.
- We develop a new benchmark dataset containing 5,168 images for performance evaluation of saliency detection methods. Experimental results on five benchmark data sets show that the proposed algorithm performs efficiently and favorably against the state-of-the-art saliency detection methods.

2 Related Work

Numerous saliency models, which exploit various cues for visual saliency and saliency detection, have been proposed. A thorough review on this topic can be found [37] and we only discuss the most related methods in this section.

Based on cognitive studies of visual search [38], Itti et al. [5] develop a saliency model based on local contrast between pixels of the foreground object and the background by computing the center-surround feature differences across multiple scales. Numerous methods have since been proposed based this principle with different measures and features. Ma and Zhang [7] propose a contrast-based saliency detection method where pixels are merged by fuzzy clustering. Klein and Frintrop [11] compute the center-surround contrast for saliency detection based on the Kullback-Leibler divergence of feature distributions. Achanta et al. [13] measure the saliency likelihood of each pixel based on its color contrast to the entire image. In [16], Cheng et al. exploit global color contrast of pixels and incorporate it with spatial compactness to extract saliency regions. A supervised method proposed by Liu et al. [23] combines a set of mid-level features which describe local and global saliency of pixels using a conditional random field model. Goferman et al. [1] propose a context-aware saliency method based on four principles of human visual attention, including local low-level clues, global considerations, visual organization rules and high-level factors. Different visual cues have also been integrated in unified energy minimization framework for saliency detection [18], Gaussian filters [17], or within the Bayesian inference framework [39]. Instead of using uniform prior distributions [39], Xie et al. [40] compute a probabilistic prior map by superpixel clusters as well as region locations, and integrate it in a Bayesian inference framework for saliency detection. In addition, Fourier spectrum [8], [41] and sparse representations [42], [43], [44] have applied to predict locations of salient objects.

A graph naturally represents relationships between image elements with affinity measures and describes the underlying geometric structure. Numerous saliency models have been proposed based on graphs with different features and affinity measures. Harel et al. [6] formulate the saliency detection problem as a random walk on a graph, in which salient regions are identified based on the frequency of node visits at equilibrium. Similarly, Wang et al. [10] present a visual saliency measure via entropy rate which denotes the average information transmitted from a node to the others during a random walk. Lu et al. [20] develop a hierarchical graph model and utilize context information to compute node affinities from which a graph is bi-partitioned for salient object detection. In [19], Chang et al. use a graphical model to integrate objectness [45] and saliency in which an iterative energy optimization process concurrently improves respective estimations through the interaction terms. On the other hand, Mai et al. [46] develop a method that combines visual cues in a conditional random field for saliency detection. More recently, Jiang and Davis [47] model saliency detection on a graph and detect salient objects by maximizing the total similarities between the hypothesized salient region centers and the contained elements represented by pixels as well as superpixels.

Gopalakrishnan et al. [12] formulate the object detection problem as a binary segmentation or labeling task on a graph. The seeds for the most salient object and background

are identified by the behavior of random walks on a complete graph and a k-regular graph. A semi-supervised graph labeling method [48] is used to infer the binary labels of the unlabeled nodes. Different from [12], the proposed saliency algorithm with manifold ranking requires only seeds from one class, which are initialized with either the boundary prior or foreground cues. The boundary prior is based on the recent findings of human fixations on images [49], which shows that humans tend to gaze at the centers of images. These priors have also been used in image segmentation and related problems [31], [34], [35] with demonstrated success. In contrast, the semi-supervised method [12] requires seeds from the salient as well as background regions, and generates a binary segmentation. Furthermore, it is difficult to determine the number and locations of salient nodes as they are generated by random walks, especially for scenes with multiple salient objects. This is a known problem with graph labeling where the results are sensitive to the selected seeds. In this work, all the background and foreground seeds can be easily generated via background priors and ranking background queries (or seeds).

3 Graph-Based Manifold Ranking

The graph-based ranking problem is described as follows. Given a node labeled as a query, the remaining nodes are ranked based on their relevances to the given query. The goal is to learn a ranking function, which defines the relevance between unlabeled nodes and the query node. In this work, the query is either a labeled node (e.g., region) from the background or foreground salient objects.

3.1 Manifold Ranking

In [50], a ranking method that exploits the intrinsic manifold structure of data points (e.g., images containing handwritten digits) for graph labeling is proposed. Given a dataset $X = \{x_1, \ldots, x_l, x_{l+1}, \ldots, x_n\} \in \mathbb{R}^{m \times n}$, some data points are labeled as queries and the rest need to be ranked according to their relevances to the queries. Let $\mathbf{f}$: $X \rightarrow \mathbb{R}^n$ denote a ranking function which assigns a value f_i to each point x_i, and $\mathbf{f}$ can be viewed as a vector $\mathbf{f} = [f_1, \ldots, f_n]^\top$. Let $\mathbf{y} = [y_1, y_2, \ldots, y_n]^\top$ denote an indication vector, in which $y_i = 1$ if x_i is a query, and $y_i = 0$ otherwise. We define a graph $G = (V, E)$ on the dataset, where the nodes V are the dataset X and the edges E are weighted by an affinity matrix $\mathbf{W} = [w_{ij}]_{n \times n}$. Given G, the degree matrix is $\mathbf{D} = \mathrm{diag}\{d_{11}, \ldots, d_{nn}\}$, where $d_{ii} = \sum_j w_{ij}$. Similar to the PageRank and spectral clustering algorithms [51], [52], [53], the optimal ranking of queries is computed by solving the following optimization problem:

$$\mathbf{f}^* = \arg\min_f \frac{1}{2}\left[\sum_{i,j=1}^{n} w_{ij}\left(\frac{f_i}{\sqrt{d_{ii}}} - \frac{f_j}{\sqrt{d_{jj}}}\right)^2 + \mu \sum_{i=1}^{n}(f_i - y_i)^2\right], \tag{1}$$

where the parameter μ controls the balance of the smoothness constraint (first term) and the fitting constraint (second term). That is, a good ranking function should not assign values such that those of nearby points (smoothness constraint) should be similar and those from the initial points should be similar to the initial assignments (fitting constraint). The minimum solution is computed by setting the derivative of the above function to be zero. The resulting ranking function can be written as

$$\mathbf{f}^* = (\mathbf{I} - \alpha \mathbf{S})^{-1}\mathbf{y}, \tag{2}$$

where $\mathbf{I}$ is an identity matrix, $\alpha = 1/(1+\mu)$ and $\mathbf{S}$ is the normalized Laplacian matrix, $\mathbf{S} = \mathbf{D}^{-1/2}\mathbf{W}\mathbf{D}^{-1/2}$ [51], [52], [53].

The ranking algorithm [50] is derived from the work on semi-supervised learning for classification [54]. Essentially, manifold ranking can be viewed as an one-class classification problem [55], where only positive examples or negative examples are required. We obtain a ranking function using the unnormalized Laplacian matrix [51] in (2):

$$\mathbf{f}^* = (\mathbf{D} - \alpha \mathbf{W})^{-1}\mathbf{y}. \tag{3}$$

Our empirical results show that salient objects can be better detected using the unnormalized Laplaican matrix and we adopt (3) in the following experiments.

3.2 Saliency Measure

Given an input image represented by a graph and queries, the saliency of each node is defined by its ranking score computed by (3) which is rewritten by

$$\mathbf{A} = (\mathbf{D} - \alpha \mathbf{W})^{-1}, \quad \mathbf{f}^* = \mathbf{A}\mathbf{y}, \tag{4}$$

for ease of presentation. The affinity matrix $\mathbf{A}$ can be considered as a learned optimal affinity matrix. The ranking score $\mathbf{f}^*(i)$ of the ith node is the inner product of the ith row of $\mathbf{A}$ and $\mathbf{y}$. As $\mathbf{y}$ is a binary indicator vector, $\mathbf{f}^*(i)$ can also be viewed as the sum of the relevances of the ith node to all the queries.

We note for some images there exist small regions with high contrast in the local surroundings. The nodes from such high contrast regions often have weak correlation with the other nodes of other parts of the same image, but strong self-correlation in the learned affinity $\mathbf{A}$. If such a node is selected as one query, its ranking value in $\mathbf{f}^*$ will contain the relevance of this query to itself, which is meaninglessly large and adversely weakens the contributions of the other queries to the ranking score (See Fig. 3). To address this issue, we set the diagonal elements of an affinity matrix $\mathbf{A}$ to 0 when computing ranking scores using (3). That is, the saliency value of each query is defined by its ranking score computed by the other queries. We note that this step has significant effects on the final results (See Fig. 10d). Finally, we denote that the saliency of nodes using the normalized ranking score $\bar{\mathbf{f}}^*$ when foreground queries are given, and $\mathbf{1} - \bar{\mathbf{f}}^*$ when background queries are presented.

4 Constructing Multi-Scale Graphs

As salient objects are likely to appear at different scales consisting of numerous perceptually heterogeneous regions in a scene, we consider multiple quantizations of the image space with multi-scale graphs. The lower layers of this hierarchical graph describe more detailed image structures, while the higher layers encode holistic visual information. The proposed multi-scale graph labeling allows visual context to be

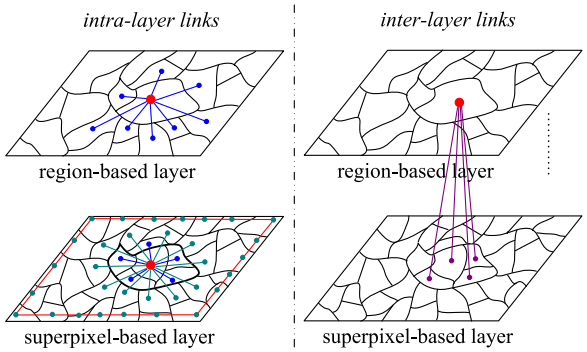


Fig. 2. Proposed multi-scale graph model. For the intra-layer links at the superpixel-based layer, we extend the scope of node connection, which is shown as the dark cyan nodes. Specifically, we define that the dark cyan nodes sharing a common boundary (shown in thick line) with the dark blue nodes are connected to the red node. In addition, the red line along the four sides indicates that all the boundary nodes are connected with each other.

incorporated at multiple quantization levels (See Fig. 2). With the connections between layers, the multi-scale saliency inference can capture long-range grouping cues for the foreground and background and encourage the nodes with similar appearance across layers to have the same label. In addition, since the parts of an object with different visual appearance can be clustered together at the higher layers, they are likely to have the same saliency values. By exploiting multi-scale graphs and interactions, the proposed algorithm can handle size and appearance variations of salient objects, and highlight the pixels therein more uniformly.

4.1 Nodes

In this work, we construct a four-layer graph $G = (V, E)$ as shown in Fig. 2, where the nodes $V = \{V_1, \ldots, V_4\}$ contain one set of superpixels V_1 and three sets of regions $\{V_2, V_3, V_4\}$. The superpixels are generated by the SLIC algorithm [56]. We group superpixels to obtain a series of region nodes at different scales by using spectral segmentation as detailed in Algorithm 1. There are fewer nodes at the higher layers, i.e., $|V_4| < |V_3| < \cdots < |V_1|$ (e.g., $|V_1| = 300$, $|V_2| = 80$, $|V_3| = 50$ and $|V_4| = 30$ respectively in our experiments).

Algorithm 1. Constructing Multi-Scale Graphs

Input: An image

1: Construct a single-scale graph $G' = (V', E')$ with superpixels as nodes V', and the undirected edges E' connect any pair of superpixels sharing a common boundary, where the edge weight is defined as $w'_{ij} = \exp(-\kappa \| \mathbf{c}_i - \mathbf{c}_j \|)$ (κ is a scaling parameter)
2: **for** $K = \{|V_2|, |V_3|, |V_4|\}$ **do**
3: Apply the K-way segmentation method [36] to group superpixels and use the results as the nodes of the multi-scale graph G.
4: **end for**

Output: Multi-Scale graphs

4.2 Edges

The edges in E are undirected links either within or between layers. As neighboring nodes are likely to share similar appearance and saliency values, we use a k-regular graph to exploit the spatial relationship. For the inter-layer links, a region node is only connected to the superpixel nodes it contains, which enforces that there exist no edges between any two region-based layers. Through the inter-layer links, we capture cross-scale grouping cues among different scales, thereby obtaining more reliable saliency results. For the intra-layer links, at the region-based layers there exists an edge if two nodes share a common boundary; and at the superpixel-based layer, each node is connected to the neighbors and the nodes sharing common boundaries with its neighboring nodes (See Fig. 2). By extending the scope of node connection with the same degree of k, we effectively utilize local smoothness cues. Furthermore, we enforce that the superpixel nodes on the four image sides are connected, i.e., any pair of boundary superpixel nodes are considered to be adjacent. Thus, we obtain a close-loop graph at the superpixel level. This close-loop constraint significantly improves the performance of the proposed algorithm as it tends to reduce the geodesic distance of similar superpixels and thereby improves the ranking results. Fig. 4 shows some examples where the ranking results with and without these constraints. We note that these constraints facilitate detecting salient objects when they appear near the image boundaries or in scenes with complex backgrounds.

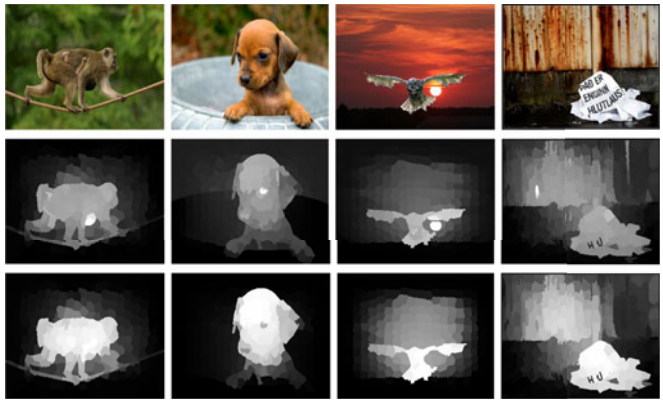


Fig. 3. Saliency measure via ranking. From top to bottom: input images, results without and with setting the diagonal elements of $\mathbf{A}$ to 0.

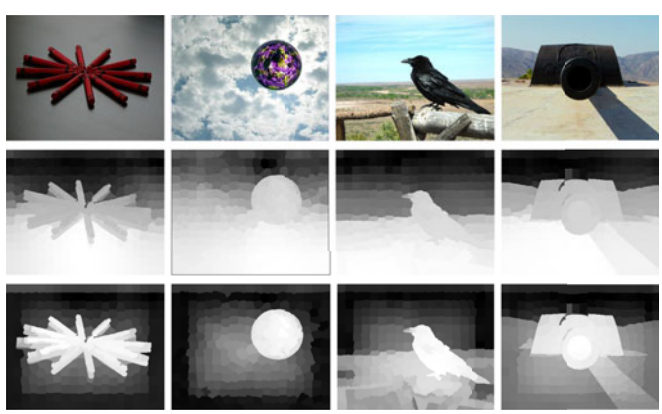


Fig. 4. Saliency maps based on top boundary prior and geodesic connection constraints of superpixels at the image boundary. Top: Input images. Middle: Results without enforcing the geodesic distance constraints. Bottom: Results with geodesic distance constraints. The geodesic constraints help detect salient objects when they appear near the image boundaries or in scenes with complex backgrounds.

4.3 Weights

With the constraints on edges, the constructed graph is sparsely connected and most elements of the affinity matrix $\mathbf{W}$ are zero. In this work, the weight between two nodes is defined by

$$w_{ij} = \begin{cases} \exp^{-\frac{\|\mathbf{c}_i - \mathbf{c}_j\|}{\sigma^2}} & \text{if } i, j \in V_l, \\ \gamma & \text{if } i \in V_1, j \in V_l, l \neq 1, \end{cases} \tag{5}$$

and satisfies $w_{ij} = w_{ji}$, where $\mathbf{c}_i$ and $\mathbf{c}_j$ denote the means of the superpixels or regions corresponding to two nodes in the CIE Lab color space, and σ is a constant that controls the affinity scale of the intra-layer weights. The weights are computed based on the distance of feature vectors in the color space as they have been shown to be effective in saliency detection [14], [57]. In (5), γ is a constant that determines the inter-layer weights.

By ranking the nodes on the constructed graph, the inverse matrix $(\mathbf{D} - \alpha\mathbf{W})^{-1}$ in (3) can be regarded as a complete affinity matrix, i.e., there exists a nonzero relevance value between any pair of nodes on the graph. This affinity matrix naturally captures spatial relationship information. The relevance between nodes is large when the spatial distance is small, which is an important cue for saliency detection [16].

5 Cascade Saliency Detection

In this section, we detail the proposed cascade scheme in three stages for bottom-up saliency detection via ranking with background and foreground queries.

5.1 Ranking with Background Queries

Based on the attention theory for visual saliency [5], we use the superpixel nodes on the image boundary as background seeds (i.e., the labeled data as query samples) to rank the relevances of all the other nodes of a graph. While this simple prior is unlikely to be correct in all images, we show that the use of several boundaries is effective and efficient. Specifically, we construct four saliency maps using boundary priors and integrate them for the final map, which is referred as the separation/combination (*SC*) approach in the following sections.

Taking the top image boundary as an example, we use the superpixel nodes on this side as the background queries and other nodes as the unlabeled data. With the indicator vector $\mathbf{y}$ for labeled and unlabeled data, all the nodes are ranked by (3) in $\mathbf{f}^*$, which is a N-dimensional vector (N is the total number of nodes of the graph). Each element in $\mathbf{f}^*$ indicates the relevance of a node to the background queries, and its complement is the saliency measure. We extract the corresponding sub-vector of $\mathbf{f}^*$ for the superpixel nodes and normalize it to the range between 0 and 1, and the saliency map using the top boundary prior, S_t can be written as

$$S_t(i) = 1 - \bar{\mathbf{f}}^*(i) \qquad i = 1, 2, \ldots, N_s, \tag{6}$$

where i indexes a superpixel node on graph, N_s is the number of superpixel nodes and $\bar{\mathbf{f}}^*$ denotes the normalized sub-vector.

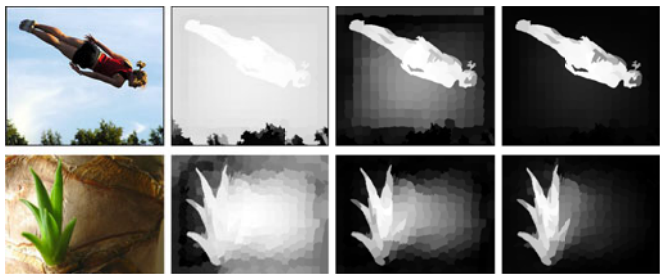


Fig. 5. Saliency maps using different queries. (a) Input images. (b) Results of using all the boundary nodes together as queries. (c) Results of integrating four maps from each side. (d) Results of ranking with foreground queries.

Similarly, we compute the other three maps S_b, S_l and S_r, using the bottom, left and right image boundary as queries respectively. We note that the saliency maps are computed with different indicator vector $\mathbf{y}$ while the weight matrix $\mathbf{W}$ and the degree matrix $\mathbf{D}$ are fixed. That is, we need to compute the matrix inverse $(\mathbf{D} - \alpha\mathbf{W})^{-1}$ only once for each image. Since the number of superpixels is small, the matrix inverse in (3) can be computed efficiently, and the overall computational load for the four maps is low. The four saliency maps are integrated by the following process:

$$S_{bq}(i) = S_t(i) \circ S_b(i) \circ S_l(i) \circ S_r(i), \tag{7}$$

where $\circ$ is an integration operator (i.e., $\{+, \times, \min, \max\}$ operators) on each node. Based on our experiments, it is shown that higher accuracy can be achieved with the product operator and adopted in this work.

There are two reasons for using the *SC* approach to generate saliency maps. First, the superpixels on different sides are often disimilar and the distance between them should be large. If we use all the boundary superpixels from four sides as queries (i.e., indicating these superpixels are similar), the labeled results are usually less optimal as these query nodes are not compatible in terms of their features. Note that the superpixels from opposite sides discussed in Section 4 can be considered as weakly labeled as only a few nodes are involved (i.e., only the superpixels from opposite sides that are close in the feature space are considered as similar) whereas the case with all superpixels from four sides can be considered as strongly labeled (i.e., all the nodes from four sides are considered as similar). Fig. 5 shows some labeling results when all nodes from four sides are considered as queries, or four query maps are constructed (using nodes from each side) and integrated. Second, it reduces the effects of imprecise queries, i.e., the ground-truth salient nodes are inadvertently selected as background queries. As shown in Fig. 6b, the saliency maps generated by using all the boundary nodes are poor. Due to the imprecise labeling results, the pixels with the salient objects have low saliency values. However, as salient objects are often compact "things" (such as a people or a car) as opposed to "stuff" (such as grass or sky) and therefore they rarely occupy three or all sides of image, the proposed *SC* approach ensures at least two saliency maps are effective (Fig. 6c). By integration of four saliency maps, some salient parts (e.g., regions away from the boundary) can be

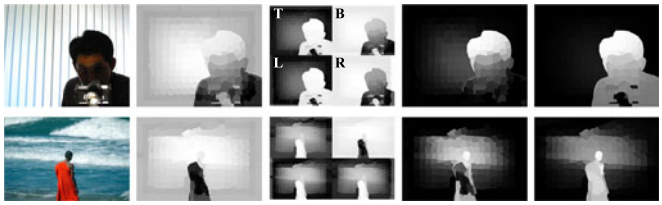


Fig. 6. Examples in which the salient objects appear at the image boundary. (a) Input images. (b) Saliency maps using all the boundary nodes together as queries. (c) Four saliency maps by different boundary priors. (d) Integration of four maps. (e) Saliency maps after the second stage.

identified, which provide sufficient cues for further processing in the second stage.

While most regions of the salient objects are highlighted in the first stage, some background nodes may not be adequately labeled with proper saliency measures (See Figs. 5 and 6). To alleviate this problem and improve the results especially when objects appear near the image boundaries, the saliency maps are further improved via ranking with foreground queries.

5.2 Ranking with Foreground Queries

The saliency map generated in the first stage indicates the confidence of each superpixel being salient. The indicator vector $\mathbf{y}$ of foreground queries in the second stage is

$$y_i = \begin{cases} S_{bq}(i) & \text{if } i \in \text{superpixel nodes} \\ 0 & \text{if } i \in \text{region nodes}, \end{cases} \tag{8}$$

where the elements have values between 0 to 1. Given $\mathbf{y}$, a new ranking vector $\mathbf{f}^*$ is computed using (3). As is carried out in the first stage, the ranking sub-vector that superpixel nodes correspond to is extracted and normalized between the range of 0 and 1 to form a saliency map by

$$S_{fq}(i) = \overline{\mathbf{f}}^*(i) \qquad i = 1, 2, \ldots, N_s, \tag{9}$$

where i indexes superpixel node on a graph, N_s is the number of superpixel nodes and $\overline{\mathbf{f}}^*$ denotes the normalized sub-vector.

We note that there are cases where nodes may be incorrectly selected as foreground queries in this stage. Despite some imprecise labeling, salient objects can be detected by the proposed algorithm as shown in Fig. 7. This can be explained as follows. The salient object regions are usually relatively compact (in terms of spatial distribution) and homogeneous in appearance (in terms of feature distribution), while background regions are the opposite. In other words, the intra-object relevance (i.e., two nodes of the salient objects) is usually much larger than the object-background and intra-background relevances, which can be inferred from the affinity matrix $\mathbf{A}$. Therefore, the sum of the relevance values of object nodes to the ground-truth salient queries is considerably larger than that of background nodes to all the queries. Thus background saliency measures can be computed effectively (Fig. 7c). In spite of the saliency maps after the first stage of Fig. 6 are not precise, salient objects can be better detected after ranking with foreground queries in the second stage.

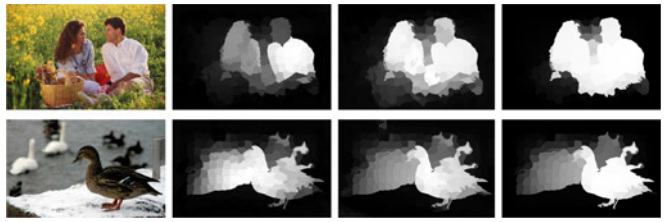


Fig. 7. Two examples in which imprecise salient queries are selected in the second stage. (a) Input images. (b) Saliency maps of the first stage. (c) Saliency maps of the second stage. (d) Final saliency maps. Although some nodes are mistakenly highlighted in the first stage, salient objects can be detected by the proposed algorithm based on intra-object relevance of foreground and background regions.

To analyze the effects of ranking results based on the affinity matrix $\mathbf{A}$, we compute the average intra-object, intra-background and object-background relevance values (referred as r_{oo}, r_{bb} and r_{ob} respectively) for each of the 300 images randomly sampled from the MSRA dataset with ground truth labels [23]. The r_{oo} value is computed by averaging all the learned relevances between any pair of object superpixel nodes in the affinity matrix $\mathbf{A}$, and r_{bb} as well as r_{ob} are computed in a similar manner. Fig. 8a shows that the intra-object relevance values are much larger than the object-background and intra-background relevance values. In addition, we also compute the sum of the relevances of each superpixel node to all other superpixel nodes, which is equivalent to computing a saliency map by (3) and (9) with all superpixel nodes (i.e., the whole image) as salient queries. This approach is essentially similar to the methods based on global contrast [16], [28]. Despite numerous queries are not correct (i.e., the background superpixels are mistakenly labeled as foreground queries), most regions of the salient objects can be detected, which also shows the learned intra-object relevances are much larger.

5.3 Re-Ranking with Mid-Level Features

A number of mid-level features have been used to estimate the likelihood of a superpixel belonging to a generic object [45]. In the third stage, we use the output of the second stage S_{fq} as mid-level features to construct a new graph where the node connections remain the same as the graph used in the first two stages. That is, the weights w_{ij} are still defined using (5), but the only differences are that $\mathbf{c}_i$ and $\mathbf{c}_j$ in (5) denote the features based on saliency measures S_{fq} rather than low-level color features.

Similar to the second stage, we define indicator vector $\mathbf{y}$ in (10), and then all nodes are re-ranked on the new graph by using (3),

$$y_i = \begin{cases} S_{fq}(i) & \text{if } i \in \text{superpixel nodes} \\ 0 & \text{if } i \in \text{region nodes}. \end{cases} \tag{10}$$

We extract the ranking scores of superpixel nodes and normalize them to obtain the final saliency map. By re-ranking on the image manifold constructed using mid-level features, better saliency detection results can be achieved in which pixels of salient objects are more uniformly highlighted and the effects of those in the background are suppressed. Figs. 7d and 9 show examples where the

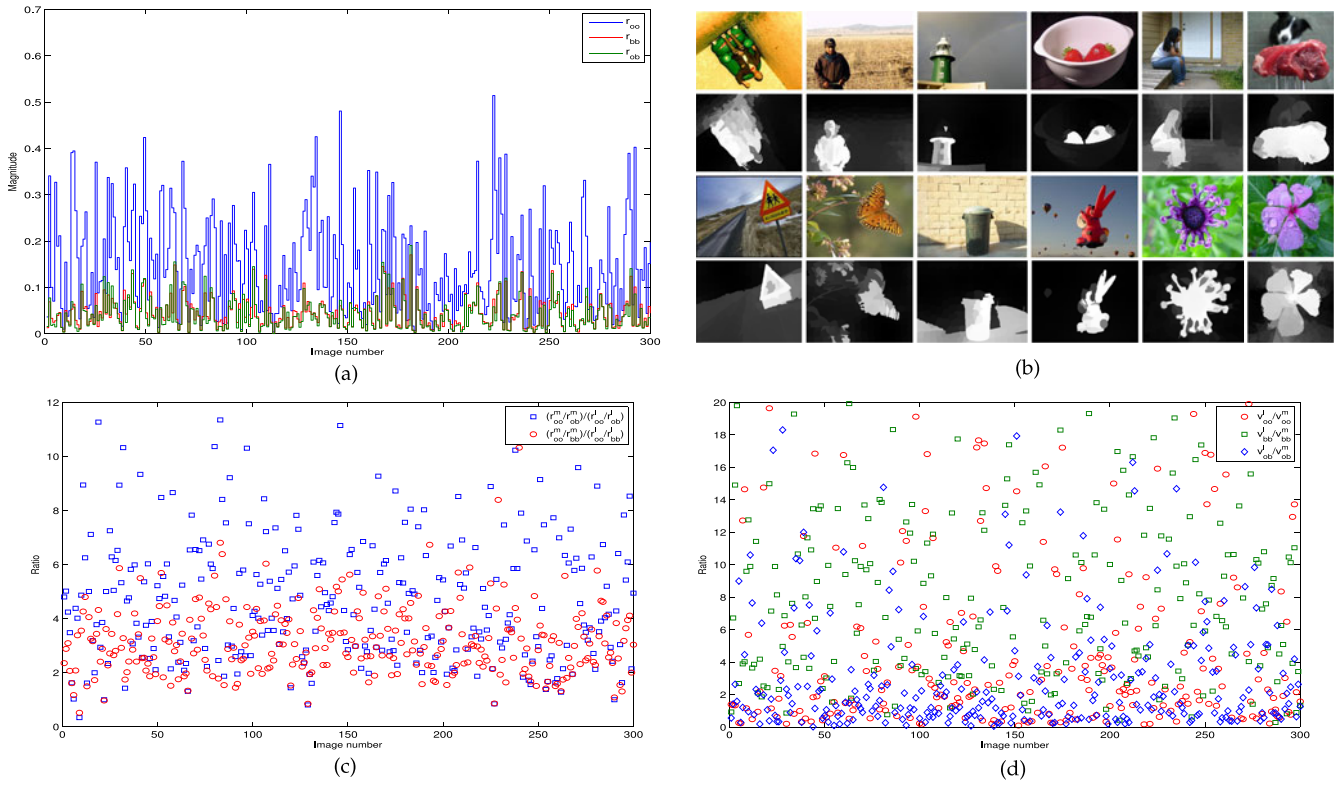


Fig. 8. Analysis of learned relevances between nodes in the affinity matrix $\mathbf{A}$. (a) Intra-object r_{oo}, intra-background r_{bb}, and object-background r_{ob} relevance values. (b) Saliency detection results using the row-sum of the sub-matrix of $\mathbf{A}$ corresponding to superpixel nodes. (c) The ratios computed with mid-level features are much larger than the corresponding ones computed with low-level features for most of the images. (d) The variances computed with mid-level features are smaller than those computed with low-level features.

saliency detection results after re-ranking are better than those obtained in the first two stages.

The use of mid-level features in ranking better describes the affinity of graph nodes and further widens the distances between the pairs of average relevance values (r_{oo}, r_{bb}) and (r_{oo}, r_{ob}). The salient regions are thus highlighted with sharper contrast from the background. For ease of presentation, we use superscript l and m to denote the ranking operations on the image manifold constructed with low-level and mid-level features. We compute four ratios of r^m_{oo}/r^m_{ob}, r^m_{oo}/r^m_{bb}, r^l_{oo}/r^l_{ob} and r^l_{oo}/r^l_{bb} for each image of the MSRA dataset, and observe that for most of the images, the former two ratios computed with mid-level features are much larger than the corresponding ones computed with low-level features (See Fig. 8c).

Ranking on graph nodes with mid-level features also decreases the variances of intra-object, intra-background, and object-background relevances in $\mathbf{A}$ (referred to as v_{oo}, v_{bb} and v_{ob}) especially for the images containing non-homogeneous regions of foreground objects and backgrounds. As a result,

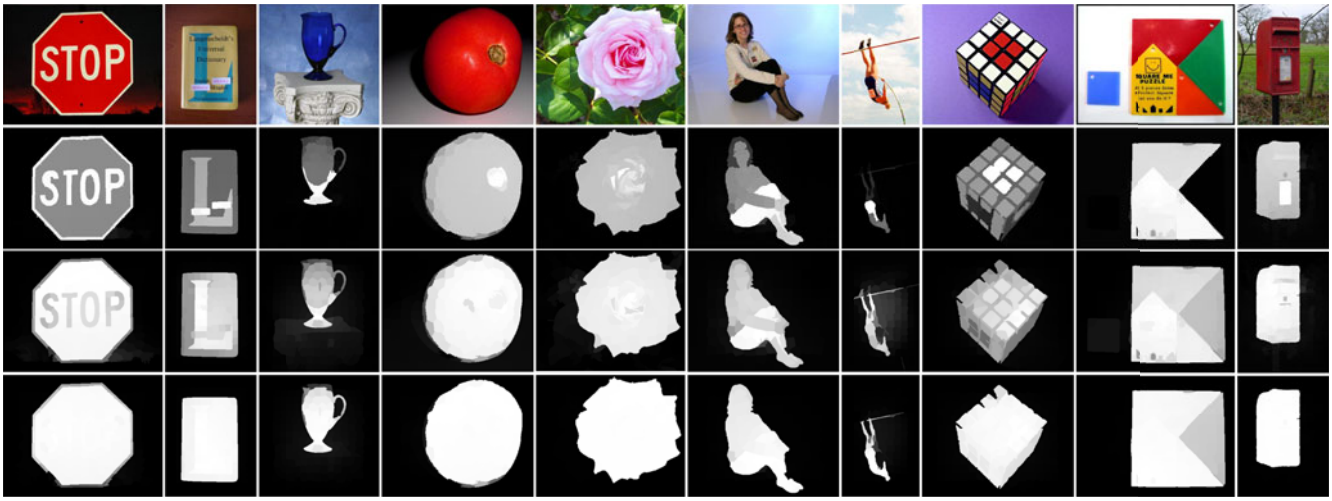
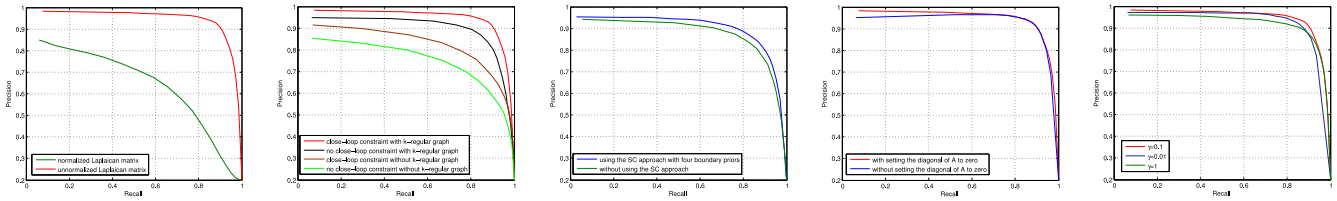


Fig. 9. Examples showing the benefits of the multi-scale analysis. From top to down: input images, results of using single-layer graph [60], results of the second stage with multi-layer graph, results of the third stage. The objects can more completely stand out from the backgrounds by the proposed method.



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

Fig. 10. Precision-recall curves on the ASD dataset by the proposed algorithm with different design options. (a) Ranking with normalized and unnormalized Laplacian matrices. (b) Graph construction with different constraints. (c) SC approach. (d) Ranking using A with zero and nonzero diagonal elements. (e) Results with different value of γ.

salient regions are more uniformly highlighted. We compute three ratios of v_{oo}^l/v_{oo}^m, v_{bb}^l/v_{bb}^m and v_{ob}^l/v_{ob}^m for each image of the MSRA dataset, and observe that the variances computed with mid-level features are smaller than those computed with low-level features for most images (See Fig. 8d). The main steps of the proposed saliency detection algorithm are summarized in Algorithm 2.

5.4 Discussions

There are significant differences between bottom-up saliency models and top-down goal-driven mechanisms. For example, human faces may not stand out from the other salient objects in a complex scene using a bottom-up model (due to size, occlusion, and contrast) but human observes are likely to fixate on the region containing faces [58]. In such cases, integrating additional prior information, such as face locations, can improve the performance of bottom-up saliency models. The proposed model can be naturally integrated with other prior for salient object detection.

In addition, existing methods (e.g., RC [16] and CB [59]) do not specifically take the label smoothness (i.e., neighboring pixels or regions with similar appearance should have the same saliency measure) into account. Consequently, pixels within salient objects are not uniformly highlighted especially near the boundaries. With the post-processing step via the proposed ranking-based model (3) and (9), similar regions are encouraged to have similar saliency measures, thereby improving the performance of existing saliency methods.

Algorithm 2. Saliency Detection via Manifold Ranking

Input: An image and default parameters.

1: Hierarchically segment the input image into superpixels and different scale regions, construct a four-layer graph G with superpixels and region as nodes, and compute its degree matrix $\mathbf{D}$ and weight matrix $\mathbf{W}$ by (5).
2: Compute $(\mathbf{D} - \alpha\mathbf{W})^{-1}$ and set its diagonal elements to 0.
3: Form an indicator vector $\mathbf{y}$ with nodes on each side of image as queries, and compute their corresponding saliency maps by (3) and (6). Compute saliency map S_{bq} by (7).
4: Obtain an indicator vector $\mathbf{y}$ by (8), and compute saliency map S_{fq} by (3) and (9).
5: Substitute $S_{fq}(i)$ for $\mathbf{c}_i$ in (5) to re-compute weight matrix $\mathbf{W}$ and degree matrix $\mathbf{D}$, and repeat Step 2.
6: Use indicator vector $\mathbf{y}$ in (10) to re-rank graph nodes by (3), and normalize the resulting ranking scores to get the final saliency map.

Output: Full-resolution saliency map.

6 Experimental Results

We evaluate the proposed algorithm on five benchmark datasets. The MSRA dataset [23] contains 5,000 images with ground-truth salient regions enclosed by bounding boxes. The ASD dataset, a subset of the MSRA set, consists of 1,000 images with segmentation masks of salient objects [14]. Each of the 10,000 images in the THUS dataset [61] contains one salient object with ground-truth region annotation. In addition, we develop the DUT-OMRON dataset, which is composed of 5,168 labeled images containing multiple objects at different scales and locations in cluttered backgrounds. We also evaluate the proposed algorithm on the MIT300 dataset [62] for eye fixation prediction. This dataset contains 300 images with eye tracking data from 39 observers. Experimental results with twenty state-of-the-art saliency detection methods, including IT [5], GB [6], MZ [7], SR [8], AC [13], Gof [1], FT [14], LC [15], RC [16], SVO [19], SF [17], CB [59], GS [31], Xie [40], DGI [63], eDN [64], SC [65], MrCNN [66], RCJ [61], and LS [25], are presented.

6.1 Experimental Setup and Evaluation Metrics

The parameter α in (3) balances the smooth and fitting constraints of the proposed manifold ranking algorithm. When the value of α is small, the initial labeling of nodes plays a more important role. On the other hand, the label consistency among neighboring nodes is more important when the α value is large. In this work, α is empirically set to be 0.99, for all the experiments. The Gaussian kernel width σ^2 of (5) is set to be 0.1 for the weights between superpixel nodes, and 1 between region nodes, respectively. In addition, the inter-layer weights γ are defined to be 0.1. Fig. 10e shows the sensitivity of the proposed algorithm to the parameter γ. Overall, the proposed algorithm is insensitive to a wide range of γ.

We evaluate salient object detection methods by pixel level precision, recall, F-measure, Area Under Curve (AUC) and Mean Absolute Error (MAE), and evaluate fixation prediction methods by AUC, Normalized Scanpath Saliency (NSS) and Similarity (S). The precision value corresponds to the ratio of salient pixels correctly assigned to all pixels of the extracted regions, while the recall value is defined as the percentage of detected salient pixels with respect to the ground-truth data. Given a saliency map with intensity values normalized to the range of 0 and 255, a number of binary maps are produced by using every possible fixed threshold in $[0, 255]$. We compute the precision/recall pairs of all the binary maps to plot the precision-recall curve. Meanwhile, we obtain true positive and false positive rates to plot the ROC curve and compute the AUC score. The

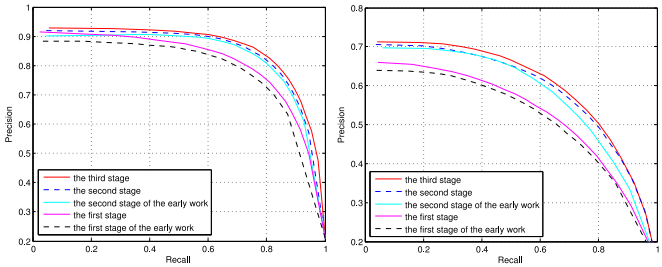


Fig. 11. Performance comparisons of each stage. (a) Results on the MSRA dataset. (b) Results on the DUT-OMRON dataset.

F-measure is the overall performance indicator computed by the weighted harmonic of precision and recall [14]. The MAE computes the average difference between the saliency map and the ground truth [17]. The average normalized salience value across all fixation locations is taken as the NSS score [67]. For the similarity metric, fixation maps are used as probability distributions from which the histogram intersection is measured [24].

6.2 Algorithmic Design Options

We first examine the design options of the proposed algorithm using the ASD dataset. The ranking results using the normalized (2) and unnormalized (3) Laplacian matrices for ranking are analyzed. Fig. 10a shows that the ranking results with the unnormalized Laplacian matrix are better, and used in all the experiments. Next, we demonstrate the effectiveness of the proposed graph construction scheme with different constraints. We compute the precision-recall curves for four cases of node connection on the graph (discussed in Section 4.2): close-loop constraint without extending the scope of node with k-regular graph, without close-loop constraint and with k-regular graph, without either close-loop constraint or k-regular graph, and close-loop constraint with k-regular graph. Fig. 10b shows that the proposed algorithm with the close-loop constraint and k-regular graph performs best.

The effects of the proposed SC approach in the first stage is also evaluated. Fig. 10c shows that our approach using the integration of saliency maps generated from different boundary priors performs better in the first stage. We further evaluate the effect of setting the diagonal elements of A to 0 in (4). Fig. 10d shows that it is more important to use null diagonal elements when the recall is low (e.g., less than 0.5), which is consistent to the examples shown in Fig. 3.

TABLE 1
Comparison of the Proposed Algorithm with Multi-Scale Graph and Early Method with Single-Scale Graph-Based Manifold Ranking (MR) [60]

Metric	Method	Dataset			
		ASD	MSRA	THUS	DUT-OMRON
AUC	MR [60]	0.962	0.927	0.930	0.845
	Ours	0.979	0.951	0.956	0.883
MAE	MR [60]	0.075	0.128	0.126	0.187
	Ours	0.063	0.108	0.101	0.177

We demonstrate the performance for each stage of the proposed algorithm and compare it with our work [60]. Fig. 11 shows that the saliency map of the second stage with foreground queries is significantly better than the results of the first stage with background queries. Likewise, the re-ranking results in the third stage are better than those from the second stage. In addition, the performance improvement from the first stage of the proposed algorithm over our early work shows the effectiveness of the multi-scale graph. The proposed algorithm consistently performs better than our early method [60] on these datasets with 2.9 and 14.2 percent improvement in terms of AUC and MAE scores as shown in Table 1.

6.3 Performance Evaluation

ASD. We first examine the proposed algorithm against fourteen state-of-the-art saliency detection methods on the ASD dataset. Fig. 12a shows the precision-recall curves of all evaluated methods. The proposed algorithm outperforms the SVO [19], Gof [1], CB [59] and RC [16] methods which are top-performing methods for saliency detection in a recent benchmark study [57]. In addition, the proposed method significantly outperforms the GS [31] model which is also based on boundary priors. The precision, recall and F-measure with an adaptive threshold and the AUC score are presented in Fig. 12b, which shows that the proposed algorithm achieves the highest precision, F-measure and AUC values. Overall, the proposed algorithm performs favorably against the state-of-the-art methods using all four evaluation metrics. Fig. 15 shows a few saliency maps generated by the evaluated methods. We note that the proposed algorithm uniformly highlights the salient regions and preserves finer object boundaries than the other methods.

MSRA. Each image of the MSRA dataset is annotated with nine bounding boxes by different users. The pixels

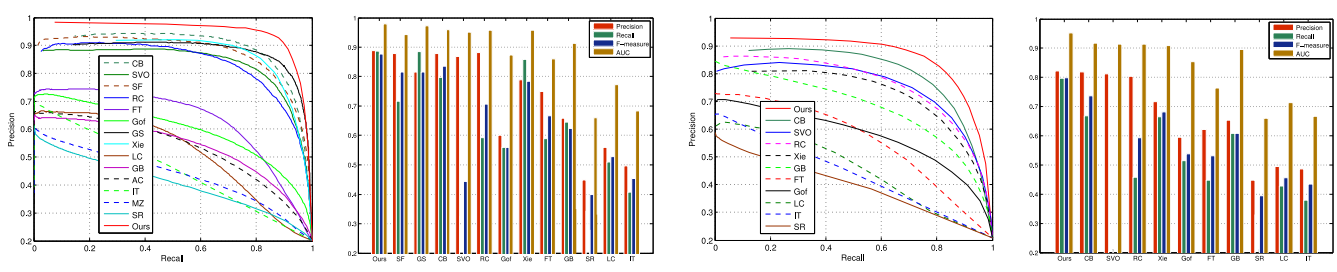


Fig. 12. Quantitative comparisons. (a), (b) Results on the ASD dataset. (c), (d) Results on the MSRA dataset. The proposed method performs well in all these metrics.

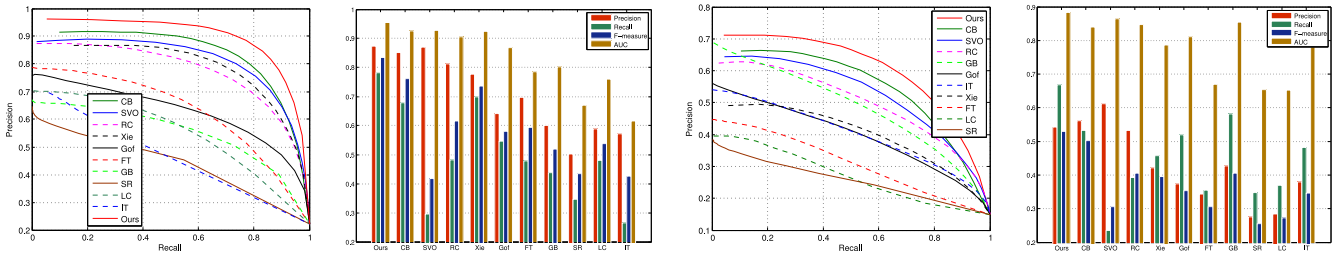


Fig. 13. Quantitative comparisons. (a), (b) Results on the THUS dataset. (c), (d) Results on the DUT-OMRON dataset. The proposed method performs well in all these metrics.

with consistency score higher than a threshold (e.g., 0.5) are considered as parts of salient regions and enclosed by bounding boxes. Jiang et al. [68] provide more accurate labeling results for the MSRA dataset, based on which we evaluate the performance of the proposed algorithm against ten state-of-the-art approaches using the original source code including the CB [59], SVO [19], RC [16], Xie [40], Gof [1], FT [14], GB [6], SR [8], LC [15] and IT [5] methods. Figs. 12c and 12d shows that the precision-recall curve of the proposed method is higher than those by the other methods on the MSRA dataset (e.g., performance gain of 8.3 and 3.7 percent over the second best method (CB [59]) in terms of the F-measure and AUC score). The reported F-measure score in [12] is 0.694 (based on bounding box ground-truth annotation). With the same ground-truth annotation, the F-measure score of the proposed method is 0.883.

THUS. The number of images in the THUS dataset is an order of magnitude larger than the ASD and MSRA datasets for evaluation of saliency detection [14]. Each of the 10,000 images in the THUS dataset contains one salient object with ground-truth pixel-wise boundary annotation. Figs. 13a and 13b shows that the proposed algorithm consistently performs well using the four evaluation criteria with improvement of 9.5 and 3.2 percent over the second best method (CB [59]) in terms of F-measure and AUC score.

DUT-OMRON. As most images in the existing datasets mainly contain single salient objects, we develop a challenging dataset with 5,168 images where each one contains one or multiple salient objects in cluttered background. Each image is resized to be no larger than 400×400 pixels and labeled by five subjects to obtain pixel-wise ground-truth annotation in a way similar to what is carried out in the MSRA dataset. We use k-means algorithm to classify the fixations from five subjects into three clusters, and then retain 90 percent of the fixations based on Euclidean distance from the cluster centers as eye-fixation ground-truth. In addition, we also provide bounding box ground-truth annotation for each image. The source images, ground-truth labels and detailed description of this dataset can be found at http://saliencydetection.net/dut-omron/.

Figs. 13c and 13d shows the results on the DUT-OMRON dataset. As this is a challenging dataset, the performance of all evaluated methods decreases significantly in terms all metrics (as opposed to the results in other datasets where there is little room for further improvement). Overall, the proposed method achieves the highest precision-recall curve, the highest F-measure and AUC values.

MIT300. This is an eye fixation benchmark dataset in which the image resolution is from $450 \times 1{,}024$ to $1{,}024 \times 1{,}024$ pixels. For computational efficiency, the images are resized to no more than 400 pixels in width and height while maintaining the aspect ratios. Similar to previous eye fixation prediction methods [25], [69], we smooth the resulting maps by Gaussian blurs with a small kernel for visualization. We evaluate the proposed algorithm against seven state-of-the-art approaches including the DGI [63], eDN [64], SC [65], MrCNN [66], RCJ [61], LS [25] methods and our early work (MR) [60]. Table 2 shows quantitative results using three metrics (some taken from the MIT saliency benchmark website [70]) in which the first four methods based on deep convolutional networks require long training time and a large number of images. Overall, no single method dominates the other approaches in all three metrics. The proposed algorithm achieves the highest S scores and performs slightly worse than the DGI [63] and eDN [64] methods in terms of the AUC score.

Features. We use the LBP, HOG and DRFI features to represent graph nodes. The DRFI feature consists of regional contrast and background cues which is used to learn a regressor for salient object detection [68]. As shown in Fig. 14, the proposed algorithm with the LAB+DRFI or LAB +HOG representations achieves comparable performance

TABLE 2
Quantitative Comparisons on the MIT300 Dataset

Metric	DGI [63]	eDN [64]	SC [65]	MrCNN [66]
AUC	**0.84**	0.82	0.80	0.79
NSS	1.22	1.14	**1.47**	1.37
S	0.39	0.41	0.45	0.48
Metric	RCJ [61]	LS [25]	MR [60]	Ours
AUC	0.79	0.78	0.75	0.80
NSS	1.18	1.02	1.12	1.25
S	0.48	0.43	0.41	**0.49**

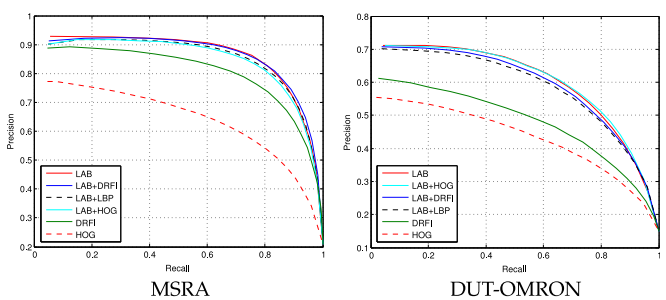
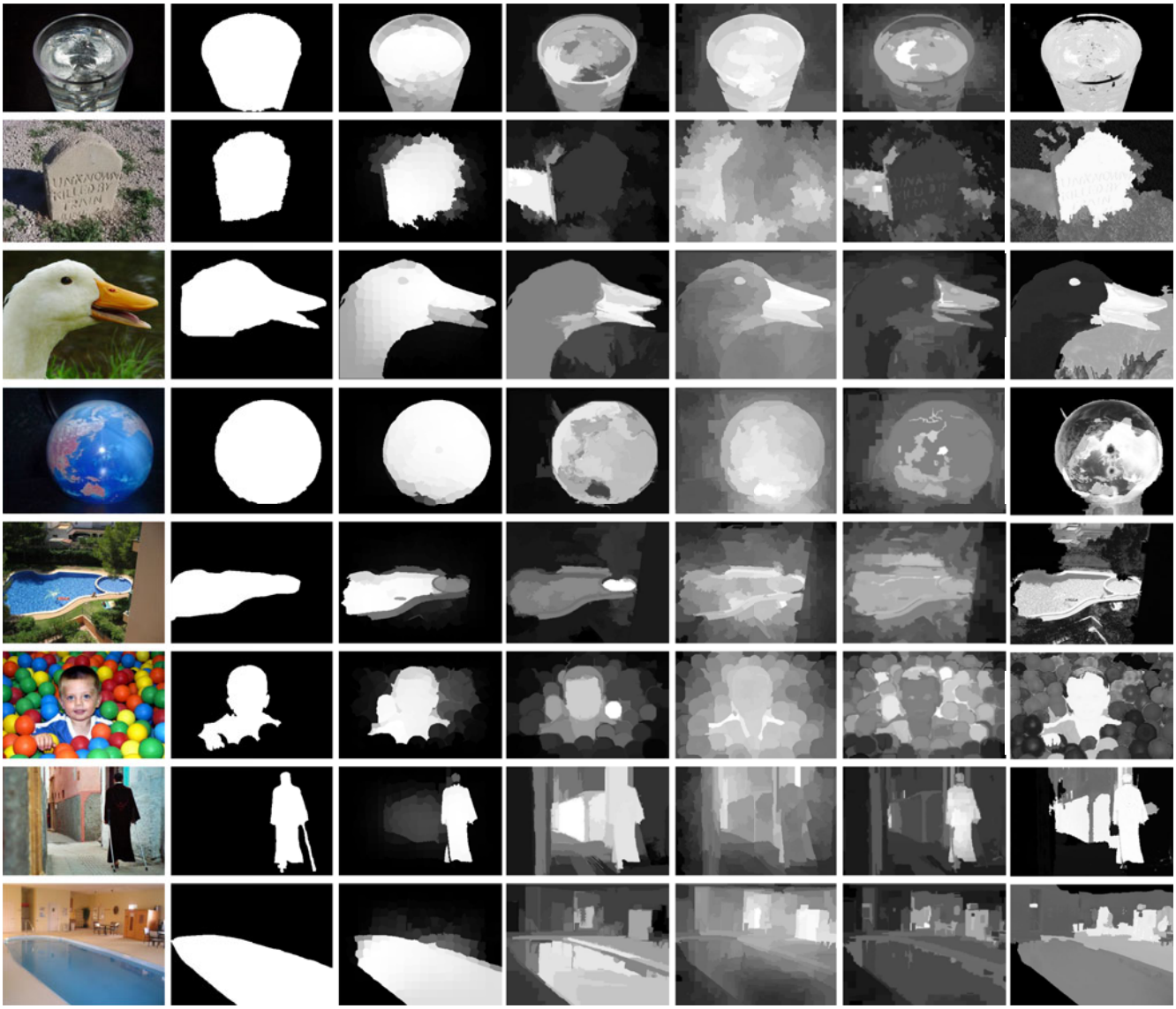


Fig. 14. Performance evaluation of the proposed algorithm with different features.



(a) Input (b) GT (c) Ours (d) CB (e) SVO (f) RC (g) Xie

Fig. 15. Salient object detection results of evaluated methods. The proposed algorithm consistently generates saliency maps close to the ground truth.

TABLE 3
Comparison of Average Execution Time (Seconds per Image)

Method	Ours	CB [59]	SVO [19]	RC [16]	Gof [1]	GB [6]	SER [10]	FT [14]	LC [15]	SR [8]	IT [5]
Code	Matlab	Matlab	Matlab	C++	Matlab	Matlab	C++	C++	C++	C++	Matlab
Time (s)	1.160	1.179	40.33	0.106	36.05	0.418	25.19	0.016	0.002	0.002	0.165

with the LAB features on the MSRA and DUT-OMRON datasets. On the other hand, the proposed algorithm with the DRFI or HOG features do not perform well. These results show that the color cues play a more important role in saliency detection for the two datasets.

6.4 Run Time Performance

The average run time of different methods on the ASD database are presented in Table 3 based on a machine with an Intel i7 3.40 GHz CPU and 32 GB RAM. Overall, the proposed algorithm performs effectively and efficiently when compared with the state-of-the-art methods. The MATLAB code of the proposed algorithm will be made available to the public.

7 Conclusion

We propose a bottom-up method to detect salient regions in images with manifold ranking on a graph which incorporates local grouping cues and boundary priors. We construct two image manifolds with low-level and mid-level features, and develop a cascade approach using background and foreground queries for ranking to generate saliency maps. We evaluate the proposed algorithm on large benchmark datasets against twenty state-of-the-art methods in salient object

detection and eye fixation prediction. In addition, we propose a large dataset for performance evaluation on saliency detection. Our future work will focus on integration of multiple features with applications to other vision problems.

Acknowledgments

L. Zhang and H. Lu were supported by the National Natural Science Foundation of China under Grant #61371157, #61472060 and #61528101. M.-H. Yang was supported in part by US National Science Foundation CAREER Grant #1149783. The authors would like to thank the reviewers and editor for the comments.

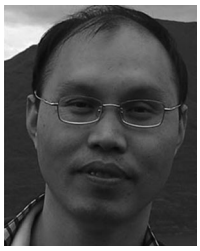
References

[1] S. Goferman, L. Zelnik-Manor, and A. Tal, "Context-aware saliency detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2010, pp. 2376–2383.
[2] U. Rutishauser, D. Walther, C. Koch, and P. Perona, "Is bottom-up attention useful for object recognition?" in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2004, vol. 2, pp. 37–44.
[3] L. Itti, "Automatic foveation for video compression using a neurobiological model of visual attention," *IEEE Trans. Image Process.*, vol. 13, no. 10, pp. 1304–1318, Oct. 2004.
[4] T. Chen, M. Cheng, P. Tan, A. Shamir, and S. Hu, "Sketch2photo: Internet image montage," *ACM Trans. Graphics.*, vol. 28, no. 5, 2009, Art. no. 124.
[5] L. Itti, C. Koch, and E. Niebur, "A model of saliency-based visual attention for rapid scene analysis," *IEEE Trans. Pattern Anal. Mach. Intell*, vol. 20, no. 11, pp. 1254–1259, Nov. 1998.
[6] J. Harel, C. Koch, and P. Perona, "Graph-based visual saliency," in *Proc. Adv. Neural Inf. Process. Syst.*, 2006, pp. 545–552.
[7] Y. Ma and H. Zhang, "Contrast-based image attention analysis by using fuzzy growing," in *Proc. 11th ACM Int. Conf. Multimedia*, 2003, pp. 374–381.
[8] X. Hou and L. Zhang, "Saliency detection: A spectral residual approach," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2007, pp. 1–8.
[9] N. Bruce and J. Tsotsos, "Saliency based on information maximization," in *Proc. Adv. Neural Inf. Process. Syst.*, 2006, pp. 155–162.
[10] W. Wang, Y. Wang, Q. Huang, and W. Gao, "Measuring visual saliency by site entropy rate," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2010, pp. 2368–2375.
[11] D. Klein and S. Frintrop, "Center-surround divergence of feature statistics for salient object detection," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2011, pp. 2214–2219.
[12] V. Gopalakrishnan, Y. Hu, and D. Rajan, "Random walks on graphs for salient object detection in images," *IEEE Trans. Image Process.*, vol. 19, no. 12, pp. 3232–3242, Dec. 2010.
[13] R. Achanta, F. Estrada, P. Wils, and S. Susstrunk, "Salient region detection and segmentation," in *Proc. Int. Conf. Comput. Vis. Syst.*, 2008, pp. 66–75.
[14] R. Achanta, S. Hemami, F. Estrada, and S. Susstrunk, "Frequency-tuned salient region detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2009, pp. 1597–1604.
[15] Y. Zhai and M. Shah, "Visual attention detection in video sequences using spatiotemporal cues," in *Proc. ACM Int. Conf. Multimedia*, 2006, pp. 815–824.
[16] M. Cheng, G. Zhang, N. Mitra, X. Huang, and S. Hu, "Global contrast based salient region detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2011, pp. 409–416.
[17] F. Perazzi, P. Krahenbuhl, Y. Pritch, and A. Hornung, "Saliency filters: Contrast based filtering for salient region detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2012, pp. 733–740.
[18] L. Wang, J. Xue, N. Zheng, and G. Hua, "Automatic salient object extraction with contextual cue," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2011, pp. 105–112.
[19] K.-Y. Chang, T.-L. Liu, H.-T. Chen, and S.-H. Lai, "Fusing generic objectness and visual saliency for salient object detection," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2011, pp. 914–921.
[20] Y. Lu, W. Zhang, H. Lu, and X. Xue, "Salient object detection using concavity context," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2011, pp. 233–240.
[21] J. Feng, Y. Wei, L. Tao, C. Zhang, and J. Sun, "Salient object detection by composition," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2011, pp. 1028–1035.
[22] J. Yang and M. Yang, "Top-down visual saliency via joint CRF and dictionary learning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2012, pp. 2296–2303.
[23] T. Liu, et al., "Learning to detect a salient object," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 2, pp. 353–367, Feb. 2011.
[24] T. Judd, K. Ehinger, F. Durand, and A. Torralba, "Learning to predict where humans look," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2009, pp. 2106–2113.
[25] A. Borji and L. Itti, "Exploiting local and global patch rarities for saliency detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2012, pp. 478–485.
[26] R. Margolin, A. Tal, and L. Zelnik-Manor, "What makes a patch distinct?" in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2013, pp. 1139–1146.
[27] M. Cheng, J. Warrell, W. Lin, S. Zheng, V. Vineet, and N. Crook, "Efficient salient region detection with soft image abstraction," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2013, pp. 1529–1536.
[28] C. Yang, L. H. Zhang, and H. C. Lu, "Graph-regularized saliency detection with convex-hull-based center prior," *IEEE Signal Process. Lett.*, vol. 20, no. 7, pp. 637–640, Jul. 2013.
[29] P. Jiang, H. Ling, J. Yu, and J. Peng, "Salient region detection by UFO: Uniqueness, focusness and objectness," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2013, pp. 1976–1983.
[30] Y. Jia and M. Han, "Category-independent object-level saliency detection," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2013, pp. 1761–1768.
[31] Y. Wei, F. Wen, W. Zhu, and J. Sun, "Geodesic saliency using background priors," in *Proc. 12th Eur. Conf. Comput. Vis.*, 2012, pp. 29–42.
[32] B. Jiang, L. Zhang, H. Lu, C. Yang, and M.-H. Yang, "Saliency detection via absorbing Markov chain," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2013, pp. 1665–1672.
[33] W. Zhu, S. Liang, Y. Wei, and J. Sun, "Saliency optimization from robust background detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2014, pp. 1–8.
[34] V. Lempitsky, P. Kohli, C. Rother, and T. Sharp, "Image segmentation with a bounding box prior," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2009, pp. 277–284.
[35] L. Grady, M. Jolly, and A. Seitz, "Segmenation from a box," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2011, pp. 367–374.
[36] T. Kim, K. Lee, and S. Lee, "Learning full pairwise affinities for spectral segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2010, pp. 2101–2108.
[37] A. Borji and L. Itti, "State-of-the-art in visual attention modeling," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 1, pp. 185–207, Jan. 2013.
[38] A. Triesman and G. Gelade, "A feature-integration theory of attention," *Cogn. Psychol.*, vol. 12, no. 1, pp. 97–136, 1980.
[39] E. Rahtu, J. Kannala, M. Salo, and J. Heikkilä, "Segmenting salient objects from images and videos," in *Proc. Eur. Conf. Comput. Vis.*, 2010, pp. 366–379.
[40] Y. Xie, H. Lu, and M.-H. Yang, "Bayesian saliency via low and mid level cues," *IEEE Trans. Image Process.*, vol. 22, no. 5, pp. 1689–1698, May 2013.
[41] C. Guo, Q. Ma, and L. Zhang, "Spatio-temporal saliency detection using phase spectrum of quaternion fourier transform," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2008, pp. 1–8.
[42] C. Lang, G. Liu, J. Yu, and S. Yan, "Saliency detection by multitask sparsity pursuit," *IEEE Trans. Image Process.*, vol. 21, no. 3, pp. 1327–1338, Mar. 2012.
[43] X. Shen and Y. Wu, "A unified approach to salient object detection via low rank matrix recovery," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2012, pp. 853–860.
[44] X. Li, H. Lu, L. Zhang, X. Ruan, and M.-H. Yang, "Saliency detection via dense and sparse reconstruction," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2013, pp. 2976–2983.
[45] B. Alexe, T. Deselares, and V. Ferrai, "What is an object?" in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2010, pp. 73–80.
[46] L. Mai, Y. Niu, and F. Liu, "Saliency aggregation: A data-driven approach," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2013, pp. 1131–1138.
[47] Z. Jiang and L. Davis, "Submodular salient region detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2013, pp. 2043–2050.

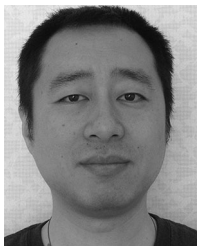
[48] X. Zhu, Z. Ghahramani, and J. Lafferty, "Semi-supervised learning using gaussian fields and harmonic functions," in *Proc. Int. Conf. Mach. Learn.*, 2003, pp. 912–919.
[49] B. Tatler, "The central fixation bias in scene viewing: Selecting an optimal viewing position independently of motor biases and image feature distributions," *J. Vis.*, vol. 7, no. 14, pp. 1–17, 2007.
[50] D. Zhou, J. Weston, A. Gretton, O. Bousquet, and B. Scholkopf, "Ranking on data manifolds," in *Proc. Adv. Neural Inf. Process. Syst.*, 2004, pp. 1–8.
[51] F.-R. Chung, *Spectral Graph Theory*. Providence, RI, USA: Amer. Math. Soc., 1997.
[52] S. Brin and L. Page, "The anatomy of a large-scale hypertextual web search engine," *Comp. Netw. ISDN Syst.*, vol. 30, no. 1, pp. 107–117, 1998.
[53] A. Y. Ng, M. I. Jordan, and Y. Weiss, "On spectral clustering: Analysis and an algorithm," in *Proc. Adv. Neural Inf. Process. Syst.*, 2002, pp. 849–856.
[54] D. Zhou, O. Bousquet, T. Lal, J. Weston, and B. Scholkopf, "Learning with local and global consistency," in *Proc. Adv. Neural Inf. Process. Syst.*, 2003, pp. 321–328.
[55] B. Scholkopf, J. Platt, J. Shawe-Taylor, A. Smola, and R. Williamson, "Estimating the support of a high-dimensional distribution," *Neural Comput.*, vol. 13, pp. 1443–1471, 2001.
[56] R. Achanta, K. Smith, A. Lucchi, P. Fua, and S. Susstrunk, "SLIC superpixels," EPFL, Lausanne, Switzerland, EPFL-REPORT-149300, 2010.
[57] A. Borji, D. Sihite, and L. Itti, "Salient object detection: A benchmark," in *Proc. Eur. Conf. Comput. Vis.*, 2012, pp. 414–429.
[58] M. Cerf, J. Harel, W. Einhauser, and C. Koch, "Predicting human gaze using low-level saliency combined with face detection," in *Proc. Adv. Neural Inf. Process. Syst.*, 2007, pp. 241–248.
[59] H. Jiang, J. Wang, Z. Yuan, T. Liu, N. Zheng, and S. Li, "Automatic salient object segmentation based on contex and shape prior," in *Proc. Brit. Mach. Vis. Conf.*, 2011, pp. 1–12.
[60] C. Yang, L. Zhang, H. Lu, X. Ruan, and M.-H. Yang, "Saliency detection via graph-based manifold ranking," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2013, pp. 3166–3173.
[61] M. Cheng, N. Mitra, X. Huang, P. 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.
[62] T. Judd, F. Durand, and A. Torralba, "A benchmark of computational models of saliency to predict human fixations," MIT, Cambridge, MA, USA, MIT-CSAIL-TR-2012-001, 2012.
[63] M. Kummerer, L. Theis, and M. Bethge, "Deep gaze I: Boosting saliency prediction with feature maps trained on imagenet," in *Proc. Int. Conf. Learn. Representations*, 2015, pp. 1–11.
[64] E. Vig, M. Dorr, and D. Co, "Large-scale optimization of hierarchical features for saliency prediction in natural images," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2014, pp. 2798–2805.
[65] J. Pan, E. Sayrol, X. G. i Nieto, K. McGuinness, and N. E. O'Connor, "Shallow and deep convolutional networks for saliency prediction," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 598–606.
[66] N. Liu, J. Han, D. Zhang, S. Wen, and T. Liu, "Predicting eye fixations using convolutional neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 362–370.
[67] R. J. Peters, A. Iyer, L. Itti, and C. Koch, "Components of bottom-up gaze allocation in natural images," *Vis. Res.*, vol. 45, no. 18, pp. 2397–2416, 2005.
[68] H. Jiang, J. Wang, Z. Yuan, Y. Wu, N. Zheng, and S. Li, "Salient object detection: A discriminative regional feature integration approach," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2013, pp. 2083–2090.
[69] X. Hou, J. Harel, and C. Koch, "Image signature: Highlighting sparse salient regions," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 1, pp. 194–201, Jan. 2012.
[70] Z. Bylinskii, et al., "MIT saliency benchmark," (2012). [Online]. Avaialble: http://saliency.mit.edu/

Lihe Zhang received the MS degree and the PhD degree in signal and information processing from Harbin Engineering University, Harbin, China, in 2001 and from Beijing University of Posts and Telecommunications, Beijing, China, in 2004, respectively. He is currently an associate professor in the School of Information and Communication Engineering, Dalian University of Technology. His current research interests include computer vision and pattern recognition.

Chuan Yang received the MS degree in signal and information processing from the Dalian University of Technology, Dalian, China, in 2013. He is currently working with Alibaba Group, Beijing, China. His current research interests include computer vision and machine learning with focus on salient object detection, image segmentation, image retrieval, and deep learning.



Huchuan Lu (SM'12) received the PhD degree in system engineering and the MS degree in signal and information processing from Dalian University of Technology (DUT), Dalian, China, in 2008 and 1998, respectively. He joined the faculty in 1998 and currently is a full professor in the School of Information and Communication Engineering, DUT. His current research interests include computer vision and pattern recognition with focus on visual tracking, saliency detection, and segmentation. He is a member of the ACM, an associate editor of the *IEEE Transactions on Cybernetics*, and a senior member of the IEEE.



Xiang Ruan received the BE degree from Shanghai Jiao Tong University, Shanghai, China, in 1997, and the ME and PhD degrees from Osaka City University, Osaka, Japan, in 2001 and 2004, respectively. He was with OMRON corporation, Kyoto Japan from 2007 to 2016 as an expert engineer. He is current co-founder and CEO of IWAKI Co., Ltd., Japan. His research interests include computer vision, machine learning, and image processing.

Ming-Hsuan Yang received the PhD degree in computer science from the University of Illinois, Urbana-Champaign, Urbana, IL, in 2000. He is an associate professor of electrical engineering and computer science with the University of California, Merced, CA. He served as an associate editor of the *IEEE Transactions on Pattern Analysis and Machine Intelligence* from 2007 to 2011, and is an associate editor of the *International Journal of Computer Vision*, the *Image and Vision Computing*, and the *Journal of Artificial Intelligence Research*. He received the US National Science Foundation CAREER Award in 2012, and the Google Faculty Award in 2009. He is a senior member of the IEEE and the ACM.