Pattern Recognition 92 (2019) 119–134

Contents lists available at ScienceDirect

Pattern Recognition

journal homepage: www.elsevier.com/locate/patcog

Salient object detection employing a local tree-structured low-rank representation and foreground consistency

Qiang Zhang[a,b], Zhen Huo[b], Yi Liu[b], Yunhui Pan[b], Caifeng Shan[d], Jungong Han[c,*]

[a] Key Laboratory of Electronic Equipment Structure Design, Ministry of Education, Xidian University, Xi'an, Shaanxi 710071, China
[b] Center for Complex Systems, School of Mechano-electronic Engineering, Xidian University, Xi'an, Shaanxi 710071, China
[c] School of Computing and Communications, InfoLab21, Lancaster University, Lancaster LA1 4YW, U.K
[d] Philips Research, Eindhoven, 5656 AE Eindhoven, The Netherlands

ARTICLE INFO

Article history:
Received 7 September 2018
Revised 27 February 2019
Accepted 23 March 2019
Available online 23 March 2019

Keywords:
Salient object detection
Structured low-rank representation
Background dictionary
Foreground consistency

ABSTRACT

We propose a local tree-structured low-rank representation (TS-LRR) model to detect salient objects under the complicated background with diverse local regions, which is problematic for most low-rank matrix recovery (LRMR) based salient object detection methods. We first impose a local tree-structured low-rank constraint on the representation coefficients matrix to capture the complicated background. Specifically, a primitive background dictionary is constructed for TS-LRR to promote its background representation ability, and thus enlarge the gap between the salient objects and the background. We then impose a group-sparsity constraint on the sparse error matrix with the intention to ensure the saliency consistency among patches with similar features. At last, a foreground consistency is introduced to identically highlight the distinctive regions within the salient object. Experimental results on three public benchmark datasets demonstrate the effectiveness and superiority of the proposed model over the state-of-the-art methods.

1. Introduction

Visual saliency aims at identifying salient regions, at which humans may fixate [1]. As an important branch of visual saliency, salient object detection focuses on uniformly highlighting the entire salient objects in a natural image, which is usually a preprocessing step of many computer vision tasks, such as object segmentation [2], image retrieval [3,4], image categorization [5] and recognition [6,7].

Recently, low-rank matrix recovery (LRMR) technique has been applied to saliency detection [8,9] as a result of its promising performance. These methods generally presume that the salient objects only occupy small proportion of the whole image, and the features of the background lie in a low-dimensional subspace. Therefore, they usually employ different LRMR techniques to decompose the feature matrix of the input image into two parts: the low-rank part and the sparse noise part (i.e., reconstruction errors), where the latter part is utilized to construct a saliency measure for the detection of salient objects within the input image. Although the preliminary results of LRMR based methods look promising, they struggle to handle more realistic scenarios, where the salient objects are usually surrounded by complex backgrounds. In this paper, a local tree-structured low-rank (TS-LRR) model is proposed for salient object detection to deal with the above situation.

A typical state-of-the-art LRMR based salient object detection method relies on the structured matrix decomposition (SMD) model [10], which formulates the task of salient object detection as a problem of LRMR and structured sparse matrix decomposition. Specifically, two structural regularizations are employed in the SMD model. One is a tree-structured sparsity-inducing regularization, enforcing patches from the same group to have similar saliency values. The other one is a Laplacian regularization, enlarging the gaps between salient objects and the background in the feature space.

However, such method suffers from several problems. First, SMD imposes a global low-rank constraint on the feature matrix to capture background, which is valid only for simple background but not for complicated one with distinctive regions. As shown in the first three rows of Fig. 1(b), some local background regions are mistakenly labeled as salient objects when directly imposing such a global low-rank constraint. This is due to the fact that these local background regions are distinct from most of the whole background regions.

* Corresponding author.
E-mail address: jungong.han@lancaster.ac.uk (J. Han).

https://doi.org/10.1016/j.patcog.2019.03.023
0031-3203/

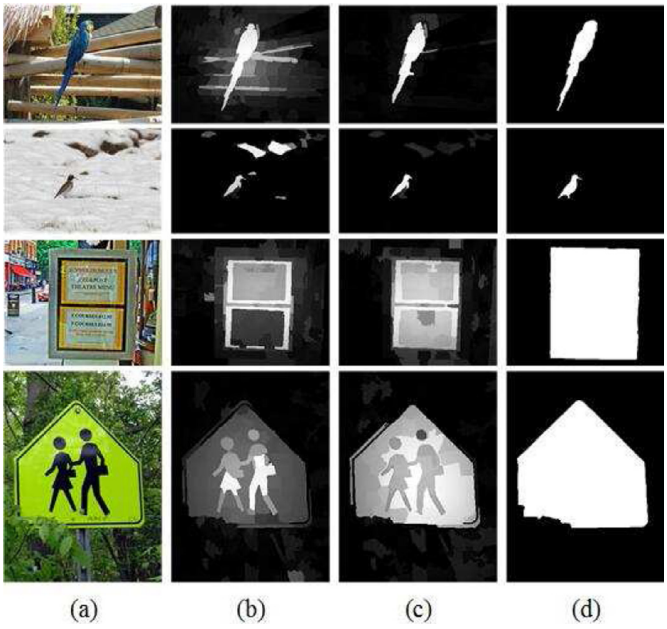


Fig. 1. Superiority of the proposed method over SMD. (a) Images; (b) Salient regions detected by SMD; (c) Detected results obtained by our method; (d) Ground Truth.

Secondly, the Laplacian regularization adopted by SMD assumes two adjacent superpixels with similar appearance to have similar saliency values. As a result, the foreground superpixels with similar appearances will possess similar saliency values. Conversely, the foreground superpixels with distinctive appearances will be assigned to different saliency values, which damages the foreground uniformity. For example, as shown in the last two rows of Fig. 1(b), distinctive regions within the same foreground object are assigned to different saliency values, thus leading to poor foreground uniformity.

Finally, from theory perspective, LRMR can be interpreted as a special case of the low-rank representation (LRR) model with an orthogonal basis dictionary. For those images with complicated scenes, such a constraint dictionary does not seem to represent their background regions appropriately. Therefore, the salient objects may not be well separated from the background.

In order to address the problems mentioned above, we present a local tree-structured low-rank representation (TS-LRR) model, which is based on LRR rather than LRMR, for salient object detection in this paper. Using TS-LRR allows us to formulate the salient object detection to a joint low-rank representation and sparsity pursuit problem. The entire algorithm is carried out in four steps.

First, a primitive background dictionary is introduced to enhance the background representation ability of TS-LRR. The proposed background dictionary is constructed by using the coarse saliency detection results from a simply modified VGG16 network [11]. Secondly, a local tree-structured low-rank constraint, instead of a global one, is imposed on the representation coefficient matrix to better capture the low-rankness of complicated background regions. Specifically, the input image is segmented into an index tree with multiple layers, each of which contains a few groups. Every group is composed of several superpixels with similar features and thus possesses better low-rankness than the entire image. Thirdly, similar to SMD, a group-sparsity constraint is imposed on the reconstruction errors matrix to ensure the saliency consistency among superpixels with similar features. Finally, a local spatial consistency, i.e., foreground consistency, is performed among potential salient patches (or superpixels) to get better uniformity and completeness of the salient objects. This is achieved by introducing a Laplacian regularization on the sparse errors matrix in the proposed model. Such regularization enforces those patches within the potential salient foregrounds to be similar in saliency values even if they have different appearances, as shown in the last two rows of Fig. 1(c). Revealed by the results, our model is capable of suppressing the background regions even for those images with complex scenes (as shown in the first two rows of Fig. 1(c)). As well, the proposed method enhances the uniformity and completeness of the detected salient objects (as shown in the last two rows of Fig. 1(c)).

In summary, the main contributions of this paper are as follows:

(1) Instead of using a global low-rank constraint as in [10], we proposed a local tree-structured low-rank representation (TS-LRR) model for salient object detection, which is able to capture the complicated background regions.
(2) The foreground consistency within potential foreground regions is considered to achieve satisfactory uniformity and completeness of the detected salient objects, which is extremely useful when those salient objects are with diverse types of regions.
(3) A primitive background dictionary is constructed for the proposed model, promoting the background representation ability of TS-LRR.

The rest of this paper is organized as follows. Section 2 reviews the most related works. Section 3 describes the proposed model in detail. Experiments are conducted on three benchmarks to evaluate the validity and superiority of our work in Section 4. Finally, Section 5 concludes the paper.

2. Related work

In the past few years, numerous methods have been proposed for salient object detection. In this section, we will review the most related works, including LRMR based and LRR based salient object detection methods. Besides, deep convolutional neural networks (CNNs) have been successfully applied for salient object detection and achieved amazing performance, which will also be reviewed in this section.

2.1. LRMR based salient object detection

LRMR have attracted more and more attention of researchers in many fields of science and engineering, such as visual classification [12] and face recognition [13]. LRMR theory was first introduced to saliency detection in [14] due to its efficiency. Since then, various extensions have been developed for salient object detection. Generally, these LRMR based methods assume that an image can be represented as a low-rank part (i.e., background) plus a sparse salient part (i.e., foreground). For example, a unified LRMR model was proposed for saliency detection by integrating higher-level knowledge and low-level features in [9]. A novel diversity induced matrix decomposition model was proposed for salient object detection in [15], where a $S_{1/2}$ regularizer was introduced to constraint the background part. Recently, some spatial relations among image regions have also been taken into account in the LRMR models. Especially, Peng et al. [10] proposed a structured matrix decomposition model, in which a tree-structured sparsity-inducing constraint was employed to detect the salient objects, and a Laplacian regularization was employed to enlarge the gaps between salient objects and background in the feature space.

2.2. LRR based salient object detection

Recently, low-rank representation (LRR) has been applied in salient object detection owing to its favorable efficiency. Unlike

LRMR that directly decomposes a matrix into two parts, LRR uses a dictionary to decompose an input image (or its feature matrix) into a low-rank part (i.e., background) and a sparse error part (i.e., foreground), where the sparse error part is usually employed to extract the salient objects. For example, Lang et al. [8] incorporated multiple types of features in a multi-task sparsity pursuit model for saliency detection. Zhao et al. [16] proposed a new low rank learning method which constructs the low rank representation matrix utilizing label information to obtain a more informative graph. In our previous work [17], a unified low-rank representation model, combined with the Laplacian sparse subspace clustering, was presented for saliency detection, where a cluster-level saliency measure and a superpixel-level one were integrated to compute the final saliency values.

2.3. CNNs based salient object detection

More recently, deep convolution neural networks (CNNs) have been successfully applied to detect salient objects from image. Wang et al. [18] proposed to detect saliency via local estimation and global search based on CNNs. Alternatively, Li and Yu [19] employed CNNs to extract multi-scale deep features from the input image for saliency detection. In [20], a CNNs-based multi-scale multi-path fusion network was presented for RGB-D salient object detection.

To sum up, most existing LRMR and LRR based methods generally work for simple scenes. However, when background regions are complicated or objects contain diverse regions, some undesirable results may be obtained, such as poor foreground uniformity and background suppression. Those CNNs based methods perform slightly better than the traditional methods, but they still suffer from the problems like blurry salient object boundaries.

3. Proposed salient object detection model

The diagram of the proposed method is shown in Fig. 2. The details of the proposed model will be elaborated below.

3.1. Feature extraction and index-tree construction

In this section, the input image is first over-segmented into a number of superpixels and features of each superpixel are extracted. Then, an index-tree structure is constructed for the input image.

3.1.1. Superpixel over-segmentation and feature extraction

In this paper, a simple iterative superpixel clustering (SLIC) algorithm [21] is adopted to over-segment an input image I into N superpixels $\{p_i|i = 1, 2, \ldots, N\}$, where N is experimentally set to 200. Following [9], the features of the input image in this paper cover the RGB color as well as the hue and saturation components (5 dimensions), steerable pyramids (12 dimensions) [22] and Gabor filter (36 dimensions) [23]. The three types of features can capture different characteristics of an image and are complementary to each other when applied to salient object detection, which will be shown in the later experimental part. The color features can well capture the appearance of an image, while the other two types pay more attention to explore the textures of an image. The features of each superpixel are obtained by averaging those of its pixels, yielding the feature matrix $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N] \in R^{m\times N}$ (here $m = 53$), where $\mathbf{x}_i$ denotes the feature vector of a superpixel p_i.

3.1.2. Index-tree generation

In many complex cases, an image may contain many regions with different types of features (or spatial appearances). To better capture these regions, we propose to perform saliency detection on an index tree with a few superpixel clusters in each layer. Similar

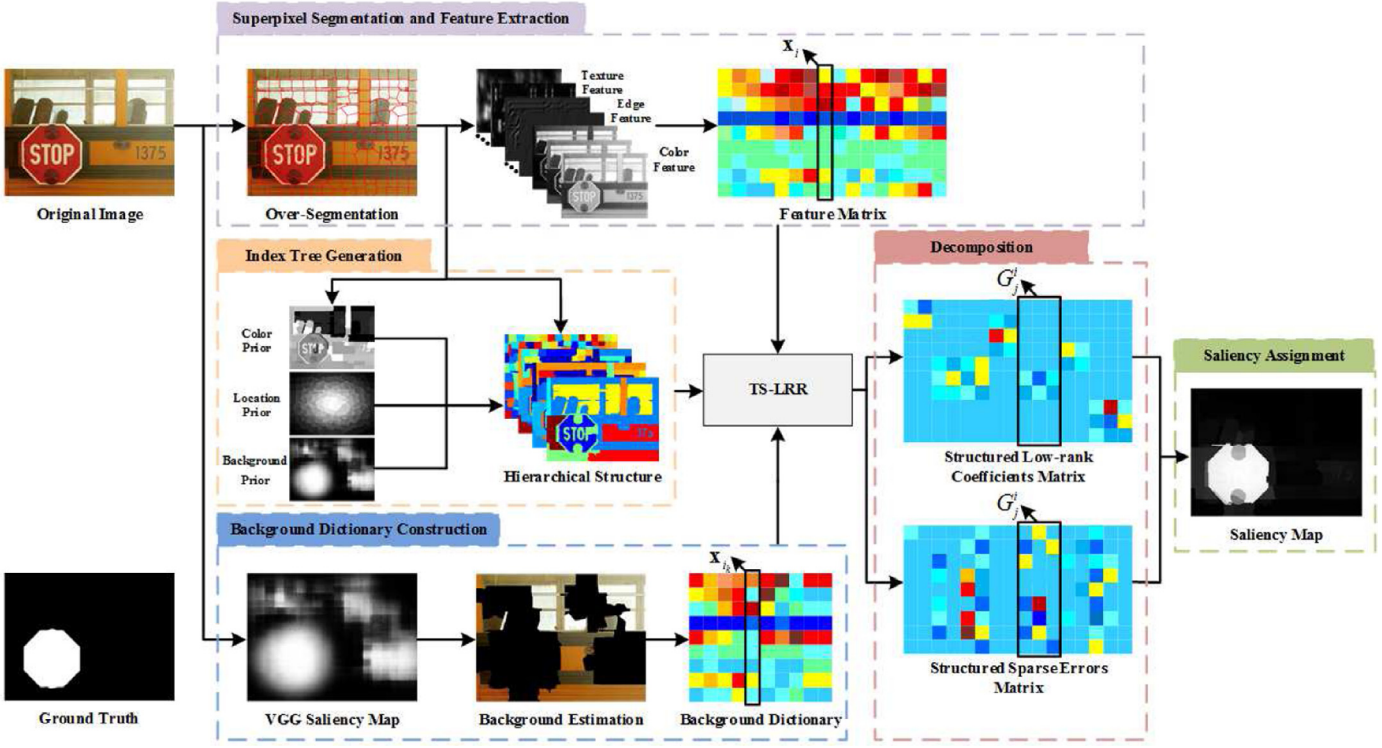


Fig. 2. Diagram of the proposed salient object detection model.

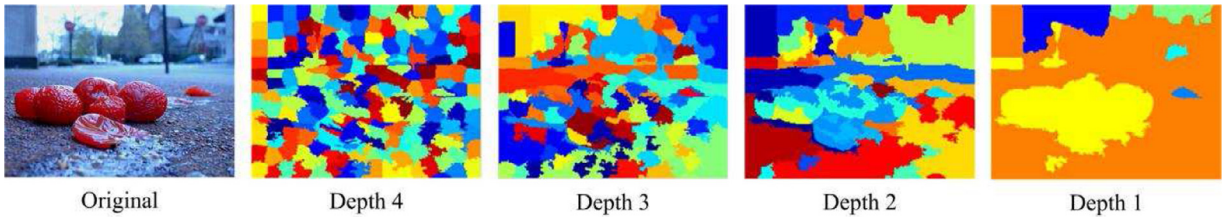


Fig. 3. Visualization of a 4-layer index tree structure. The regions that are marked by the same color represent one node in the index tree.

to that in [10], an index-tree of an input image can be generated as follows.

Specifically, given the superpixels $\{p_i|i=1,2,\ldots,N\}$, an affinity of each adjacent superpixel pair is first computed by

$$\omega_{i,j}^{tree}=\begin{cases}\exp\left(-\frac{\|\mathbf{x}_i-\mathbf{x}_j\|^2}{2\sigma^2}\right), & \text{if}\left(p_i,p_j\right)\in\Omega\\ 0, & \text{otherwise}\end{cases}, \tag{1}$$

where Ω denotes the set of spatially adjacent superpixel pairs. σ is experimentally set to 0.05. According to the affinity computed by Eq. (1), a graph-based image segmentation algorithm [24] is then applied for the input image to produce a sequence of layers containing scale-increasing groups. In each layer, a group corresponds to one node at the corresponding layer. The scale is controlled by a layer-dependent threshold θ_i. Finally, a hierarchical fine-to-coarse segmentation $\{G_j^i\}(i=1,2,\ldots,d;\ j=1,2,\ldots,n_i)$ of the input image is obtained. Here, d denotes the depth of the index tree. n_i denotes the total number of nodes at the ith layer in the index tree. G_j^i is the jth node at the ith layer, which contains a set of similar superpixels. Fig. 3 shows a visualized example of a 4-layer index tree for an input image.

3.2. Proposed salient object detection model

As discussed in Section 3.1, the superpixels within the sub-region G_j^i have similar features and thus their feature matrices should have high low-rankness. As a result of this observation, a local low-rank constraint on the sub-region G_j^i, rather than a global low-rank constraint on the entire image, will better capture the complicated background with distinctive sub-regions.

As well, under the same dictionary, the superpixels within the sub-region G_j^i should have similar reconstruction errors and thus have similar saliency values. As discussed in [10], this may be achieved by imposing a local consistency among the superpixels within the same group via a group-sparsity constraint on the sparse error matrix, which enforces the superpixels within G_j^i to identically have similar errors. Unfortunately, those foreground superpixels with different appearances will be easily assigned to different saliency values by just considering the local consistency among the superpixels within the same group, which in turn will damage the foreground uniformity.

In fact, all superpixels within the same object are expected to have similar saliency values even though they may have different features (or spatial appearances). For that, a foreground consistency is further introduced to ensure the foreground uniformity in this paper. This can be achieved by incorporating another Laplacian regularization on the sparse error matrix in the proposed salient object detection model.

Based on the above observations, the proposed salient object detection problem can be solved by the following local tree-structured low-rank representation (TS-LRR) model:

$$\begin{aligned}&\min_{\mathbf{Z},\mathbf{E}}\sum_{i=1}^{d}\sum_{j=1}^{n_i}\left\|\mathbf{Z}_{G_j^i}\right\|_*+\alpha\sum_{i=1}^{d}\sum_{j=1}^{n_i}\upsilon_j^i\left\|\mathbf{E}_{G_j^i}\right\|_\infty+\beta Tr\left(\mathbf{ELE}^T\right),\\ &s.t.\quad \mathbf{X}=\mathbf{DZ}+\mathbf{E},\end{aligned} \tag{2}$$

where $\mathbf{Z}\in R^{M\times N}$ and $\mathbf{E}\in R^{m\times N}$ are the representation coefficient matrix and sparse error matrix, respectively. $\mathbf{Z}_{G_j^i}$ and $\mathbf{E}_{G_j^i}$ denote the sub-matrix of $\mathbf{Z}$ and $\mathbf{E}$ corresponding to the sub-region G_j^i, respectively. $\|\mathbf{Z}_{G_j^i}\|_*$ denotes the nuclear norm of the matrix $\mathbf{Z}_{G_j^i}$ and is defined as the sum of the singular values of the matrix $\mathbf{Z}_{G_j^i}$. $\|\mathbf{E}_{G_j^i}\|_\infty$ refers to the ℓ_∞-norm of the matrix $\mathbf{E}_{G_j^i}$ and is defined as the maximum absolute error value of regions within the group G_j^i. $\mathbf{D}\in R^{m\times M}$ is a primitive background dictionary, which will be discussed in the following Section 3.2.1. α and β are two positive trade-off parameters and are experimentally set to 1 and 0.1, respectively.

The first term $\sum_{i=1}^{d}\sum_{j=1}^{n_i}\|\mathbf{Z}_{G_j^i}\|_*$ in Eq. (2) involves a local tree-structured low-rank constraint to describe the intrinsic low-rankness within each group G_j^i, which helps to capture the complicated background. In order to improve the background representation ability of the model, a primitive background dictionary $\mathbf{D}$ is constructed for the proposed TS-LRR model.

The second term $\sum_{i=1}^{d}\sum_{j=1}^{n_i}\upsilon_j^i\|\mathbf{E}_{G_j^i}\|_\infty$ is a structured group sparsity regularization on the sparse error matrix, in which the ℓ_∞-norm is employed to enforce the superpixels within the same group to possess identical saliency values as in [10]. $\upsilon_j^i(\upsilon_j^i\geq 0)$ refers to the weight associated to each sub-region G_j^i and is computed based on some high-level priors, including location, color and background priors as in [10].

Finally, the Laplacian regularization term $Tr(\mathbf{ELE}^T)$ is introduced to enforce the foreground consistency within potential foreground regions, which will be further discussed in the following Section 3.2.2. Jointly engaging the above three terms ensures foreground uniformity and background suppression in complex scenes.

In the following contents, we will discuss the construction of the primitive background dictionary $\mathbf{D}$ and the Laplacian regularization term $Tr(\mathbf{ELE}^T)$ in detail.

3.2.1. Construction of the primitive background dictionary

In order to improve the background representation ability of TS-LRR, a primitive background dictionary is employed in the proposed model, which is constructed by using the coarse detection results from the VGG16 network [11].

VGG16 is an important backbone for deep learning based computer vision tasks, such as image classification [11]. Recently, it has been utilized for salient object detection [25–27] and achieved an amazing improvement. Here, we simply modify the VGG16 architecture by removing the last three fully connected layers and adding five deconvolution layers with strides of 2 to increase the resolution of the saliency map. As shown in Fig. 4(b), the

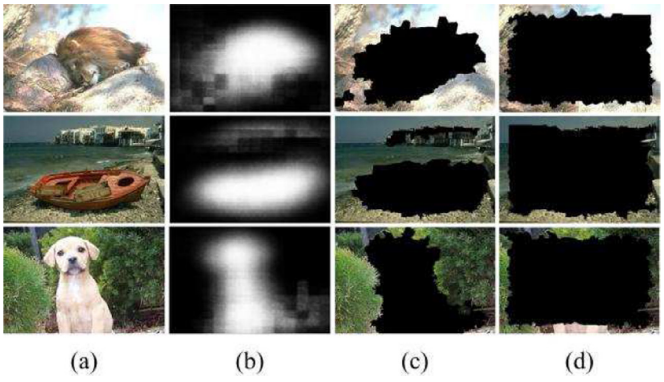


Fig. 4. Illustrations of the proposed primitive background dictionary. (a) Images; (b) Coarse saliency maps obtained by the simply modified VGG16 network; (c) Proposed primitive background dictionaries; (d) Traditional background dictionaries based on the boundary prior [28].

simply modified VGG16 produces a coarse saliency map, which can accurately locate most of foregrounds. The coarse saliency score $F_{VGG}(p_i)$ for each superpixel p_i is obtained by averaging the saliency values of those pixels within the superpixel p_i. Here, $F_{VGG}(p_i)$ is normalized to[0, 1] by using a linear normalization method.[1] Accordingly, the background probability of each superpixel is obtained by

$$B_{VGG}(p_i) = 1 - F_{VGG}(p_i). \tag{3}$$

A higher value of $B_{VGG}(p_i)$ indicates that the superpixel p_i is more likely to be a background region. As shown in Fig. 4(c), those superpixels whose background probabilities are larger than a pre-defined threshold T_{bg} (is experimentally set to 0.8 in this paper) are selected to construct the primitive background dictionary, i.e., $\mathbf{D} = [\mathbf{x}_{i_1}, \mathbf{x}_{i_2}, ..., \mathbf{x}_{i_k}, ..., \mathbf{x}_{i_M}] \in R^{m \times M}$, where M denotes the number of selected superpixels. $\mathbf{x}_{i_k}$ denotes the feature vector of superpixel p_{i_k} with $B_{VGG}(p_{i_k}) \geq T_{bg}$. The remaining superpixels will be seen as the potential foreground ones in this paper, resulting in the potential foreground regions.

The proposed primitive background dictionary is different from the traditional background dictionary, which is assumed to be the image boundaries [28]. As illustrated in the first two rows of Fig. 4(d), those traditional dictionaries based on image boundaries generally cover parts of the background regions. As shown in the last row of Fig. 4(d), some foreground regions touching image borders will also be mistaken as background dictionary atoms. Differently, the proposed primitive background dictionaries can more accurately locate most background regions, even for those images with foreground regions touching boundaries, which is depicted in the last row of Fig. 4(c).

3.2.2. Foreground consistency

In most cases, there exist diverse types of regions with different appearances within the same salient object. These regions will have different reconstruction errors under the same pre-defined dictionary. Thus, they are likely to be assigned with different saliency values, leading to poor foreground uniformity. In order to address this problem, we involve a Laplacian regularization, i.e., the last term in Eq. (2), in the objective function for the foreground consistency. Doing so ensures the superpixels within the potential foreground regions to have identically high saliency values even if they possess different appearances. The Laplacian regularization in Eq. (2) is specifically defined as

$$Tr\left(\mathbf{ELE}^T\right) = \frac{1}{2}\sum_{i,j=1}^{N} \left\|\mathbf{e}_i - \mathbf{e}_j\right\|_2^2 \omega_{ij}^{fc}, \tag{4}$$

$$\omega_{ij}^{fc} = \begin{cases} 1, & \text{if } p_i \text{ and } p_j \text{ are both potential foreground superpixles} \\ 0, & \text{otherwise} \end{cases}, \tag{5}$$

where $\mathbf{e}_i$ denotes the ith column of the sparse error matrix $\mathbf{E}$. ω_{ij}^{fc} represents the similarity between the superpixels p_i and p_j. In Eq. (4), the Laplacian matrix $\mathbf{L}$ is defined by $\mathbf{L} = \mathbf{F}^{fc} - \mathbf{W}^{fc}$, where $\mathbf{F}^{fc} \in R^{N \times N}$ is a diagonal degree matrix with its ith diagonal element $f_{i,i}^{fc} = \sum_j \omega_{ij}^{fc}$, and the affinity matrix $\mathbf{W}^{fc} \in R^{N \times N}$ is constructed with its (i, j)th entry as ω_{ij}^{fc}.

As illustrated in Fig. 5, each superpixel is connected with all the other ones within the potential foreground regions. Suppose that the superpixels p_i and p_j have similar reconstruction errors, they will be assigned to similar saliency values by setting ω_{ij} to 1 if they both belong to the potential foreground regions. That is to say, all superpixels belonging to the potential foreground will be given similar saliency values under our consistency constraint, no matter if the foreground is formed by diverse regions. As a result, the proposed foreground consistency can achieve better uniformity for the salient objects, especially for those salient objects containing diverse regions, than the traditional spatially-adjacent consistency [29,30]. This will be verified in the later experimental part.

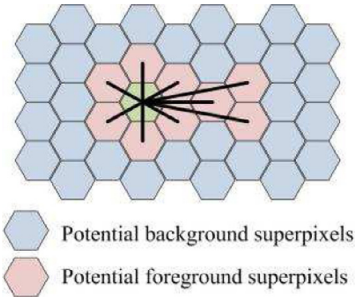


Fig. 5. Superpixel connections within the potential foreground regions. Each potential foreground superpixel is connected with all the other potential ones.

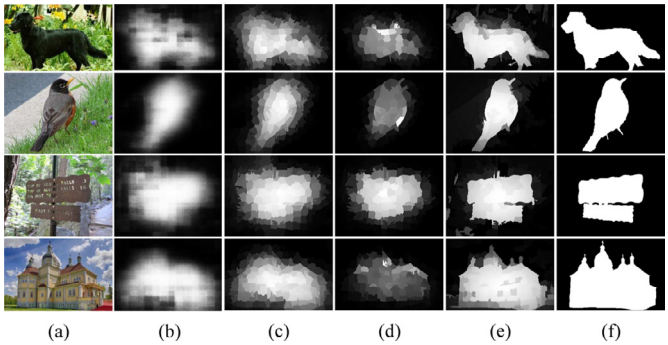


Fig. 6. Detection results obtained by the modified VGG16 network and different saliency measures in our proposed method. (a) Images; (b) Modified VGG16 network; (c) Saliency measure based on representation coefficients; (d) Saliency measure based on reconstruction errors; (e) Optimized saliency measure; (f) Ground Truth.

[1] Given a set of data Ω and a data instance $x \in \Omega$, its normalized value x' is computed by $x' = (x - x_{min})/(x_{max} - x_{min})$, where x_{max} and x_{min} are the maximum and minimum values in the data set Ω, respectively. It is noted that this normalization is used throughout the whole paper.

Algorithm 1
Optimization of the proposed model using ADM.

Input: Data matrix $\mathbf{X} \in R^{m \times N}$, dictionary $\mathbf{D} \in R^{m \times M}$, parameters α and β, index tree $\{G_j^i\}(i = 1, 2, \ldots, d; j = 1, 2, \ldots, n_i)$, and layer node weights υ_j^i.
Output: $\mathbf{Z}$ and $\mathbf{E}$.
Initialization: $\mathbf{Z}^0 = \mathbf{0}$, $\mathbf{E}^0 = \mathbf{0}$, $\mathbf{J}^0 = \mathbf{0}$, $\mathbf{H}^0 = \mathbf{0}$, $\mathbf{Y}_1^0 = \mathbf{Y}_2^0 = \mathbf{Y}_3^0 = \mathbf{0}$, $\mu^0 = 10^{-6}$, $\mu_{\max} = 10^{30}$, $\varepsilon = 10^{-8}$, $\rho = 1.1$, $k = 0$.
While not converged **do**
(1) Fix the others and update $\mathbf{J}$ using Eq. (A3)
(2) Fix the others and update $\mathbf{Z}$ using Eq. (A5)
(3) Fix the others and update $\mathbf{H}$ using Eq. (A7)
(4) Fix the others and update $\mathbf{E}$ using Eq. (A9);
(5) Update the multipliers $\mathbf{Y}_1$, $\mathbf{Y}_2$ and $\mathbf{Y}_3$:
$\mathbf{Y}_1^{k+1} = \mathbf{Y}_1^k + \mu^k(\mathbf{X} - \mathbf{DZ}^{k+1} - \mathbf{E}^{k+1})$, $\mathbf{Y}_2^{k+1} = \mathbf{Y}_2^k + \mu^k(\mathbf{Z}^{k+1} - \mathbf{J}^{k+1})$, $\mathbf{Y}_3^{k+1} = \mathbf{Y}_3^k + \mu^k(\mathbf{E}^{k+1} - \mathbf{H}^{k+1})$;
(6) Update μ:
$\mu^{k+1} = \min(\rho\mu^k, \mu_{\max})$;
(7) Update k:
$k = k + 1$;
(8) Check the convergence conditions:
$\|\mathbf{X} - \mathbf{DZ} - \mathbf{E}\|_\infty < \varepsilon$ & $\|\mathbf{Z} - \mathbf{J}\|_\infty < \varepsilon$ & $\|\mathbf{E} - \mathbf{H}\|_\infty < \varepsilon$;
End while

3.3. Optimization of the proposed model

The proposed model in Eq. (2) is a convex optimization problem, and can be efficiently solved by using the alternating direction method (ADM) [31]. To this end, we introduce two auxiliary variable $\mathbf{J}$ and $\mathbf{H}$ to make the objective function separable. Then, Eq. (2) is recast as

$$\min_{\mathbf{Z},\mathbf{E}} \sum_{i=1}^{d}\sum_{j=1}^{n_i} \left\|\mathbf{J}_{G_j^i}\right\|_* + \alpha \sum_{i=1}^{d}\sum_{j=1}^{n_i} \upsilon_j^i \left\|\mathbf{E}_{G_j^i}\right\|_\infty + \beta Tr\left(\mathbf{HLH}^T\right) \quad s.t. \quad \mathbf{X} = \mathbf{DZ} + \mathbf{E}, \quad \mathbf{Z} = \mathbf{J}, \quad \mathbf{E} = \mathbf{H} \tag{6}$$

After introducing three Lagrangian multipliers $\mathbf{Y}_1$, $\mathbf{Y}_2$ and $\mathbf{Y}_3$ to remove the equality constraints, the optimization model in Eq. (6) can be solved by minimizing the following augmented Lagrangian function L

$$\begin{aligned} &L(\mathbf{Z}, \mathbf{E}, \mathbf{J}, \mathbf{H}, \mathbf{Y}_1, \mathbf{Y}_2, \mathbf{Y}_3, \mu) \\ &= \sum_{i=1}^{d}\sum_{j=1}^{n_i} \left\|\mathbf{J}_{G_j^i}\right\|_* + \alpha \sum_{i=1}^{d}\sum_{j=1}^{n_i} \upsilon_j^i \left\|\mathbf{E}_{G_j^i}\right\|_\infty + \beta Tr\left(\mathbf{HLH}^T\right) \\ &\quad + \langle \mathbf{Y}_1, \mathbf{X} - \mathbf{DZ} - \mathbf{E}\rangle + \langle \mathbf{Y}_2, \mathbf{Z} - \mathbf{J}\rangle + \langle \mathbf{Y}_3, \mathbf{E} - \mathbf{H}\rangle \\ &\quad + \frac{\mu}{2}\|\mathbf{X} - \mathbf{DZ} - \mathbf{E}\|_F^2 + \frac{\mu}{2}\|\mathbf{Z} - \mathbf{J}\|_F^2 + \frac{\mu}{2}\|\mathbf{E} - \mathbf{H}\|_F^2, \end{aligned} \tag{7}$$

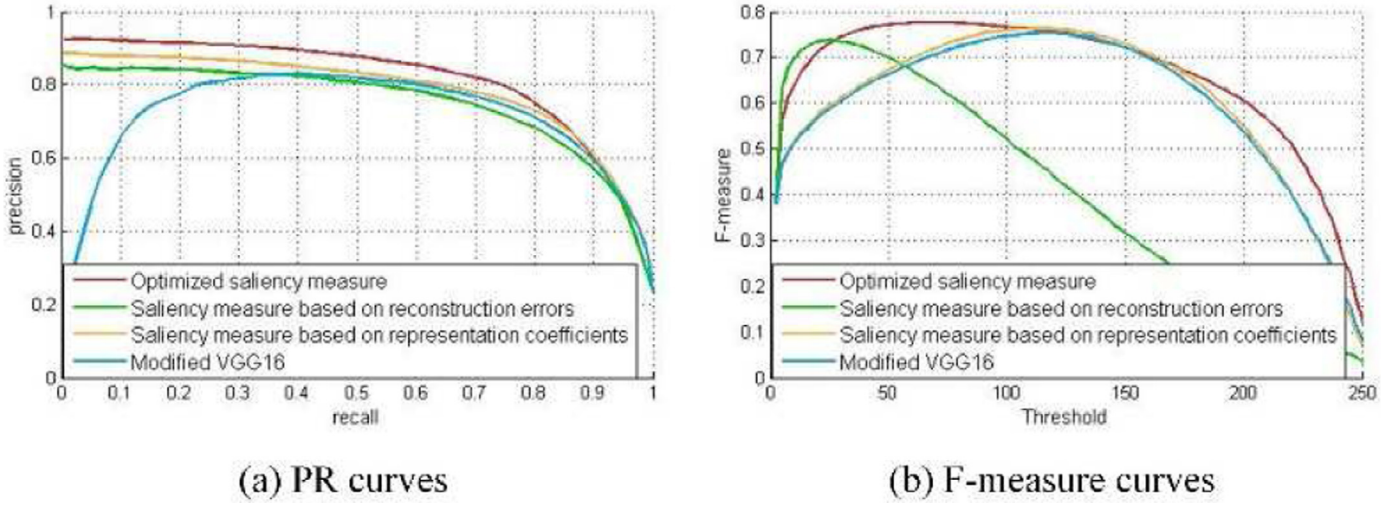
where $\mu > 0$ is a penalty parameter. $\langle \mathbf{A}, \mathbf{B}\rangle$ denotes the Euclidean inner product of matrices $\mathbf{A}$ and $\mathbf{B}$. Clearly, this problem becomes unconstrained, and can be alternately minimized with respect to $\mathbf{Z}$, $\mathbf{E}$, $\mathbf{J}$ and $\mathbf{H}$. Algorithm 1 summarizes the optimization of Eq. (7). More details can be seen in Appendix A.

3.4. Saliency map generation

The sparse error matrix $\mathbf{E}$ and representation coefficient matrix $\mathbf{Z}$ obtained by solving the model in Eq. (2) may contain certain salient information of each superpixel. As discussed in [16], the representation coefficients $\mathbf{Z}(:, i)$ (i.e., the ith column in $\mathbf{Z}$) reveal the similarity between the superpixel p_i and the atoms in the primitive background dictionary $\mathbf{D}$. Accordingly, the sparse errors $\mathbf{E}(:, i)$ (i.e., the ith column in $\mathbf{E}$) indicate the difference between the superpixel p_i and the atoms in $\mathbf{D}$ to some extent. Therefore, given the primitive background dictionary $\mathbf{D}$, a background superpixel will have high absolute values of representation coefficients and low sparse errors. On the contrary, a foreground superpixel will have low absolute values of representation coefficients and high sparse errors. Therefore, in this paper, two saliency measures are first defined based on the representation coefficient matrix $\mathbf{Z}$ and the sparse error matrix $\mathbf{E}$, respectively. Then, these two measures are integrated via an optimization framework such that the final saliency score for each superpixel can be achieved.

3.4.1. Saliency measure based on the representation coefficients

Under the pre-defined primitive background dictionary, the representation coefficients $\mathbf{Z}(:, i)$ measure the similarity between the



(a) PR curves (b) F-measure curves

Fig. 7. PR and F-measure curves obtained by the simply modified VGG16 network and different saliency measures on ECSSD.

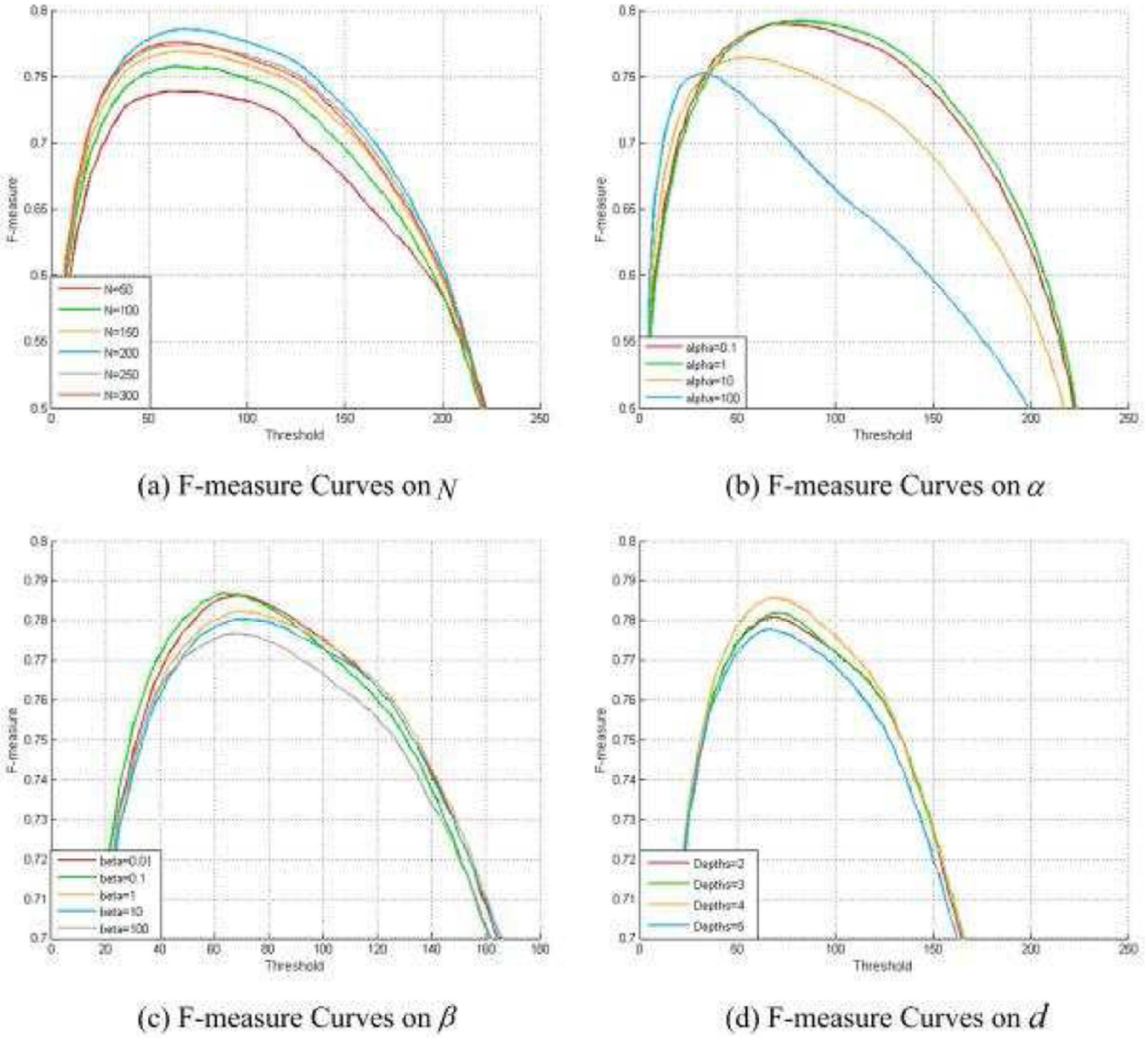


Fig. 8. Illustrations of parameters settings.

superpixel p_i and the background dictionary. Therefore, in this paper, the ℓ_1-norm of the vector $\mathbf{Z}(:, i)$, i.e., $\|\mathbf{Z}(:,i)\|_1 = \sum_{j=1}^{M} |\mathbf{Z}(j,i)|$, is adopted to compute the background probability of the superpixel p_i, i.e., $B_{\mathbf{Z}}(p_i) = \|\mathbf{Z}(:,i)\|_1$. As well, considering the coarse background probability $B_{VGG}(p_i)$ in Eq. (3) obtained by the simply modified VGG16 network, the background probability $B(p_i)$ of the superpixel p_i is further computed by

$$B(p_i) = \lambda B'_{\mathbf{Z}}(p_i) + (1-\lambda) B_{VGG}(p_i), \tag{8}$$

where λ is a balance parameter and is experimentally set to 0.5 in this paper. $B'_{\mathbf{Z}}(p_i)$ is the normalized version of $B_{\mathbf{Z}}(p_i)$ with the range of values in [0, 1]. Therefore, the representation coefficients based saliency measure $Sal_{\mathbf{Z}}(p_i)$ for superpixel p_i are defined as

$$Sal_{\mathbf{Z}}(p_i) = 1 - B(p_i). \tag{9}$$

3.4.2. Saliency measure based on reconstruction errors

Generally, a superpixel will be more salient if it has lager reconstruction errors under the same background dictionary. Considering that, the saliency measure $Sal_{\mathbf{E}}(p_i)$ of each superpixel p_i based on the reconstruction errors may be simplified to

$$Sal_{\mathbf{E}}(p_i) = \|\mathbf{E}(:,i)\|_1. \tag{10}$$

Similarly, the obtained saliency values of all the superpixels are also normalized to be in the range of [0, 1]. Higher $Sal_{\mathbf{E}}(p_i)$ indicates that the superpixel p_i is more likely to be a salient one.

3.4.3. Optimized saliency measure

Given the two superpixel-level based saliency measures $Sal_{\mathbf{Z}}(p_i)$ and $Sal_{\mathbf{E}}(p_i)$, the optimized saliency values $\{M_{sp}(p_i)|i = 1, 2, \ldots, N\}$

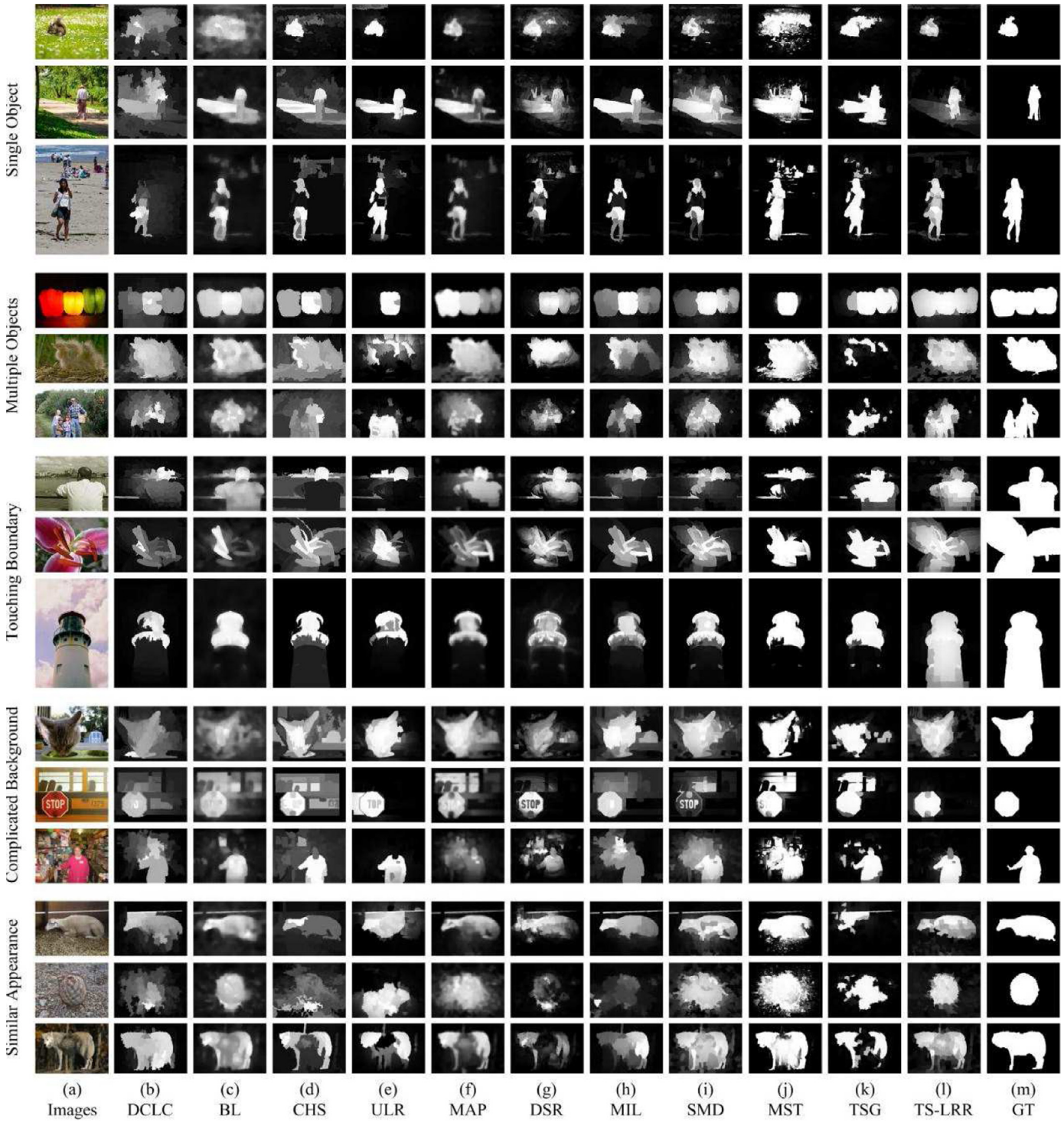


Fig. 9. Visual comparisons of different methods.

of all the superpixels are finally obtained by minimizing the following optimization model as in [30]

$$\sum_{i=1}^{N}(1-Sal_{\mathbf{Z}}(p_i))(M_{sp}(p_i))^2+\sum_{i=1}^{N}Sal_{\mathbf{E}}(p_i)(1-M_{sp}(p_i))^2 + \sum_{i,j} w_{ij}^{sal}\left(M_{sp}(p_i)-M_{sp}(p_j)\right)^2. \tag{11}$$

The weight ω_{ij}^{sal} is defined as

$$\omega_{ij}^{sal}=\exp\left(\frac{\left\|\mathbf{x}_i-\mathbf{x}_j\right\|_2^2}{2\sigma^2}\right)+\mu, \tag{12}$$

where σ is set to 0.05, which is the same as Eq. (1). It should be also noted that the parameter μ is a small constant (empirically set to 0.1) to regularize the optimization for cluttered image

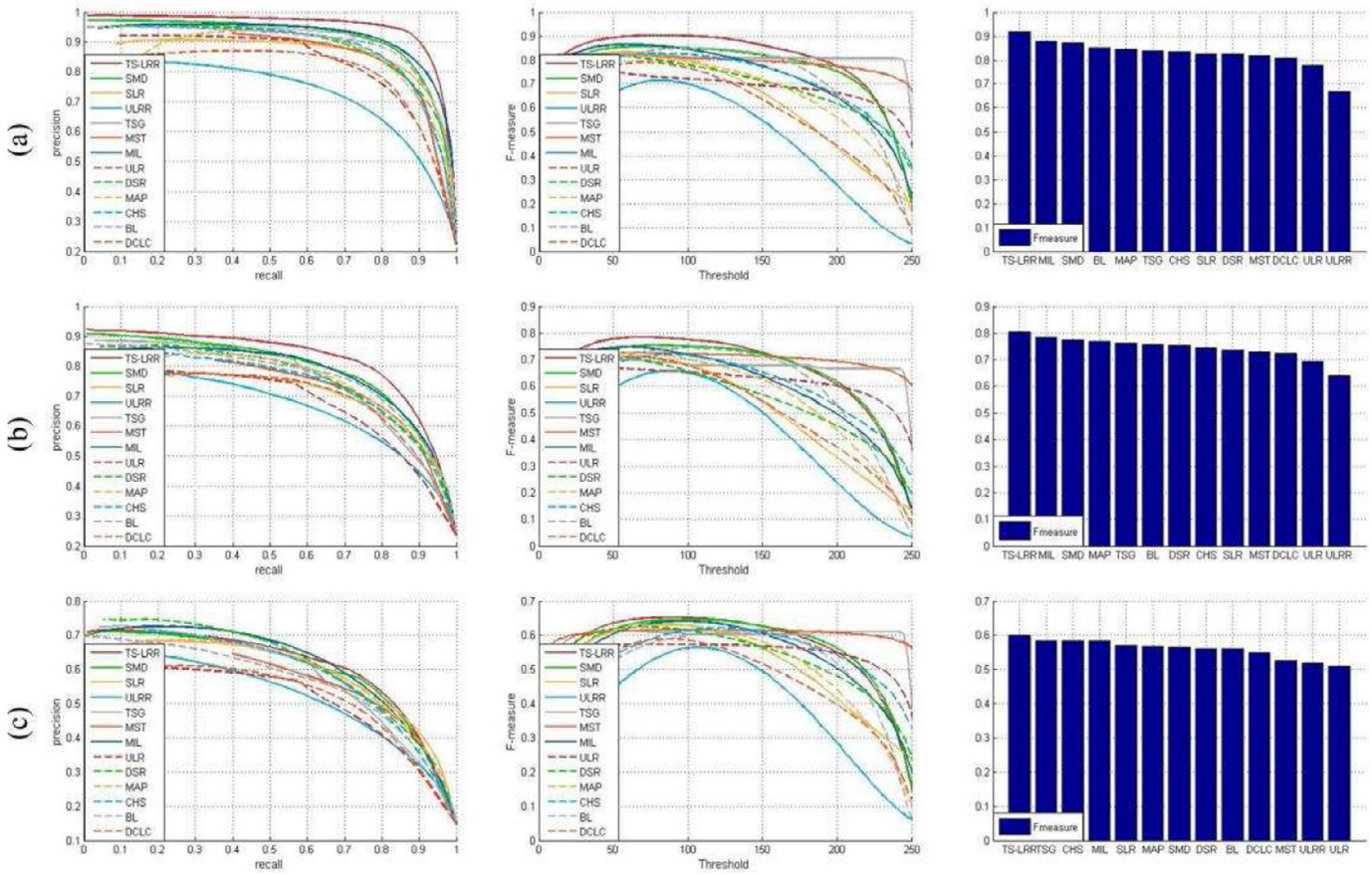


Fig. 10. PR curves, F-measure curves and average F-measure bars for (a) MSRA10K; (b) ECSSD; (c) DUT-OMRON.

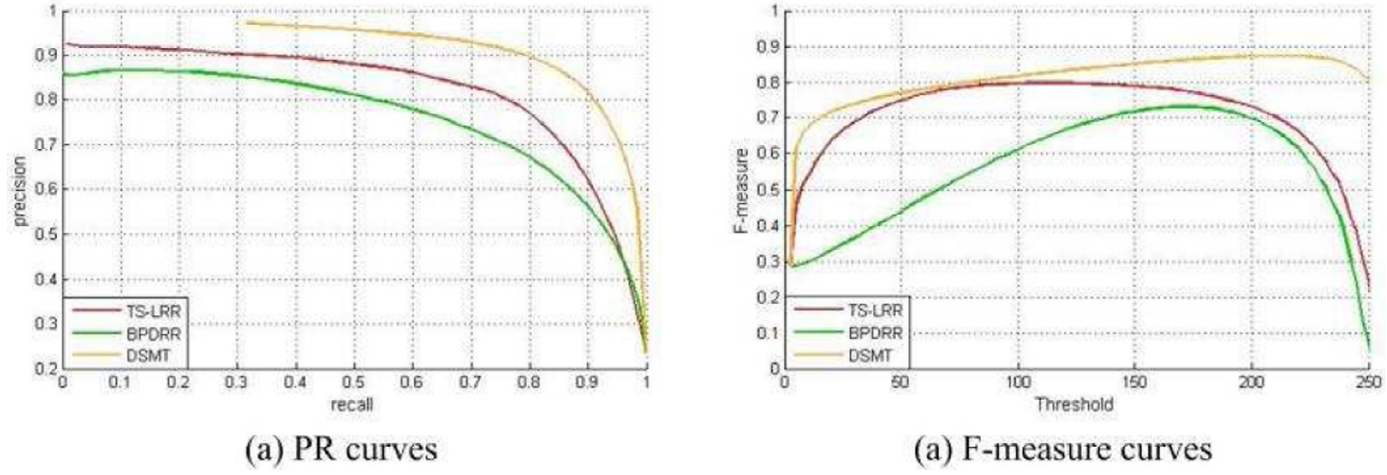


Fig. 11. Quantitative comparisons with two deep learning based methods on ECSSD.

regions. As discussed in [32], the introduction of μ is useful to erase small noise for both background and foreground regions.

Considering Eq. (9), Eq. (11) can be recast as

$$\sum_{i=1}^{N} B(p_i)(M_{sp}(p_i))^2 + \sum_{i=1}^{N} Sal_{\mathbf{E}}(p_i)(1 - M_{sp}(p_i))^2 + \sum_{i,j} \omega_{ij}^{sal}\left(M_{sp}(p_i) - M_{sp}(p_j)\right)^2. \quad (13)$$

Let $\mathbf{M}_{sp} = [M_{sp}(p_1), M_{sp}(p_2), \ldots, M_{sp}(p_N)]^T$ be the optimized saliency values of an image. The objective function in Eq. (11) is quadratic on $\mathbf{M}_{sp}$. Therefore, the optimization model is simply solved by a closed-form solution as

$$\mathbf{M}_{sp} = \left(\mathbf{D}_{diag}^{sal} - \mathbf{W}^{sal} + \mathbf{B}_{diag} + \mathbf{E}_{diag}\right)^{-1}\left(\mathbf{E}_{diag} * \mathbf{I}^{c}\right), \quad (14)$$

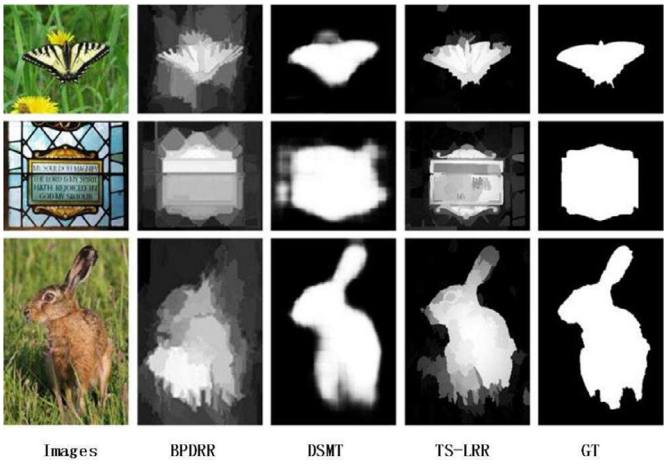


Fig. 12. Visual comparisons with two deep learning based methods on ECSSD.

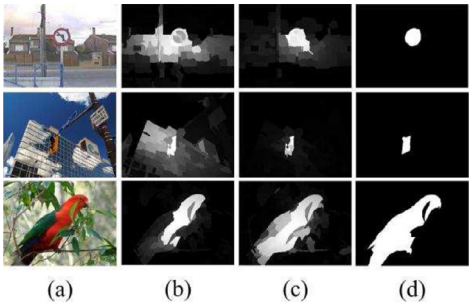


Fig. 13. Superiority of TS-LRR over LRR. (a) Images; (b) Saliency maps based on the traditional LRR; (c) Saliency maps based on the proposed TS-LRR; (d) Ground Truth.

where

$$\begin{cases} \mathbf{D}_{diag}^{sal} = diag\left\{d_1^{sal}, d_2^{sal}, \ldots, d_N^{sal}\right\}, \quad d_i^{sal} = \sum_j \omega_{ij}^{sal} \\ \mathbf{W}^{sal}(i,j) = \omega_{ij}^{sal} \\ \mathbf{B}_{diag} = diag\{B(p_1), B(p_2), \ldots, B(p_N)\} \\ \mathbf{E}_{diag} = diag\{Sal_{\mathbf{E}}(p_1), Sal_{\mathbf{E}}(p_2), \ldots, Sal_{\mathbf{E}}(p_N)\} \end{cases}. \tag{15}$$

In Eq. (14), $\mathbf{I}^c \in R^N$ is a column vector with N elements, which are all 1.

Given the saliency value $M_{sp}(p_i)$ of each superpixel, the saliency value $M_{pixel}(\mathbf{q})$ of a pixel $\mathbf{q}$ is directly obtained by

$$M_{pixel}(\mathbf{q}) = M_{sp}(p_i), \text{if} \mathbf{q} \in p_i. \tag{16}$$

Fig. 6 illustrates the validity of the optimized saliency measure. It can be easily observed that the optimized saliency measure in Eq. (14) can accurately detect the salient objects with better foreground uniformity and background suppression than the saliency measure in Eq. (9) and the saliency measure in Eq. (10). Similarly, Fig. 7[2] also indicates that the optimized saliency measure achieves better performance than the saliency measures based on the reconstruction errors and representation coefficients, respectively.

As well, it can be seen from Fig. 6 that the modified VGG16 can well locate the salient objects (Fig. 6(b)) but the boundaries of the detected objects are very blurry. However, the other saliency measures (Fig. 6(c)–(e)) can better capture the object boundaries. Especially, the saliency measure based on representation coefficients and the optimized saliency measure perform better than the modified VGG16, as shown in Fig. 7.

[2] The metrics about PR and F-measure curves will be explained in the later experimental part.

3.5. Computational complexity analysis

Suppose that the data matrix $\mathbf{X}$ and dictionary $\mathbf{D}$ are with the sizes of $m \times N$ and $m \times M$ $(1 \le M \le N)$, respectively. Then, the coefficients matrix $\mathbf{Z}$ has size of $M \times N (1 \le M \le N)$. The main computational cost of Algorithm 1 is updating $\mathbf{Z}$, which may require computing the product of three matrices. Therefore, the computational complexity of Algorithm 1 is about $O(rmM^2N)$, where r is the number of iterations until convergence. It demonstrates that the number of dictionary atoms M impacts significantly on the computational complexity of the proposed method than the other parameters. In the proposed method, M is about 50, and is far smaller than the total number of superpixels N (about 200). As a result of that, our model is computationally efficient.

4. Experiments and analysis

In this section, we conduct a series of experiments on three public benchmark datasets: MSRA10K [33], ECSSD [34] and DUT-OMRON [35]. MSRA10K [33] contains 10,000 images with relatively simple scenes and high contrast. ECSSD [34] has 1000 images with multiple objects and structurally complex scenes. DUT-OMRON [35] includes 5168 images with different-size objects in complex scenes. Apparently, DUT-OMRON [35] is more challenging for salient object detection.

In order to validate the effectiveness and superiority of the proposed method, another 10 state-of-the-art methods, including BL [36], CHS [34], DSR [28], SMD [10], DCLC [37], ULR [16], MAP [38], MIL [39], MST [40] and TSG [41], are also performed on the same datasets in this paper. In addition, we also consider two deep learning based algorithms, i.e., BPDRR [42] and DSMT [43], for comparison. For fair comparisons, we directly use the detection results (BL [36], CHS [34], DSR [28], SMD [10], DCLC [37], ULR [16], MIL [39], MST [40], TSG [41], BPDRR [42] and DSMT [43]) or detection results obtained by the codes (MAP [38]) published by their corresponding authors for comparisons.

Moreover, following [43,44], we use multiple widely used metrics to evaluate different methods objectively, including precision-recall curve, F-measure curve, average F-measure and mean absolute error (MAE).

Precision-Recall (PR). Given a continuous saliency map S, we convert it to a binary mask B using a threshold. Then, its precision and recall are computed as $precision = |B \cap S|/|B|$ and $recall = |B \cap S|/|S|$, respectively, where $|\cdot|$ accumulates the non-zero entries in a mask. The average precision/recall pairs on all the binary maps are computed with different thresholds to plot the precision-recall curve.

F-measure. The F-measure is formulated by a weighted combination of Precision and Recall

$$F_{measure} = \frac{\left(1+\beta^2\right) precision \times recall}{\beta^2 \times precision + recall}. \tag{17}$$

Here, we set β^2 to 0.3 to emphasize the precision over recall. A lot of F-measure values under different thresholds are computed to plot the F-measure curve.

Mean absolute error (MAE). MAE reflects the average pixel-wise absolute difference between the saliency map and Ground Truth.

$$MAE = \frac{1}{W \times H} \sum_{x=1}^{W} \sum_{y=1}^{H} |S(x,y) - B(x,y)|. \tag{18}$$

Here, W and H are the height and width of the image, respectively.

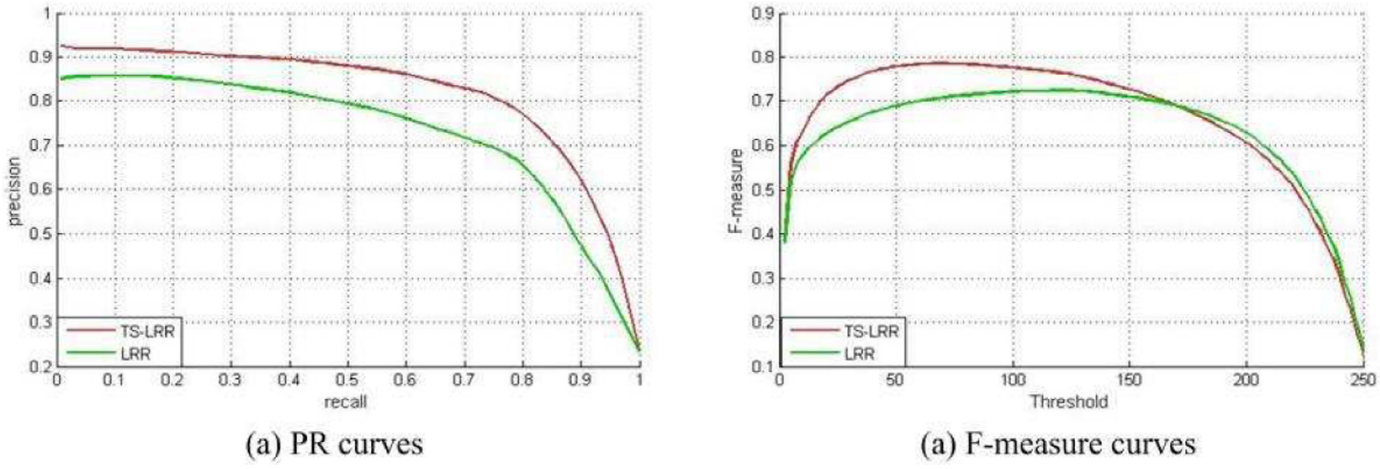


Fig. 14. Performance comparisons of TS-LRR and LRR.

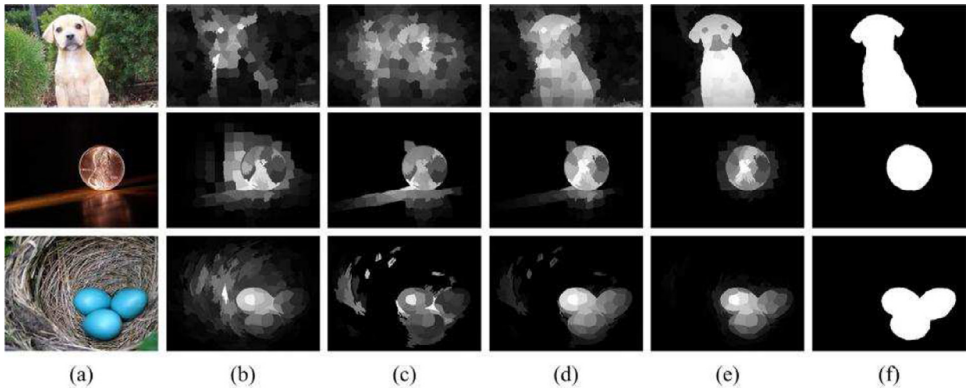


Fig. 15. Illustrations of LRR models with different dictionaries. (a) Images; (b) Orthogonal basis (employed by LRMR); (c) Data themselves; (d) Boundary background dictionary; (e) The proposed primitive background dictionary; (f) Ground Truth.

4.1. Implementation

4.1.1. Parameters settings

There are several important parameters in the proposed method, including the number of superpixels N, the parameters α and β in Eq. (2), and index-tree depth d. We set these important parameters based on a series of experiments on ECSSD. Fig. 8 illustrates the selections of different parameters. It can be easily seen from Fig. 8(a) that the proposed method achieves the best performance when N is set to 200. Therefore, N is set to 200 in this paper. Similarly, as illustrated in Fig. 8(b), (c), and (d), α, β, and d are set to 1, 0.1, and 4, respectively.

4.1.2. Training the simply modified VGG16 network

The MSRA10K dataset that contains 10,000 images with high contrast is employed to train the modified VGG16 network. We adopt the data augmentation technique by flipping all the training images horizontally. The corresponding weights in the modified VGG16 are initialized with the pretrained model of VGG16 [11]. The weights in the other newly added layers are initialized randomly with a truncated normal ($\sigma = 0.01$) and the biases are initialized to 0. The stochastic gradient descent (SGD) algorithm is adopted to train the modified VGG16 model with an initial learning rate of 10^{10} by using the "fixed" learning policy. The momentum parameter is set to 0.9 and the weight decay is 0.0005. The trained model is used for all the three public test datasets.

4.2. Comparisons with state-of-the-art methods

Fig. 9 shows the visual comparisons of different methods for those images with a single object, multiple objects, objects touching boundaries, complicated background, and similar appearance, respectively. It can be found from Fig. 9 that our proposed method is able to accurately detect the complete salient objects, while most of the others either just detect parts of the salient objects or mistakenly label some backgrounds as salient regions. Specifically, for those images with a single object, the proposed method can accurately extract the complete salient object with satisfactory uniformity. It can also suppress the background well. For those images with multiple objects, most of the state-of-the-art methods just detect parts of salient objects, as opposed to it, the proposed method can completely highlight all of the salient objects. For those images with objects touching boundaries, the foreground parts touching image boundaries are still completely detected by the proposed method, but are usually missed by the compared methods. For those images with complicated background, most of state-of-the-art methods fail to identify the salient objects, while our model accurately recognizes them. For those images with similar appearance, the proposed method successfully separates the salient objects from the similar background, which is difficult for the other methods.

Fig. 10 provides quantitative comparisons of different methods. As shown in Fig. 10(a) and (b), it can be easily found that the

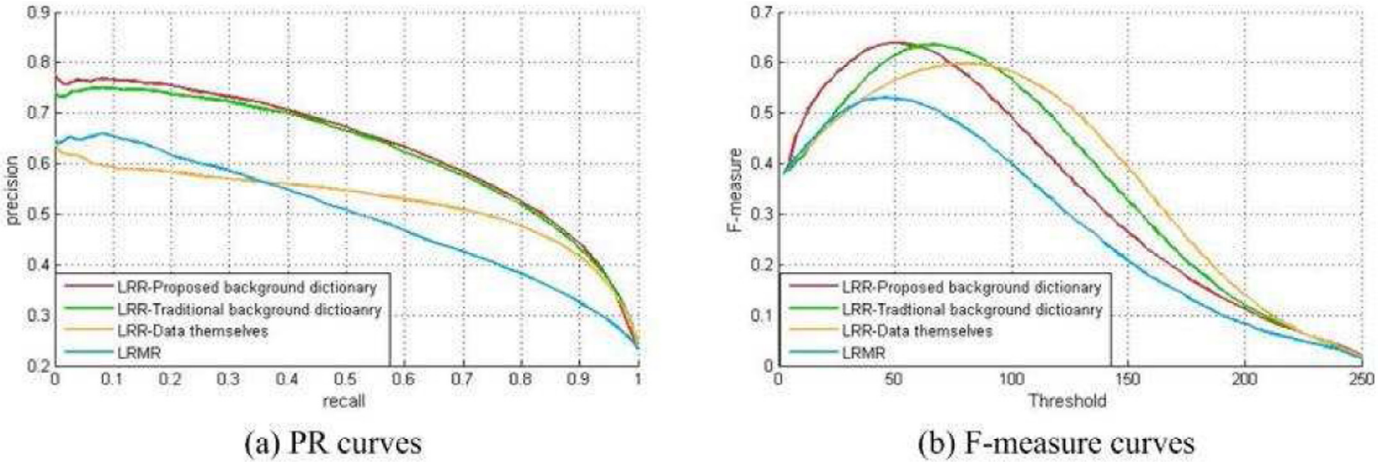


Fig. 16. Performance of LRR models with different dictionaries.

proposed method performs the best on MSRA10K and ECSSD in terms of the PR curves, F-measure curves, and average F-measure bars. From Fig. 10(c), it is clear that the proposed method is competitive with TSG [41] and DSR [28], and performs better than the other methods on DUT-OMRON.

4.3. Comparisons with deep learning based methods

In this section, we compare the proposed algorithm with two deep learning based methods, i.e., BPDRR [42] and DSMT [43]. Especially, the latter is based on CNNs. Figs. 11 and 12 show the quantitative comparisons and visual comparisons, respectively. According to Fig. 11, the proposed method performs better than BPDRR but worse than DSMT, which demonstrates that CNNs have great potential for salient object detection. However, as shown in Fig. 12, DSMT often produces blurry object boundaries, which can be well addressed by the proposed method. Moreover, the proposed method is an unsupervised one, while both DSMT and BPDRR are supervised algorithms, which require a large labeled dataset for training.

4.4. Experimental analysis of the proposed method

4.4.1. Superiority of TS-LRR over LRR

Fig. 13 illustrates the superiority of the proposed tree-structured low-rank representation (TS-LRR) over the traditional low-rank representation (LRR) for salient object detection.[3] For those images with complicated background regions (e.g., the first two rows of Fig. 13), the traditional LRR easily gets confused, which can be well solved by the proposed TS-LRR. Furthermore, as illustrated in the last row of Fig. 13, the proposed TS-LRR can even accurately extract the foreground regions that are similar to the background regions, while the traditional LRR wrongly label some foreground regions as the non-salient ones (e.g., the wings of the bird in the last row of Fig. 13(b)). Similarly, it can be easily seen from Fig. 14 that the proposed TS-LRR achieves much better performance than the traditional LRR.

4.4.2. Validity of the primitive background dictionary

To evaluate the effectiveness of the proposed primitive background dictionary, we compare a few versions of LRR models with different dictionaries, including orthogonal basis (employed by LRMR), data themselves, boundary background dictionary [28] and the proposed primitive background dictionary. It can be viewed from Fig. 15 that the dictionary using data themselves and the two background dictionaries (Fig. 15(c), (d) and (e)) performs better than the dictionary using orthogonal basis (Fig. 15(b)). The two background dictionaries (Fig. 15(d) and (e)) get better background suppression than the dictionary using data themselves (Fig. 15(c)). Furthermore, it can be easily found from Fig. 15(e) and (d) that the primitive background dictionary (Fig. 15(e)) achieves better background suppression and foreground prominence than the traditional background dictionary (Fig. 15(d)). Fig. 16 can also arrive at similar conclusions.

4.4.3. Superiority of the proposed foreground consistency over the traditional spatially-adjacent consistency

Fig. 17 illustrates the superiority of the proposed foreground consistency over the traditional spatially-adjacent consistency [10]. It can be easily seen that the introduction of the local consistency (Fig. 17(c) and d) is helpful for foreground uniformity. Compared with the traditional spatially-adjacent consistency (Fig. 17(c)), the proposed foreground consistency (Fig. 17(d)) achieves better foreground uniformity and salient object prominence. Similarly, Fig. 18 also demonstrates the superiority of the proposed foreground consistency over the traditional spatially-adjacent consistency.

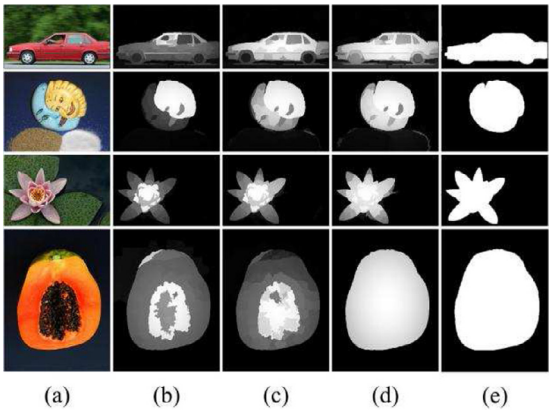


Fig. 17. Superiority of the proposed foreground consistency over the traditional spatially-adjacent consistency. (a) Images; (b) Saliency maps without local consistency; (c) Saliency maps using the traditional spatially-adjacent consistency; (d) Saliency maps using the proposed foreground consistency; (e) Ground Truth.

[3] For a fair comparison, we replace the TS-LRR term (i.e., the first term) in the optimization model in Eq. (2) with the traditional LRR term to obtain the saliency maps based on the traditional LRR.

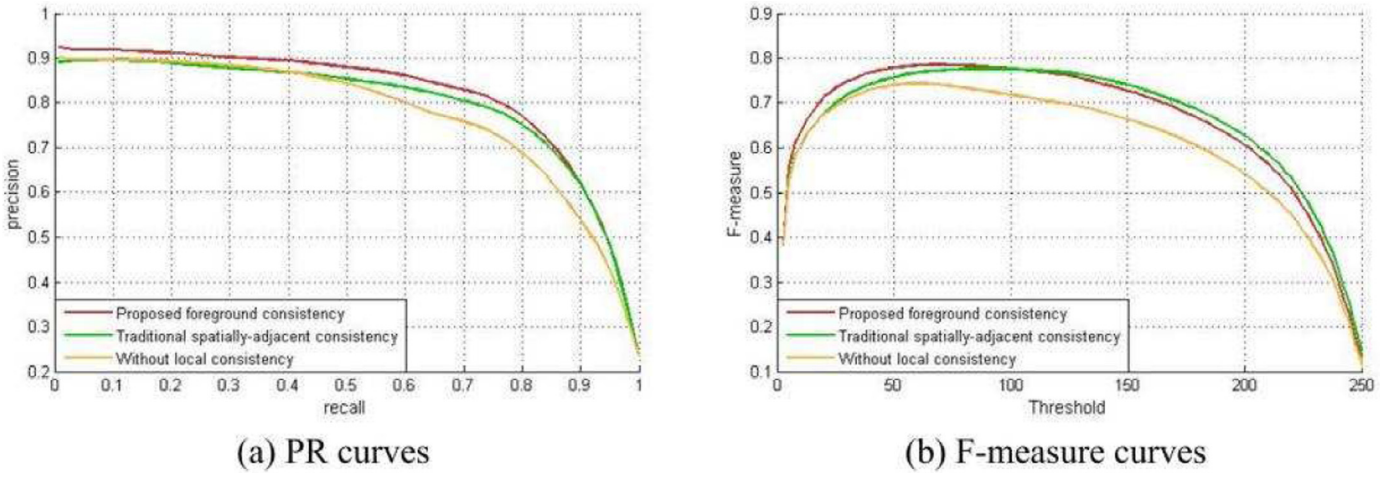


Fig. 18. Quantitative performance of different local consistency.

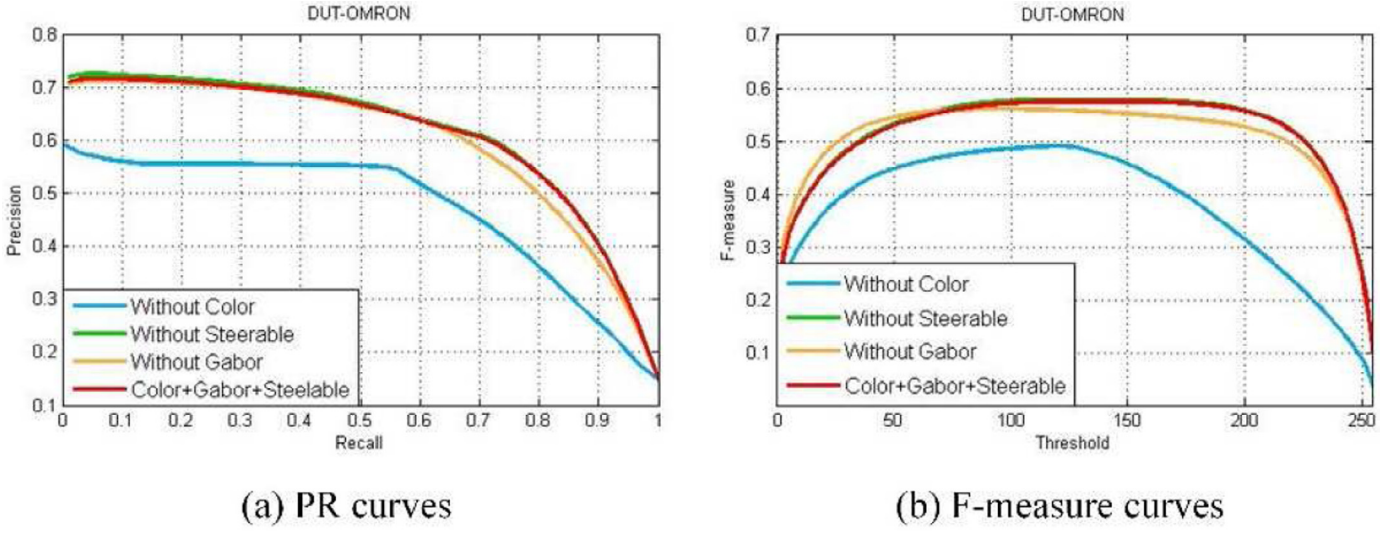


Fig. 19. PR and F-measure curves of different combinations of features on DUT-OMRON.

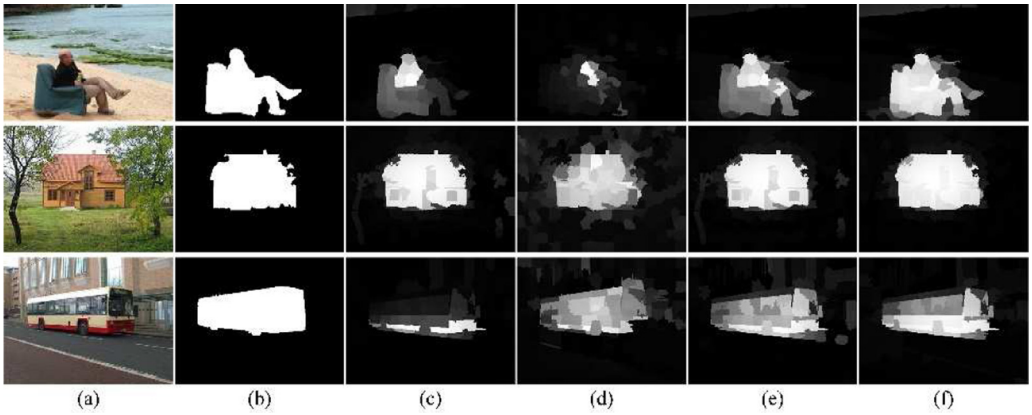


Fig. 20. Visual comparisons of different combinations of features. (a) Images; (b) Ground Truth; (c) Without Color; (d) Without Steerable; (e) Without Gabor; (f) Color + Gabor + Steerable.

4.4.4. Impacts of different types of features

As discussed in the previous Section 3.1.1, three types of features are employed in our proposed method, including color (the RGB color as well as the hue and the saturation components (5 dimensions)), steerable pyramids (12 dimensions), and Gabor filters (36 dimensions). In this subsection, we employ different types of features in our proposed method to test the impact of each type of features on the saliency detection results. Figs. 19 and 20 show the visual and quantitative saliency detection results on DUT-OMRON obtained by the proposed method but with different types of features, respectively. It can be easily found from Fig. 19 that the color features make a large contribution to the performance. Although the steerable pyramid features do not improve the quantitative performance to some extent, these features help to suppress those

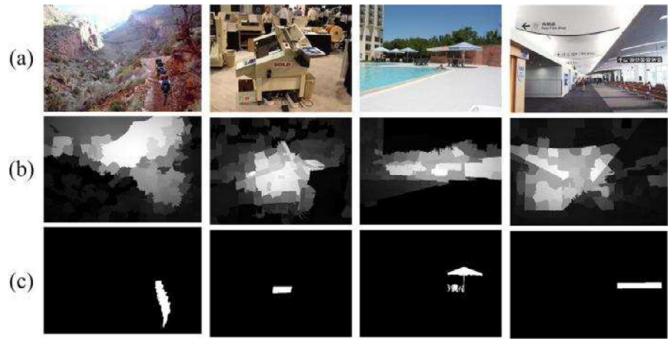


Fig. 21. Some failure cases of our method. (a) Images; (b) TS-LRR; (c) Ground Truth.

backgrounds with complicated textures, as shown in Fig. 20(d). Especially, as shown in Fig. 20(f), the combination of three types of features achieves better uniformity and completeness of the salient objects.

4.4.5. Failure cases

Fig. 21 displays some failure examples for the proposed method. These images in Fig. 21 contain much abundant background. Moreover, the salient objects in these images are too small to be recognized from the confused background. Considering its powerful potential, deep learning will be adopted to improve the performance of the proposed method in the future.

5. Conclusions

In this paper, we have proposed a new salient object detection model for those images with complicated backgrounds and diverse local salient regions. Specifically, a local tree-structured low-rank constraint is employed to capture the complicated background. A foreground consistency is exploited to promote the foreground uniformity among the diverse local salient regions. However, the detection of those salient objects that are too small or surrounded by a large amount of complicated backgrounds seems to be a challenge for our proposed method. To address this problem, we will integrate the deep features learned from CNNs in our proposed method or integrate the idea of our proposed method in the CNNs architecture in the future. Another possible future work is to apply our saliency detector to industrial applications, such as object tracking [45–47] and instance-level object retrieval [48–50].

Acknowledgments

This work is supported by the National Natural Science Foundation of China under Grant no. 61773301, and by the Fundamental Research Funds for the Central Universities under Grant no. JBZ170401.

Appendix A

In this appendix, the update scheme for solving Eq. (7) in the text is described in detail.

(1) Update $\mathbf{J}$

$$\begin{aligned}\mathbf{J}^{k+1} &= \arg\min_{\mathbf{J}} \sum_{i=1}^{d}\sum_{j=1}^{n_i} \left\|\mathbf{J}_{G_j^i}\right\|_* + \left\langle \mathbf{Y}_2^k, \mathbf{Z}^k - \mathbf{J}\right\rangle + \tfrac{\mu^k}{2}\left\|\mathbf{Z}^k - \mathbf{J}\right\|_F^2 \\ &= \arg\min_{\mathbf{J}} \lambda \sum_{i=1}^{d}\sum_{j=1}^{n_i} \left\|\mathbf{J}_{G_j^i}\right\|_* + \tfrac{1}{2}\left\|\mathbf{J} - \mathbf{X}_{\mathbf{J}}\right\|_F^2\end{aligned}, \tag{A1}$$

where $\lambda = 1/\mu^k$, and $\mathbf{X}_{\mathbf{J}} = \mathbf{Z}^k + \mathbf{Y}_2^k/\mu^k$. The solution of each sub-region G_j^i is recast as follows:

$$\mathbf{J}_{G_j^i}^{k+1} = \arg\min_{\mathbf{J}_{G_j^i}} \lambda \left\|\mathbf{J}_{G_j^i}\right\|_* + \frac{1}{2}\left\|\mathbf{J}_{G_j^i} - \mathbf{X}_{\mathbf{J}_{G_j^i}}\right\|_F^2. \tag{A2}$$

The solution to Eq. (A2) can be derived as [31]:

$$\mathbf{J}_{G_j^i}^{k+1} = \mathbf{U} diag\left(\left\{(\sigma_i - \tau)_+\right\}\right)\mathbf{V}^T, \tag{A3}$$

where $(\mathbf{U}, \mathbf{\Sigma}, \mathbf{V}^T) = SVD(\mathbf{X}_{\mathbf{J}_{G_j^i}})$.

(1) Update $\mathbf{Z}$

$$\begin{aligned}\mathbf{Z}^{k+1} &= \arg\min_{\mathbf{Z}} \left\langle \mathbf{Y}_1^k, \mathbf{X} - \mathbf{DZ} - \mathbf{E}^k\right\rangle + \left\langle \mathbf{Y}_2^k, \mathbf{Z} - \mathbf{J}^{k+1}\right\rangle \\ &\quad + \frac{\mu^k}{2}\left\|\mathbf{X} - \mathbf{DZ} - \mathbf{E}^k\right\|_F^2 + \frac{\mu^k}{2}\left\|\mathbf{Z} - \mathbf{J}^{k+1}\right\|_F^2 \\ &= \arg\min_{\mathbf{Z}} \frac{\mu^k}{2}\left\|\mathbf{X} - \mathbf{DZ} - \mathbf{E}^k + \frac{1}{\mu^k}\mathbf{Y}_1^k\right\|_F^2 \\ &\quad + \frac{\mu^k}{2}\left\|\mathbf{Z} - \mathbf{J}^{k+1} + \frac{1}{\mu^k}\mathbf{Y}_2^k\right\|_F^2.\end{aligned} \tag{A4}$$

Taking the derivative of the objective function in Eq. (A4), we have [31]:

$$\mathbf{Z}^{k+1} = \left(\mathbf{D}^T\mathbf{D} + \mathbf{I}\right)^{-1}\left(\mathbf{D}^T\left(\mathbf{X} - \mathbf{E}^k\right) + \mathbf{J}^{k+1} + \frac{1}{\mu^k}\left(\mathbf{D}^T\mathbf{Y}_1^k - \mathbf{Y}_2^k\right)\right). \tag{A5}$$

(1) Update $\mathbf{H}$

$$\begin{aligned}\mathbf{H}^{k+1} &= \arg\min_{\mathbf{H}} \beta \mathrm{Tr}\left(\mathbf{HLH}^T\right) + \left\langle \mathbf{Y}_3^k, \mathbf{E}^k - \mathbf{H}\right\rangle + \tfrac{\mu^k}{2}\left\|\mathbf{E}^k - \mathbf{H}\right\|_F^2 \\ &= \arg\min_{\mathbf{H}} \beta \mathrm{Tr}\left(\mathbf{HLH}^T\right) + \tfrac{\mu^k}{2}\left\|\mathbf{E}^k - \mathbf{H} + \tfrac{1}{\mu^k}\mathbf{Y}_3^k\right\|_F^2\end{aligned}. \tag{A6}$$

The solution of $\mathbf{H}^{k+1}$ in Eq. (A6) has a closed-form as in [31]:

$$\mathbf{H}^{k+1} = \left(\mu^k\mathbf{E}^k + \mathbf{Y}_3^k\right)\left(2\beta\mathbf{L} + \mu^k\mathbf{I}\right)^{-1}. \tag{A7}$$

(1) Update $\mathbf{E}$

$$\begin{aligned}\mathbf{E}^{k+1} &= \arg\min_{\mathbf{E}} \alpha \sum_{i=1}^{d}\sum_{j=1}^{n_i} \upsilon_j^i \left\|\mathbf{E}_{G_j^i}\right\|_\infty + \left\langle \mathbf{Y}_1^k, \mathbf{X} - \mathbf{DZ}^{k+1} - \mathbf{E}\right\rangle \\ &\quad + \left\langle \mathbf{Y}_3^k, \mathbf{E} - \mathbf{H}^{k+1}\right\rangle + \frac{\mu^k}{2}\left\|\mathbf{X} - \mathbf{DZ}^{k+1} - \mathbf{E}\right\|_F^2 + \frac{\mu^k}{2}\left\|\mathbf{E} - \mathbf{H}^{k+1}\right\|_F^2 \\ &= \arg\min_{\mathbf{E}} \eta \sum_{i=1v}^{d}\sum_{j=1}^{n_i} \upsilon_j^i \left\|\mathbf{E}_{G_j^i}\right\|_\infty + \frac{1}{2}\left\|\mathbf{E} - \mathbf{X}_E\right\|_F^2,\end{aligned} \tag{A8}$$

where $\eta = \alpha/(2\mu^k)$ and $\mathbf{X}_E = (\mathbf{X} - \mathbf{DZ}^{k+1} + \mathbf{H}^{k+1} + (\mathbf{Y}_1^k - \mathbf{Y}_3^k)/\mu^k)/2$. The above problem can be solved by the hierarchical proximal operator [23]:

$$\mathbf{E}_{G_j^i}^{k+1} = \begin{cases} \dfrac{\left\|\mathbf{E}_{G_j^i}\right\|_1 - \eta\upsilon_j^i}{\left\|\mathbf{E}_{G_j^i}\right\|_1}\mathbf{E}_{G_j^i}, & \text{if } \left\|\mathbf{E}_{G_j^i}\right\|_1 > \eta\upsilon_j^i \\ 0, & \text{otherwise} \end{cases}. \tag{A9}$$

References

[1] T. Liu, Z. Yuan, J. Sun, N.N. Zheng, Learning to detect a salient object, IEEE Trans. Pattern Anal. Mach. Intell. 33 (2) (2011) 353–367.
[2] X. Zhi, H. Shen, Saliency driven region-edge-based top down level set evolution reveals the asynchronous focus in image segmentation, Pattern Recognit. 80 (2018) 241–255.
[3] Y. Gao, M. Wang, D. Tao, R. Ji, Q. Dai, 3-D object retrieval and recognition with hypergraph analysis, IEEE Trans. Image Process. 21 (9) (2012) 4290–4303.
[4] H. Zhang, T. Zhang, W. Pedrycz, C. Zhao, D. Miao, Improved adaptive image retrieval with the use of shadowed sets, Pattern Recognit. 90 (2019) 390–403.
[5] C. Siagian, L. Itti, Rapid biologically-inspired scene classification using features shared with visual attention, IEEE Trans. Pattern Anal. Mach. Intell. 29 (2) (2007) 300–312.
[6] D. Gao, S. Han, N. Vasconcelos, Discriminant saliency, the detection of suspicious coincidences, and applications to visual recognition, IEEE Trans. Pattern Anal. Mach. Intell. 31 (6) (2009) 989–1005.
[7] A. Abdulmunem, Y.-K. Lai, X. Sun, Saliency guided local and global descriptors for effective action recognition, Comput. Vis. Media 2 (1) (2016) 97–106.
[8] C. Lang, G. Liu, J. Yu, S. Yan, Saliency detection by multitask sparsity pursuit, IEEE Trans. Image Process. 21 (3) (2012) 1327–1338.
[9] X. Shen, Y. Wu, A unified approach to salient object detection via low rank matrix recovery, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2012, pp. 853–860.
[10] H. Peng, B. Li, H. Ling, W. Hu, W. Xiong, S.J. Maybank, Salient object detection via structured matrix decomposition, IEEE Trans. Pattern Anal. Mach. Intell. 39 (4) (2017) 818–832.
[11] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, in: Proceedings of the International Conference on Learning Representations (ICLR), 2015.
[12] J. Tao, D. Song, S. Wen, W. Hu, Robust multi-source adaptation visual classification using supervised low-rank representation, Pattern Recognit. 61 (2017) 47–65.
[13] Z. Zheng, M. Yu, J. Jia, Fisher discrimination based low rank matrix recovery for face recognition, Pattern Recognit. 47 (11) (2014) 3502–3511.
[14] J. Yan, M. Zhu, H. Liu, Visual saliency detection via sparsity pursuit, IEEE Signal Process. Lett. 17 (8) (2010) 739–742.
[15] X. Sun, Z. He, C. Xu, X. Zhang, W. Zou, G. Baciu, Diversity induced matrix decomposition model for salient object detection, Pattern Recognit. 66 (2017) 253–267.
[16] M. Zhao, L. Jiao, W. Ma, Classification and saliency detection by semi-supervised low-rank representation, Pattern Recognit. 51 (2016) 281–294.
[17] Q. Zhang, Y. Liu, S. Zhu, J. Han, Salient object detection based on super-pixel clustering and unified low-rank representation, Comput. Vis. Image Understand. 161 (2017) 51–64.
[18] L. Wang, H. Lu, X. Ruan, M.H. Yang, Deep networks for saliency detection via local estimation and global search, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 3183–3192.
[19] G. Li, Y. Yu, Visual saliency based on multiscale deep features, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 5455–5463.
[20] H. Chen, Y. Li, D. Su, Multi-modal fusion network with multi-scale multi–path and cross-modal interactions for RGB-D salient object detection, Pattern Recognit. 86 (2019) 376–385.
[21] R Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, S. Süsstrunk, SLIC superpixels compared to state-of-the-art superpixel methods, IEEE Trans. Pattern Anal. Mach. Intell. 34 (11) (2012) 2274–2282.
[22] E.P. Simoncelli, W.T. Freeman, The steerable pyramid: a flexible architecture for multi-scale derivative computation, in: Proceedings of the IEEE International Conference on Image Processing (ICIP), 1995, pp. 23–26.
[23] H.G. Feichtinger, T. Strohmer, Gabor Analysis and Algorithms: Theory and Applications, Birkhäuser, Boston, 1998.
[24] P.F. Felzenszwalb, D.P. Huttenlocher, Efficient graph-based image segmentation, Int. J. Comput. Vis. 59 (2) (2004) 167–181.
[25] G. Lee, Y.-W. Tai, J. Kim, ELD-Net: an efficient deep learning architecture for accurate saliency detection, IEEE Trans. Pattern Anal. Mach. Intell. 40 (7) (2018) 1599–1610.
[26] L. Wang, L. Wang, H. Lu, Salient object detection with recurrent fully convolutional networks, IEEE Trans. Pattern Anal. Mach. Intell. (2018) 1, doi:10.1109/TPAMI.2018.2846598.
[27] G. Li, Y. Yu, Deep contrast learning for salient object detection, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 478–487.
[28] H. Lu, X. Li, L. Zhang, X. Ruan, Dense and sparse reconstruction error based saliency descriptor, IEEE Trans. Pattern Anal. Mach. Intell. 25 (4) (2016) 1592–1603.
[29] P. Krahenbuhl, Saliency filters: contrast based filtering for salient region detection, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2012, pp. 733–740.
[30] L. Huo, S. Yang, L. Jiao, S. Wang, Local graph regularized sparse reconstruction for salient object detection, Neurocomputing 194 (2016) 348–359.
[31] Z. Lin, R. Liu, Z. Su, Linearized alternating direction method with adaptive penalty for low-rank representation, in: Proceedings of the Neural Information Processing Systems (NIPS), 2011, pp. 612–620.
[32] W. Zhu, S. Liang, Y. Wei, J. Sun, Saliency optimization from robust background detection, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 2814–2821.
[33] M.-M. Cheng, N.J. Mitra, X. Huang, Global contrast based salient region detection, IEEE Trans. Pattern Anal. Mach. Intell. 37 (3) (2015) 569–582.
[34] J. Shi, Q. Yan, L. Xu, J. Jia, Hierarchical image saliency detection on extended CSSD, IEEE Trans. Pattern Anal. Mach. Intell. 38 (4) (2016) 717–729.
[35] C. Yang, L. Zhang, H. Lu, X. Ruan, M.-H. Yang, Saliency detection via graph-based manifold ranking, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 3166–3173.
[36] N. Tong, H. Lu, X. Ruan, M.-H. Yang, Salient object detection via bootstrap learning, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 1884–1892.
[37] L. Zhou, Z. Yang, Q. Yuan, Z. Zhou, D. Hu, Salient region detection via integrating diffusion-based compactness and local contrast, IEEE Trans. Image Process. 24 (11) (2015) 3308–3320.
[38] J. Sun, H. Lu, X. Liu, Saliency region detection based on markov absorption probabilities, IEEE Trans. Image Process. 24 (5) (2015) 1639–1649.
[39] F. Huang, J. Qi, H. Lu, L. Zhang, X. Ruan, Salient object detection via multiple instance learning, IEEE Trans. Image Process. 26 (4) (2017) 1911–1922.
[40] W.-C. Tu, S. He, Q. Yang, S.-Y. Chien, Real-time salient object detection with a minimum spanning tree, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (CVPR), 2016, pp. 2334–2342.
[41] Y. Liu, J. Han, Q. Zhang, L. Wang, Salient object detection via two-stage graphs, IEEE Trans. Circuits Syst. Video Technol. (2018) 1, doi:10.1109/TCSVT.2018.2823769.
[42] J. Han, D. Zhang, X. Hu, L. Guo, J. Ren, F. Wu, Background prior-based salient object detection via deep reconstruction residual, IEEE Trans. Circuits Syst. Video Technol. 25 (8) (2015) 1309–1321.
[43] X. Li, L. Zhao, L. Wei, DeepSaliency: multi-task deep neural network model for salient object detection, IEEE Trans. Image Process. 25 (8) (2016) 3919–3930.
[44] C. Aytekin, A. Iosifidis, M. Gabbouj, Probabilistic saliency estimation, Pattern Recognit. 74 (2018) 359–372.
[45] J. Han, E. Pauwels, P. de Zeeuw, P. de With, Employing a RGB-D sensor for real-time tracking of humans across multiple re-entries in a smart environment, IEEE Trans. Consum. Electron. 58 (2) (2016) 255–263.
[46] C. Yan, H. Xie, J. Chen, Z. Zha, X. Hao, Y. Zhang, Q. Dai, A fast Uyghur text detection for complex background images, IEEE Trans. Multimed. 20 (12) (2018) 3389–3398.
[47] G. Ding, W. Chen, S. Zhao, J. Han, Q. Liu, Real-time scalable visual tracking via quadrangle kernelized correlation filters, IEEE Trans. Intell. Transp. Syst. 19 (1) (2018) 140–150.
[48] C. Yan, H. Xie, D. Yang, J. Yin, Y. Zhang, Q. Dai, Supervised hash coding with deep neural network for environment perception of intelligent vehicles, IEEE Trans. Intell. Transp. Syst. 19 (1) (2018) 284–295.
[49] G. Wu, J. Han, Y. Guo, L. Liu, G. Ding, Q. Ni, L. Shao, Unsupervised deep video hashing via balanced code for large-scale video retrieval, IEEE Trans. Image Process. 28 (4) (2019) 1993–2007.
[50] G. Wu, J. Han, Z. Lin, G. Ding, B. Zhang, Q. Ni, Joint image-text hashing for fast large-scale cross-media retrieval using self-supervised deep learning, IEEE Trans. Ind. Electron., in press (2019), doi:10.1109/TIE.2018.2873547.

Qiang Zhang received the B.S. degree in automatic control, the M.S. degree in pattern recognition and intelligent systems, and the Ph.D. degree in circuit and system from Xidian University, China, in 2001, 2004, and 2008, respectively. He was a Visiting Scholar with the Center for Intelligent Machines, McGill University, Canada. He is currently a professor with the Automatic Control Department, Xidian University, China. His current research interests include image processing, computer vision and pattern recognition.

Zhen Huo received the B.S. degree in Automatic Control from Xi'an Polytechnic University, Xi'an, China, in 2015. He is currently pursuing his M.S. degree in Control Engineering at Xidian University, Xi'an, China. His current research interests include computer vision and salient object detection.

Yi Liu received the B. S. degree from Nanjing Institute of Technology, Nanjing, China, in 2012, and the M. S. degree from the Dalian University, Dalian, China, in 2015. He is currently working towards the Ph.D. degree in Control Theory and Control Engineering at Xidian University, Xian. His current research interests include computer vision and salient object detection.

Yunhui Pan received the B.S. degree in Automation from Xidian University, Xi'an, China, in 2014. She is currently pursuing his M.S. degree in Control Engineering at Xidian University, Xi'an, China. Her current research interests include computer vision and salient object detection.

Caifeng Shan is currently a Senior Scientist and Project Leader with Philips Research, Eindhoven, The Netherlands. He is also a part-time researcher at Eindhoven University of Technology. He received the PhD degree in computer vision from Queen Mary, University of London. His research interests include computer vision, pattern recognition, image and video analysis, machine learning, bio-medical imaging, and related applications. He has authored more than 80 scientific papers and 50 patent applications. He has been Associate Editor and Guest Editor of many

scientific journals including IEEE Transactions on Circuits and Systems for Video Technology (T-CSVT), IEEE Transactions on Multimedia (T-MM), IEEE Journal of Biomedical and Health Informatics (J-BHI), Journal of Visual Communication and Image, Signal Processing (Elsevier), Machine Vision and Applications, and Journal of Real-Time Image Processing. He has organized several international conferences and workshops, and served as Program Committee Member and Reviewer for numerous international conferences and journals. He is a Senior Member of IEEE.

Jungong Han is a tenured Associate Professor with the School of Computing and Communications at Lancaster University, Lancaster, UK. Previously, he was a faculty member with the Department of Computer and Information Sciences at Northumbria University, UK. His research interests include computer vision, image processing, machine learning, and artificial intelligence.