Salient Object Detection Via Two-Stage Graphs

Yi Liu, Jungong Han, Qiang Zhang, and Long Wang

***Abstract*—Despite recent advances made in salient object detection using graph theory, the approach still suffers from accuracy problems when the image is characterized by a complex structure, either in the foreground or background, causing erroneous saliency segmentation. This fundamental challenge is mainly attributed to the fact that most of existing graph-based methods take only the *adjacently spatial consistency* among graph nodes into consideration. In this paper, we tackle this issue from a coarse-to-fine perspective and propose a two-stage-graphs approach for salient object detection, in which *two graphs having the same nodes but different edges* are employed. Specifically, a weighted joint robust sparse representation (WJRSR) model, rather than the commonly used manifold ranking model, helps to compute the saliency value of each node in the first-stage graph, thereby providing a saliency map at the coarse level. In the second-stage graph, along with the *adjacently spatial consistency*, a new *regionally spatial consistency* among graph nodes is considered in order to refine the coarse saliency map, assuring uniform saliency assignment even in complex scenes. Particularly, the second stage is generic enough to be integrated in existing salient object detectors, enabling to improve their performance. Experimental results on benchmark datasets validate the effectiveness and superiority of the proposed scheme over related state-of-the-art methods.**

***Index Terms*—Salient object detection, two-stage graphs, robust sparse representation, manifold ranking**

I. Introduction

SALIENCY detection aims to find the regions or objects catching human eye attention in a scene for further processing [1]. During the past two decades, research in this field has grown in two pathways: eye fixation prediction in human vision [2], [3] and salient object detection in computer vision [4]–[8]. The former topic focuses on identifying the fixation points of a human viewer at the first glance [9], [10], whereas the latter topic tends to locate or/and segment the most conspicuous objects from the scene [11]. Because of its low computational cost, salient object detection has emerged as a powerful image pre-processing tool in image segmentation [12], object recognition [13], image retrieval [14], image fusion [15], etc.

Recently, graph theory has been adopted in salient object detection [6]–[8], [16]–[19] due to its simplicity and efficiency. A typical graph-based saliency detection usually consists of three algorithmic components. First, a graph is constructed, including graph nodes[1], e.g., pixels [20], patches [19], or superpixels [6]–[8], and graph edges, i.e., node connections. Secondly, some seed nodes (i.e., background or foreground seed nodes) are selected to determine the initial saliency values of nodes [6]–[8], [19], [20]. Finally, saliency values of nodes are computed based on the initial saliency values via some methods, such as manifold ranking [6]–[8], [21], random walks ranking [21], etc.

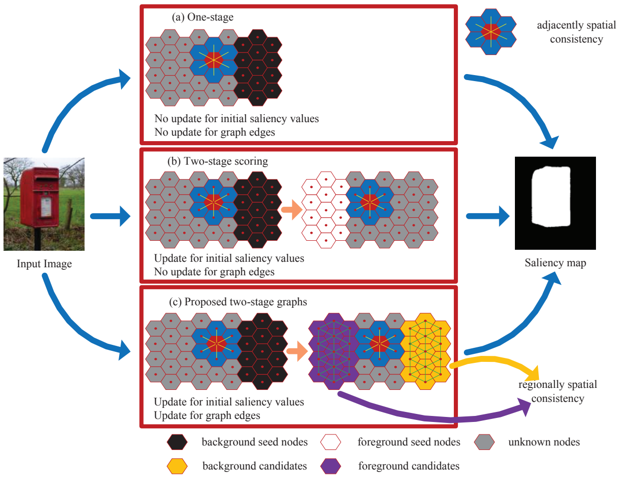

(a) One-stage
No update for initial saliency values
No update for graph edges
adjacently spatial consistency
Input Image
(b) Two-stage scoring
Update for initial saliency values
No update for graph edges
Saliency map
(c) Proposed two-stage graphs
Update for initial saliency values
Update for graph edges
regionally spatial consistency
background seed nodes
foreground seed nodes
unknown nodes
background candidates
foreground candidates

Fig. 1: Illustrations of the proposed two-stage graphs against the previous graph-based methods. Details are described in the text.

In the graph-based methods, graph construction is a vital issue, especially the graph edges that bridge the nodes. Most graph-based methods adopt a regular graph constructed by connecting each node with its neighbors [6]–[8], [19], in which the spatial consistency within a local neighborhood is adequately considered. Recently, the global contrast has been taken into account by connecting each node with the boundary nodes [8]. Moreover, any pairs of boundary nodes are connected to achieve a close-loop graph [6]–[8], [19]. Another important issue for the graph-based methods is the initial saliency values of nodes (also called selection of seed nodes). To this end, background seed nodes are usually abstracted to determine the initial saliency values of nodes from the boundary regions based on the boundary prior [6]–[8], [19]. While for other methods, the initial saliency values of nodes rely on the coarse detection results [6], [21].

Existing graph-based salient object detection methods can be divided into two categories: one-stage and two-stage scoring, as shown in Fig. 1(a) and (b). In the first category, as shown in Fig. 1(a), saliency is propagated via a one-stage process [8], [19]. The initial saliency values of nodes are determined merely by selecting background seed nodes,

Yi Liu and Qiang Zhang are with Key Laboratory of Electronic Equipment Structure Design, Ministry of Education, Xidian University, Xi'an Shaanxi 710071, China, and also with Center for Complex Systems, School of Mechano-Electronic Engineering, Xidian University, Xi'an Shaanxi 710071, China. Email: yLiu_89@stu.xidian.edu.cn, qzhang@xidian.edu.cn. Qiang Zhang is the corresponding author.

Jungong Han is with School of Computing and Communications, Lancaster University, Lancashire, LA1 4YW, U.K.. Email: jungonghan77@gmail.com.

Long Wang is with Center for Systems and Control, College of Engineering, Peking University, Beijing 100871, China. Email: longwang@pku.edu.cn.

[1] In this paper, we use node and superpixel interchangeably when we discuss the graph or the image.

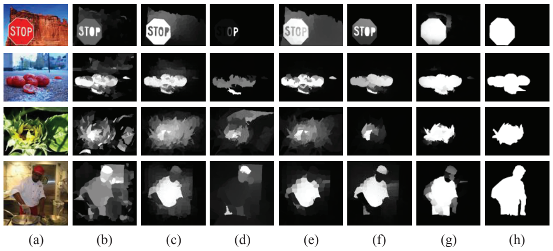


Fig. 2: Detection results of different graph-based salient object detection methods. (a) Original images; (b) GS; (c) BSCA; (d) RW_MR; (e) MR; (f) TLLT; (g) OUR; (h) Ground truth. (b) and (c) are one-stage based methods. (d) - (f) are two-stage scoring based methods.

which makes the whole system sensitive to the initial saliency values. This can easily end up mislabeling some backgrounds as foregrounds. Such an exemplar can be found in the last row of Fig. 2(b) and (e), where some backgrounds are wrongly detected as foregrounds.

In the second category, as shown in Fig. 1(b), saliency is computed via a two-stage scoring process [6], [21]. In these methods, the initial saliency values of graph nodes at the second stage are updated according to the coarse detection results obtained from the first stage, which certainly enhances the robustness of the initial saliency values (as shown in the last row in Fig. 2(d) and (f)).

More importantly, those graph-based methods only consider the *adjacently spatial consistency*, i.e., each node is connected to its local neighbors[2]. It is very clear that the two-stage scoring based methods employ only the *adjacently spatial consistency* at both stages, and also, the graph edges are not updated in the second-stage graph, which means essentially only a single graph is employed in the two-stage scoring based methods. This degrades the uniformity of the detected foregrounds and inevitably generates some “holes” in the detected salient objects. For instance, as shown in the first row of Fig. 2(b)-(f), the one-stage and two-stage scoring based methods achieve poor foreground uniformity in the nonhomogeneous regions. Especially, as shown in the second row of Fig. 2, although the salient objects have almost the same appearance within the inner regions, nonuniformity could still be encountered when such methods are applied. Furthermore, it is hard for the one-stage and two-stage scoring based methods to separate the foreground from the background completely and uniformly in the complex scene. For example, these methods fail to detect the foreground due to the similar appearance between foreground and background (as displayed in the third row of Fig. 2(b)-(f)) or complicated background (as displayed in the last row of Fig. 2(b)-(f)). Such undesirable detection results are attributed to the fact that the *adjacently spatial consistency* can reflect the relationships between nodes in the simple cases, but may fail in the nonhomogeneous regions or complex scenes.

In this paper, we tackle the above-mentioned problems from a coarse-to-fine perspective and propose a two-stage-graphs-based salient object detection method, in which *two graphs having the same nodes but different edges* are employed, as illustrated in Fig. 1(c). In the first-stage graph, which is analogous to most of existing graph-based methods, the *adjacently spatial consistency* is considered such that the spatial consistency within a local neighborhood can be preserved, based on which the saliency maps at the coarse level can be obtained. Once the coarse detection results are obtained, the graph nodes can be divided into three categories, i.e., potential foreground nodes, potential background nodes, and uncertain nodes. Therefore, the major task in the second-stage graph is to further determine the property of each graph node. To this end, a novel graph structure is presented, in which any pairs of potential foreground nodes (not necessarily neighboring superpixels) are connected, and any pairs of potential background nodes are likewise connected (we call this *regionally spatial consistency* among graph nodes). In other words, any pairs of potential foreground nodes are treated as neighbors, and any pairs of potential background nodes are treated as neighbors. In addition, each node is connected to its spatial neighbors (i.e., the *adjacently spatial consistency*). Consequently, in the second-stage graph, along with the *adjacently spatial consistency*, a new *regionally spatial consistency* among graph nodes is considered so as to refine the coarse saliency map, facilitating saliency detection in complex scenes. This obviously differs from the two-stage scoring based methods, in which only the *adjacently spatial consistency* among graph nodes is considered. In essence, two-stage scoring achieves a coarse-to-fine perspective by simply calculating the saliency values of nodes twice through two stages on the same graph. Differently, our two-stage graphs approach offers a novel coarse-to-fine perspective that employs coarse node connections in the first-stage graph followed by a node-connections refinement in the second-stage graph. Due to the introduction of *regionally spatial consistency*, our second-stage graph specifically promotes the foreground uniformity and background suppression, as can be seen in Fig. 2.

In our approach, the initial saliency values and graph edges (i.e., node connections) in the second-stage graph are mainly determined by the coarse detection results from the first-stage graph. Therefore, the computation of each node’s saliency value for the first-stage graph plays an important role in our proposed method. To this end, we propose a weighted joint robust sparse representation (WJRSR) model to compute the saliency value of each node in the first-stage graph, which is more robust to the initial saliency values of nodes (also called background dictionary in the sparse representation based methods) than the commonly used manifold ranking model [6], [21].

In short, the contributions of this paper are summarized as follows:

(1) Unlike existing graph-based methods that employ only a single graph, our major contribution lies in a two-stage-graphs-based salient object detection method, in which *two graphs having the same nodes but different edges* are employed. More importantly, in the second-stage graph of our proposed method, the *regionally spatial consistency* and *adjacently spatial consistency* among graph nodes are simultaneously

[2] It is noted that, in some graphs, each node is not only connected to the neighboring nodes, but also connected to the nodes sharing common boundaries with its neighboring nodes. We still call this type of node connections as the *adjacently spatial consistency*.

considered, thus facilitating saliency detection in complex scenes.

(2) The second contribution is a WJRSR model, which replaces the commonly used manifold ranking model [6], [21] to compute the saliency value of each node in the first-stage graph.

(3) Especially, the second stage in our proposed method is generic enough to be integrated in existing salient object detectors to improve their performance.

The reminder of this paper is organized as follows. Section II reviews the most related works. The proposed salient object detection model is described in Section III in detail. In Section IV, experiments are conducted to validate the effectiveness and superiority of the proposed method. Finally, Section V concludes the paper.

II. Related work

In the literature, a growing body of research has been devoted to salient object detection [4]–[8], [22]–[30]. In this section, we will review the works most related to ours, including sparse representation based methods and graph-based methods for salient object detection. Besides, deep convolutional neural networks (CNNs) based salient object detection has been a research hotspot recently, which will also be reviewed in this section.

A. Sparse Representation Based Salient Object Detection

Sparse representation (SR) theory has been applied in salient object detection due to its efficiency. SR based methods first construct an over-complete dictionary. Then, the input image is sparsely reconstructed by the dictionary. Saliency is measured according to the coding length or reconstruction errors. In [31], [32], the center patch was sparsely reconstructed by its surroundings, and saliency was measured by the coding length or residual. These methods usually assigned higher saliency values to the object boundaries, as the surroundings were already included in the dictionary. Afterwards, the image boundary regions were extracted as the background templates to sparsely reconstruct the image [4], [5]. Recently, a Laplacian regularization term was imposed on the sparse representation coefficients to take the local spatial consistency into account in [33]. In [22], a compact background dictionary was learned for sparse reconstruction, such that the background regions could be well reconstructed and could be discriminated from the foreground regions.

B. Graph-based Salient Object Detection

Graph theory is another important theory that was successfully applied in salient object detection. In [34], saliency values were computed based on the equilibrium distribution over map locations. Afterwards, saliency was measured by averaging the transmitted information in view of information maximization [35]. In [36], the authors improved visual attention by integrating the saliency and objectness into a graphical model. The salient object was segmented via a hierarchical model which efficiently utilized the concavity cue [16]. Recently, salient regions were detected by optimizing a submodular objective function that integrated the similarity and “facility” costs [17]. Saliency was computed via a Conditional Random Field aggregation model [37]. In [6], the image elements were ranked according to their similarities with the background and foreground cues. Saliency was propagated by using the teaching-to-learn and learning-to-teach strategies [18]. Alternatively, saliency was measured on a graph based on jointly considering the local consistency and global contrast [8], and saliency detection was conducted by a two-stage scoring scheme [21].

C. Deep Convolutional Neural Networks Based Salient Object Detection

Recently, deep convolutional neural networks (CNNs) have achieved many successes in salient object detection. In [38], the authors presented a neural network architecture, which had fully connected layers on top of CNNs responsible for feature extraction at three different scales. In [39], the authors proposed a CNNs-based salient object detection architecture working in a global-to-local and coarse-to-fine manner. In [40], a pixel-level fully convolutional stream and a segment-wise spatial pooling stream were designed to complement each other. In general, these methods achieve better performance than traditional methods.

III. Proposed Salient Object Detection

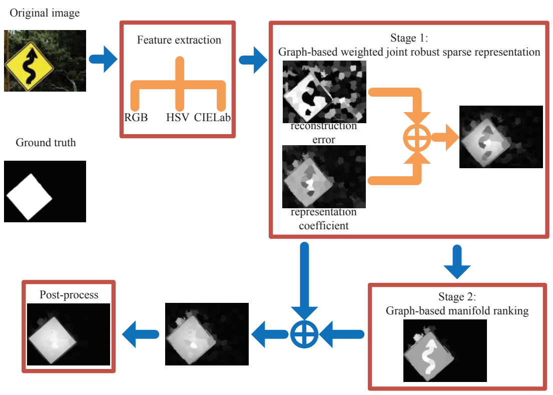
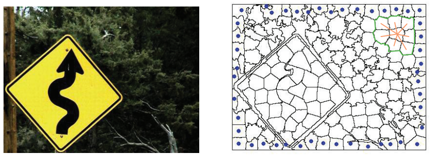

Original image
Feature extraction
RGB
HSV
CIELab
Stage 1:
Graph-based weighted joint robust sparse representation
reconstruction error
representation coefficient
Ground truth
Post-process
Stage 2:
Graph-based manifold ranking

Fig. 3: Diagram of the proposed salient object detection method.

In this section, we will describe the proposed salient object detection system. The diagram of the proposed method is illustrated in Fig. 3. The proposed system consists of four components: feature extraction; graph-based weighted joint robust sparse representation; graph-based manifold ranking; and post processing. Each part is elaborated below.

A. Feature Extraction

In our proposed method, we consider the superpixels instead of pixels as the image elements. The input image I is initially over-segmented into N superpixels by the simple linear iterative clustering (SLIC) algorithm [41] due to its simplicity and efficiency. For each superpixel, a feature vector $x_i \in R^m$ of dimension $m = 9$ is constructed, which covers the color



(a) (b)

Fig. 4: Graph $\mathbb{G}_1$ at the first stage. (a) Original image. (b) Graph $\mathbb{G}_1$. Any node is connected to its neighbors, as shown in the area delineated by the green curve. Besides, the boundary nodes are selected as the background seed nodes, as shown by the nodes marked by blue color in (b).

features of RGB, HSV, and CIELab. Traditionally, per-pixel feature vector is converted to per-superpixel feature vector through averaging. However, this appears to perform well only if the scene is characterized by simple color information and texture, whereas in nonhomogeneous regions and complex scenes it fails to maintain robustness. Instead, we obtain the per-superpixel feature vector by

$$x_i = \frac{1}{C}\sum_{j=1}^{n_i} w_{ji} f_j, \tag{1}$$

where $f_j \in R^{m\times 1}$ is the feature vector of the pixel j within the superpixel i. n_i is the number of pixels in the superpixel i. C is a normalization constant. w_{ji} is the distance weight and is computed by

$$w_{ji} = \exp\left(\frac{\|p_j - p_i\|_2^2}{2 * \sigma_p^2}\right), \tag{2}$$

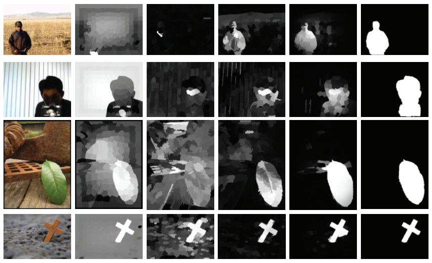
where p_i and p_j are the positions of the centers of the superpixel i and the pixel j within the superpixel i, respectively. σ_p is a scalar, and is set to $\sqrt{2}$.

Finally, horizontally stacking the feature vectors of all superpixels produces the feature matrix $X \in R^{m\times N}$ for the input image, i.e., $X = [x_1, x_2, \ldots, x_N] \in R^{m\times N}$.

B. Stage 1: Graph-based Weighted Joint Robust Sparse Representation

In this section, we will describe the first stage of the proposed method, which consists of three parts: graph construction, weighted joint robust sparse representation model, and saliency measure.

1) Graph Construction: At the first stage, an undirected regular graph is constructed $\mathbb{G}_1 = (\mathbb{V}_1, \mathbb{E}_1)$, where the nodes $\mathbb{V}_1$ are the superpixels. As illustrated in Fig. 4(b), each node is connected to its neighboring ones, which considers a spatial consistency within a local neighborhood, i.e., the *adjacently spatial consistency*, in light of the observation that a superpixel and its neighbors are likely to share similar appearance and thereby similar saliency values. In addition, the image boundary nodes, as shown in Fig. 4(b), are selected as the background seed nodes according to the boundary prior [19].



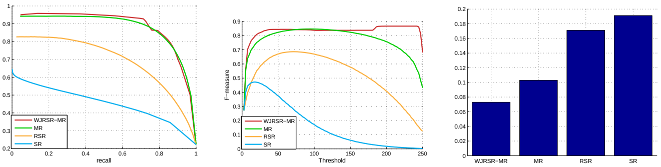
(a) (b) (c) (d) (e) (f)

Fig. 5: Illustrations of the proposed WJRSR against the traditional manifold ranking based and SR based methods. (a) Original images; (b) MR; (c) SR; (d) RSR; (e) Proposed WJRSR; (f) Ground truth. The boundary regions are selected as the background seed nodes (background dictionary).

The edge weights are defined as

$$w_{ij}^{G_1} = \begin{cases} \exp\left(-\frac{\|x_i - x_j\|_2^2}{2\sigma_{G_1}^2}\right), \; if\ i\ and\ j\ are\ connected \\ 0,\ otherwise. \end{cases}, \tag{3}$$

where σ_{G_1} is a scalar and is experimentally set to 5.



(a) PR Curves (b) F-measure Curves (c) MAE

Fig. 6: Quantitative comparisons of the proposed WJRSR model with MR, RSR, and SR models on MSRA10K. (a) PR Curves; (b) F-measure Curves; (c) Mean Absolute Error (MAE). The MSRA10K dataset, PR, F-measure, and MAE evaluation metrics will be discussed in Section VI.

2) Weighted Joint Robust Sparse Representation Model for the Computing of Saliency Values: Given the graph $\mathbb{G}_1$, the way the saliency values of nodes are calculated is of great importance. Most graph-based saliency detection methods adopt manifold ranking to compute the saliency value of each node [6]–[8], [21]. However, manifold ranking is sensitive to the initial saliency values of nodes. It becomes unreliable when the seed nodes are mixed with noise, producing undesirable detection results. For example, as shown in the first two rows of Fig. 5(b), when parts of foregrounds reach the image boundary, manifold ranking fails at identifying the foreground objects.

Instead, we apply a weighted joint robust sparse representation (WJRSR) model to compute the saliency values of nodes. More specifically, the proposed WJRSR model is based on the robust sparse representation (RSR) rather than the traditional sparse representation (SR). Compared to the traditional SR model, RSR model replaces the least squared errors with the sparse reconstruction errors [42], thus allowing the RSR model to be less sensitive to the selection of background seed nodes (also called background dictionary in the RSR model). As shown in the first two rows of Fig. 5(d) and (e), the RSR model and the proposed WJRSR model can identify most foreground regions and background regions, even if the salient object appears around the image boundary. Besides, when applied to the detection of salient objects, the RSR model possesses higher distinctiveness between the foreground objects and their backgrounds than the SR model. As shown in the last two rows of Fig. 5(d) and (e), the salient objects can be more uniformly highlighted and the background noise can also be better suppressed by the RSR model, as opposed to the SR model. It can be also found from Fig. 6 that the RSR model achieves better performance than the traditional SR model. Besides, the proposed WJRSR model ~~performs better than~~ the previous MR and RSR models. In the following, we will describe the WJRSR model in detail.

Joint Robust Sparse Representation (JRSR) Model. Based on the *adjacently spatial consistency* defined by the graph $\mathbb{G}_1$, the superpixel to be tested and its neighbors share similar appearances and thereby will have more or less the same saliency values. This implies that their representation coefficients when sparsely reconstructed using the same dictionary will look similar. As a result, a row-sparsity constraint is imposed on the representation coefficients of the superpixel to be tested and its neighbors, so that only few rows of the representation coefficients matrix are zero. Besides, the background seed nodes are taken as the background dictionary in the RSR model, which promotes the global contrast.

We horizontally stack the feature vectors of each superpixel to be tested and its neighbors, i.e., $X^i = [x_i, x_{i-1}, x_{i-2}, \ldots, x_{i-N_i}]$, where $x_{i-1}, x_{i-2}, \ldots, x_{i-N_i}$ are the feature vectors of the neighboring superpixels belonging to the superpixel i. Based on the above discussions, we formulate a joint robust sparse representation (JRSR) model for computing the saliency value of the superpixel i as:

$$\begin{aligned}&\min_{Z^i,E^i} \left\|Z^i\right\|_{1,2} + \lambda\left\|E^i\right\|_{2,1}\\ &s.t.\ X^i = DZ^i + E^i\end{aligned}. \tag{4}$$

Here, D is the background dictionary, i.e., the background seed nodes. Z^i and E^i are the representation coefficients matrix and reconstruction errors matrix, respectively. $\left\|Z^i\right\|_{1,2}$ is the $l_{1,2}$- norm of Z^i, and defined as $\left\|Z^i\right\|_{1,2} = \sum\limits_j \left\|Z^i(j,:)\right\|_2$, where $Z^i(j,:)$ is the jth row of Z^i. $\left\|Z^i(j,:)\right\|_2$ is the l_2-norm of $Z^i(j,:)$. $\left\|E^i\right\|_{2,1}$ is the $l_{2,1}$- norm of E^i, and defined as $\left\|E^i\right\|_{2,1} = \sum\limits_k \sqrt{\sum\limits_j \left(E^i(j,k)\right)^2}$, where $E^i(j,k)$ is the (j,k)th entry of E^i. Similar to X^i, Z^i and E^i are the matrices which horizontally stack the representation coefficients vectors and reconstruction errors vectors of the superpixel i and its neighbors, respectively, i.e., $Z^i = [z_i, z_{i-1}, z_{i-2}, \ldots, z_{i-N_i}]$ and $E^i = [e_i, e_{i-1}, e_{i-2}, \ldots, e_{i-N_i}]$. $\left\|Z^i\right\|_{1,2}$ achieves the *adjacently spatial consistency*, and enforces the superpixel to be tested and its adjacent superpixels to be with similar representation coefficients under the same dictionary.

Weighted JRSR (WJRSR) Model. As defined by Eq. (4), the representation coefficients for each superpixel and its neighboring ones are assumed to be similar if the "row-sparsity" constraint is directly imposed on the coefficient matrix in the JRSR model. This assumption seems reasonable for the homogeneous regions, but it is no longer valid for those nonhomogeneous regions, especially for the object boundaries. For example, as shown in Fig. 7(b), some backgrounds (foregrounds) are mistakenly labeled as foregrounds (backgrounds) at the object boundary. To achieve both the diversity for the nonhomogeneous regions and the consistency for the homogeneous regions, we introduce a weight matrix Q^i, leading to a weighted JRSR (WJRSR) model:

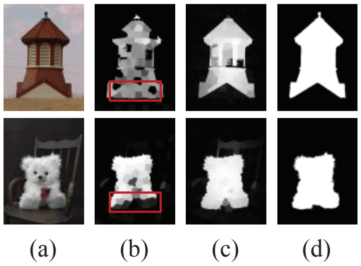


Fig. 7: Illustrations of the superiority of the WJRSR model over the JRSR model. (a) Original images; (b) Proposed JRSR model in Eq. (4); (c) Proposed WJRSR model in Eq. (5); (d) Ground truth.

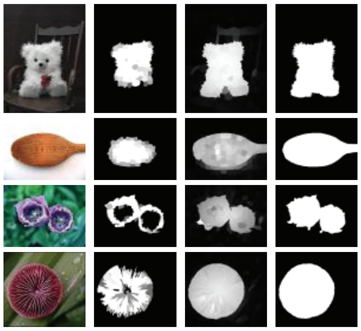
$$\begin{aligned}&\min_{Z^i,E^i} \left\|Z^iQ^i\right\|_{1,2} + \lambda\left\|E^i\right\|_{2,1}\\ &s.t.\ X^i = DZ^i + E^i\end{aligned}, \tag{5}$$

where Q^i is defined as

$$Q^i = \begin{bmatrix} 1 & 0 & \cdots & 0 \\ 0 & q_{i-1} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & q_{i-N_i} \end{bmatrix}. \tag{6}$$

Here, $q_{i-k}, k = 1, 2, \ldots, N_i$ is the feature similarity measure of the superpixel i and its neighboring superpixel k, and is computed by Eq. (3). Therefore, the row-sparsity constraint enforces the representation coefficients associated with each superpixel to be tested and its neighboring ones to be similar in case they share similar appearance, but not vice versa. This not only imposes the consistency for the homogeneous regions, but also preserves the diversity for the nonhomogeneous regions. This is different from the JRSR model defined by Eq. (4). Thanks to this scheme, we can observe from Fig. 7(c) that the WJRSR model accurately detects the foreground regions and background regions even at the object boundary.

Moreover, the proposed WJRSR model is different from WSC [43]. First, WSC [43] is based on the SR model, while our proposed WJRSR model is based on the RSR model, which is superior to the SR model for salient object detection (as discussed in Fig. 5). Secondly, the penalty weights in WSC [43] are defined to be inversely proportional to the appearance similarities between the superpixels to be tested and the dictionary atoms. In contrast, our proposed WJRSR model defines the weights as the appearance similarities between the superpixels to be tested and their adjacently spatial neighbors. Thirdly, WSC [43] computes the saliency value of each superpixel independently, while our proposed WJRSR model considers the *adjacently spatial consistency* among superpixels. The aforementioned differences make the proposed WJRSR model more powerful than WSC [43]. For example, as shown in the first two rows of Fig. 8, WSC [43]



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

Fig. 8: Illustrations of the superiority of the proposed WJRSR model over WSC [43]. (a) Original images; (b) WSC; (c) WJRSR; (d) Ground truth.

obtains incomplete detection results, whereas our proposed WJRSR model gets more wholeness of the salient objects. It can also be obviously found from the last two rows of Fig. 8 that our proposed WJRSR model achieves better foreground uniformity than WSC [43] does.

Optimization. For the sake of clarity, we remove the subscript index of the matrices in Eq. (4), and the WJRSR model in Eq. (5) can be rewritten as

$$\begin{aligned} &\min_{\tilde{Z},\tilde{E}} \left\|\tilde{Z}\tilde{Q}\right\|_{1,2} + \lambda\left\|\tilde{E}\right\|_{2,1} \\ &s.t.\ \tilde{X} = D\tilde{Z} + \tilde{E} \end{aligned} . \tag{7}$$

This optimization model is convex and can be solved efficiently. We first convert it to the following equivalent problem:

$$\begin{aligned} &\min_{J,\tilde{Z},\tilde{E}} \|J\|_{1,2} + \lambda\left\|\tilde{E}\right\|_{2,1} \\ &s.t.\ \tilde{X} = D\tilde{Z} + \tilde{E}, \\ &\quad\ \tilde{Z}\tilde{Q} = J \end{aligned} . \tag{8}$$

In this paper, to optimize the objective function defined in Eq. (8), we adopt the ADMM method [44] which minimizes the following augmented Lagrange function:

$$\begin{aligned} L = {} & \|J\|_{1,2} + \lambda\left\|\tilde{E}\right\|_{2,1} + \left\langle Y_1, \tilde{X} - D\tilde{Z} - \tilde{E}\right\rangle \\ & + \left\langle Y_2, \tilde{Z}\tilde{Q} - J\right\rangle + \frac{\mu}{2}\left(\left\|\tilde{X} - D\tilde{Z} - \tilde{E}\right\|_F^2 + \left\|\tilde{Z}\tilde{Q} - J\right\|_F^2\right) \end{aligned}, \tag{9}$$

where Y_1 and Y_2 are Lagrange multipliers, and $\mu > 0$ is a penalty parameter.

The optimization procedure is outlined in Algorithm 1. The detailed solving process is shown in Supplementary Material.

Through Algorithm 1, we can obtain the optimal representation coefficients matrix Z^{i^*} and reconstruction errors matrix E^{i^*} for X^i. Then, the optimal z_i^* and e_i^* are extracted from Z^{i^*} and E^{i^*}, respectively. Similarly, we are able to get the optimal representation coefficients vectors and reconstruction errors vectors corresponding to the other superpixels. Thus, we can obtain the optimal representation coefficients matrix $Z^* = [z_1^*, z_2^*, \ldots, z_N^*]$ and the optimal reconstruction errors matrix $E^* = [e_1^*, e_2^*, \ldots, e_N^*]$ for the input image.

Algorithm 1 Solving the optimization model in Eq. (9).

```
Input: Feature matrix $\tilde{X}$, weight matrix $\tilde{Q}$, and parameter $\lambda$.
Output: $\tilde{Z}$ and $\tilde{E}$.
intialize: $\tilde{Z} = \mathbf{0}$, $\tilde{E} = \mathbf{0}$, $Y_1 = \mathbf{0}$, $Y_2 = \mathbf{0}$, $\mu = 1$, $\mu_{\max} = 10^{10}$, and $\rho = 1.1$.
repeat
    Fix the others and update $J$:
        $J = \arg\min_J \frac{1}{\mu}\|J\|_{1,2} + \frac{1}{2}\left\|J - (\tilde{Z}\tilde{Q} + \frac{1}{\mu}Y_2)\right\|_F^2$.
    Fix the others and update $\tilde{Z}$ by updating each column of $\tilde{Z}$:
        $\tilde{z}_k = (A + \tilde{Q}_{i,i}I)^{-1}c_i$,
    where $A = \tilde{Q}^{-1}D^TD$, $\tilde{Q}_{i,i}$ is the $(i,i)$ entry of $\tilde{Q}$, and $c_i$ is the $i$th column of the matrix $C$, which is formulated as
        $C = \tilde{Q}^{-1}\left[D^T\left(\tilde{X} - \tilde{E} + \frac{Y_1}{\mu}\right) + \tilde{Q}(J - \frac{Y_2}{\mu})\right]$.
    Fix the others and update $\tilde{E}$:
        $\tilde{E} = \arg\min_{\tilde{E}} \frac{\lambda}{\mu}\|E\|_{2,1} + \frac{1}{2}\left\|\tilde{E} - \left(\tilde{X} - D\tilde{Z} + \frac{Y_1}{\mu}\right)\right\|_F^2$.
    Update the multipliers:
        $Y_1 = Y_1 + \mu\left(\tilde{X} - D\tilde{Z} - \tilde{E}\right)$.
        $Y_2 = Y_2 + \mu\left(\tilde{Z}\tilde{Q} - J\right)$.
    Update the parameter $\mu$: $\mu = \min(\rho\mu, \mu_{\max})$
until Convergence: $\tilde{X} - D\tilde{Z} - \tilde{E} \to \mathbf{0}$ and $\tilde{Z}\tilde{Q} - J \to \mathbf{0}$
```

3) Saliency Measure: In this part, we will describe how we define the saliency measures based on the reconstruction errors and representation coefficients.

Saliency Measure Based on Reconstruction Errors. Given a background dictionary D, each column of the optimal sparse errors matrix E^* may contain the salient information of each superpixel that is distinct from the background. Generally, a superpixel will be more salient if it has larger reconstruction errors with respect to the background dictionary. Hence, we define the reconstruction errors based saliency measure $sal_E(i)$ for the superpixel i as

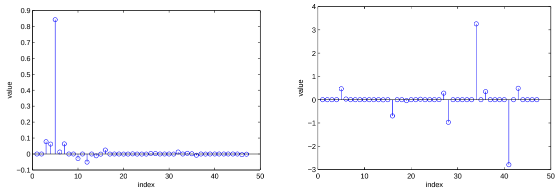
$$sal_E(i) = 1 - \exp\left(-\frac{\|E^*(:,i)\|_2^2}{2\sigma_E^2}\right), \tag{10}$$

where σ_E is a scalar parameter and is experimentally set to 1.

Saliency Measure Based on Representation Coefficients. In addition to the reconstruction errors, the saliency value of each superpixel can also be determined by its representation coefficients to some extent. For example, as shown in Fig. 9, when sparsely reconstructed by the same background dictionary, a background superpixel gains its representation coefficients with low energy, while a foreground superpixel gains its representation coefficients with high energy. This is because the background dictionary has lower contrast with the background superpixel but higher contrast with the foreground superpixel. Based on the observations, we define the representation coefficients based saliency measure $sal_Z(i)$ for the superpixel i as

$$sal_Z(i) = 1 - \exp\left(-\frac{\|Z^*(:,i)\|_2^2}{2\sigma_Z^2}\right), \tag{11}$$

where σ_Z is a scalar and is experimentally set to $\sqrt{2}$.



(a) (b)

Fig. 9: Comparisons of representation coefficients between (a) background superpixel and (b) foreground superpixel.

The two saliency measures sal_E and sal_Z are integrated, resulting in the saliency measure for the superpixel i

$$sal^{E\_Z}(i) = \alpha * sal_E(i) + (1-\alpha) * sal_Z(i), \tag{12}$$

where $\alpha \in (0,1)$ is a balance weight and is experimentally set to 0.2.

The pixel-level saliency map is obtained by equaling the saliency value of each pixel to that of its corresponding superpixel. In order to suppress the noise, the object-based Gaussian model [4], [5] is further applied to refine the saliency detection results. In the subsequent processing at the second stage, we again transform the pixel-level saliency map to superpixel-level saliency map. Each superpixel achieves its saliency value by averaging the saliency values of the pixels within it. The refined results are denoted as sal^{G_1}.

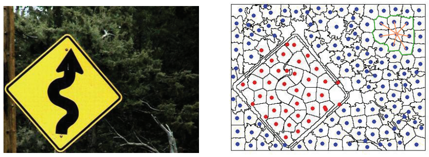
C. Stage 2: Graph-based Manifold Ranking

In this section, we will describe the second stage of the proposed method, which consists of graph construction and graph-based manifold ranking.

1) Graph Construction: The coarse detection results obtained from the first stage help to locate the potential background region R_B and foreground region R_F if a low threshold $thre_{low}$ and a high threshold $thre_{high}$ are set. Those superpixels with saliency values lower than $thre_{low}$ are labeled as background ones, while those supeprixels with saliency values higher than $thre_{high}$ are labeled as foreground ones. To ensure the accuracy of the potential background and foreground regions detection, we set $thre_{low} = 0.8 * mean(sal^{G_1})$ and $thre_{high} = 2 * mean(sal^{G_1})$. Based on the potential background and foreground regions, we construct a novel undirected graph $\mathbb{G}_2 = (\mathbb{V}_2, \mathbb{E}_2)$, where nodes are the superpixels. As shown in Fig. 10(b), the edges of this graph are composed of three parts:

$\mathbb{E}_2^1$: Each node is connected to its neighboring nodes.

$\mathbb{E}_2^2$: Any pairs of the potential background nodes are connected.



(a) (b)

Fig. 10: Graph $\mathbb{G}_2$ at the second stage. (a) Original image. (b) Graph $\mathbb{G}_2$. The nodes marked by red color represent the foreground superpixels and the nodes marked by blue color represent background superpixels. In this graph, each node is connected to its neighboring nodes, as shown in the area delineated by the green curve. Besides, any pairs of the red nodes are connected, and any pairs of the blue nodes are connected, which are not marked for clarity.

$\mathbb{E}_2^3$: Any pairs of the potential foreground nodes are connected.

$\mathbb{E}_2^1$ considers the local consistency within a local neighborhood, i.e., the *adjacently spatial consistency*, which plays the same role as that employed in the graph $\mathbb{G}_1$ at the first stage. $\mathbb{E}_2^2$ treats any pairs of R_B to be adjacent, which enforces a consistency among the potential background nodes, resulting in the background uniformity. $\mathbb{E}_2^3$ treats any pairs of R_F to be adjacent, which enforces consistency among the potential foreground nodes, leading to the foreground uniformity. $\mathbb{E}_2^2$ and $\mathbb{E}_2^3$ impose the *regionally spatial consistency* within the background candidates and within the foreground candidates, respectively. Furthermore, combining $\mathbb{E}_2^2$ and $\mathbb{E}_2^3$ can additionally enhance the discrimination between background and foreground, thus improving the separation of the foreground from the background. Similar to $\mathbb{G}_1$, the edge weight between two nodes on $\mathbb{G}_2$ is defined as

$$w_{ij}^{G_2} = \begin{cases} \exp\left(-\frac{\|x_i - x_j\|_2^2}{2\sigma_{G_2}^2}\right), & if \quad i \ and \ j \ are \ connected \\ 0, \ otherwise. \end{cases} . \tag{13}$$

Here, σ_{G_2} is a scalar and is experimentally set to $\sqrt{8}$, which is different from the first stage.

2) Graph-based Manifold Ranking: At this stage, we initialize the saliency value of each node $y = [y_1, y_2, \ldots, y_N]^T$ with the saliency value of the coarse detection conducted in the first stage, i.e.,

$$y_i = sal^{G_1}(i), \ i = 1, 2, \ldots, N. \tag{14}$$

Given the weight matrix $W^{G_2} = \left[w_{ij}^{G_2}\right]_{N \times N}$, we can obtain the degree matrix $D^{G_2} = diag\,(d_1, d_2, \ldots, d_N)$, where $d_i = \sum\limits_i w_{ij}^{G_2}$. Let f be the ranking function assigning rank values $f = [f_1, f_2, \ldots, f_N]^T$. This can be obtained by solving the following minimization problem:

$$\begin{aligned} f^* = \arg\min_f \frac{1}{2} \sum_{i,j=1}^{N} w_{ij}^{G_2} \left\| \frac{f_i}{\sqrt{d_i}} - \frac{f_j}{\sqrt{d_j}} \right\|_2^2, \\ + \frac{\beta}{2} \sum_{i=1}^{N} \|f_i - y_i\|_2^2 \end{aligned} \tag{15}$$

where β is a controlling parameter. The optimized solution is given in [45] as:

$$f^* = \left(D^{G_2} - \gamma W^{G_2}\right)^{-1} y, \tag{16}$$

where $\gamma = \frac{1}{1+\beta}$, and γ is set to 0.99.

Then, the saliency value of the superpixel i at the second stage is

$$sal^{G_2}(i) = f^*(i). \tag{17}$$

Furthermore, sal^{G_1} and sal^{G_2} may contain noise in both foreground and background regions, we integrate the detection results of the two stages to complement each other by

$$sal^{G_1\_G_2} = \frac{sal^{G_1} + sal^{G_2}}{2}. \tag{18}$$

D. Post Process

The pixel-level saliency map is first obtained from $sal^{G_1\_G2}$ by setting the saliency value of each superpixel to that of the pixels within the superpixel. An enhanced pixel-level saliency map $sal^{enhance}$ is then obtained by using the following enhancement function:

$$g(x) = x + sgn(x - \varepsilon) * \exp(-\frac{x - \varepsilon}{2\sigma^2_{enhance}}), \tag{19}$$

where ε is an adaptive threshold based on the Otsus binary threshold method [46]. $\sigma_{enhance}$ is a predefined parameter to control the level of contrast, and is set to 1. $sgn(\cdot)$ is a sign function. Here, x denotes the saliency value of a pixel.

Generally, salient object detection is essentially a binary segmentation problem [47] that extracts the entire salient objects from the background. To advance the binary segmentation, we apply the Max-Flow method [48] on $sal^{enhance}$ to generate a foreground mask sal^{MF}. Similarly, considering that the binary saliency map sal^{MF} may also contain noise in both foreground and background regions, we get the final saliency map as formulated by

$$sal^{post} = \frac{sal^{enhance} + sal^{MF}}{2}. \tag{20}$$

It should be noted that the post process can actually improve the performance of the proposed method to some extent. However, ~~this~~ depends on the pre-detection results of the proposed two-stage graphs before the post process. In other words, the post process will improve the performance in case the pre-detection results are good, but will ~~degrade the performance~~ otherwise. For example, as shown in Fig. 11, the post-process operation improves the performance when the proposed two-stage graphs achieve good detection results (as shown in the first two rows of Fig. 11), but ~~degrade~~ the performance when the proposed two-stage graphs achieve unsatisfactory detection results (as shwon in the last two rows of Fig. 11).

Moreover, compared with the post process in [20], [49], the proposed post-process can better improve the performance of salient object detection. As shown in Fig. 12, some foreground regions are suppressed instead of being promoted to some extent by the post-process in [20], [49]. While the foreground regions and the background regions are further promoted and suppressed, respectively, by our proposed post-process.

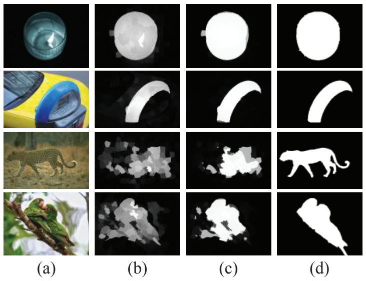


Fig. 11: Illustrations of the improvements of the post-process for the performance of the proposed method. (a) Original images; (b) Saliency maps obtained by the proposed two-stage graphs without post-process; (c) Saliency maps refined by the post-process; (d) Ground truth.

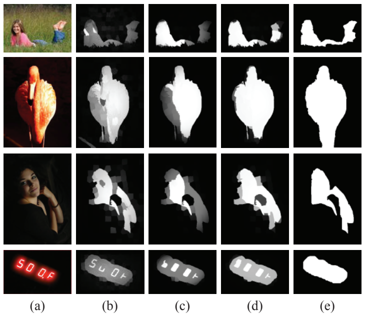


Fig. 12: Illustrations of superiority of the proposed post-process. (a) Original images; (b) Saliency maps obtained by the proposed two-stage graphs without post-process; (c) Saliency maps refined by the post-process in [20], [49]; (d) Saliency maps refined by the proposed post-process (e) Ground truth.

E. Summary

To recapitulate, the proposed salient object detection method is summarized as follows:

(1) Extract features for each superpixel by Eq. (1);

(2) Compute the coarse saliency map sal^{G_1} at the first stage;

(3) Compute the saliency map sal^{G_2} at the second stage;

(4) Compute the integrated saliency map $sal^{G_1\_G_2}$;

(5) Obtain the final saliency map sal^{post} via post process operations.

Fig. 13 illustrates the saliency detection results obtained by the main phases of the proposed method. As shown in Fig. 13(d), most background and foreground regions are identified in the first stage, and they become more uniform and discriminative through the refinement in the second stage (See the example in Fig 13(e)).

F. Complexity Analysis

Firstly, the computational complexity at the first stage can be analyzed as follows: suppose the data matrix X and dictionary D are with the sizes of $m \times N$ and $m \times K$, respectively. Then, the coefficients matrix Z has size of $K \times N$. As discussed in [42], the computational complexity of Algorithm 1 at the first stage is mainly determined by the computation burden of updating the matrix Z. Theoretically, the computational complexity of Algorithm 1 is $O(rK^3N)$, where r is the number of iterations needed for convergence[3]. It demonstrates

[3] Note that we assume that all the iterations for updating $Z^i, i = 1, 2, \ldots, N$ are approximately equal to r here.

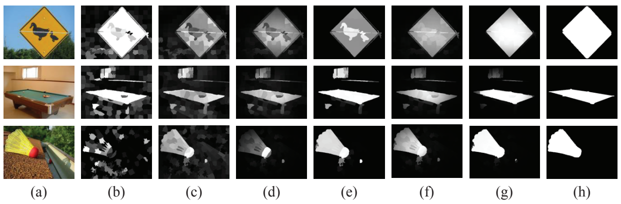


Fig. 13: Saliency maps obtained by the main phases of the proposed method. (a) Original image; (b)-(d) Saliency maps obtained from the first stage: (b) based on the reconstruction errors; (c) based on the representation coefficients; (d) by fusing (b) and (c); (e) Saliency maps obtained from the second stage; (f) Saliency maps by fusing (d) and (e); (g) Final saliency maps via post process operations; (h) Ground truth.

that the number of dictionary atoms K has a greater impact on the computational complexity of the proposed Algorithm 1 than the other parameters. In the proposed method, K is set to the number of boundary superpixels (about 49) and is far smaller than the total number of superpixels N (about 200). This makes the computational cost of the proposed Algorithm 1 acceptable.

Next, we will discuss the computational complexity at the second stage. The computational complexity of the manifold ranking model at the second stage mainly depends on the matrix inverse operation in Eq. (16). The matrices D^{G_2} and W^{G_2} are both with the size of $N \times N$. The computational complexity of the manifold ranking model is thus $O(N^3)$. Therefore, the total computational complexity of our proposed method is $O(rK^3N) + O(N^3)$.

IV. Experiments and Analysis

In this section, a number of experiments are conducted to validate the effectiveness and superiority of the proposed salient object detection method.

Datasets. We evaluate the proposed method on three benchmark datasets, including MSRA10K [50], ECSSD [51], and DUT-OMRON [6], [7]. MSRA10K [50] contains 10000 images with simple scene, most of which contain a single object with high contrast to background. ECSSD [51] contains 1000 images with structurally complex scene. Most images in this dataset contain multiple objects belonging to various categories. DUT-OMRON [6], [7] contains 5168 images with cluttered background, most of which have one or more objects with different scales and locations.

Evaluation metrics. Multiple widely used evaluation metrics are used to evaluate the proposed method, including precision-recall curve [52], F-measure [52], mean absolute error (MAE) [53]. Here, the precision value is defined as the ratio of salient pixels correctly assigned to all pixels of the extracted regions, while the recall value refers to the percentage of detected salient pixels with respect to the ground truth data. For a saliency map, we generate a set of binary images by using different thresholding values in the range of $[0, 1]$. The precision/recall pairs of all the binary maps are computed to plot the precision-recall curve [52]. F-measure is used as the overall performance measure:

$$F = \frac{(1+\beta^2) * precision * recall}{\beta^2 * precision + recall}, \tag{21}$$

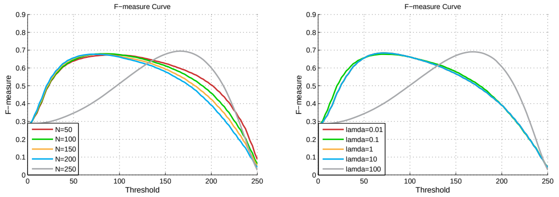
where $\beta^2 = 0.3$ as suggested in [52] to emphasize the precision. The MAE computes the average difference between the saliency map and the ground truth [53].

A. Parameters Setup

The superpixel number N has an important impact on the performance of the proposed method. Besides, the parameter λ in Eq. (4) balances the two constraints of the proposed WJRSR model. We set the two important parameters by fixing one and tuning the other on ECSSD within the first stage of our proposed method.

It can be seen from Fig. 14(a) that WJRSR gets good performance when $N = 50, 100, 150$, and 200. However, the F-measure curve gets high values over a suddenly narrow range when $N = 250$. This indicates that the detection algorithm poorly distinguishes the foreground from the background, which would result in inaccurate location of potential foreground and background regions. Considering that more superpixels are beneficial to detecting the salient object in the case of nonhomogeneous regions and complex scene, we set $N = 200$.

From Fig. 14(b), it can be viewed that WJRSR performs similarly when $\lambda = 0.01, 0.1, 1$ and 10. But the F-measure curve is obviously poor when $\lambda = 100$, which degrades the localization accuracy of potential foreground and background regions. In the following experiments, we set $\lambda = 0.1$.



(a) F-measure Curves on N(b) F-measure Curves on λ

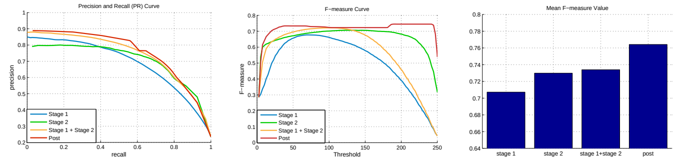
Fig. 14: Illustrations of parameters setting. It is noted that the F-measure curves for $\lambda = 0.01$, $\lambda = 1$, and $\lambda = 10$ overlap in (b).

B. Performance Comparisons on Each Stage

Fig. 15 provides the detection results of each stage in our proposed method on ECSSD. It is obvious from Fig. 15 that compared with “Stage 1”, “Stage 2” achieves higher PR curve when recall value is greater than 0.4 (Fig. 15(a)), much higher and wider F-measure curve (Fig. 15(b)), and higher mean F-measure value (Fig. 15(c)). Besides, since ~~that there exists a complementarity between “Stage 1” and “Stage 2”,~~ the performance is further improved by integrating the detection results of the two stages. And thus “Stage 1 + Stage 2” obtains better performance than “Stage 1” and “Stage 2”, which can be easily seen from Fig. 15. Finally, a simple but effective post process is performed on “Stage 1 + Stage 2” to further promote performance.

Fig. 16 gives some visual examples of each stage in our proposed method. It is obvious from Fig. 16(b) that “Stage 1” can accurately locate the foreground objects but with unsatisfactory uniformity. This is well addressed by “Stage

2” with much better foreground uniformity and background suppression, which is easily seen from Fig. 16(c). Besides, it can be found from Fig. 16(d) that “Stage 1 + Stage 2” makes “Stage 1” and “Stage 2” complement each other to achieve more accurate saliency maps. The detection results are more close to ground truth with the help of the post process.



(a) PR Curves (b) F-measure Curves (c) Mean F-measure

Fig. 15: Performance comparisons of each stage on ECSSD. “Stage 1” represents the saliency map sal^{G_1} obtained from the first stage. “Stage 2” represents the saliency map sal^{G_2} obtained from the second stage. “Stage 1 + Stage 2” represents the integrated saliency map $sal^{G_1\text{-}G_2}$. “Post” represents the final saliency map sal^{post}. Mean F-measure value is computed with an adaptive threshold, i.e., $thre = \frac{2}{W \times H} \sum_{i=1}^{W} \sum_{j=1}^{H} S(i,j)$, where $S(i,j)$ represents the saliency value of the (i,j)-th pixel in the saliency map S.

C. *Comparisons with State-of-the-art Methods*

In this section, we validate the effectiveness and superiority of the proposed method via visual and quantitative comparisons with 20 state-of-the-art methods, including MST [20], TLLT [18], BSCA [54], DSR [5], WSC [43], MBD [49], MR [7], RBD [55], HS [51], PCA [56], TD [57], GC [58], DCLC [59], RW_MR [21], MAP [60], GS [19], HCT [61], BL [62], DRFI [63], and MILPS [64]. Among these methods, GS [19], BSCA [54], RBD [55], MST [20], RW_MR [21], MR [7], TLLT [18], MAP [60], and DCLC [59] are graph based state-of-the-art methods. Specifically, GS [19], BSCA [54], RBD [55], and MST [20] are one-stage based methods. RW_MR [21], MR [7], TLLT [18], MAP [60], and DCLC [59] are two-stage scoring based methods. The other methods, i.e., MBD [49], HS [51], WSC [43], DSR [5], HCT [61], BL [62], DRFI [63], MILPS [64], are other state-of-the-art methods.

1) Visual Comparisons on Several Types of Images: To efficiently validate the effectiveness and superiority of the proposed method, we compare the proposed method with the state-of-the-art methods on several types of images. Fig. 17 - Fig. 22 show the visual comparisons on different methods for those images with a single object, multiple objects, large object, object touching the image borders, similar appearance between background and foreground, and complex scene, respectively. Most methods deliver good results in the simple cases, such as those images with a single object (in Fig. 17), but fail to produce satisfactory results in more complex cases. In contrast, the proposed method can not only extract the salient object accurately for those images with a single object, but also offers pretty good detection in complex scenes. Especially, it can be found that the proposed two-stage graphs achieves much better performance in foreground uniformity as well as background suppression.

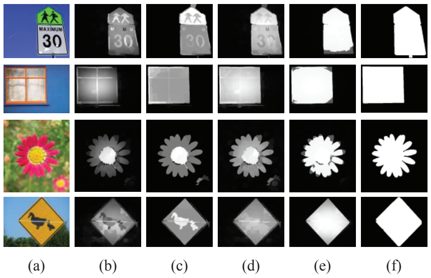


Fig. 16: Detection results of each stage. (a) Original images; (b) Stage 1; (c) Stage 2; (d) Stage 1 + Stage 2; (e) Post; (f) Ground truth. Please refer to Fig. 15 for the explanations of “Stage 1”, “Stage 2”, “Stage 1 + Stage 2”, and “Post”.

For those images with multiple objects (in Fig. 18), our proposed method can extract all the salient objects. Especially, as shown in the second and fourth rows of Fig. 18, those multiple salient objects can also be well separated from background by using our proposed method. For those images with large objects (in Fig. 19), our proposed method is able to detect the entire salient object, whereas most of the other methods just detect parts of the salient objects. For those images with salient object touching the image borders (in Fig. 20), it is difficult to segment the entire object, especially for those boundary prior based methods, i.e., BSCA [54], RW_MR [21], MR [7], RBD [55], MBD [49], WSC [43], DSR [5], and MST [20]. In contrast, our proposed method can still extract the entire object, which may owe to the WJRSR model at our first stage. For those images with similar appearance between background and foreground (in Fig. 21), our proposed method can successfully separate the foreground from the background. On the contrary, such similar appearances confuse most of the other algorithms. For those images with complex scene (in Fig. 22), our proposed method can identify the salient object pretty accurately, but most of the other methods fail, especially in the second row of Fig. 22.

2) Quantitative Comparisons: Fig. 23 provides the PR and F-measure curves on different methods. From Fig. 23, it is clear that the proposed method is competitive with DRFI [63] and MILPS [64], and performs better than the other methods in terms of the PR curves for MSRA10K and ECSSD. It also demonstrates that the proposed method scores the best for the three benchmark datasets in terms of F-measure curves based on the fact that the proposed method obtains the highest F-measure values over the widest range for the three datasets. This also indicates that the saliency values for the foreground regions obtained by the proposed method are relatively larger, while those for the background regions are smaller. As a result, the separation of the foreground regions from the background regions is more robust to the thresholding values by using the proposed method than other methods. Moreover, as shown in Table I, it is obvious that the proposed method achieves the smallest MAE score among all the aforementioned methods for the three benchmark datasets, which implies that the detection results by the proposed method are the closest to the ground truth. Especially, it is obvious that the proposed method outperforms other graph-based ones. This also efficiently verifies the superiority of the two-stage graphs over the previous one-stage process and two-stage scoring.

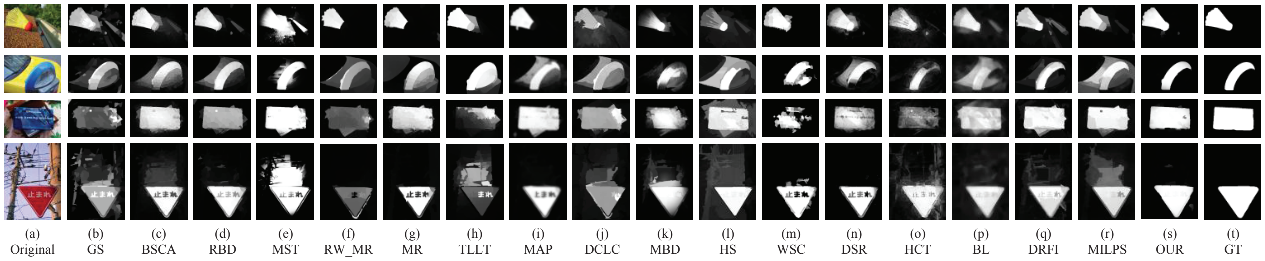


Fig. 17: Visual comparisons on different methods for those images with a single object. (b)-(e) are one-stage based methods. (f)-(j) are two-stage scoring based methods. (k)-(r) are other state-of-the-art methods.

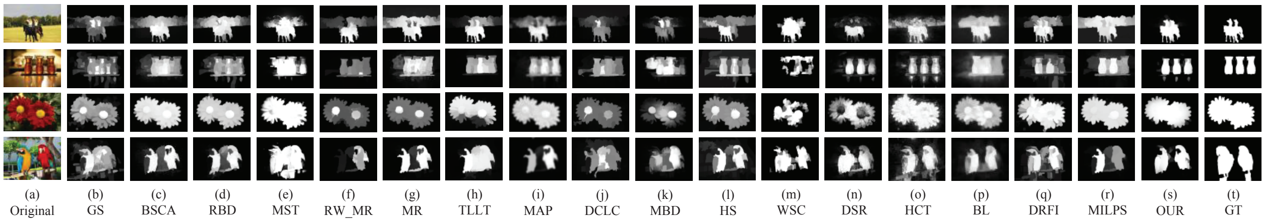


Fig. 18: Visual comparisons on different methods for those images with multiple objects. (b)-(e) are one-stage based methods. (f)-(j) are two-stage scoring based methods. (k)-(r) are other state-of-the-art methods.

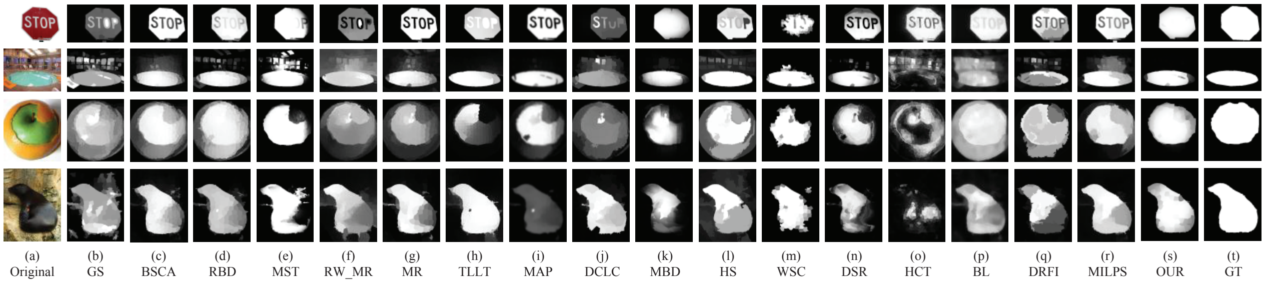


Fig. 19: Visual comparisons on different methods for those images with large object. (b)-(e) are one-stage based methods. (f)-(j) are two-stage scoring based methods. (k)-(r) are other state-of-the-art methods.

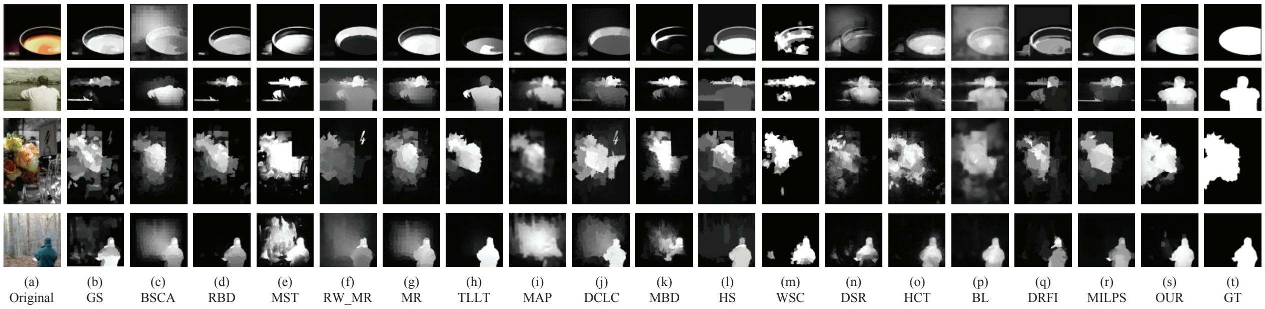


Fig. 20: Visual comparisons on different methods for those images with object touching the image borders. (b)-(e) are one-stage based methods. (f)-(j) are two-stage scoring based methods. (k)-(r) are other state-of-the-art methods.

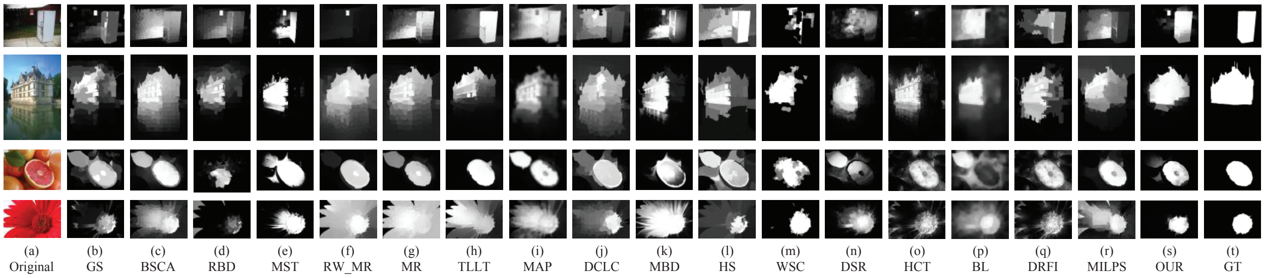


Fig. 21: Visual comparisons on different methods for those images with similar appearance between background and foreground. (b)-(e) are one-stage based methods. (f)-(j) are two-stage scoring based methods. (k)-(r) are other state-of-the-art methods.

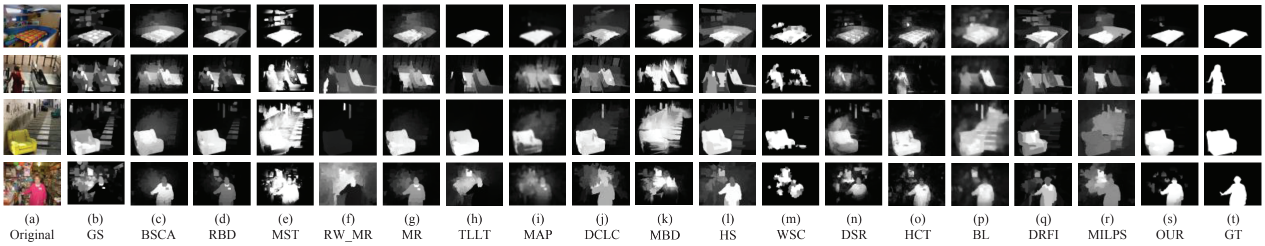


Fig. 22: Visual comparisons on different methods for those images with complex scene. (b)-(e) are one-stage based methods. (f)-(j) are two-stage scoring based methods. (k)-(r) are other state-of-the-art methods.

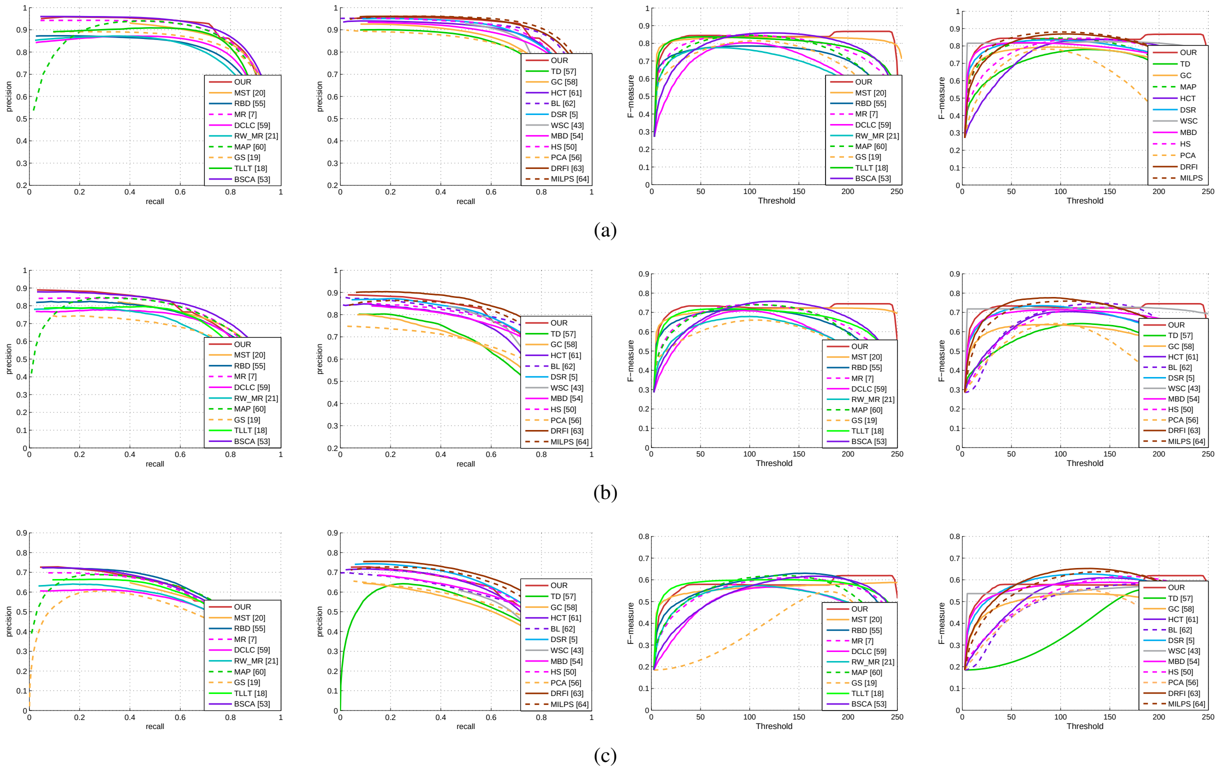


Fig. 23: PR and F-measure curves on different methods for (a) MSRA10K, (b) ECSSD, and (c) DUT-OMRON.

D. Improvement of State-of-the-art Methods

Fig. 24 illustrates some improved results by integrating our second stage into different state-of-the-art salient object detection methods. The coarse detection results used in our proposed second stage are the detection results of the original methods. It is obvious that those improved methods achieve higher F-measure values over a wider range and smaller MAE scores than their corresponding original methods.

For better understanding, Fig. 25 shows some visual examples to illustrate the improved results of our proposed second stage on some state-of-the-art methods. More visual examples can be found in the Supplementary Material. It can be easily found that our second stage improves the state-of-the-art methods with much better foreground uniformity and background suppression. The improved results are more close to ground truth.

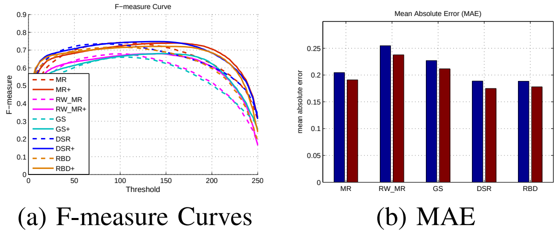


Fig. 24: Illustrations of the improvement of our proposed second stage on some state-of-the-art methods. On the F-measure curves, "+" represents the improved results by integrating the original method with our proposed second stage. For example, "MR+" represents the improved results by integrating the original "MR" method with our second stage. On the MAE bars, blue bars represent the results obtained by the original methods, and red bars represent the improved results by our second stage. For example, on the first group of MAE bars, the left blue bar represents the MAE of the original "MR" method, and the right red bar represents the improved result by integrating the original "MR" with our second stage.

E. Comparisons with Deep Learning Based Methods

In this section, we compare the proposed method with some deep learning based methods, including BPDRR [65] and

TABLE I: The MAE comparisons on different methods for (a) MSRA10K, (b) ECSSD, and (c) DUT-OMRON.

(a) MSRA10K

Methods	Ours	MST [20]	TLLT [18]	BSCA [54]	DRFI [63]	MILPS [64]	DSR [5]	WSC [43]	MBD [49]	MR [7]	RBD [55]
MAE	**0.0730**	0.0800	0.0939	0.1028	0.0941	0.0906	0.0994	0.0866	0.1032	0.1030	0.1080
Methods	Ours	HCT [61]	BL [62]	HS [51]	PCA [56]	TD [57]	GC [58]	DCLC [59]	RW_MR [21]	MAP [60]	GS [19]
MAE	**0.0730**	0.1181	0.1300	0.1224	0.1525	0.1327	0.1143	0.1408	0.1387	0.1044	0.1140

(b) ECSSD

Methods	Ours	MST [20]	TLLT [18]	BSCA [54]	DRFI [63]	MILPS [64]	DSR [5]	WSC [43]	MBD [49]	MR [7]	RBD [55]
MAE	**0.1632**	0.1723	0.1895	0.2000	0.1754	0.1946	0.1887	0.1670	0.1917	0.2046	0.1184
Methods	Ours	HCT [61]	BL [62]	HS [51]	PCA [56]	TD [57]	GC [58]	DCLC [59]	RW_MR [21]	MAP [60]	GS [19]
MAE	**0.1632**	0.2176	0.2413	0.2505	0.2720	0.2495	0.2348	0.2356	0.2547	0.2030	0.2270

(c) DUT-OMRON

Methods	Ours	MST [20]	TLLT [18]	BSCA [54]	DRFI [63]	MILPS [64]	DSR [5]	WSC [43]	MBD [49]	MR [7]	RBD [55]
MAE	**0.1316**	0.1610	0.1443	0.1903	0.1375	0.1673	0.1387	0.1359	0.1567	0.1868	0.1437
Methods	Ours	HCT [61]	BL [62]	HS [51]	PCA [56]	TD [57]	GC [58]	DCLC [59]	RW_MR [21]	MAP [60]	GS [19]
MAE	**0.1316**	0.1638	0.2377	0.2265	0.2054	0.2042	0.1964	0.2111	0.2037	0.1761	0.1731

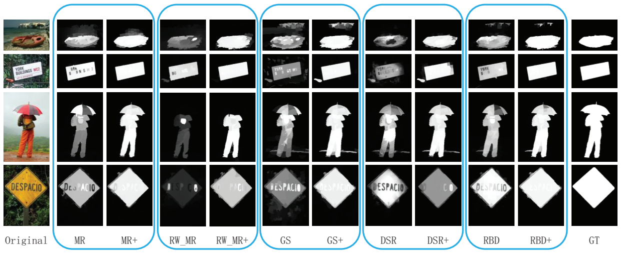


Fig. 25: Visual examples to illustrate the improvement of our proposed second stage on some state-of-the-art methods. Please refer to Fig. 24 for the explanation of "+".

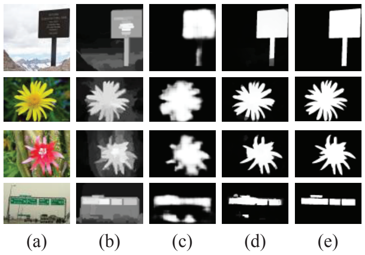
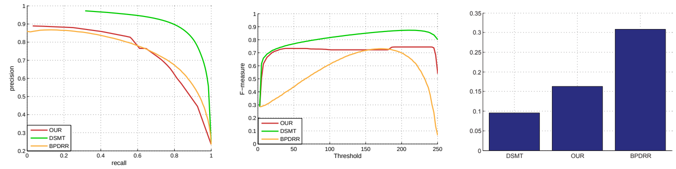


Fig. 27: Visual comparisons with some deep learning based methods on ECSSD. (a) Original images; (b) BPDRR; (c) DSMT; (d) OUR; (e) Ground truth.

DSMT [66]. More specifically, BPDRR [65] is based on the autoencoders, and DSMT [66] is based on the convolutional neural networks (CNNs). Fig. 26 and Fig. 27 show the quantitative comparisons and visual comparisons, respectively. More visual comparisons can be found in the Supplementary Material. It can be noticed from Fig. 26 that the proposed method performs better than BPDRR [65] but worse than DSMT [66]. This demonstrates that deep CNNs have great potential for salient object detection. However, it can also be seen from Fig. 27 that DSMT [66] obtains blurry object boundaries while the proposed method achieves more accurate foreground objects, especially at the object boundaries. The results of our proposed method are the closest to the ground truth.



(a) PR Curves (b) F-measure Curves (c) MAE

Fig. 26: Quantitative comparisons with some deep learning based methods on ECSSD. (a) PR curve; (b) F-measure curve; (c) Mean absolute error (MAE).

F. *Computational Complexity Comparison*

Here, we list the average execution time of several state-of-the-art methods and our proposed method on the MSRA10K dataset [50]. These methods are all run on a PC with an Intel(R) Core(TM) i7-4790 3.60 GHz CPU. As shown in Table II, it will take about 2 seconds for our proposed method to process an image of size 400×300, which is faster than TLLT [18], DSR [5], WSC [43], and PCA [56]. Besides, the total running time of the proposed method for the MSRA10K [50], ECSSD [51], and DUT-OMRON [6], [7] datasets is about 5.7 hours, 0.5 hours, and 2.94 hours, respectively.

V. Conclusion

In this paper, we perform salient object detection via two-stage graphs. This is clearly different from most of existing graph-based methods, which employ only a single graph. As a result, the proposed method is shown to be superior to the state-of-the-art methods in terms of the uniform detection of foreground salient objects as well as the suppression of background noise. In particular, the second stage is generic enough to be integrated in existing salient object detectors to improve their performance. In the future, we will integrate the *regionally spatial consistency* and *adjacently spatial consis-*

TABLE II: Average execution time of several methods (seconds per image).

Methods	Ours	TLLT	BSCA	DSR	WSC	MR	DCLC	RW_MR	PCA
Time (s)	2.052	2.374	1.353	2.806	4.069	0.756	1.055	1.001	2.4589

tency in the deep CNNs architecture to further improve the performance of our proposed method.

Acknowledgment

This work is supported by the National Natural Science Foundation of China under Grant No.61773301, by Natural Science Basic Research Plan in Shaanxi Province of China (Program No. 2016JM6008), and by the Fundamental Research Funds for the Central Universities under Grant No. JBZ170401.

References

[1] T. Liu, Z. Yuan, J. Sun, J. Wang, N. Zheng, X. Tang, and H.-Y. Shum, "Learning to detect a salient object," *IEEE Transactions on Pattern analysis and machine intelligence*, vol. 33, no. 2, pp. 353–367, 2011.

[2] Q. Zhao and C. Koch, "Learning saliency-based visual attention: A review," *Signal Processing*, vol. 93, no. 6, pp. 1401–1407, 2013.

[3] A. Borji, D. N. Sihite, and L. Itti, "Quantitative analysis of human-model agreement in visual saliency modeling: A comparative study," *IEEE Transactions on Image Processing*, vol. 22, no. 1, pp. 55–69, 2013.

[4] X. Li, H. Lu, L. Zhang, X. Ruan, and M.-H. Yang, "Saliency detection via dense and sparse reconstruction," in *Proceedings of IEEE International Conference on Computer Vision*, 2013, pp. 2976–2983.

[5] H. Lu, X. Li, L. Zhang, R. Xiang, and M. H. Yang, "Dense and sparse reconstruction error based saliency descriptor," *IEEE Transactions on Image Processing*, vol. 25, no. 4, pp. 1592–1603, 2016.

[6] C. Yang, L. Zhang, H. Lu, X. Ruan, and M.-H. Yang, "Saliency detection via graph-based manifold ranking," in *Proceedings of IEEE conference on computer vision and pattern recognition*, 2013, pp. 3166–3173.

[7] L. Zhang, C. Yang, H. Lu, R. Xiang, and M. H. Yang, "Ranking saliency," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 9, pp. 1892–1904, 2017.

[8] Q. Wang, W. Zheng, and R. Piramuthu, "Grab: Visual saliency via novel graph model and background priors," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 535–543.

[9] J. Shi, Q. Yan, L. Xu, and J. Jia, "Hierarchical image saliency detection on extended cssd," *IEEE transactions on pattern analysis and machine intelligence*, vol. 38, no. 4, pp. 717–729, 2016.

[10] Q. Fan and C. Qi, "Saliency detection based on global and local short-term sparse representation," *Neurocomputing*, vol. 175, pp. 81–89, 2016.

[11] H. Peng, B. Li, H. Ling, W. Hu, W. Xiong, and S. J. Maybank, "Salient object detection via structured matrix decomposition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 4, pp. 818–832, 2017.

[12] J. Han, K. N. Ngan, M. Li, and H.-J. Zhang, "Unsupervised extraction of visual attention objects in color images," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 16, no. 1, pp. 141–145, 2006.

[13] D. Gao, S. Han, and N. Vasconcelos, "Discriminant saliency, the detection of suspicious coincidences, and applications to visual recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, no. 6, pp. 989–1005, 2009.

[14] P. Wang, J. Wang, G. Zeng, J. Feng, H. Zha, and S. Li, "Salient object detection for searched web images via global saliency," in *Proceedings of IEEE Confernece on Computer Vision and Pattern Recognition*. IEEE, 2012, pp. 3194–3201.

[15] J. Han, E. J. Pauwels, and P. De Zeeuw, "Fast saliency-aware multi-modality image fusion," *Neurocomputing*, vol. 111, pp. 70–80, 2013.

[16] Y. Lu, W. Zhang, H. Lu, and X. Xue, "Salient object detection using concavity context," in *Proceedings of IEEE International Conference on Computer Vision*. IEEE, 2011, pp. 233–240.

[17] Z. Jiang and L. S. Davis, "Submodular salient region detection," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2013, pp. 2043–2050.

[18] C. Gong, D. Tao, W. Liu, S. J. Maybank, M. Fang, K. Fu, and J. Yang, "Saliency propagation from simple to difficult," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 2531–2539.

[19] Y. Wei, F. Wen, W. Zhu, and J. Sun, "Geodesic saliency using background priors," *Proceedings of European Conference on Computer Vision*, pp. 29–42, 2012.

[20] W.-C. Tu, S. He, Q. Yang, and S.-Y. Chien, "Real-time salient object detection with a minimum spanning tree," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 2334–2342.

[21] Y. Liu, Q. Cai, X. Zhu, J. Cao, and H. Li, "Saliency detection using two-stage scoring," in *Proceedings of IEEE International Conference on Image Processing*. IEEE, 2015, pp. 4062–4066.

[22] S. Wang, M. Wang, S. Yang, and K. Zhang, "Salient region detection via discriminative dictionary learning and joint bayesian inference," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. PP, no. 99, pp. 1–14, 2017.

[23] D. Zhang, D. Meng, and J. Han, "Co-saliency detection via a self-paced multiple-instance learning framework," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 5, pp. 865–878, 2016.

[24] D. Zhang, J. Han, C. Li, J. Wang, and X. Li, "Detection of co-salient objects by looking deep and wide," *International Journal of Computer Vision*, vol. 120, no. 2, pp. 215–232, 2016.

[25] D. Zhang, J. Han, J. Han, and L. Shao, "Image adaptation and dynamic browsing based on two-layer saliency combination," *IEEE Transactions on Broadcasting*, vol. 59, no. 4, pp. 602–613, 2013.

[26] ——, "Cosaliency detection based on intrasaliency prior transfer and deep intersaliency mining," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 27, no. 6, pp. 1163–1176, 2015.

[27] D. Zhang, J. Han, L. Jiang, S. Ye, and X. Chang, "Revealing event saliency in unconstrained video collection," *IEEE Transactions on Image Processing*, vol. 26, no. 4, pp. 1746–1758, 2017.

[28] X. Shen and Y. Wu, "A unified approach to salient object detection via low rank matrix recovery," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2012, pp. 853–860.

[29] J. Chen, H. Zhao, Y. Han, and X. Cao, "Visual saliency detection based on photographic composition," in *International Conference on Internet Multimedia Computing and Service*, 2013, pp. 13–16.

[30] Q. Zhang, Y. Liu, S. Zhu, and J. Han, "Salient object detection based on super-pixel clustering and unified low-rank representation," *Computer Vision and Image Understanding*, vol. 51, pp. 51–64, 2017.

[31] Y. Li, Y. Zhou, L. Xu, X. Yang, and J. Yang, "Incremental sparse saliency detection," in *Proceedings of IEEE International Conference on Image Processing*. IEEE, 2009, pp. 3093–3096.

[32] B. Han, H. Zhu, and Y. Ding, "Bottom-up saliency based on weighted sparse coding residual," in *Proceedings of ACM International Conference on Multimedia*. ACM, 2011, pp. 1117–1120.

[33] L. Huo, S. Yang, L. Jiao, S. Wang, and S. Wang, "Local graph regularized sparse reconstruction for salient object detection," *Neurocomputing*, vol. 194, pp. 348–359, 2016.

[34] J. Harel, C. Koch, P. Perona *et al.*, "Graph-based visual saliency," in *NIPS*, vol. 1, no. 2, 2006, pp. 545–552.

[35] W. Wang, Y. Wang, Q. Huang, and W. Gao, "Measuring visual saliency by site entropy rate," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2010, pp. 2368–2375.

[36] K.-Y. Chang, T.-L. Liu, H.-T. Chen, and S.-H. Lai, "Fusing generic objectness and visual saliency for salient object detection," in *Proceedings of IEEE International Conference on Computer Vision*. IEEE, 2011, pp. 914–921.

[37] L. Mai, Y. Niu, and F. Liu, "Saliency aggregation: A data-driven approach," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2013, pp. 1131–1138.

[38] G. Li and Y. Yu, "Visual saliency based on multiscale deep features," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 5455–5463.

[39] N. Liu and J. Han, "Dhsnet: Deep hierarchical saliency network for salient object detection," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 678–686.

[40] G. Li and Y. Yu, "Deep contrast learning for salient object detection," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 478–487.

[41] R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, and S. Süsstrunk, "Slic superpixels compared to state-of-the-art superpixel methods," *IEEE transactions on pattern analysis and machine intelligence*, vol. 34, no. 11, pp. 2274–2282, 2012.

[42] Q. Zhang and M. D. Levine, "Robust multi-focus image fusion using multi-task sparse representation and spatial context," *IEEE Transactions on Image Processing*, vol. 25, no. 5, pp. 2045–2058, 2016.

[43] N. Li, B. Sun, and J. Yu, "A weighted sparse coding framework for saliency detection," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 5216–5223.

[44] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein, "Distributed optimization and statistical learning via the alternating direction method of multipliers," *Foundations and Trends® in Machine Learning*, vol. 3, no. 1, pp. 1–122, 2011.

[45] D. Zhou, J. Weston, A. Gretton, O. Bousquet, and B. Schölkopf, "Ranking on data manifolds," in *NIPS*, 2003, pp. 169–176.

[46] N. Otsu, "A threshold selection method from gray-level histograms," *IEEE Transactions on systems, man, and cybernetics*, vol. 9, no. 1, pp. 62–66, 1979.

[47] A. Borji, M.-M. Cheng, H. Jiang, and J. Li, "Salient object detection: A benchmark," *IEEE Transactions on Image Processing*, vol. 24, no. 12, pp. 5706–5722, 2015.

[48] Y. Boykov and V. Kolmogorov, "An experimental comparison of min-cut/max-flow algorithms for energy minimization in vision," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 26, no. 9, pp. 1124–1137, 2004.

[49] J. Zhang, S. Sclaroff, Z. Lin, X. Shen, B. Price, and R. Mech, "Minimum barrier salient object detection at 80 fps," in *Proceedings of IEEE International Conference on Computer Vision*, 2015, pp. 1404–1412.

[50] M.-M. Cheng, N. J. Mitra, X. Huang, P. H. Torr, and S.-M. Hu, "Global contrast based salient region detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 37, no. 3, pp. 569–582, 2015.

[51] Q. Yan, L. Xu, J. Shi, and J. Jia, "Hierarchical saliency detection," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2013, pp. 1155–1162.

[52] R. Achanta, S. Hemami, F. Estrada, and S. Susstrunk, "Frequency-tuned salient region detection," in *Proceedings of IEEE Ocnference on Computer Vision and Pattern Recognition*. IEEE, 2009, pp. 1597–1604.

[53] F. Perazzi, P. Krähenbühl, Y. Pritch, and A. Hornung, "Saliency filters: Contrast based filtering for salient region detection," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2012, pp. 733–740.

[54] Y. Qin, H. Lu, Y. Xu, and H. Wang, "Saliency detection via cellular automata," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 110–119.

[55] W. Zhu, S. Liang, Y. Wei, and J. Sun, "Saliency optimization from robust background detection," in *Proceedings of IEEE conference on computer vision and pattern recognition*, 2014, pp. 2814–2821.

[56] R. Margolin, A. Tal, and L. Zelnik-Manor, "What makes a patch distinct?" in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2013, pp. 1139–1146.

[57] C. Scharfenberger, A. Wong, K. Fergani, J. S. Zelek, and D. A. Clausi, "Statistical textural distinctiveness for salient region detection in natural images," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2013, pp. 979–986.

[58] M.-M. Cheng, J. Warrell, W.-Y. Lin, S. Zheng, V. Vineet, and N. Crook, "Efficient salient region detection with soft image abstraction," in *Proceedings of IEEE International Conference on Computer vision*, 2013, pp. 1529–1536.

[59] L. Zhou, Z. Yang, Q. Yuan, Z. Zhou, and D. Hu, "Salient region detection via integrating diffusion-based compactness and local contrast," *IEEE Transactions on Image Processing*, vol. 24, no. 11, pp. 3308–3320, 2015.

[60] J. Sun, H. Lu, and X. Liu, "Saliency region detection based on markov absorption probabilities," *IEEE Transactions on Image Processing*, vol. 24, no. 5, pp. 1639–1649, 2015.

[61] J. Kim, D. Han, Y.-W. Tai, and J. Kim, "Salient region detection via high-dimensional color transform," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 883–890.

[62] N. Tong, H. Lu, X. Ruan, and M.-H. Yang, "Salient object detection via bootstrap learning," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 1884–1892.

[63] H. Jiang, J. Wang, Z. Yuan, Y. Wu, N. Zheng, and S. Li, "Salient object detection: A discriminative regional feature integration approach," in *Proceedings of IEEE conference on computer vision and pattern recognition*, 2013, pp. 2083–2090.

[64] F. Huang, J. Qi, H. Lu, L. Zhang, and X. Ruan, "Salient object detection via multiple instance learning," *IEEE Transactions on Image Processing*, vol. 26, no. 4, pp. 1911–1922, 2017.

[65] J. Han, D. Zhang, X. Hu, L. Guo, J. Ren, and F. Wu, "Background prior-based salient object detection via deep reconstruction residual," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 25, no. 8, pp. 1309–1321, 2015.

[66] X. Li, L. Zhao, L. Wei, M. H. Yang, F. Wu, Y. Zhuang, H. Ling, and J. Wang, "Deepsaliency: Multi-task deep neural network model for salient object detection," *IEEE Transactions on Image Processing A Publication of the IEEE Signal Processing Society*, vol. 25, no. 8, pp. 3919–3930, 2015.