Visual Saliency Detection via Kernelized Subspace Ranking with Active Learning

Lihe Zhang, Jiayu Sun, Tiantian Wang, Yifan Min and Huchuan Lu, *Senior Member, IEEE*

***Abstract*—Saliency detection task has witnessed a booming interest for years, due to the growth of the computer vision community. In this paper, we introduce a new saliency model that performs active learning with kernelized subspace ranker (KSR) referred to as KSR-AL. This pool-based active learning algorithm ranks the informativeness of unlabeled data by considering both uncertainty sampling and information density, thereby minimizing the cost of labeling. The informative images are selected to train the KSR iteratively and incrementally. The learning model of this algorithm is designed on object-level proposals and region-based convolutional neural network (R-CNN) features, by jointly learning a Rank-SVM classifier and a subspace projection. When the active learning process meets its stopping criteria, the saliency map of each image is generated by a weight fusion of its top-ranked proposals, whose ranking scores are graded by the learned ranker. We show that the KSR-AL achieves a reduction in annotation, as well as improvement in performance, compared with the supervised learning scheme. Besides, the proposed algorithm also outperforms the state-of-the-art methods. These improvements are demonstrated by extensive experiments on six publicly available benchmark datasets.**

***Index Terms*—Saliency detection, active learning, subspace ranking, support vector machines, feature projection.**

I. Introduction

THE task of saliency detection is to make computers mimic human visual attention mechanism - the mechanism that automatically decides which parts of an image are most attractive and informative. With the rapid development of the computer vision community, the task of saliency detection has become increasingly popular as an important pre-processing step to reduce computational complexity. It has been widely applied to numerous computer vision tasks, such as image segmentation [?], object recognition [?], image compression [?], and content-based image retrieval [?] etc.

Some pioneering work of saliency detection takes full advantage of low-level properties of pixels within image, such as color, intensity or orientation to detect the salient objects. Later, mid-level structure property of image regions is introduced, which can integrate some contextual information to measure the saliency for each region. To keep the semantic completeness of salient object, these pixel-wise and region-wise methodologies have to consider the image elements from both local and global view in combination. Recently, we propose an object-wise saliency method KSR [?], which considers the object characteristics of region proposals and formulates a weighted combination of them to generate saliency maps.

In the previous work [?], we take both advantage of object-level proposals and region-based convolutional neural network (R-CNN) features. The object proposal techniques are used to significantly reduce the search space for salient object detection. Some good proposals contain important object-level prior knowledge, such as rich boundary shape cues and the encapsulation of object visual information. It is helpful for salient object detection to sort out and employ these good proposals containing a part of the object or an entire object segment. Moreover, R-CNN has already shown its powerful representation capability in recent works, since it can give a good description of both low-level and high-level image cues. The object-wise method jointly learns a ranker of proposals based on a Rank-SVM model and a feature projection to select region proposals on the R-CNN features.

In the process of model training, we previously apply supervised learning manner, which needs numerous labeled data for training to achieve appropriate results. Nevertheless, the process of obtaining fully-labeled data, especially the ones used in saliency task, is quite complicated and costly. This is because of the subjectivity of labeling salient object by different people and the difficulty of pixel-level data acquisition. In this case, it always needs lots of people to use expensive eye trackers for a long time to complete the labeling. Besides, a large amount of labeled training data may bring into some redundancy for training, which often decreases the performance [?], [?], [?]. Therefore, it is necessary to reduce the labeling cost appropriately.

In this work, we utilize the active learning mechanism to improve the model performance by using a small quantity of valuable training samples. Because the width of the projection matrix in the following subspace learning is equal to the number of training samples, reducing the training data can naturally boost the efficiency of saliency computation. Active learning [?] aims at achieving good learning performance without demanding too many labeled examples. It is able to interactively pose queries to oracle and obtain the desired outputs at new data points rather than just using unlabeled examples. Active learning is also a statistic concept, it is known as “experiment design” in the statistics literature.

The key to designing an active learning algorithm is to select the most informative unlabeled data. Thus, it is important

This work was supported by Dalian Science and Technology Innovation Foundation #2019J12GX039 and the National Natural Science Foundation of China #61876202, #61829102, #61751212, #61725202. L. Zhang, J. Sun, Y. Min, and H. Lu are with School of Information and Communication Engineering, Dalian University of Technology, Dalian, China. E-mail: zhanglihe@dlut.edu.cn; jiayusun666@gmail.com; fnminyf@mail.dlut.edu.cn; lhchuan@dlut.edu.cn

T. Wang is with Department of Electrical Science and Computer Engineering, University of California, Merced. E-mail: tiantianwang.ice@gmail.com

Manuscript received March 29, 2019

to quantify the informativeness of the unlabeled data and select the most informative ones for further model training. Since the most informative data always own a relatively high uncertainty towards the trained model, the main framework of the proposed algorithm is an uncertainty sampling framework. In this work, the uncertainty of each unlabeled image can be measured based on the proportion of its "paradoxical proposals", where the "paradoxical proposal" represents the proposal that is hard to tell whether it is a positive proposal or a negative proposal. The "paradoxicality" of the proposal is defined on the ranking score graded by the KSR model. By uncertainty sampling, a set of image candidates has been picked out and ready for annotation. However, it has been proven that the set of image candidates is somewhat rough if only uncertainty sampling strategy is used. To avoid redundant labeling and only query the most representative instances in the underlying subspace, we further consider the information density of image candidates by density-based clustering.

The contributions of this work are listed as follows:

(1) We combine the kernelized subspace ranker into an active learning framework to improve the effectiveness of the ranker for saliency detection.

(2) We introduce a novel pool-based active learning algorithm, which is based on object proposal techniques, to overcome the burden of data annotation. This active learning algorithm considers both uncertainty sampling and information density.

(3) It is demonstrated that the proposed algorithm performs favorably against the state-of-the-art saliency methods on the ECSSD, PASCAL-S, SOD, HKU-IS, DUT-OMRON, and SOC-val benchmark datasets.

The remainder of this paper is organized as follows: Section II discusses some previous works related to this paper. In Section III, kernelized subspace ranking model with active learning method is proposed. Section IV demonstrates and analyzes the experimental results. Finally, the whole paper is concluded in Section V.

II. Related Work

An increasing number of algorithms have been proposed and shown good performance in saliency detection [?], which will be briefed in this section. Our algorithm is object-wise saliency model with pool-based active learning. Therefore, the two aspects are reviewed, saliency models with different supervision schemes, as well as active learning scenarios.

A. Saliency Detection

Unsupervised approaches usually define visual saliency based on visual rarity or distinctness. To simulate saliency detection, Itti *et al.* [?] first propose a visual attention model, in which some center-surround contrasts on some basic visual characters in multiple scales are exploited. These basic visual characters refer to the color, intensity, and orientation properties of images. Then, Han *et al.* [?] use Itti's model as a prior and employ a Markov random field model to integrate the attention value and the low-level features for salient object detection. Later, some researchers, such as Cheng *et al.* [?] and Perazzi *et al.* [?] etc., calculate the global contrast with the combination of appearance difference and spatial coherence. While some other researchers, such as Goferman *et al.* [?] and Wang *et al.* [?] etc., take both local and global perspectives into consideration. Moreover, domain model [?], discrimination features [?], speed strategy [?], and distance metric [?] have also been integrated into some saliency detection algorithms with unsupervised manner. Latterly, some researchers have also proposed some unsupervised method based on CNNs [?].

Most semi-supervised saliency detection algorithms are developed upon a kind of pair-wised label propagation from labeled instances to unlabeled instances. Harel *et al.* [?] apply ergodic Markov chain on the multiscale feature maps of Itti's model [?]. They propose a graph-based saliency detection model by formulating saliency labeling as a random walking problem. Gopalakrishnan *et al.* [?] promote the above graph-based model by introducing the equilibrium-hitting time of the ergodic Markov chain to detect the most salient node in the graph and use seeded salient region identification mechanism for saliency detection. There are also some other promising semi-supervised approaches to detect salient objects. Jiang *et al.* [?] using absorbing time of an absorbing Markov chain to measure saliency. Yang *et al.* [?] introduce manifold ranking into saliency detection. Thus saliency results can be acquired through label propagation from the given seeds or queries. The idea of both random walking and manifold ranking are soon extended to other works, such as Li *et al.* [?] propose a regularized random walking approach by the combination of them to jointly considers local image data and prior estimation.

There are also some methods based on supervised learning for salient object detection. Jiang and Davis [?] introduce a supervised learning scheme to learn prior knowledge. Some researchers, such as Liu *et al.* [?] and Mai *et al.* [?] etc., propose supervised algorithms learning to combine multiple saliency features with the use of conditional random field models. While Zhao and Koch [?] learn saliency maps using the regression models. Moreover, Li *et al.* [?] introduce an adaptive metric learning method to depict training set distribution, but it requires tons of labeled data. Thus, Tong *et al.* [?] try to replace labeled data with some pseudo ones determined by some priors.

Furthermore, what is worth to mention is the emerging of deep learning based saliency detection methods, where they automatically extract multi-scale deep features, instead of using hand-crafted features. In [?], Han *et al.* model the background with deep structures and then separate salient objects from the background. Li and Yu [?] introduce a deep contrast network consisted of a pixel-level fully convolutional stream and a segment-wise spatial pooling stream, where the former predicts pixel-level saliency, whereas the latter estimate segment-level saliency. Lee *et al.* [?] propose a unified deep learning framework, where it takes both advantage of high-level features and a low-level distance map. Liu and Han [?] provide a hierarchical recurrent convolutional neural network, which does a gradually local and fine refinement by integrating local context on a global but coarse prediction. This initial prediction is automatically learned by some global structured saliency cues. Wang *et al.* [?] propose a feed-forward neural network for saliency detection with a multistage refinement mechanism, which progressively encodes local context information for finer prediction, as well as a pyramid pooling module, which is used to gather global context information. Besides, Li *et al.* [?] recently convert a well-trained contour detection model into a saliency detection model. It is a two-branch network, where a contour branch and a saliency branch mutually feedback with each other, through a transferring method and an alternating training pipeline. Recurrent fully convolutional networks (RFCNs) has also been used in saliency detection tasks. Wang *et al.* [?] use a kind of recurrent architecture to incorporate saliency prior knowledge for accurate saliency inference. Wang *et al.* [?] introduce a block-wise recurrent network, which repeatedly combines output and input features of each block to incorporate contextual knowledge. Moreover, Hou *et al.* [?] introduce short connections to the skip-layer structures within the Holistically-Nested Edge Detector (HED) architecture for saliency detection. Chen *et al.* [?] introduce reverse attention to guide side-output residual learning.

Recently, Wu *et al.* [?] introduce a mutual learning method. They formulate saliency detection, foreground contour detection and edge detection tasks in an end-to-end network. In [?], Zhao and Wu introduce PFA network for saliency detection, which consists of the context-aware pyramid feature extraction module, the channel-wise module, and the spatial attention module. It is trained under the edge-preservation supervision. Qin *et al.* [?] propose a boundary-aware network, which uses a residual refinement module after prediction under the supervision of a three-level hybrid loss.

In the same vein, co-saliency detection is another rising saliency task [?], [?], [?]. Co-saliency detection, unlike single-image saliency detection, aims at discovery the common saliency on multiple images. Fu *et al.* [?] propose a two-layers cluster-based co-saliency method, which generates its co-saliency maps by fusing together three cluster-level saliency cues. Cao *et al.* [?] introduce a co-saliency method, where each saliency map that participates in the fusion process is self-adaptively weighted under the rank constraints. In [?], Zhang *et al.* exploit the deep and wide information for co-saliency detection. They propose a Bayesian framework to integrate intra-image contrast and intra-group consistency metrics. There are plenty of applications based on Co-saliency detection, such as image co-segmentation [?], common pattern discovery [?], [?], and object co-localization [?], [?].

In this work, we use active learning instead of supervised learning to train the ranker and the feature projection, in order to reduce laborious annotation and improve the performance.

B. Active Learning

According to different ways of instances selecting, the selection strategy of active learning largely falls into three scenarios: membership query synthesis, stream-based selective sampling, and pool-based sampling. The membership query synthesis, first developed by Augluin [?]. In this scenario, the model itself controls the generation of unlabeled data for querying. The performance of the model can be increased through training repeatedly. It has shown high efficiency and effectiveness in some specific domains [?], [?], especially where labels are gotten from some experiments instead of human annotators. However, there are also some shortcomings of this scenario, e.g., Lang and Baum [?] indicate that query learning works poorly when human annotators are needed for labeling, as the model may generate some unrecognizable symbols without semantic meaning.

To overcome the shortcoming of the aforementioned scenario, researchers have brought some selective strategies, such as steam-based selecting strategies and pool-based selecting strategies [?], [?]. In stream-based or sequential selecting strategies, unlabeled instances are judged in order on the model to decide whether the instance should be queried for labeling or not. To make the decision, Dagan and Engelson [?] introduce an informativeness measure approach, where instances are evaluated and the more informative instances are more likely to be queried. Cohn *et al.* [?] propose another approach where they explicit an instance subspace as the region of uncertainty and only query the instance falls into it. This region of uncertainty is decided by a minimum threshold of informativeness measure or the version space of current training data [?].

Last but not least, the pool-based strategy is the most widely-used scenario, as numerous unlabeled data can be easily gathered at once in the real world. Unlike the steam-based strategies, there is no need to make the query decision of each instance individually and sequentially in a pool-based scenario. This is because, in pool-based strategies, there is a set of labeled data, along with a pool of unlabeled data. The whole pool of unlabeled data can be evaluated and ranked at once, thus the most informative instances are chosen and queried for annotation together. In this work, we propose a pool-based active learning scenario for saliency detection.

III. Kernelized Subspace Ranking with Active Learning

In this section, we first overall introduce the active learning strategy combined with subspace ranking model. Then, the kernelized subspace ranking model is detailedly designed, which is trained across positive and negative proposals of these actively selected images. Finally, the acquisition function that actively determines training images is designed in detail.

A. Pool-based Active Learning Strategy

In this work, we follow a pool-based strategy to train an active learner for saliency detection. Figure 1 shows the overall workflow of active learner. This active learner consists of three components, which are the kernelized subspace ranker f, the two-class acquisition function q based on uncertainty sampling and information density, and the current set of labeled data $\mathcal{L}$. $\mathcal{L} = \{(I_i, G_i)\}_{i=1}^{N}$ contains N training images, where G_i is the pixel-wise ground-truth of image I_i.

The full process of active learning is shown as follows: An unlabeled data pool $\mathcal{U} = \{U_i\}_{i=1}^{M}$ is given at the start, which includes M unlabeled images. On the first iteration, the KSR model f is trained on the set of labeled data $\mathcal{L}$, which is a set of randomly selected labeled data. Then, the

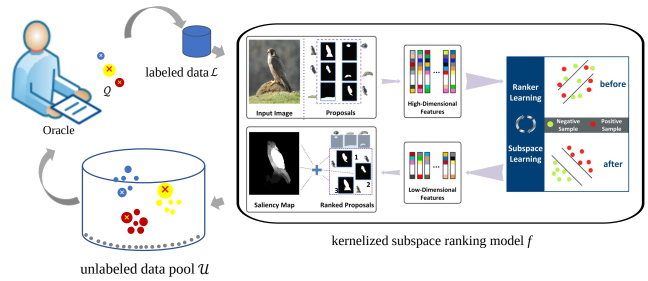


Fig. 1: The overview of the proposed algorithm.

learner actively chooses the most informative unlabeled images from the pool $\mathcal{U}$ based on the designed acquisition function q. The set $\mathcal{L}$ is updated by adding those chosen images, and f is re-trained on the updated pool $\mathcal{L}$. The process repeats until the stopping criterion is reached. The whole procedure is illustrated in Algorithm 1.

Algorithm 1: Active Learning Strategy

```
Input: $\mathcal{L}$ (a set of labeled data); $\mathcal{U}$ (unlabeled data pool);
Output: kernelized subspace ranking model $f$;
repeat
    Train $f$ on $\mathcal{L}$;
    Select the set $\mathcal{Q}$ using acquisition function $q$ for annotation;
    Query the oracle for labelling;
    Update $\mathcal{U}$ and $\mathcal{L}$: $\mathcal{U} = \mathcal{U} \backslash \mathcal{Q}, \mathcal{L} = \mathcal{L} \bigcup \mathcal{Q}$;
until iteration stopping criterion is reached
```

B. Kernelized Subspace Ranking

The Kernelized Subspace Ranking (KSR) model f is composed of three main stages: (i) segment an image into object proposals and extract their deep features. (ii) determine positive and negative proposals and 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.

1) Object Proposal: In this stage, the geodesic object proposal algorithm [**?**] is employed to generate region proposals and these proposals are taken as basic processing units. Region proposals can model the appearance of objects and shapes with a well-defined closed boundary. For each proposal, the CNN features are extracted using the pre-trained model provided in [**?**]. Compared with the hand-crafted features, the CNN feature 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 help to handle the relative simple scenes and the higher-level features more easily detects the complex semantic objects.

2) Ranking in Primal Space and Kernelized Subspace: Most candidate objects 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 casts saliency detection as a ranking problem to sort out the object proposals with high segmentation precision and recall to detect salient objects via a weighted fusion of them.

Ranking in primal space. We investigate a primal-based Rank-SVM (PRSVM) proposed by Chapelle and Keerthi [**?**] as it makes the training for large amounts of imbalanced positive and negative samples available. The PRSVM is also used as a core model to measure the informativeness of each proposal in the acquisition algorithm.

Assume there exists a set of candidate objects $\boldsymbol{X} = [\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n]$ with relevance ranks $\boldsymbol{x}_n \succ \cdots \succ \boldsymbol{x}_i \succ \boldsymbol{x}_j \succ \cdots \succ \boldsymbol{x}_1$, where $\succ$ denotes the order and $\boldsymbol{x}_k \in \mathbb{R}^d$ is the feature vector of the k-th instance. In a Rank-SVM problem, instances ranking ahead could have higher scores than the behind ones, which is described in the following formula:

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

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 satisfy $\mathcal{P} = \{(i,j)|y_i > y_j\}$, and $y_i \in \{-1, +1\}$ is the label of the i-th training instance. Additionally, the preference pairs are only defined on the between-class instances (i.e., positive and negative instances) for computation efficiency.

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

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

To determine positive and negative instances, the confidence measure has been taken into consideration, which is an overall performance measurement weighted by the accuracy score A and coverage score C as mentioned in [**?**]. 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 balancing the weight between accuracy score and coverage score. In our experiments, the instances with confidence score higher than 0.9 are regarded as positive samples, whereas instances with confidence score lower than 0.4 are treated as negative ones. We uses all possible positive samples but only a fraction of the negative ones, as the amount of positive samples is much larger than 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. To address this issue, we learn a feature projection matrix to project high-dimensional features into a low-dimensional subspace with a kernel approach. aiming at learning a linear projection matrix that maps data points into a low-dimensional subspace.

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 [?] firstly propose pairwise constrained component analysis (PCCA) for learning this transformation matrix with similarity and dissimilarity constraints. Xiong *et al.* [?] further improve it by incorporating a regularization model. We simultaneously considers the ranker and subspace learning in a unified formula:

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

where $\ell_\delta(x) = \frac{1}{\delta}\log(1 + e^{\delta x})$ is the generalized logistic loss function as mentioned in [?]. $||\cdot||$ represents the Euclidean distance and $||\cdot||_F$ is the Frobenius norm of matrix. $\psi(\boldsymbol{x}_i)$ is the feature of instance $\boldsymbol{x}_i$ through kernel projection. p is the number of constraints for the instance pairs $\boldsymbol{x}_{i_n}$ and $\boldsymbol{x}_{j_n}$, 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. $\boldsymbol{L} \in \mathbb{R}^{l\times d} (l < d)$ is the learned projection matrix and μ is the regularization parameter. The first two terms are the Rank-SVM formula defined in the subspace, which encourages 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 $\boldsymbol{L}$.

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

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

where $Tr(\cdot)$ denotes the trace of a matrix. As shown in Figure 2, 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.

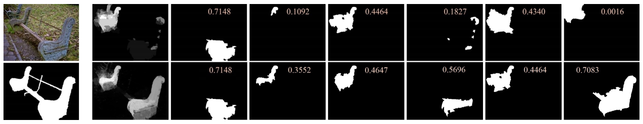


Fig. 2: Ranking results in different feature spaces. Top: results ranked in the primal space. Bottom: results ranked in the kernelized subspace. The decimals in yellow font denote the corresponding confidence scores.

Joint ranker and subspace learning. We learn the Rank-SVM model coefficient $\mathbf{w}$ and projection matrix $\boldsymbol{P}$ jointly by optimizing Equation 4. The optimization problem can be efficiently solved using the alternating optimization method.

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

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

and the Hessian matrix is,

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

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

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

where η is found by line search.

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

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

where $\boldsymbol{T}_n = (\boldsymbol{e}_{i_n} - \boldsymbol{e}_{j_n})(\boldsymbol{e}_{i_n} - \boldsymbol{e}_{j_n})^\top$. $\sigma_\delta(x)$ denotes the value of $(1 + e^{-\delta x})^{-1}$ and $\boldsymbol{e}_k$ is the k-th vector of the canonical 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 $\boldsymbol{K}^{-1}$, the projection matrix $\boldsymbol{P}$ is computed by iteratively solving the following problem

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

where $\boldsymbol{Q}_{ij}^t = (\mathbf{w}^\top \boldsymbol{P}(\boldsymbol{k}_i - \boldsymbol{k}_j) - 1)\mathbf{w}(\boldsymbol{k}_i - \boldsymbol{k}_j)^\top$,$\boldsymbol{A}_n^t = y_n\sigma_\delta(y_n(||\boldsymbol{P}(\boldsymbol{k}_{i_n} - \boldsymbol{k}_{j_n})||^2 - 1))$ at the t-th iteration and α represents the learning rate. The joint learning algorithm is summarized in Algorithm 2.

Algorithm2: Kernelized Subspace Ranking

```
Input: $\boldsymbol{K}$ (kernel matrix); $\lambda$ (trade-off parameter); $y_n$ (label of instance pairs); $\mathcal{P}$ (preference pairs); $\mu$ (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 $\boldsymbol{P}$.
repeat
  • Update the ranking weight $\mathbf{w}$ with fixed $\boldsymbol{P}$
  repeat
    Evaluate the ranking gradient $\boldsymbol{g}$ by Equ. 5;
    Compute the ranking Hessian matrix $\boldsymbol{H}$ by Equ. 6;
    Update the ranking weight $\mathbf{w}$ by Equ. 7;
  until Convergence
  • Update the projection matrix $\boldsymbol{P}$ with fixed $\mathbf{w}$
  repeat
    Solve the derivative $\partial E/\partial \boldsymbol{P}$ by Equ. 8;
    Update the projection matrix $\boldsymbol{P}$ by Equ. 9;
  until Convergence
until iteration stopping criterion is reached
```

Saliency computation. The ranking score of the i-th proposal is computed by given the Rank-SVM coefficient $\mathbf{w}$ and projection matrix $\boldsymbol{P}$ as

$$s_i = \mathbf{w}^\top \boldsymbol{P}\boldsymbol{k}_i. \tag{10}$$

We considers 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 combines the top-ranked proposals weighted by their ranking scores to compute the saliency score for each pixel:

$$S(x) = \sum_{i=1}^{K} \exp(2 \times s_i) \times m_i(x), \tag{11}$$

where $m_i(x)$ is 1 if pixel x is included in the i-th proposal, and 0 otherwise. We normalize the saliency scores of all pixels to obtain the initial map by the min-max strategy.

C. Active Acquisition Algorithm

The acquisition algorithm q is designed to evaluate the informativeness of unlabeled instances from two aspects: uncertainty and information density.

Uncertainty sampling. We define an uncertainty metric β to grade each image instance based on the feature distribution of its object proposals towards the decision boundary of the KSR model. SVM based active learning strategy is first proposed by Tong *et al.* [**?**]. However, our method is different from existing works, such as [**?**], [**?**] and [**?**], that directly query the instances closed to the boundary. In this work, we query the high-uncertainty images, which have large proportion of "paradoxical proposals" to their respective proposal pools. "Paradoxical proposals" refer to the proposals lying in the proximity of the decision boundary of model f.

Specifically, the uncertainty sampling process is shown as follows: At first, each unlabeled image $\boldsymbol{U}$ belonging to the current unlabeled data pool $\mathcal{U}$ are tested using the current KSR model f, which computes an ranking score s_i for each proposal $\boldsymbol{x}_i \in \mathcal{X}$ of image $\boldsymbol{U}$. The proposals with higher scores are more likely to be foreground proposals, whereas ones with lower scores are taken as background ones. Those in the middle are "paradoxical", which are usually more informative and benefit model training. Then, the scores of all proposals are normalized using min-max normalization, that is

$$s_i^{norm} = \frac{s_i - s_{min}}{s_{max} - s_{min}}, \tag{12}$$

where s_{min} and s_{max} denote the lowest and the highest ranking scores in $\mathcal{X}$, respectively. We take the candidate object, whose normalized score s^{norm} falls into the range between 0.4 and 0.9, as the "paradoxical" proposal $\boldsymbol{x}_p$. The "paradoxical" proposals are grouped as

$$\mathcal{X}_p = \{\boldsymbol{x}_i \in \mathcal{X} | \forall i : 0.4 < s_i^{norm} < 0.9\}. \tag{13}$$

Furthermore, we calculate the proportion of the "paradoxical" proposals as the uncertainty score β, where

$$\beta = \frac{card\,(\mathcal{X}_p)}{card\,(\mathcal{X})}. \tag{14}$$

$card$ represents the cardinality of a set. Assume that $\mathcal{B}$ denotes the set of uncertainty scores corresponding to the unlabeled data $\mathcal{U}$. We select the images with high uncertainty as candidates to ask an external oracle for labelling. Specifically, the selected candidate set is formulated as follows,

$$\mathcal{Q}_{uc} = \{\boldsymbol{U}_i \in \mathcal{U} \mid \forall i : \beta_i > \gamma + \rho\nu\}, \tag{15}$$

where γ is the mean value of $\mathcal{B}$, ν is its standard deviation and ρ is a trade-off parameter. That is, the threshold is adaptively determined based on the dispersion of $\mathcal{B}$.

Information density. Since informative instances should be both uncertain to the discriminative model and representative in the underlying feature distribution, we further take diversity sampling into account. We use a density-based algorithm [**?**] to cluster the subspace feature vectors of image candidates $\mathcal{Q}_{uc}$, and the candidates with the highest uncertainty in each cluster are selected. After clustering, we can obtain some high-density clusters $\{\mathcal{C}_i\}_{i=1}^n$ and some outliers $\{\boldsymbol{O}_i\}_{i=1}^m$, that is

$$\mathcal{Q}_{uc} = \{\mathcal{C}_i, i = 1, ..., n\} \bigcup \{\boldsymbol{O}_i, i = 1, ..., m\}. \tag{16}$$

We then successively choose a representative candidate $\boldsymbol{U}_t$ from each high-density cluster $\mathcal{C}_t$, which is the one with the highest uncertainty:

$$\boldsymbol{U}_t = \underset{\boldsymbol{U} \in \mathcal{C}_t}{\operatorname{argmax}}\, \beta(\boldsymbol{U}). \tag{17}$$

Besides, we also include the candidates that lie alone to enrich the distribution diversity,

$$\mathcal{Q} = \{\boldsymbol{U}_t, t = 1, ..., n\} \bigcup \{\boldsymbol{O}_i, i = 1, ..., m\}, \tag{18}$$

where the generated set $\mathcal{Q}$ represents the most informative instances for the current time-step KSR model f. They are chosen to ask an external oracle for labelling. The overview of acquisition algorithm is shown in Algorithm 3. After sample acquisition, the selected images $\mathcal{Q}$ are combined into the labeled data set $\mathcal{L}$ by the active learner, as we mentioned above in Section III-A. In addition, the proposed active acquisition algorithm exploits object-level proposals to formulate both image-level uncertainty and image-level diversity. This thought can be easily and naturally applied in instance segmentation and salient instance detection.

Algorithm3: Active Acquisition Algorithm

```
Input: $\mathcal{U}$ (unlabeled data pool); $\rho$ (trade-off parameter); $r$ (distance parameter).
Output: Query set $\mathcal{Q}$ for annotation.
for each image $\boldsymbol{U}$ in the unlabeled data $\mathcal{U}$ do
    Generate a set of candidate proposals $\mathcal{X}$
    for each proposal $\boldsymbol{x}_i$ do
        Compute the ranking score $s_i$ using f by Equ. 10
    end
    Compute the uncertainty score $\beta$ of image $\boldsymbol{U}$ by Equ. 12, 13 and 14
end
Rank all uncertainty scores in the set $\mathcal{B}$
Obtain the uncertain image set $\mathcal{Q}_{uc}$ by Equ. 15
Cluster in $\mathcal{Q}_{uc}$ and choose a representative image from each cluster by Equ. 16 and 17
Yield the actively selected image set $\mathcal{Q}$ by Equ. 18
```

IV. Experiments

We extensively evaluate the proposed algorithm on six representative benchmark datasets and compare it with seventeen state-of-the-art saliency methods, including the GC [?], PCA [?], HS [?], UFO [?], RR [?], wCtr [?], MAP [?], BL [?], HDCT [?], MR [?], LEGS [?], DRFI [?], RS [?], HCCH [?], RCRR [?] and two end-to-end CNN-based methods UCF [?] and DGRL [?]. We get the saliency results of these competitors by either executing their source codes or using saliency maps provided by the authors. In this section, some details will be covered, including datasets, parameter settings, evaluation criterions, examinations of design options, and the comparison with some state-of-the-art models. Especially, we conduct experimental comparisons between this work and the previous work [?] in Section IV-D.

A. Datasets

Seven datasets are used in the experiments, including ECSSD, PASCAL-S, SOD, HKU-IS, DUT-OMRON, SOC-val, and MSRA datasets. All datasets provide accurate human-labeled pixel-wise ground-truth masks. The ECSSD dataset [?] contains $1,000$ structurally complex images acquired from the Internet. The PASCAL-S dataset [?], one of the most challenging saliency datasets, is composed of 850 natural images. Another challenging dataset, SOD dataset [?], contains 300 images from the Berkeley segmentation dataset, where many images contain multiple salient objects either with low contrast or overlapping with the image boundary. The HKU-IS dataset [?] contains $4,447$ images. The DUT-OMRON dataset [?] contains $5,168$ images. It is also challenging since most of the images contains multiple objects at different scales and location in complex backgrounds. Moreover, MSRA dataset provided by Liu *et al.* [?] and Jiang *et al.* [?] is widely used for salient object detection. It has $5,000$ images and most of the images contain only one salient object. A recently-built dataset SOC [?] includes $3,000$ salient images and $3,000$ non-salient objects from daily object categories. We use its validation set as one of the testing sets.

TABLE I
Quantitative comparisons of different design options in terms of F-measure score.

	KSR	KSR-AL
ECSSD	0.771	0.791
PASCAL	0.676	0.710
SOD	0.662	0.681
HKU-IS	0.747	0.758
DUT-OMRON	0.590	0.602
SOC-val	0.289	0.289

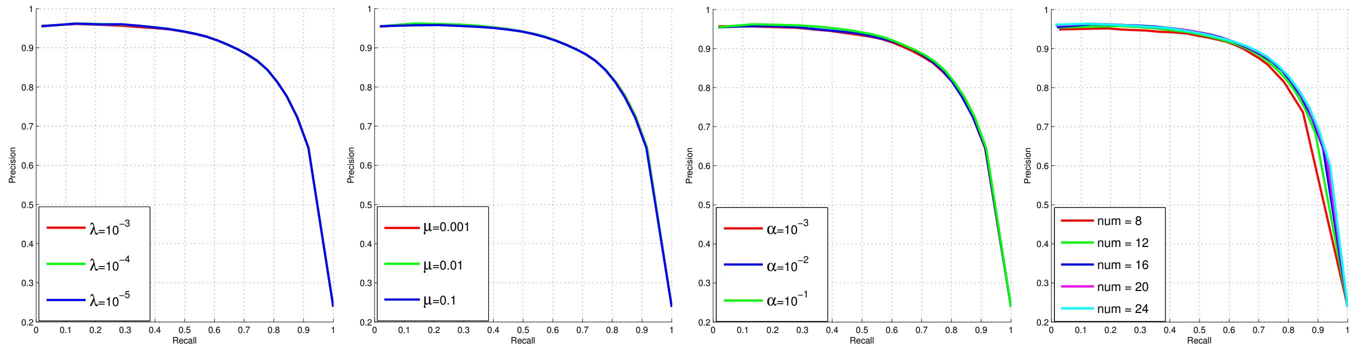
In our previous work, we randomly choose $2,000$ images from MSRA dataset for training. In this work, instead of using all $2,000$ training image at a time, we randomly select 500 images from the MSRA training subset as an initial training set, and then gradually expand the training set by active learning. We finally choose 858 images as the training set, which shows a great reduction comparing with our previous work as well as some state-of-the-art saliency methods. Rather than training a model for each dataset, we train the model only on the MSRA dataset and test it on others, because we actually learn a category-independent ranker to rank the proposals according to their objectness without using any knowledge about object categories.

B. Evaluation Criterions

We use precision and recall (P-R) curve, F-measure, S-measure [?] and Weighted F-measure [?] to qualify the detection results. The precision value is defined as the ratio of salient pixels correctly assigned to all pixels of the extracted regions, while the recall value corresponds to the ratio 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. The F-measure is the overall performance measurement computed by the weighted harmonic of precision and recall: $F_\gamma = \frac{(1+\gamma^2)\times Precision \times Recall}{\gamma^2 Precision + Recall}$, where γ^2 is set to be 0.3 to weigh more on precision as suggested in [?].

C. Parameter Settings

In KSR, 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 Equation 3 are set to be 10^{-4} and 0.01, respectively. In the following Figure 3(a) and 3(b), it explicitly shows that the saliency results are insensitive to the value of the λ and μ, since the corresponding model is trained using the same training set but tenfold increases or decreases the value of these two parameters. This tendency is consistent in terms of the F-measure, S-measure [?], and E-measure [?]. The



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

Fig. 3: Precision-recall curves on the ECSSD dataset by the proposed algorithm with different parameter values. (a) Results with different values of λ in Equ. 3. (b) Results with different values of μ in Equ. 3. (c) Results with different values of α in Equ. 9. (d) Results with different numbers of fusion proposals.

TABLE II
Quantitative comparisons of different methods. Two end-to-end CNN-based methods are listed at the bottom.

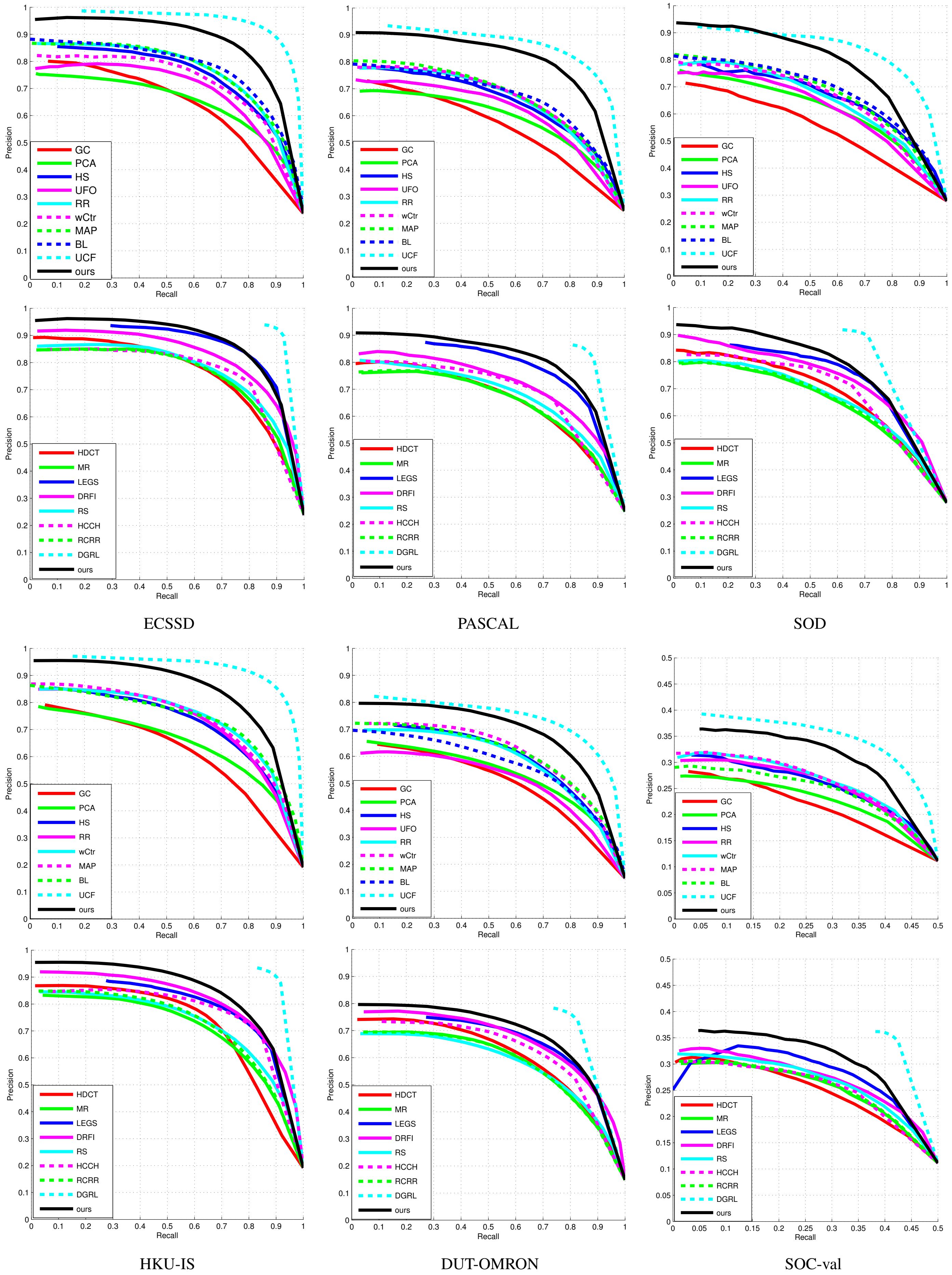
*	ECSSD			PASCAL			SOD			HKU-IS			DUT-OMRON			SOC-val		
	F_β	S_m	F_β^w	F_β	S_m	F_β^w	F_β	S_m	F_β^w	F_β	S_m	F_β^w	F_β	S_m	F_β^w	F_β	S_m	F_β^w
Ours	0.791	0.761	0.622	0.710	0.717	0.568	0.681	0.633	0.496	0.758	0.728	0.576	0.602	0.712	0.487	0.289	0.714	-
GC	0.573	0.610	0.447	0.486	0.552	0.375	0.464	0.522	0.362	0.560	0.613	0.419	0.417	0.604	0.359	0.193	0.610	-
PCA	0.580	0.625	0.367	0.530	0.576	0.330	0.537	0.564	0.338	0.578	0.637	0.350	0.462	0.613	0.287	0.212	0.668	-
HS	0.636	0.685	0.457	0.528	0.623	0.404	0.521	0.599	0.406	0.636	0.673	0.421	0.519	0.633	0.350	0.209	0.604	-
UFO	0.644	0.632	0.424	0.550	0.567	0.356	0.548	0.543	0.355	-	-	-	0.494	0.617	0.332	-	-	-
RR	0.658	0.684	0.500	0.587	0.626	0.417	0.567	0.584	0.399	0.667	0.680	0.460	0.529	0.648	0.384	0.237	0.649	-
wCtr	0.677	0.686	0.511	0.597	0.642	0.453	0.598	0.585	0.419	0.677	0.706	0.515	0.527	0.682	0.428	0.241	0.726	-
MAP	0.701	0.699	0.495	0.586	0.627	0.412	0.583	0.599	0.403	0.659	0.683	0.453	0.522	0.657	0.386	0.238	0.646	-
BL	0.683	0.714	0.463	0.567	0.650	0.410	0.572	0.628	0.409	0.660	0.698	0.419	0.499	0.624	0.318	0.220	0.629	-
HDCT	0.690	0.683	0.453	0.582	0.615	0.379	0.611	0.605	0.400	0.658	0.702	0.445	0.528	0.665	0.357	0.228	0.702	-
MR	0.693	0.692	0.499	0.588	0.621	0.416	0.570	0.583	0.401	0.655	0.669	0.444	0.526	0.646	0.381	0.236	0.647	-
LEGS	0.785	0.786	0.690	0.695	0.722	0.598	0.649	0.659	0.554	0.723	0.742	0.615	0.592	0.714	0.523	0.264	0.804	-
DRFI	0.734	0.750	0.543	0.616	0.670	0.454	0.603	0.648	0.465	0.722	0.740	0.506	0.550	0.698	0.408	0.243	0.719	-
RS	0.661	0.704	0.503	0.598	0.636	0.425	0.580	0.616	0.415	0.660	0.687	0.458	0.524	0.651	0.366	0.242	0.673	-
HCCH	0.721	0.706	0.583	0.613	0.639	0.488	0.634	0.599	0.467	0.723	0.734	0.603	0.571	0.701	0.498	0.244	0.741	-
RCRR	0.714	0.693	0.501	0.663	0.626	0.419	0.576	0.589	0.406	0.666	0.680	0.461	0.527	0.649	0.384	0.238	0.653	-
UCF	0.841	0.880	0.788	0.735	0.803	0.679	0.699	0.754	0.645	0.808	0.866	0.751	0.621	0.760	0.574	0.280	0.658	-
DGRL	0.905	0.899	0.887	0.825	0.836	0.800	0.799	0.771	0.739	0.890	0.895	0.876	0.733	0.806	0.709	0.351	0.811	-

saliency results are also insensitive to the learning rate α in Equation 9 according to Figure 3(c). This learning rate α is fixed to be 0.01. In addition, we uses the Gaussian RBF kernel $k(\boldsymbol{x}, \boldsymbol{x}') = \exp(-\|\boldsymbol{x} - \boldsymbol{x}'\|/\sigma^2)$. The kernel parameter σ is equal to the first quantile of all distances [?].

The proposal algorithm [?] 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. Hence, we compute the percentage of the area of the proposals with respect to the whole image and remove the oversized ($> 70\%$) and undersized ones ($< 2\%$). Besides, we remove the proposals touching four boundaries of an image.

To compute the acquisition function q, we specific the trade-off parameter ρ to be 1.145 in Equation 15. We fuse the top-16 candidates to compute the final saliency map. Although it is shown in Figure 3(d) that our results are insensitive to the number of the fusion candidates when the number is fluctuating, the fusion of the top-16 ones is the optimal choice.

We also show quantitative comparisons under different training datasets and different sizes of initial training set to validate the generalization ability of KSR-AL in Table III. Specifically, we use 200 images as the initial training set. When the active learning process ends, the training set consists of 864 images. Table III shows that the performance of using 200 initial set is slightly better than using 500 initial set on two large datasets (HKU-IS and DUT-OMRON). However, the performance decays on three small datasets (ECSSD, PASCAL and SOD). It is worth to mention that we get the same performance on the SOC-val dataset. Moreover, the DUT-S dataset [?] is one of the most popular training sets for deep saliency model. We compare the performance of using different training datasets with 500 images as initial set. The result shows using the DUT-S as training set performs slightly worse than using the MSRA. The reason may be the quality of the proposals generated from the DUT-S is not as good as the MSRA. In Table IV, the performance and the number of training annotations at each active learning iteration are presented. The performance gradually increases and finally reaches the optimal result.



ECSSD

PASCAL

SOD

HKU-IS

DUT-OMRON

SOC-val

Fig. 4: Quantitative comparisons of different methods on six datasets.

TABLE III
Quantitative comparisons of different training datasets and different number of initial training annotations in terms of F-measure score.

	MSRA-init500	MSRA-init200	DUT-S-init500
ECSSD	0.791	0.789	0.781
PASCAL	0.710	0.708	0.702
SOD	0.681	0.677	0.672
HKU-IS	0.758	0.765	0.746
DUT-OMRON	0.602	0.608	0.588
SOC-val	0.289	0.289	0.285

TABLE IV
Quantitative comparisons of different learning iteration on DUT-OMRON dataset

times	size of training set	F-measure
1	500	0.584
2	612	0.587
3	757	0.590
4	858	0.602

D. Examinations of Design Options

We examine each design options of the proposed algorithm on six datasets, including the ECSSD, PASCAL-S, SOD, HKU-IS, DUT-OMRON, and SOC-val datasets. The results are shown in Table I, where the first column KSR shows the results of our previous work [?], which is in a supervised learning scheme and using $2,000$ training images. The second column KSR-AL represents the results of using the active-learning scheme. As clearly indicated in Table I, there are significant performance increases of the KSR-AL method on all six datasets comparing with the KSR. In addition, the quantity of training data significantly reduces to 858 in the active learning scheme, which is less than half of the previous scale. This reduction directly causes the improvement of testing efficiency. The running time of KSR-AL is 9.96s per image, less than one-third of the time of KSR, which is 33.72s per image. Therefore, the KSR-AL method shows its profound improvement in both efficiency and effectiveness.

E. Comparisons with Other Models

We report the F-measure, S-measure, and Weighted F-measure values of different methods in Table II. Two end-to-end CNN-based methods are listed at the bottom of Table II. Although the CNN-based models perform better in most datasets, we surpass UCF [?] on the SOC-val in terms of both F-measure and S-measure. We do not present the Weighted F-measure values on the SOC-val dataset since this metric cannot evaluate the images without foreground. We miss the performance of UFO based [?] results on the HKU-IS and SOC-val dateset, since it is unavailable. Figure 4 shows the precision-recall curves and F-measure of different methods on five datasets, including ECSSD, PASCAL-S, SOD, HKU-IS, and DUT-OMRON. The figure clearly shows that the precision-recall curve of the proposed method outperforms other competitors. What is more, Figure 5 shows a few saliency maps generated by the evaluated methods. It uniformly highlights the salient regions with well-defined contours, even when the backgrounds are cluttered or the objects and backgrounds have a similar appearance. The source code and its experiment instruction will be available to the public.

V. Conclusions

In this paper, we propose a saliency detection algorithm via kernelized subspace ranking with active learning. The proposal-based kernelized subspace ranker jointly learns an SVM ranker and a distance metric, in which the distance metric maps the high-dimensional R-CNN features into a low-dimensional subspace using kernel projection. Besides, a pool-based active learning algorithm is integrated into kernelized subspace ranker, which considers both uncertainty sampling and information density. This article has conveyed that kernelized subspace ranker with active learning introduces a reduction in annotation, as well as improvement in performance. It has also shown that the proposed method performs favorably against fifteen state-of-the-art methods on six public datasets by conducting extensive experiments.

References

[1] H. Xiao, Y. Wei, Y. Liu, M. Zhang, and J. Feng, "Transferable Semi-supervised Semantic Segmentation," *arXiv e-prints*, p. arXiv:1711.06828, Nov 2017.

[2] Z. Ren, S. Gao, L. Chia, and I. W. Tsang, "Region-based saliency detection and its application in object recognition," *IEEE TCSVT*, vol. 24, no. 5, pp. 769–779, May 2014.

[3] F. Zünd, Y. Pritch, A. Sorkine-Hornung, S. Mangold, and T. Gross, "Content-aware compression using saliency-driven image retargeting," in *ICIP*, Sep. 2013, pp. 1845–1849.

[4] G.-H. Liu, J.-Y. Yang, and Z. Li, "Content-based image retrieval using computational visual attention model," *Pattern Recognition*, vol. 48, no. 8, pp. 2554 – 2566, 2015.

[5] T. Wang, L. Zhang, H. Lu, C. Sun, and J. Qi, "Kernelized subspace ranking for saliency detection," in *ECCV*, 2016, pp. 450–466.

[6] D. Angluin, "Queries and concept learning," *Machine Learning*, vol. 2, no. 4, pp. 319–342, 1988.

[7] Z. Zhou, J. Y. Shin, L. Zhang, S. R. Gurudu, M. B. Gotway, and J. Liang, "Fine-tuning convolutional neural networks for biomedical image analysis: Actively and incrementally." in *CVPR*, 2017, pp. 4761–4772.

[8] D. Yoo and I. S. Kweon, "Learning loss for active learning," in *CVPR*, June 2019.

[9] D. A. Cohn, Z. Ghahramani, and M. I. Jordan, "Active learning with statistical models," *JAIR*, vol. 4, pp. 129–145, 1996.

[10] W. Wang, Q. Lai, H. Fu, J. Shen, and H. Ling, "Salient Object Detection in the Deep Learning Era: An In-Depth Survey," *arXiv e-prints*, p. arXiv:1904.09146, Apr 2019.

[11] L. Itti, C. Koch, and E. Niebur., "A model of saliency-based visual attention for rapid scene analysis," *IEEE TPAMI*, vol. 20, no. 11, pp. 1254–1259, 1998.

[12] J. Han, K. N. Ngan, Mingjing Li, and Hong-Jiang Zhang, "Unsupervised extraction of visual attention objects in color images," *IEEE TCSVT*, vol. 16, no. 1, pp. 141–145, Jan 2006.

[13] M. Cheng, G. Zhang, N. Mitra, X. Huang, and S. Hu., "Global contrast based salient region detection," in *CVPR*, 2011, pp. 409–416.

[14] F. Perazzi, P. Krahenbuhl, Y. Pritch, and A. Hornung., "Saliency filters: Contrast based filtering for salient region detection," in *CVPR*, 2012, pp. 733–740.

[15] S. Goferman, L. Zelnik-Manor, and A. Tal., "Context-aware saliency detection," in *CVPR*, 2010, pp. 2376–2383.

[16] M. Wang, J. Konrad, P. Ishwar, K. Jing, and H. Rowley, "Image saliency: From intrinsic to extrinsic context," in *CVPR*, 2011, pp. 417–424.

[17] X. Hou and L. Zhang., "Saliency detection: A spectral residual approach," in *CVPR*, 2007, pp. 1–8.

[18] X. Li, H. Lu, L. Zhang, X. Ruan, and M.-H. Yang, "Saliency detection via dense and sparse reconstruction," in *ICCV*, 2013, pp. 2976–2983.

[19] J. Zhang, S. Sclaroff, Z. Lin, X. Shen, B. Price, and R. Mech, "Minimum barrier salient object detection at 80 fps," in *CVPR*, 2015, pp. 1404–1412.

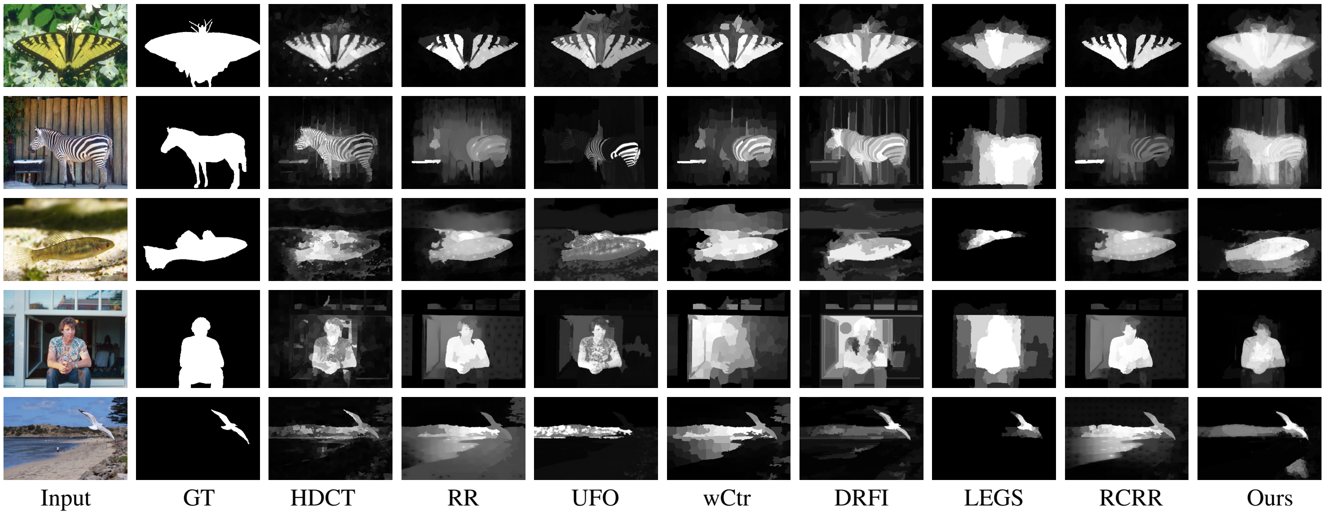


Fig. 5: Visual comparisons with seven state-of-the-art methods.

[20] D. Klein and S. Frintrop., "Center-surround divergence of feature statistics for salient object detection," in *ICCV*, 2011, pp. 2214–2219.

[21] D. Zhang, J. Han, Y. Zhang, and D. Xu, "Synthesizing supervision for learning deep saliency network without human annotation," *IEEE TPAMI*, pp. 1–1, Feb. 2019.

[22] J. Harel, C. Koch, and P. Perona, "Graph-based visual saliency," in *NIPS*, 2006, pp. 545–552.

[23] V. Gopalakrishnan, Y. Hu, and D. Rajan., "Random walks on graphs for salient object detection in images," *IEEE TIP*, vol. 19, no. 12, pp. 3232–3242, 2010.

[24] B. Jiang, L. Zhang, H. Lu, C. Yang, and M.-H. Yang, "Saliency detection via absorbing markov chain," in *ICCV*, 2013, pp. 1665–1672.

[25] C. Yang, L. Zhang, H. Lu, X. Ruan, and M.-H. Yang, "Saliency detection via graph-based manifold ranking," in *CVPR*, 2013, pp. 3166–3173.

[26] C. Li, Y. Yuan, W. Cai, Y. Xia, and D. D. Feng, "Robust saliency detection via regularized random walks ranking," in *CVPR*, 2015, pp. 2710–2717.

[27] Z. Jiang and L. Davis, "Submodular salient region detection," in *CVPR*, 2013, pp. 2043–2050.

[28] T. Liu, S. Jian, N. N. Zheng, X. Tang, and H. Y. Shum, "Learning to detect a salient object," *IEEE TPAMI*, vol. 33, no. 2, 2011.

[29] L. Mai, Y. Niu, and F. Liu, "Saliency aggregation: A data-driven approach," in *CVPR*, 2013, pp. 1131–1138.

[30] Q. Zhao and C. Koch., "Learning a saliency map using fixated locations in natural scenes," *JoV*, vol. 11, no. 3, 2011.

[31] S. Li, H. Lu, Z. Lin, X. Shen, and B. Price, "Adaptive metric learning for saliency detection," *IEEE TIP*, vol. 24, no. 11, pp. 3321–3331, 2015.

[32] N. Tong, H. Lu, X. Ruan, and M.-H. Yang, "Salient object detection via bootstrap learning," in *CVPR*, 2015, pp. 1884–1892.

[33] J. Han, D. Zhang, X. Hu, L. Guo, J. Ren, and F. Wu, "Background prior-based salient object detection via deep reconstruction residual," *IEEE TCSVT*, vol. 25, no. 8, pp. 1309–1321, Aug 2015.

[34] G. Li and Y. Yu, "Deep contrast learning for salient object detection," in *CVPR*, 2016, pp. 478–487.

[35] G. Lee, Y.-W. Tai, and J. Kim, "Deep saliency with encoded low level distance map and high level features," in *CVPR*, 2016.

[36] N. Liu and J. Han, "Dhsnet: Deep hierarchical saliency network for salient object detection," in *CVPR*, 2016, pp. 678–686.

[37] T. Wang, A. Borji, L. Zhang, P. Zhang, and H. Lu, "A stagewise refinement model for detecting salient objects in images," in *ICCV*, 2017, pp. 4019–4028.

[38] X. Li, F. Yang, H. Cheng, W. Liu, and D. Shen, "Contour knowledge transfer for salient object detection," in *ECCV*, 2018, pp. 370–385.

[39] L. Wang, L. Wang, H. Lu, P. Zhang, and X. Ruan, "Saliency detection with recurrent fully convolutional networks," in *ECCV*, 2016, pp. 825–841.

[40] T. Wang, L. Zhang, S. Wang, H. Lu, G. Yang, X. Ruan, and A. Borji, "Detect globally, refine locally: A novel approach to saliency detection," in *CVPR*, 2018, pp. 3127–3135.

[41] Q. Hou, M.-M. Cheng, X. Hu, A. Borji, Z. Tu, and P. H. Torr, "Deeply supervised salient object detection with short connections," in *CVPR*, 2017, pp. 3203–3212.

[42] S. Chen, X. Tan, B. Wang, and X. Hu, "Reverse attention for salient object detection," in *ECCV*, 2018, pp. 236–252.

[43] R. Wu, M. Feng, W. Guan, D. Wang, H. Lu, and E. Ding, "A mutual learning method for salient object detection with intertwined multi-supervision," in *CVPR*, June 2019.

[44] T. Zhao and X. Wu, "Pyramid feature attention network for saliency detection," in *CVPR*, June 2019.

[45] X. Qin, Z. Zhang, C. Huang, C. Gao, M. Dehghan, and M. Jagersand, "Basnet: Boundary-aware salient object detection," in *CVPR*, June 2019.

[46] D. Zhang, J. Han, Chao Li, and J. Wang, "Co-saliency detection via looking deep and wide," in *CVPR*, June 2015, pp. 2994–3002.

[47] D. Zhang, D. Meng, and J. Han, "Co-saliency detection via a self-paced multiple-instance learning framework," *IEEE TPAMI*, vol. 39, no. 5, pp. 865–878, May 2017.

[48] D. Zhang, H. Fu, J. Han, A. Borji, and X. Li, "A review of co-saliency detection algorithms: Fundamentals, applications, and challenges," *ACM Trans. Intell. Syst. Technol.*, vol. 9, no. 4, pp. 38:1–38:31, Jan. 2018.

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

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

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

[52] Z. Tao, H. Liu, H. Fu, and Y. Fu, "Multi-view saliency-guided clustering for image cosegmentation," *TIP*, vol. 28, no. 9, pp. 4634–4645, Sep. 2019.

[53] Hung-Khoon Tan and Chong-Wah Ngo, "Common pattern discovery using earth mover's distance and local flow maximization," in *ICCV*, 2005.

[54] J. Yuan and Y. Wu, "Spatial random partition for common visual pattern discovery," in *ICCV*, 2007.

[55] M. Cho, S. Kwak, C. Schmid, and J. Ponce, "Unsupervised object discovery and localization in the wild: Part-based matching with bottom-up region proposals," in *CVPR*, June 2015.

[56] K. Tang, A. Joulin, L.-J. Li, and L. Fei-Fei, "Co-localization in real-world images," in *CVPR*, June 2014.

[57] D. Angluin, "Queries revisited," in *ALT*, 2001, pp. 12–31.

[58] R. D. King, K. E. Whelan, F. M. Jones, P. G. K. Reiser, C. H. Bryant, S. H. Muggleton, D. B. Kell, and S. G. Oliver, "Functional genomic hypothesis generation and experimentation by a robot scientist." *Nature*, vol. 427, no. 6971, pp. 247–252, 2004.

[59] E. B. Baum and K. Lang, "Query learning can work poorly when a human oracle is used," in *IJCNN*, 1992, pp. 335–340.

[60] L. E. Atlas, D. A. Cohn, and R. E. Ladner, "Training connectionist networks with queries and selective sampling," in *NIPS*, 1990.

[61] D. Cohn, L. Atlas, and R. Ladner, "Improving generalization with active learning," *Machine Learning*, vol. 15, no. 2, pp. 201–221, 1994.

[62] I. Dagan and S. P. Engelson, "Committee-based sampling for training probabilistic classifiers," *ICML*, pp. 150–157, 1995.

[63] T. M. Mitchell, "Generalization as search," *Artificial Intelligence*, vol. 18, no. 2, pp. 203–226, 1982.

[64] P. Krähenbühl and V. Koltun, "Geodesic object proposals," in *ECCV*, 2014, pp. 725–739.

[65] O. Chapelle and S. S. Keerthi, "Efficient algorithms for ranking with svms," *Information Retrieval*, vol. 13, no. 3, pp. 201–215, 2010.

[66] L. Wang, H. Lu, X. Ruan, and M.-H. Yang, "Deep networks for saliency detection via local estimation and global search," in *CVPR*, 2015, pp. 3183–3192.

[67] A. Mignon and F. Jurie, "Pcca: A new approach for distance learning from sparse pairwise constraints," in *CVPR*, 2012, pp. 2666–2672.

[68] F. Xiong, M. Gou, O. Camps, and M. Sznaier, "Person re-identification using kernel-based metric learning methods," in *ECCV*, 2014, pp. 1–16.

[69] T. Zhang and F. J. Oles, "Text categorization based on regularized linear classification methods," *Information Retrieval*, vol. 4, no. 1, pp. 5–31, 2001.

[70] S. Tong and D. Koller, "Support vector machine active learning with applications to text classification," *JMLR*, vol. 2, no. 11, pp. 45–66, 2001.

[71] S. Tong and E. Chang, "Support vector machine active learning for image retrieval," in *ACM MM*, 2001, pp. 107–118.

[72] G. Schohn and D. Cohn, "Less is more: Active learning with support vector machines," in *ICML*, 2000, pp. 839–846.

[73] M. Ester, H.-P. Kriegel, J. Sander, X. Xu *et al.*, "A density-based algorithm for discovering clusters in large spatial databases with noise." in *KDD*, vol. 96, no. 34, 1996, pp. 226–231.

[74] M. Cheng, J. Warrell, W. Lin, S. Zheng, V. Vineet, and N. Crook., "Efficient salient region detection with soft image abstraction," in *ICCV*, 2013, pp. 1529–1536.

[75] R. Margolin, A. Tal, and L. Zelnik-Manor, "What makes a patch distinct?" in *CVPR*, 2013, pp. 1139–1146.

[76] Q. Yan, L. Xu, J. Shi, and J. Jia, "Hierarchical saliency detection," in *CVPR*, 2013, pp. 1155–1162.

[77] P. Jiang, H. Ling, J. Yu, and J. Peng, "Salient region detection by ufo: Uniqueness, focusness and objectness," in *ICCV*, 2013, pp. 1976–1983.

[78] W. Zhu, S. Liang, Y. Wei, and J. Sun, "Saliency optimization from robust background detection," in *CVPR*, 2014, pp. 2814–2821.

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

[80] J. Kim, D. Han, Y.-W. Tai, and J. Kim, "Salient region detection via high-dimensional color transform," in *CVPR*, 2014, pp. 883–890.

[81] H. Jiang, J. Wang, Z. Yuan, Y. Wu, N. Zheng, and S. Li, "Salient object detection: A discriminative regional feature integration approach," in *CVPR*, 2013, pp. 1–8.

[82] L. Zhang, C. Yang, H. Lu, X. Ruan, and M.-H. Yang, "Ranking saliency," *IEEE TPAMI*, vol. 39, no. 9, pp. 1892–1904, 2017.

[83] Q. Liu, X. Hong, B. Zou, J. Chen, Z. Chen, and G. Zhao, "Hierarchical contour closure-based holistic salient object detection," *IEEE TIP*, vol. 26, no. 9, pp. 4537–4552, 2017.

[84] Y. Yuan, C. Li, J. Kim, W. Cai, and D. D. Feng, "Reversion correction and regularized random walk ranking for saliency detection," *IEEE TIP*, vol. 27, no. 3, pp. 1311–1322, 2018.

[85] P. Zhang, D. Wang, H. Lu, H. Wang, and B. Yin, "Learning uncertain convolutional features for accurate saliency detection," in *ICCV*, Oct 2017.

[86] Y. Li, X. Hou, C. Koch, J. M. Rehg, and A. L. Yuille, "The secrets of salient object segmentation," in *CVPR*, 2014, pp. 280–287.

[87] V. Movahedi and J. Elder, "Design and perceptual validation of performance measures for salient object segmentation," in *CVPRW*, 2010, pp. 49–56.

[88] G. Li and Y. Yu, "Visual saliency based on multiscale deep features," in *CVPR*, 2015, pp. 5455–5463.

[89] D.-P. Fan, M.-M. Cheng, J.-J. Liu, S.-H. Gao, Q. Hou, and A. Borji, "Salient objects in clutter: Bringing salient object detection to the foreground," in *ECCV*, September 2018.

[90] D.-P. Fan, M.-M. Cheng, Y. Liu, T. Li, and A. Borji, "Structure-measure: A New Way to Evaluate Foreground Maps," in *ICCV*, 2017.

[91] R. Margolin, L. Zelnik-Manor, and A. Tal, "How to evaluate foreground maps," in *CVPR*, June 2014, pp. 248–255.

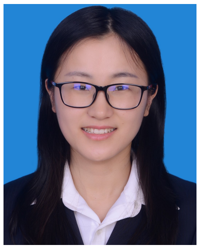
[92] R. Achanta, S. Hemami, F. Estrada, and S. Susstrunk., "Frequency-tuned salient region detection," in *CVPR*, 2009, pp. 1597–1604.

[93] D.-P. Fan, C. Gong, Y. Cao, B. Ren, M.-M. Cheng, and A. Borji, "Enhanced-alignment measure for binary foreground map evaluation," in *IJCAI*, 2018.

[94] L. Wang, H. Lu, Y. Wang, M. Feng, D. Wang, B. Yin, and X. Ruan, "Learning to detect salient objects with image-level supervision," in *CVPR*, 2017.

Lihe Zhang received the M.S. degree and the Ph.D. degree in Signal and Information Processing from Harbin Engineering University, Harbin, China, in 2001 and from Beijing University of Posts and Telecommunications, Beijing, China, in 2004, respectively. He is currently an Associate Professor with the School of Information and Communication Engineering, Dalian University of Technology (DUT). His current research interests include computer vision and pattern recognition.

Jiayu Sun is currently pursuing the Ph.D. degree with Dalian University of Technology. She received her B.E. degree from Northeastern University, Shenyang, China and M.S. degree from the University of Queensland, Brisbane, Australia in 2016 and 2018 respectively. Her research interests include computer vision and machine learning.



Tiantian Wang received the M.S. degree in Signal and Information Processing from the Dalian University of Technology (DUT), Dalian, China, in 2018. She is currently pursuing the Ph.D. degree with the Department of Electrical Science and Computer Engineering, University of California, Merced, USA. Her current research interest includes deep learning, machine learning and their applications in computer vision.

Yifan Min received the B.E. degree in communication engineering from the Dalian Minzu University, in 2017, and the M.S. degree in Signal and Information Processing from the Dalian University of Technology (DUT), Dalian, China, in 2019, respectively. His research interest is in saliency detection.

Huchuan Lu (SM'12) received the Ph.D. degree in System Engineering and the M.S. degree in Signal and Information Processing from Dalian University of Technology (DUT), Dalian, China, in 2008 and 1998, respectively. He joined the faculty in 1998 and currently is a Full Professor of the School of Information and Communication Engineering, DUT. His current research interests include computer vision and pattern recognition with focus on visual tracking, saliency detection, and segmentation. He is a member of the ACM and an Associate Editor of the IEEE Transactions on Cybernetics.