Kernelized Subspace Ranking for Saliency Detection

Tiantian Wang, Lihe Zhang(✉), Huchuan Lu, Chong Sun, and Jinqing Qi

School of Information and Communication Engineering,
Dalian University of Technology, Dalian, China
tiantianwang.ice@gmail.com, {zhanglihe,lhchuan,Jinqing}@dlut.edu.cn,
waynecool@mail.dlut.edu.cn

Abstract. In this paper, we propose a novel saliency method that takes advantage of object-level proposals and region-based convolutional neural network (R-CNN) features. We follow the learning-to-rank methodology, and solve a ranking problem satisfying the constraint that positive samples have higher scores than negative ones. As the dimensionality of the deep features is high and the amount of training data is low, ranking in the primal space is suboptimal. A new kernelized subspace ranking model is proposed by jointly learning a Rank-SVM classifier and a subspace projection. The projection aims to measure the pairwise distances in a low-dimensional space. For an image, the ranking score of each proposal is assigned by the learnt ranker. The final saliency map is generated by a weighted fusion of the top-ranked candidates. Experimental results show that the proposed algorithm performs favorably against the state-of-the-art methods on four benchmark datasets.

Keywords: Saliency detection · Subspace ranking · Feature projection

1 Introduction

The task of saliency detection is to identify the most attractive and informative regions in images and videos. It has gained much popularity in recent years, owing to its series of important applications in computer vision, such as adaptive compression, context-aware image editing and image resizing. An effective saliency model can save lots of unnecessary human labour in vision tasks. Although much progress in saliency detection has been made in recent years, it remains a challenging problem.

The early works [19,21] exploit the low-level image properties of pixels, such as intensity, color, orientation, texture and motion, to compute saliency. Numerous region-wise saliency methods [10,13,43] are proposed subsequently, which investigate the mid-level structure properties of image regions and incorporate the contextual information to measure the saliency for each region. The aforementioned works, either in the pixel-wise or region-wise fashion, have to fully consider the relationship between image elements from overall and local perspectives to guarantee the semantic completeness of salient objects. In this work, we

B. Leibe et al. (Eds.): ECCV 2016, Part VIII, LNCS 9912, pp. 450–466, 2016.
DOI: 10.1007/978-3-319-46484-8_27

explore the category-independent object characteristics of region proposals, and propose a principled framework to weight and combine these region candidates, thereby highlighting the salient instances.

Object proposals technique has been widely applied to many vision fields. It generally produces either bounding box proposals [2,8] which inevitably aggregate visual information from objects and background clutter, or region proposals [4,11,32] that shape an informative and well-defined contour. This technique, striving to find instances of all categories, usually produces thousands of object candidates which significantly reduce the search space of salient object detection. Furthermore, good proposals encapsulate the visual information of objects and have informative boundary shape cues, which provide the necessary object-level prior knowledge for saliency detection. However, a very large proportion of proposals contain very few object regions, even may contain only the backgrounds. Owing to the diversity of object categories and backgrounds as well as the varied shape and size of proposals, it is extremely difficult to select the proposals most similar to the ground truth. Therefore, the proposal-based salient object detection remains a very challenging task. Recent works [22,25,36] simply integrate the bounding box proposals weighted by their objectness scores [2] as a feature map to help saliency detection. Since the computed score in [2] is inaccurate, the feature map only coarsely indicates the location of the objects in the image. In this work, we aim to sort out some good region proposals that contain a part of the objects even a complete object instance, and employ them to detect salient objects. Figure 1 shows some examples of object proposal. Actually, background proposals (i.e., contain much more background pixels than foreground ones) are in majority in the proposal pool compared with foreground proposals.

Recent progress on metric learning models [20,44,47] reveals the effectiveness of an optimal distance metric, which can significantly narrow the distances between similar samples and simultaneously expand the gaps of dissimilar samples. Besides, the performance of ranking models highly depends on the pairwise similarity/dissimilarity constraints. Therefore, we combine the two learning mechanisms together to propose a novel ranking model for proposal selection. The core idea of the proposed model is to learn a category-independent ranker upon distance metric learning of object proposals with a joint learning approach, thereby obtaining the optimal orderings of object proposals and linearly combining the top-ranked candidates weighted by their ranking scores. By using the projection obtained in distance learning, data points (i.e. object proposals) are mapped into a low-dimensional subspace, and further ranked on the data manifold constructed by the learnt distances. Thus the positive and negative pairs of data can be more easily separated. Different from superpixels that contain the low-level and mid-level image information, object proposals carry more higher-level and object-level cues. Therefore, the hand-crafted features used to represent the superpixels are not suitable for the proposals. To overcome this problem, we adopt the region-based convolutional neural network (R-CNN) features, which depict both the low-level and high-level image cues and demonstrate very powerful representation capability, as witnessed in recent works [12,33].

Fig. 1. Several examples of object proposals. From left to right: (a) input, (b) foreground proposals (blue color) and background proposals (red color). (c) ground truth. (d) saliency map generated by ranking in the primal space. (e) saliency map generated by ranking in the kernelized subspace. (Color figure online)

The contributions of this work are listed as follows:

- We jointly learn a ranker and a distance metric with a kernel approach to formulate salient object detection as a subspace ranking issue.
- We propose a object-wise saliency model, which purely exploits object candidates represented with R-CNN features to achieve saliency detection. The deep features can capture the high-level saliency cues of object candidates.
- It is demonstrated that the proposed algorithm performs favorably against the state-of-the-art saliency detection methods on MSRA-5000, ECSSD, PASCAL-S and SOD benchmark datasets.

2 Related Work

Numerous saliency models and algorithms have emerged recently, which can be roughly classified into unsupervised, semi-supervised and supervised schemes. We refer the readers to a comprehensive review on this topic in [3,50].

Unsupervised approaches usually heuristically characterize visual rarity or distinctness to define image saliency. The most common way to quantify rarity is to calculate the difference between various visual elements. Itti *et al.* [21] integrate the center-surround contrasts on multiple feature channels and scales to estimate visual saliency. Achanta *et al.* [1] compute the different between Gaussian blurred features and mean image features to define rarity. Some methods combine appearance difference and spatial coherence to calculate the global contrast with different image abstraction representations [9,10,48]. While Goferman *et al.* [13] and Wang *et al.* [53] investigate image saliency from both local and global perspectives. Wei *et al.* [55] compute the shortest distance of each patch to image boundary as saliency measure. In addition, much effort has been made to design discriminative features [37,43,53], domain models [16,19], distance metrics [31], speed strategy [59].

Generally, semi-supervised approaches achieve saliency detection by label propagation from the labeled elements to the unlabeled ones based on their

pairwise affinities. Harel *et al.* [17] formulate saliency labeling as a random walk problem, construct an ergodic Markov chain and use the equilibrium distribution to define image saliency. Wang *et al.* [54] and Gopalakrishnan *et al.* [15] respectively introduce the entropy rate and the hitting time based on ergodic Markov chains. Different from the aforementioned methods, Jiang *et al.* [23] construct an absorbing Markov chain and use the absorbed time to measure the saliency. While Yang *et al.* [58] cast saliency detection into a manifold ranking problem, which is also a propagation-based method. Recently, Li *et al.* [34] combine random walks and manifold ranking to propose the regularized random walks ranking with a newly defined constraint to consider local image data and prior estimation. Li and Yu [35] use quadratic energy models to refine the initial saliency results generated by deep convolutional neural networks. There are some similar methods [5,14,26,41,49,57,63]. They use a semi-supervised learning mechanism to assign saliency labels based on various initial labeling.

Supervised learning is also applied in saliency detection. Jiang *et al.* [27] learn the prior knowledge in a supervised manner. Some methods learn to combine multiple saliency features using the conditional random field models [39,42] and the regression model [61]. Lu *et al.* [40] use the large-margin formulation to learn the optimal set of salient seeds for saliency propagation. While some other methods train the support vector machines [28,29] and the regressor [24] to distinguish salient regions from the backgrounds. Li *et al.* [36] learn a generic distance metric to depict the global distribution of the whole training set. The aforementioned methods require a large number of annotated images to train the parameter models. Recently, Tong *et al.* [51] uses the pseudo labels rather than human annotated ones to train saliency models. They integrates three priors to determine the pseudo positive and pseudo negative samples. In this work, we jointly learn a ranker and a feature projection in a supervised manner to select region proposals on top of the R-CNN features. The proposed model makes full use of the object-level information to improve salient object detection.

3 Kernelized Subspace Ranking

The proposed approach can be divided into three main stages: (i) segment an image into object proposals and extract the deep features. (ii) Learn to rank in the kernelized subspace by jointly optimizing the Rank-SVM and distance metric objectives. (iii) Compute saliency map by a weighted fusion of the top-ranked proposals. The overview of the proposed algorithm is shown in Fig. 2.

3.1 Object Proposal

We employ the geodesic object proposal algorithm [32] to generate region proposals and take region proposals as basic processing units. Region proposals can model the appearance of objects and shapes with a well-defined closed boundary. For each proposal, we extract the CNN features using the pre-trained model provided in [18]. Compared with the hand-crafted features, the CNN feature

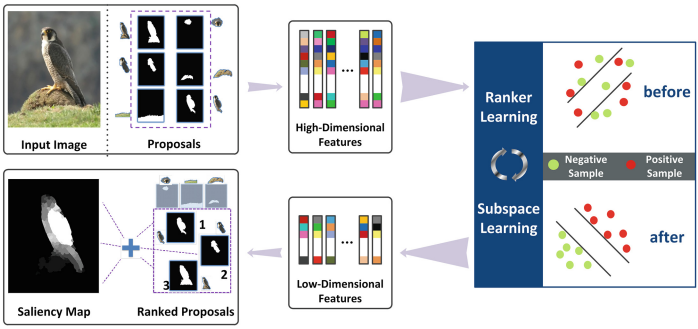


Fig. 2. The overview of the proposed algorithm

can capture richer structure information including low-level visual information (extracted in the earlier layers) and higher-level semantic information (extracted in the latter layers). Features in different layers serve as complementary ones, because the low-level features helps to handle the relative simple scenes and the higher-level features more easily detects the complex semantic objects.

3.2 Problem Formulation

Ranking in primal space. Most candidate objects generated by existing algorithms cannot exactly detect the contours and shapes of salient objects. In order to separate foreground proposals from background ones and obtain accurate saliency result, we cast saliency detection as a ranking problem. We wish to sort out the object proposals with high segmentation precision and recall to detect salient objects via a weighted fusion of them.

We investigate a primal-based Rank-SVM (PRSVM) proposed by Chapelle and Keerthi [6] as it makes the training for large amounts of imbalanced positive and negative samples available. Assume there exists a set of candidate objects $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n]$ with relevance ranks $\mathbf{x}_n \succ \cdots \succ \mathbf{x}_i \succ \mathbf{x}_j \succ \cdots \succ \mathbf{x}_1$, where $\succ$ denotes the order and $\mathbf{x}_k \in \mathbb{R}^d$ is the feature vector of the k-th instance. In a Rank-SVM problem, we wish that instances ranking ahead have higher scores than the behind ones, which can be described in the following formula:

$$\begin{aligned} &\min_{\mathbf{w},\varepsilon} \frac{1}{2}||\mathbf{w}||^2 + \lambda \sum_{(i,j)\in\mathcal{P}} \varepsilon_{ij} \\ &s.t. \mathbf{w}^T(\mathbf{x}_i - \mathbf{x}_j) \geq 1 - \varepsilon_{ij}, \varepsilon_{ij} \geq 0, \end{aligned} \tag{1}$$

where $\mathbf{w}$ corresponds to a weight vector which indicates the importance of each feature. The parameter λ is a trade-off for the regularization and loss term and ε_{ij} is the slack variable. $\mathcal{P}$ represents the preference pairs that satisfies $\mathcal{P} = \{(i,j)|y_i > y_j\}$, and $y_i \in \{-1, +1\}$ is the label of the i-th training instance. Note that, for computation efficiency, our preference pairs are only defined on the between-class instances (i.e., positive and negative instances).

The above function can be rewritten as an unconstrained optimization problem by exploiting the hinge loss function using the L2-loss:

$$\min_{\mathbf{w}} \frac{1}{2}||\mathbf{w}||^2 + \lambda \sum_{(i,j)\in\mathcal{P}} \max(0, 1 - \mathbf{w}^T(\mathbf{x}_i - \mathbf{x}_j))^2. \tag{2}$$

To determine positive and negative instances, we mainly consider the confidence measure, which is an overall performance measurement weighted by the accuracy score A and coverage score C as mentioned in [52]. A_i and C_i can be computed as $A_i = \frac{|O_i \bigcap G|}{|O_i|}, C_i = \frac{|O_i \bigcap G|}{|G|}$. O_i, G respectively represent the i-th proposal and the corresponding ground truth with binary annotation. The notation $|\cdot|$ denotes the number of matrix elements equal to 1.

The accuracy score A_i measures the percentage of the i-th proposal pixels correctly assigned to the salient object, while the coverage score C_i is defined as the ratio of the corresponding ground truth area overlapped with the i-th proposal.

The confidence score is given by $conf_i = \frac{(1+\xi) \times A_i \times C_i}{\xi A_i + C_i}$, where ξ is used to balance the weight between accuracy score and coverage score. The instances with confidence score higher than 0.9 are regarded as positive samples, and instances with confidence score lower than 0.6 are treated as negative ones. In this paper, we use all possible positive samples but only a fraction of the negative ones.

Kernelized subspace ranking. Although R-CNN features have many good properties as described above, they usually have much redundant information in very high-dimensional space. This may reduce the reliability of the ranking problem. To address this issue, we learn a feature projection matrix to project high-dimensional features into a low-dimensional subspace with a kernel approach.

Several methods are proposed in literature aiming at learning a linear projection matrix that maps data points into a low-dimensional subspace. Mignon and Jurie [45] firstly propose pairwise constrained component analysis (PCCA) for learning this transformation matrix with similarity and dissimilarity constraints. Xiong *et al.* [56] further improve it by incorporating a regularization model.

In this work, we simultaneously consider the ranker and subspace learning in a unified formula:

$$\begin{aligned}\min_{\mathbf{w},\mathbf{L}} E = \frac{1}{2}||\mathbf{w}||^2 + \lambda \sum_{(i,j)\in\mathcal{P}} \max(0, 1 - \mathbf{w}^T\mathbf{L}(\psi(\mathbf{x}_i) - \psi(\mathbf{x}_j)))^2 \\ + \sum_{n=1}^{p} \ell_\delta(y_n(||\mathbf{L}(\psi(\mathbf{x}_{i_n}) - \psi(\mathbf{x}_{j_n}))||^2 - 1)) + \mu||\mathbf{L}||_F^2,\end{aligned} \tag{3}$$

where $\ell_\delta(x) = \frac{1}{\delta}\log(1+e^{\delta x})$ is the generalized logistic loss function as mentioned in [60]. $||\cdot||$ represents the Euclidean distance and $||\cdot||_F$ is the Frobenius norm of matrix. $\psi(\mathbf{x}_i)$ is the feature of instance $\mathbf{x}_i$ through kernel projection. p is the number of constraints for the instance pairs $\mathbf{x}_{i_n}$ and $\mathbf{x}_{j_n}$, where (i_n, j_n) indicates

the indices of two instances for the n-th constraint. $y_n \in \{-1, +1\}$ indicates whether the instances belong to the same class or not. $\mathbf{L} \in \mathbb{R}^{l \times d}(l < d)$ is the learnt projection matrix and μ is the regularization parameter. The first two terms are the Rank-SVM formula defined in the subspace, which encourage that foreground proposals should have higher ranking scores than background proposals. The third term acts as a loss function encouraging that the intra-class instances have smaller distances than the inter-class instances, while the fourth term is the regularization term for the projection matrix $\mathbf{L}$.

To further handle the problem where some instances are linearly inseparable, we apply a feature projection matrix $\mathbf{P} \in \mathbb{R}^{l \times N}$ to project primal features into a kernel subspace, where N is the number of training instances. Specially, we let $\mathbf{L} = \mathbf{P}\psi^T(\mathbf{X})$. Then $\mathbf{L}\psi(\mathbf{x}_i) = \mathbf{P}\psi^T(\mathbf{X})\psi(\mathbf{x}_i) = \mathbf{P}\mathbf{k}_i$, where $\mathbf{k}_i = \psi^T(\mathbf{X})\psi(\mathbf{x}_i)$ is the i-column of the kernel matrix $\mathbf{K} = \psi^T(\mathbf{X}) \times \psi(\mathbf{X})$. Equation 3 can be rewritten as

$$\begin{aligned} \min_{\mathbf{w},\mathbf{P}} E = \frac{1}{2}||\mathbf{w}||^2 + \lambda \sum_{(i,j)\in\mathcal{P}} \max(0, 1 - \mathbf{w}^T\mathbf{P}(\mathbf{k}_i - \mathbf{k}_j))^2 \\ + \sum_{n=1}^{p} l_\delta(y_n(||\mathbf{P}(\mathbf{k}_{i_n} - \mathbf{k}_{j_n})||^2 - 1)) + \mu Tr(\mathbf{P}\mathbf{K}\mathbf{P}^T), \end{aligned} \tag{4}$$

where $Tr(\cdot)$ denotes the trace of a matrix. As shown in Fig. 3, the object proposals are sorted using our ranker in descending order. The decimals in yellow font denote the corresponding confidence scores computed by using ground truth. The figure shows that the overall confidence scores of the top-ranked proposals in kernelized subspace are higher than that of the top-ranked proposals in primal space.

The recently proposed methods HARF [64], MCDL [62] and LEGS [52] employ deep features directly or indirectly. Our method is significantly different from these methods in the following aspects: (i) HARF casts saliency detection as a regression problem and works in region-wise manner. This method segments an image into multi-level regions, then compute regional deep features and handcrafted features and feed them to a regressor for saliency prediction. Our method treats saliency detection as a subspace ranking problem and works in object-wise manner. Object proposals carry rich higher-level and object-level structural information, which guarantees the semantic completeness of salient objects. (ii) Both MCDL and LEGS treat saliency detection as a binary classification problem. They train existing deep neural networks to predict the probabilities of pixels (or superpixels) as their saliency values, respectively. We extract deep features and propose a joint subspace ranking framework, and obtain saliency map by weighted combination of the top-ranked proposals.

3.3 Joint Ranker and Subspace Learning

In this section, we aim to learn the Rank-SVM model coefficient $\mathbf{w}$ and projection matrix $\mathbf{P}$ jointly by optimizing Eq. 4. The proposed optimization problem can be efficiently solved using the alternating optimization method.

Fig. 3. Ranking results in different feature spaces. Top: results ranked in the primal space. Bottom: results ranked in the kernelized subspace. The decimals in yellow font denote the corresponding confidence scores. (Color figure online)

Update the ranking coefficient w. Given the estimated projection matrix $\mathbf{P}$, Eq. 4 becomes a Rank-SVM problem, and we use the Truncated Newton optimization similar to [6] to solve it efficiently. The gradient of the objective (4) with respect to $\mathbf{w}$ is,

$$\mathbf{g} := \mathbf{w} + 2\lambda \sum_{(i,j)\in\mathcal{SV}} (\mathbf{w}^T\mathbf{P}(\mathbf{k}_i - \mathbf{k}_j) - 1) \cdot \mathbf{P}(\mathbf{k}_i - \mathbf{k}_j), \tag{5}$$

and the Hessian matrix is,

$$\mathbf{H} := \mathbf{I} + 2\lambda \sum_{(i,j)\in\mathcal{SV}} (\mathbf{P}(\mathbf{k}_i - \mathbf{k}_j))(\mathbf{P}(\mathbf{k}_i - \mathbf{k}_j))^T, \tag{6}$$

where $\mathcal{SV}$ is the set of "support vector pairs" with $\mathcal{SV} = \{(i,j)|(i,j) \in \mathcal{P}, \mathbf{w}^T\mathbf{P}(\mathbf{k}_i - \mathbf{k}_j) < 1\}$. $\mathbf{I}$ is the identify matrix. The ranking coefficient $\mathbf{w}$ is iteratively computed by

$$\mathbf{w}_{t+1} = \mathbf{w}_t - \eta \cdot \mathbf{H}^{-1}\mathbf{g}, \tag{7}$$

where η is found by line search.

Update the projection matrix *P*. Given the ranking coefficient $\mathbf{w}$, Eq. 4 becomes a metric learning problem with kernel trick. We handle the problem directly using gradient descent algorithm. The derivative of the Eq. 4 with respect to $\mathbf{P}$ is

$$\begin{aligned}\frac{\partial E}{\partial \mathbf{P}} =& 2\mathbf{P}\sum_{n=1}^{p} y_n\sigma_\delta(y_n(||\mathbf{P}(\mathbf{k}_{i_n} - \mathbf{k}_{j_n})||^2 - 1))\mathbf{K}\mathbf{T}_n\mathbf{K} + 2\mu PK \\ &+ 2\lambda \sum_{(i,j)\in\mathcal{SV}} (\mathbf{w}^T\mathbf{P}(\mathbf{k}_i - \mathbf{k}_j) - 1)\mathbf{w}(\mathbf{k}_i - \mathbf{k}_j)^T,\end{aligned} \tag{8}$$

where $\mathbf{T}_n = (\mathbf{e}_{i_n} - \mathbf{e}_{j_n})(\mathbf{e}_{i_n} - \mathbf{e}_{j_n})^T$. $\sigma_\delta(x)$ denotes the value of $(1 + e^{-\delta x})^{-1}$ and $\mathbf{e}_k$ is the k-th vector of the cannonical basis, with 1 located in the k-th element and 0 in others.

By multiplying the first two terms of the above computed gradient matrix with preconditioner $\mathbf{K}^{-1}$, the kernel projection matrix $\mathbf{P}$ is computed by iteratively solving the following problem

$$\mathbf{P}_{t+1} = \mathbf{P}_t - 2\alpha \left(\mathbf{P} \sum_{n=1}^{p} \mathbf{A}_n^t \mathbf{K} \mathbf{T}_n + \mu \mathbf{P} + \lambda \sum_{(i,j) \in \mathcal{SV}} \mathbf{Q}_{ij}^t \right), \tag{9}$$

where $\mathbf{Q}_{ij}^t = (\mathbf{w}^T \mathbf{P}(\mathbf{k}_i - \mathbf{k}_j) - 1)\mathbf{w}(\mathbf{k}_i - \mathbf{k}_j)^T$, $\mathbf{A}_n^t = y_n \sigma_\delta(y_n(||\mathbf{P}(\mathbf{k}_{i_n} - \mathbf{k}_{j_n})||^2 - 1))$ at the t-th iteration. And α represents the learning rate.

The proposed joint learning algorithm is summarized in Algorithm 1.

Algorithm1: Kernelized Subspace Ranking

Input: $\mathbf{K}$ (kernel matrix); λ (trade-off parameter); y_n (label of instance pairs); $\mathcal{P}$ (preference pairs); μ (regularization parameter); (i_n, j_n) (indices of instance pairs for the n-th constraint, $n = 1, \cdots, p$)
Output: ranking coefficient $\mathbf{w}$ and projection matrix $\mathbf{P}$.

repeat
  • Update the ranking weight $\mathbf{w}$ with fixed $\mathbf{P}$,
  repeat
    Evaluate the ranking gradient $\mathbf{g}$ by Equation 5;
    Compute the ranking Hessian matrix $\mathbf{H}$ by Equation 6;
    Update the ranking weight $\mathbf{w}$ **by** Equation 7;
      $\mathbf{w}_{t+1} = \mathbf{w}_t - \eta \cdot \mathbf{H}^{-1}\mathbf{g}$ (η found by line search);
  until Convergence
  • Update the projection matrix $\mathbf{P}$ with fixed $\mathbf{w}$,
  repeat
    Solve the derivative $\frac{\partial E}{\partial \mathbf{P}}$ by Equation 8;
    Update the projection matrix $\mathbf{P}$ by Equation 9;
      $\mathbf{P}_{t+1} = \mathbf{P}_t - 2\alpha(\mathbf{P} \sum_{n=1}^{p} \mathbf{A}_n^t \mathbf{K} \mathbf{T}_n + \mu \mathbf{P} + \lambda \sum_{(i,j) \in \mathcal{SV}} \mathbf{Q}_{ij}^t)$;
  until Convergence
until iteration stopping criterion is reached

3.4 Saliency Map

Given the Rank-SVM coefficient $\mathbf{w}$ and projection matrix $\mathbf{P}$, we compute the ranking score of the i-th proposal as $s_i = \mathbf{w}^T \mathbf{P} \mathbf{k}_i$. We consider the top-ranked object candidates to contain salient objects with high precision and recall. As the proposals may cover each other, in order to highlight salient regions, we combine the top-ranked proposals weighted by their ranking scores to compute the saliency score for each pixel:

$$S(x) = \sum_{i=1}^{K} \exp(2 \times s_i) \times m_i(x), \tag{10}$$

where $m_i(x)$ is 1 if pixel x is included in the i-th proposal, and 0 otherwise. After the saliency scores of all pixels are computed, the final saliency map is obtained by a min-max normalization.

4 Experiments

We extensively evaluate the proposed algorithm on four representative benchmark datasets, and compare it with thirteen state-of-the-art saliency methods, including the BL [51], AMC [23], DRFI [24], DSR [37], GC [7], HDCT [30], HS [57], LEGS [52], MR [58], PCA [43], RR [34], UFO [25] and wCtr [63]. We get the saliency results of these competitors either by running the source codes or directly using the saliency maps provided by the authors. Subsequently, we detail the datasets, parameter settings, quantitative and qualitative comparison.

Datasets: We use the MSRA-5000, ECSSD, PASCAL-S and SOD datasets. The MSRA-5000 dataset contains 5,000 images with a large variety of image contents, which is constructed by Liu *et al.* [39]. They exclude very large salient objects and label the ground truth with a bounding box. Afterwards, Jiang *et al.* [24] provide more accurate pixel-wise annotations for saliency evaluation. The ECSSD dataset [57] contains 1,000 images with structurally complex foregrounds and cluttered backgrounds. The PASCAL-S dataset [38] contains 850 natural images surrounded by cluttered backgrounds, which ascends from the validation set of the PASCAL VOC 2012 segmentation challenge. This dataset is one of the most challenging saliency datasets without various design biases (e.g., center bias and color contrast bias). The SOD dataset [46] contains 300 images from the challenging Berkeley segmentation dataset. Some of the images include multiple salient objects with various sizes.

Parameter Settings: The confidence score parameter ξ is 0.3 in the implementation to emphasize the impact of the accuracy score on the final confidence. The trade-off parameters λ and μ in Eq. 3 are set to be 10^{-4} and 0.01, respectively. The learning rate α in Eq. 9 is fixed to be 0.01, similar to [56]. In addition, we use the Gaussian RBF kernel $k(\mathbf{x}, \mathbf{x}') = \exp(-\|\mathbf{x} - \mathbf{x}'\|/\sigma^2)$. The kernel parameter σ is equal to the first quantile of all distances [56]. We fuse the top-16 candidates to compute the final saliency map.

The proposal algorithm [32] roughly produces 1,000 candidate segments for each image. There are many too small or too large candidates which make little contribution to saliency detection. Therefore, we compute the percentage of the area of the proposals with respect to the whole image and remove the oversized proposals (>70 %) and undersized ones (<2 %). Besides, we remove the proposals which touch four boundaries of an image. We randomly sample 2,000 images from the MSRA-5000 dataset to train our model and treat the rest images as the testing dataset. Rather than training a model for each dataset, we use the model trained from the MSRA-5000 dataset and test it over others. Because we actually learn a category-independent ranker to rank the proposals according to their objectness without using any knowledge about object categories.

Quantitative Comparison: We use the precision-recall curve, F-measure and Area Under Curve (AUC) to quantitatively evaluate the experimental results. The precision value corresponds to the ratio of salient pixels correctly assigned to all pixels of the extracted regions, while the recall value is defined as the

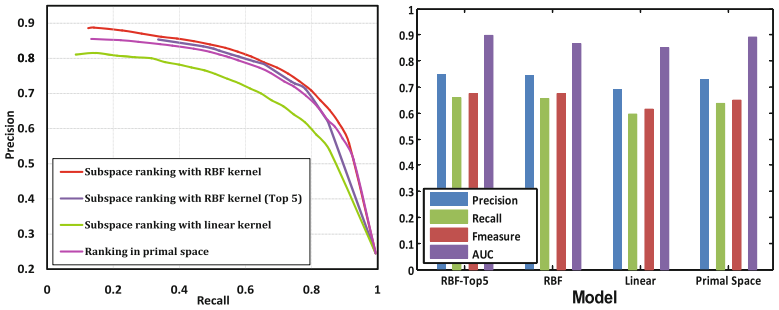


Fig. 4. Performance comparison on the PASCAL-S dataset by the proposed algorithm with different design options.

percentage of detected salient pixels with respect to the ground-truth data. Given a saliency map with intensity values normalized to the range of 0 and 255, a number of binary maps are produced by using every possible fixed threshold in [0; 255]. We compute the precision/recall pairs of all the binary maps to plot the precision-recall curve. Meanwhile, we obtain true positive and false positive rates to plot the ROC curve and AUC score. Similar to existing methods, we also compute the precision, recall and F-measure with an adaptive threshold, defined as twice the mean saliency of an input image [1]. The F-measure is the overall performance indicator computed by the weighted harmonic of precision and recall as follows: $F_\gamma = \frac{(1+\gamma^2)\times Precision \times Recall}{\gamma^2 Precision + Recall}$, where γ^2 is set to be 0.3 to weigh more on precision as suggested in [1]. Figure 6 shows the precision-recall curves and F-measure of different methods on all four datasets. Because the DRFI [24] and LEGS [52] methods also randomly select the images from the MSRA-5000 dataset to train their models, where the former learns a random forest regressor and the latter trains a convolutional neural network, and both of them only provide the pre-training models, we don't give the experimental results of the two methods on the MSRA-5000 dataset for a fair comparison. As shown in Fig. 6, we can see that the precision-recall curve of the proposed algorithm significantly performs better than other methods on the SOD dataset and slightly better than the second best method (LEGS [52]) on the ECSSD and PASCAL-S datasets. In terms of F-measure score, the proposed algorithm outperforms other methods on the SOD and MSRA-5000 datasets, and is the second best on the PASCAL-S and ECSSD datasets, slightly worse than the LEGS [52], as shown in Table 1. Table 2 shows the AUC values of all evaluated methods. The proposed algorithm consistently performs better than these competitors on the SOD, MSRA-5000 and PASCAL-S datasets and slightly poorly compared to the DRFI [24] on the ECSSD dataset.

In addition, we evaluate the performance of subspace ranking with different kernels and primal space ranking (i.e., without feature projection matrix $\mathbf{P}$) on the PASCAL-S dataset. We utilize the performance of directly averaging the top 5 proposals ranked in the kernelized subspace. The results are shown in Fig. 4, we can see that the performance is significantly improved by using the

Table 1. Quantitative comparisons in terms of F-measure score. The best and second best results are shown in red color and blue color respectively.

Datasets	AMC	BL	DSR	GC	HDCT	HS	MR
MSRA	0.7575	0.7328	0.7440	0.6575	0.7364	0.7115	0.7510
SOD	0.5888	0.5723	0.5968	0.4642	0.6108	0.5210	0.5697
PASCAL	0.5987	0.5668	0.5513	0.4861	0.5824	0.5278	0.5881
ECSSD	0.7002	0.6825	0.6636	0.5726	0.6897	0.6363	0.6932
	PCA	RR	UFO	wCtr	DRFI	LEGS	Ours
MSRA	0.6723	0.7575	0.7265	0.7437	-	-	0.7763
SOD	0.5370	0.5665	0.5480	0.5978	0.6031	0.6492	0.6622
PASCAL	0.5298	0.5873	0.5502	0.5972	0.6159	0.6951	0.6760
ECSSD	0.5796	0.6577	0.6442	0.6774	0.7337	0.7852	0.7705

Table 2. Quantitative comparison in terms of AUC score. The best and second best results are shown in red color and blue color respectively

Datasets	AMC	BL	DSR	GC	HDCT	HS	MR
MSRA	0.9292	0.9360	0.9247	0.8398	0.9318	0.9043	0.9044
SOD	0.8391	0.8503	0.8210	0.7046	0.8504	0.8145	0.7899
PASCAL	0.8616	0.8633	0.8079	0.7321	0.8582	0.8330	0.8205
ECSSD	0.9067	0.9147	0.8604	0.7848	0.9039	0.8821	0.8820
	PCA	RR	UFO	wCtr	DRFI	LEGS	Ours
MSRA	0.9248	0.9089	0.8950	0.9169	-	-	0.9376
SOD	0.8212	0.7888	0.7840	0.8014	0.8464	0.8117	0.8510
PASCAL	0.8371	0.8251	0.8088	0.8433	0.8913	0.8857	0.8970
ECSSD	0.8737	0.8283	0.8587	0.8779	0.9391	0.9239	0.9257

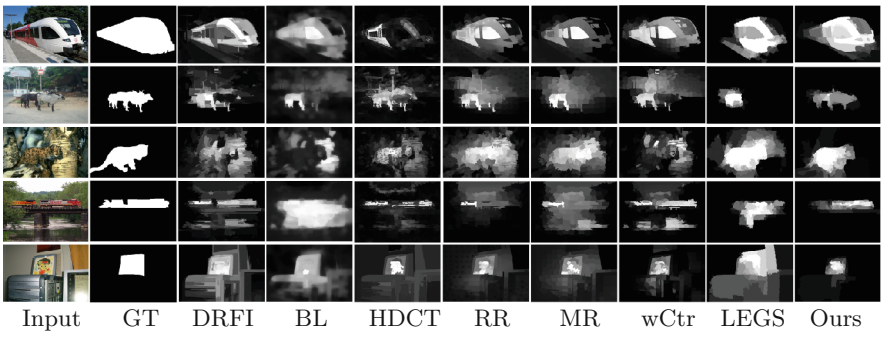
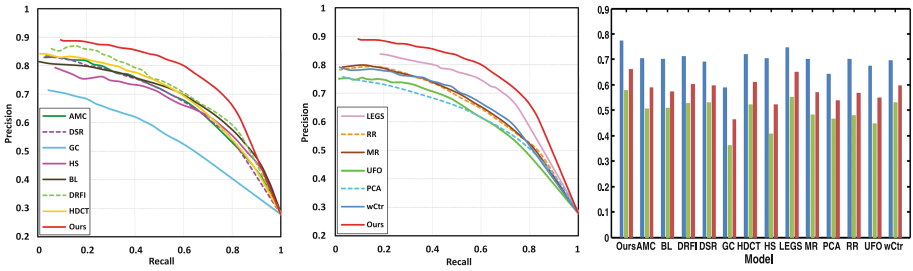
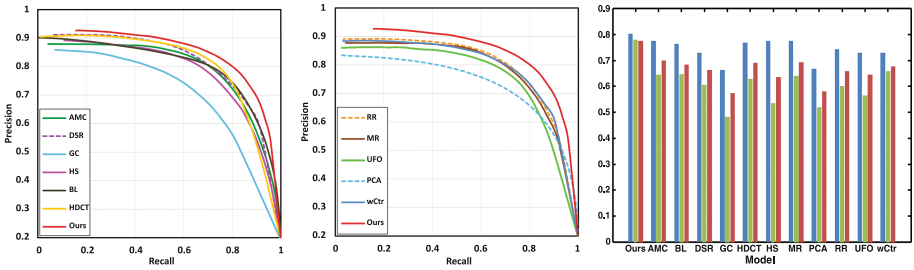


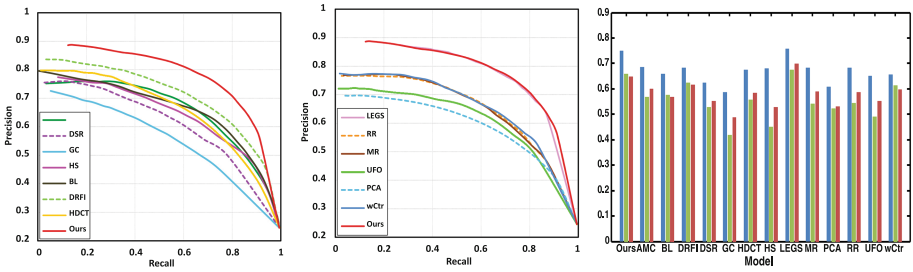
Fig. 5. Visual comparison with seven state-of-the-art methods.



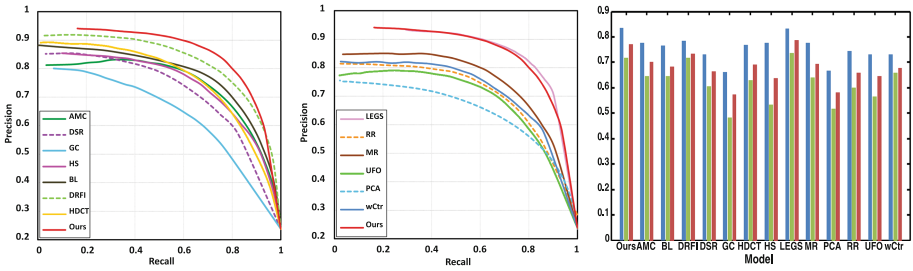
(a) SOD dataset



(b) MSRA dataset



(c) PASCAL-S dataset



(d) ECSSD dataset

Fig. 6. Quantitative comparison of different methods on all four datasets.

Gaussian-RBF kernel projection, compared to the linear kernel projection and primal space ranking. Moreover, the weighted fusion by the ranking scores performs better than non-weighted fusion.

Qualitative Comparison: Figure 5 shows a few saliency maps generated by the evaluated methods. We note that the proposed algorithm uniformly highlights the salient regions with well-defined contours. Owing to the contribution of the subspace ranking and kernel projection, the proposed method can detect salient objects accurately when the backgrounds are cluttered or the objects and backgrounds have similar appearance. More results can be found in the supplementary material[1].

5 Conclusions

In this paper, we explore a novel and effective ranking based approach for saliency detection by jointly learning a SVM ranker and a distance metric in a unified framework. The learnt metric uses kernel projection matrix to map the high-dimensional R-CNN features into a low-dimensional subspace, thereby removing the redundancy of feature channels and prompting sample pairs more separable. Different from existing methods that concentrate on the pixel or superpixel level to detect salient objects, we present a proposal-based saliency approach, which exploits the object-level saliency cues of proposals to increase the completeness of detected results and avoid fitting to locally salient parts. Specifically, we rank the object candidates and compute the saliency map by a weighted fusion of the top-ranked candidates according to their ranking scores. We conduct extensive experiments on four benchmark datasets and demonstrate favorable performance against thirteen state-of-the-art methods.

Acknowledgments. The work was supported by the National Natural Science Foundation of China under Grant #61371157, Grant #61472060 and Grant #61528101.

References

1. Achanta, R., Hemami, S., Estrada, F., Susstrunk., S.: Frequency-tuned salient region detection. In: CVPR, pp. 1597–1604 (2009)
2. Alexe, B., Deselaers, T., Ferrari, V.: Measuring the objectness of image windows. IEEE TPAMI **34**(11), 2189–2202 (2012)
3. Borji, A., Itti, L.: State-of-the-art in visual attention modeling. IEEE Trans. Pattern Anal. Mach. Intell. **35**(1), 185–207 (2013)
4. Carreira, J., Sminchisescu, C.: CPMC: automatic object segmentation using constrained parametric min-cuts. IEEE TPAMI **34**(7), 1312–1328 (2012)
5. Chang, K.Y., Liu, T.L., Chen, H.T., Lai., S.H.: Fusing generic objectness and visual saliency for salient object detection. In: ICCV, pp. 914–921 (2011)
6. Chapelle, O., Keerthi, S.S.: Efficient algorithms for ranking with SVMs. Inf. Retrieval **13**(3), 201–215 (2010)

[1] http://ice.dlut.edu.cn/lu/index.html.

7. Cheng, M.M., Warrell, J., Lin, W.Y., Zheng, S., Vineet, V., Crook, N.: Efficient salient region detection with soft image abstraction. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 1529–1536 (2013)
8. Cheng, M.M., Zhang, Z., Lin, W.Y., Torr, P.: Bing: binarized normed gradients for objectness estimation at 300 fps. In: CVPR (2014)
9. Cheng, M., Warrell, J., Lin, W., Zheng, S., Vineet, V., Crook., N.: Efficient salient region detection with soft image abstraction. In: ICCV, pp. 1529–1536 (2013)
10. Cheng, M., Zhang, G., Mitra, N., Huang, X., Hu., S.: Global contrast based salient region detection. In: CVPR, pp. 409–416 (2011)
11. Endres, I., Hoiem, D.: Category independent object proposals. In: Daniilidis, K., Maragos, P., Paragios, N. (eds.) ECCV 2010, Part V. LNCS, vol. 6315, pp. 575–588. Springer, Heidelberg (2010)
12. Girshick, R., Donahue, J., Darrell, T., Malik, J.: Heterogeneous metric learning with joint graph regularization for cross-media retrieval. In: CVPR (2014)
13. Goferman, S., Zelnik-Manor, L., Tal, A.: Context-aware saliency detection. In: CVPR, pp. 2376–2383 (2010)
14. Gong, C., Tao, D., Liu, W., Maybank, S.J., Fang, M., Fu, K., Yang, J.: Saliency propagation from simple to difficult. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2531–2539 (2015)
15. Gopalakrishnan, V., Hu, Y., Rajan, D.: Random walks on graphs for salient object detection in images. IEEE TIP **19**(12), 3232–3242 (2010)
16. Guo, C., Ma, Q., Zhang., L.: Spatio-temporal saliency detection using phase spectrum of quaternion fourier transform. In: CVPR (2008)
17. Harel, J., Koch, C., Perona, P.: Graph-based visual saliency. In: NIPS, pp. 545–552 (2006)
18. Hariharan, B., Arbeláez, P., Girshick, R., Malik, J.: Simultaneous detection and segmentation. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) ECCV 2014, Part VII. LNCS, vol. 8695, pp. 297–312. Springer, Heidelberg (2014)
19. Hou, X., Zhang., L.: Saliency detection: a spectral residual approach. In: CVPR, pp. 1–8 (2007)
20. Huang, Z., Wang, R., Shan, S., Chen, X.: Learning Euclidean-to-Riemannian metric for point-to-set classification. In: CVPR, pp. 1677–1684 (2014)
21. Itti, L., Koch, C., Niebur, E.: A model of saliency-based visual attention for rapid scene analysis. IEEE TPAMI **20**(11), 1254–1259 (1998)
22. Jia, Y., Han., M.: Category-independent object-level saliency detection. In: ICCV, pp. 1761–1768 (2013)
23. Jiang, B., Zhang, L., Lu, H., Yang, C., Yang, M.H.: Saliency detection via absorbing markov chain. In: ICCV, pp. 1665–1672 (2013)
24. Jiang, H., Wang, J., Yuan, Z., Wu, Y., Zheng, N., Li, S.: Salient object detection: a discriminative regional feature integration approach. In: CVPR (2013)
25. Jiang, P., Ling, H., Yu, J., Peng, J.: Salient region detection by UFO: Uniqueness, Focusness and Objectness. In: ICCV, pp. 1976–1983 (2013)
26. Jiang, P., Vasconcelos, N., Peng, J.: Generic promotion of diffusion-based salient object detection. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 217–225 (2015)
27. Jiang, Z., Davis, L.: Submodular salient region detection. In: CVPR, pp. 2043–2050 (2013)
28. Judd, T., Ehinger, K., Durand, F., Torralba., A.: Learning to predict where humans look. In: ICCV, pp. 2106–2113 (2009)
29. Kienzle, W., Wichmann, F.A., Schölkopf, B., Franz, M.O.: A nonparametric approach to bottom-up visual saliency. In: NIPS, pp. 417–424 (2007)

30. Kim, J., Han, D., Tai, Y.W., Kim, J.: Salient region detection via high-dimensional color transform. In: CVPR, pp. 883–890 (2014)
31. Klein, D., Frintrop., S.: Center-surround divergence of feature statistics for salient object detection. In: ICCV, pp. 2214–2219 (2011)
32. Krähenbühl, P., Koltun, V.: Geodesic object proposals. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) ECCV 2014, Part V. LNCS, vol. 8693, pp. 725–739. Springer, Heidelberg (2014)
33. Krizhevsky, A., Sutskever, I., Hinton, G.: Imagenet classification with deep convolutional neural networks. In: NIPS (2012)
34. Li, C., Yuan, Y., Cai, W., Xia, Y., Feng, D.D.: Robust saliency detection via regularized random walks ranking. In: CVPR, pp. 2710–2717 (2015)
35. Li, G., Yu, Y.: Visual saliency based on multiscale deep features. In: CVPR (2015)
36. Li, S., Lu, H., Lin, Z., Shen, X., Price, B.: Adaptive metric learning for saliency detection. IEEE TIP **24**(11), 3321–3331 (2015)
37. Li, X., Lu, H., Zhang, L., Ruan, X., Yang, M.H.: Saliency detection via dense and sparse reconstruction. In: ICCV, pp. 2976–2983 (2013)
38. Li, Y., Hou, X., Koch, C., Rehg, J.M., Yuille, A.L.: The secrets of salient object segmentation. In: CVPR, pp. 280–287 (2014)
39. Liu, T., Yuan, Z., Sun, J., Wang, J., Zheng, N., Tang, X., Shum, H.: Learning to detect a salient object. IEEE Trans. Pattern Anal. Mach. Intell. **33**(2), 353–367 (2011)
40. Lu, S., Mahadevan, V., Vasconcelos, N.: Learning optimal seeds for diffusion-based salient object detection. In: CVPR, pp. 2790–2797 (2014)
41. Lu, Y., Zhang, W., Lu, H., Xue, X.: Salient object detection using concavity context. In: ICCV, pp. 233–240 (2011)
42. Mai, L., Niu, Y., Liu, F.: Saliency aggregation: a data-driven approach. In: CVPR, pp. 1131–1138 (2013)
43. Margolin, R., Tal, A., Zelnik-Manor, L.: What makes a patch distinct? In: CVPR, pp. 1139–1146 (2013)
44. Mcfee, B., Lanckriet, G., Jebara, T.: Learning multi-modal similarity. JMLR **12**(2), 491–523 (2011)
45. Mignon, A., Jurie, F.: PCCA: a new approach for distance learning from sparse pairwise constraints. In: 2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2666–2672 (2012)
46. Movahedi, V., Elder, J.: Design and perceptual validation of performance measures for salient object segmentation. In: CVPR Workshop, pp. 49–56 (2010)
47. Peng, Y., Xiao, J.: Heterogeneous metric learning with joint graph regularization for cross-media retrieval. In: AAAI, pp. 1198–1204 (2013)
48. Perazzi, F., Krahenbuhl, P., Pritch, Y., Hornung., A.: Saliency filters: contrast based filtering for salient region detection. In: CVPR, pp. 733–740 (2012)
49. Qin, Y., Lu, H., Xu, Y., Wang, H.: Saliency detection via cellular automata. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 110–119 (2015)
50. Toet, A.: Computational versus psychophysical bottom-up image saliency: a comparative evaluation study. IEEE TPAMI **33**(11), 2131–2146 (2011)
51. Tong, N., Lu, H., Ruan, X., Yang, M.H.: Salient object detection via bootstrap learning. In: CVPR, pp. 1884–1892 (2015)
52. Wang, L., Lu, H., Ruan, X., Yang, M.H.: Deep networks for saliency detection via local estimation and global search. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3183–3192 (2015)

53. Wang, M., Konrad, J., Ishwar, P., Jing, K., Rowley, H.: Image saliency: from intrinsic to extrinsic context. In: CVPR, pp. 417–424 (2011)
54. Wang, W., Wang, Y., Huang, Q., Gao., W.: Measuring visual saliency by site entropy rate. In: CVPR, pp. 2368–2375 (2010)
55. Wei, Y., Wen, F., Zhu, W., Sun, J.: Geodesic saliency using background priors. In: Fitzgibbon, A., Lazebnik, S., Perona, P., Sato, Y., Schmid, C. (eds.) ECCV 2012, Part III. LNCS, vol. 7574, pp. 29–42. Springer, Heidelberg (2012)
56. Xiong, F., Gou, M., Camps, O., Sznaier, M.: Person re-identification using kernel-based metric learning methods. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) ECCV 2014, Part VII. LNCS, vol. 8695, pp. 1–16. Springer, Heidelberg (2014)
57. Yan, Q., Xu, L., Shi, J., Jia, J.: Hierarchical saliency detection. In: CVPR, pp. 1155–1162 (2013)
58. Yang, C., Zhang, L., Lu, H., Ruan, X., Yang, M.H.: Saliency detection via graph-based manifold ranking. In: CVPR, pp. 3166–3173 (2013)
59. Zhang, J., Sclaroff, S., Lin, Z., Shen, X., Price, B., Mech, R.: Minimum barrier salient object detection at 80 fps. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 1404–1412 (2015)
60. Zhang, T., Oles, F.J.: Text categorization based on regularized linear classification methods. Inf. Retrieval **4**(1), 5–31 (2001)
61. Zhao, Q., Koch, C.: Learning a saliency map using fixated locations in natural scenes. JoV **11**(3), 1–15 (2011)
62. Zhao, R., Ouyang, W., Li, H., Wang, X.: Saliency detection by multi-context deep learning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1265–1274 (2015)
63. Zhu, W., Liang, S., Wei, Y., Sun, J.: Saliency optimization from robust background detection. In: CVPR, pp. 2814–2821 (2014)
64. Zou, W., Komodakis, N.: HARF: hierarchy-associated rich features for salient object detection. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 406–414 (2015)