Front.Comput.Sci.
DOI

RESEARCH ARTICLE

Joint Salient Object Detection and Existence Prediction

Huaizu Jiang [2]* **Ming-Ming Cheng** (✉)[1]* **Shi-Jie Li**[2] **Ali Borji**[3] **Jingdong Wang**[4]

[1]CCCE, Nankai University [2]University of Massachusetts Amherst
[3]Center for Research in Computer Vision, University of Central Florida [4]Microsoft Research

Abstract Recent advances in supervised salient object detection modeling has resulted in significant performance improvements on benchmark datasets. However, most of the existing salient object detection models assume that at least one salient object exists in the input image. Such an assumption often leads to less appealing saliency maps on the background images with no salient object at all. Therefore, handling those cases can reduce the false positive rate of a model. In this paper, we propose a supervised learning approach for jointly addressing the salient object detection and existence prediction problems. Given a set of background-only images and images with salient objects, as well as their salient object annotations, we adopt the structural SVM framework and formulate the two problems jointly in a single integrated objective function: saliency labels of superpixels are involved in a classification term conditioned on the salient object existence variable, which in turn depends on both global image and regional saliency features and saliency labels assignments. The loss function also considers both image-level and region-level mis-classifications. Extensive evaluation on benchmark datasets validate the effectiveness of our proposed joint approach compared to the baseline and state-of-the-art models. Source code and data is available at http://mmcheng.net/salexist/.

Keywords Salient object detection, existence prediction, joint inference, saliency detection

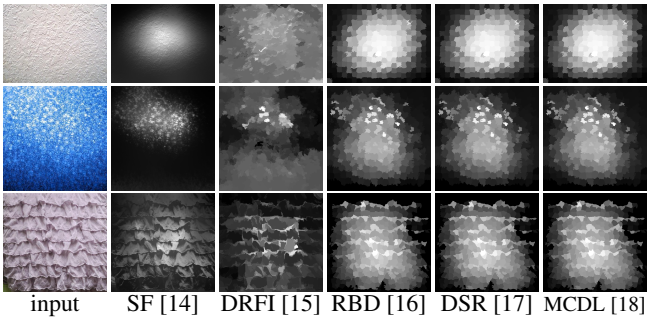


Fig. 1: Saliency maps produced by state-of-the-art models on the background images.

1 Introduction

Salient object detection, deviating from classical human fixation prediction [1–3], aims to detect and segment out the entire salient object(s) that attract most of humans' attention in a scene [4]. Driven by several applications of saliency detection in computer vision and graphics, such as content-aware image resizing [5], image montage [6, 7], sketch based image retrieval [8], photo collection visualization [9], action recognition [10, 11], and Internet image processing [12, 13], many computational models have been proposed in the past decade.

There are two main concerns regarding existing salient object detection methods. First, recent advances in supervised salient object detection has resulted in significant performance improvements on benchmark datasets [15, 18, 19]. Compared with traditional heuristic approaches [14, 16, 17], discriminative learning based on hand-crafted features [15] achieves better performance. By contrast, learned representations via Deep Convolutional Neural Networks (CNNs) [20,

Received month dd, 2017; accepted May 03, 2017. First two authors contribute to this paper equally.

E-mail: cmm.thu@qq.com

21] lead to even more superior results [18, 19]. Second, the majority of existing salient object detection algorithms assume that there is at least one salient object in the input image (See [22]). As a counter example, we show in Fig. 1, some *background images* [23] with no (salient) objects in them. Due to this impractical assumption, all three state-of-the-art approaches [14–16] produce unsatisfactory saliency maps on background images. To this end, we study the problem of salient object existence prediction. Given a set of background images and salient object images, as well as their salient object annotations, our goal is to predict both image-level existence labels and pixel-level saliency values. Note that salient object annotations of background images are effortless (*i.e.*, no salient objects in a background image).

We propose a supervised learning approach to jointly deal with salient object detection and existence prediction problems. The input image is first segmented into a set of superpixels[1)]. Image-level existence labels (*i.e.*, background image vs. salient object image) and region-level saliency labels (*i.e.*, foreground vs. background) are then integrated in a structural SVM framework, where inference can be solved efficiently using the graph cut algorithm [24]. By jointly studying these two problems, we hope they are mutually beneficial in contrast to their separate formulations. The training problem is built upon the large-margin learning framework, where the loss function takes into account both image-level and region-level mis-classifications.

Our main contribution therefore is two folds: (*i*) we propose a supervised learning approach based on the structural SVM framework, for joint salient object detection and existence prediction, where they can benefit from each other, and (*ii*) compared with previous approaches, our model is aware of salient object existence, leading to less false positive detections on background images. Experimental results validate our contributions by comparing our approach to baseline models. Moreover, our approach performs better than most of unsupervised salient object detection models and is comparable with the best supervised approaches.

2 Related Work

In this section, we briefly introduce related works in two areas: salient object detection (and segmentation) and salient object existence prediction.

Salient object detection. We refer readers to [25–27] for a comprehensive review of salient object detection models. Here, we briefly introduce some of the most related works.

Visual saliency is usually related to the uniqueness, distinctiveness, and disparity of items in a scene. Consequently, most of existing works focus on designing models to capture the unique items in the scene. The uniqueness can be computed for each pixel in the frequency domain [28], by comparing a patch to its most similar ones [29], learning complementary saliency priors [30] or discriminative subspace [31], or by comparing a patch to the average patch of the input image in the principal components space [32]. Benefiting from image segmentation algorithms, several approaches try to compute the regional uniqueness in a global manner [14, 33–35], based on multi-scale [36] and hierarchical segmentations [37, 38] of the image [39]. Moreover, several priors about a salient object have been developed in recent years. Since a salient object is more likely to be placed near the center of the image to attract more attention (*i.e.*, photographer bias), it is natural to assume that the narrow border of the image belongs to the background. Such a background prior is widely studied and utilized [17, 40–42]. It has been recently extended to the background connectivity prior assuming that a salient object is less likely to be connected to the border area [16, 43]. In addition, generic objectness prior is also utilized for salient object detection [44–46]. Other priors include spatial distribution [14, 47] and focusness [45].

Some models segment salient objects in a *supervised* manner. The Conditional Random Field [4, 48] and Large-Margin framework [49] are adopted to learn the fusion weights of saliency features. Integration of saliency features can also be discovered based on the training data using Random Forest [15], Boosted Decision Trees (BDT) [50, 51], and mixture of Support Vector Machines [52]. Recent works [18, 19, 53] based on deeply learned representations via CNNs demonstrate even better performance.

Our approach also uses deeply learned CNN representations. Compared with previous models, our proposed salient object detection approach (Sec. 3) is capable of jointly addressing the salient object existence and detection problems.

Salient object existence prediction. In [23], the salient object existence prediction problem is studied as a standard binary classification problem based on global saliency features of thumbnail images. Zhang *et al.* [54] investigate not only existence but also counting the number of salient objects based on holistic cues. In this paper, we focus on recognizing salient object existence. By incorporating superpixels' saliency labels, better performance than [23] can be achieved.

[1)] Here, we use the terms "superpixel" and "region" interchangeably.

3 Joint Salient Object Detection and Existence Prediction

In this section, we first present a joint approach for salient object detection and existence prediction based on the structural SVM framework (Sec. 1). We then introduce our saliency features for the two tasks (Sec. 2).

3.1 A Structural SVM Formulation

In this paper, we are interested in learning a model that not only can predict whether there exist salient objects in the input image but also where they are (if they exist). Our training data is composed of a set of images and their ground-truth annotations in terms of both image-level salient object existence labels (*i.e.*, salient object image *vs.* background image) as well as region-level saliency labels (*i.e.* foreground *vs.* background).

Let I denote the input image consisting of N superpixels $\{r_i\}_{i=1}^N$. Salient object existence label is represented by a binary label $y \in \mathcal{Y} = \{-1, 1\}$, denoting existence of a salient object in the image (-1 for no existence). Regional saliency labels of the image are denoted as $\mathbf{s} = [s_i]_{i=1}^N$, where $s_i \in \mathcal{S} = \{-1, 1\}$ indicates the saliency label for the superpixel r_i (-1 is for background and 1 is for foreground).

Given a set of training samples $\{(I_m, y_m, \mathbf{s}_m)\}_{m=1}^M$, our goal is to learn a model that can be used to predict the salient object existence label y as well as regional saliency labels $\mathbf{s}$ of an unseen test image. To this end, we learn a discriminative function

$$f_{\mathbf{w}} : \mathcal{I} \times \mathcal{Y} \times \mathcal{S}^N \to \mathbb{R} \tag{1}$$

over the image I, its salient object existence label y, and regional saliency labels $\mathbf{s}$, where $\mathbf{w}$ are the parameters. During testing, we use $f_{\mathbf{w}}$ to make predictions of the input image as

$$(y^*, \mathbf{s}^*) = \arg \max_{y \in \mathcal{Y}, \mathbf{s} \in \mathcal{S}^N} f_{\mathbf{w}}(I, y, \mathbf{s}). \tag{2}$$

We consider the global features $\Phi^{ext}(I)$ of the input image I to capture the salient object existence. Additionally, each superpixel r_i is represented by a saliency feature vector $\Phi_i^{sal}(I)$. Their detailed definitions are introduced in Sec. 2. To account for the spatial constraints of two adjacent superpixels that tend to share the same saliency labels, we construct an undirected graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$. The vertex $j \in \mathcal{V}$ corresponds to the saliency configuration of the superpixel r_j and $(j, k) \in \mathcal{E}$ indicates the spatial constraints of superxpixles r_j and r_k. Finally, $f_{\mathbf{w}}(I, y, \mathbf{s}) = \langle \mathbf{w}, \Psi(I, y, \mathbf{s}) \rangle$ is defined as follows,

$$\begin{aligned} \langle \mathbf{w}, \Psi(I, y, \mathbf{s}) \rangle = & \frac{y}{2} \langle \mathbf{w}^{ext}, \Phi^{ext}(I) \rangle \\ & + \sum_{a \in \mathcal{Y}} \sum_{j \in \mathcal{V}} \delta(y = a) \frac{s_j}{2} \langle \mathbf{w}_a^{sal}, \Phi_j^{sal}(I) \rangle \\ & + \sum_{a \in \mathcal{Y}} \sum_{j \in \mathcal{V}, b \in \mathcal{S}} \delta(y = a) \delta(s_j = b) w_{ab}^{sal} \\ & - \sum_{a \in \mathcal{Y}} \sum_{(j,k) \in \mathcal{E}} \delta(y = a) \delta(s_j \neq s_k) w_a^{sm} \cdot v_{jk}. \end{aligned} \tag{3}$$

The model parameters $\mathbf{w}$ are the concatenation of the parameters of all factors in the above equation, *i.e.*, $\mathbf{w} = [\mathbf{w}^{ext}, \mathbf{w}_a^{sal}, w_{ab}^{sal}, w_a^{sm}]_{a \in \mathcal{Y}, b \in \mathcal{S}}$, where w_{ab}^{sal} is a prior term for each region to be foreground and background, respectively. w_a^{sm} is a weight indicating the smoothness term.

Salient object existence prediction. In the above formulation, both salient object detection and existence prediction problems are modeled together in a single integrated objective function. Salient object existence label does not only depend on the global image features $\Phi^{ext}(I)$ as in a standard classification term, but also on the regional saliency labels $\mathbf{s}$ and features $\Phi_j^{sal}(I)$. Compared with existing supervised models of predicting salient object existence labels [23, 54], regional saliency labels are taken into consideration in our approach, leading to better performance.

Salient object detection. In turn, regional saliency labels $\mathbf{s}$ are dependent on the salient object existence label y as well. We learn the prior term w_{ab}^{sal} to model the influence of salient object existence label y on the salient object detection $\mathbf{s}$. Moreover, we learn the last smoothness term for salient object detection on salient object images and background images, respectively, encouraging adjacent regions to take the same saliency label. v_{jk} captures the similarity of two neighboring regions r_j and r_k. It is defined as

$$v_{jk} = \exp(-\frac{|\mathbf{c}_j - \mathbf{c}_k|^2}{2\sigma_c^2}), \tag{4}$$

where $\mathbf{c}_j$ is the average color vector of the superpixel r_j and parameter σ_c is set manually.

3.2 Features

With the availability of huge amount of labeled training data (*i.e.*, ImageNet [55]) and powerful computational resources (*e.g.*, GPUs), CNNs have demonstrated impressive performance on image classification [21]. Deeply learned rep-

resentations can also be used for other visual tasks including object detection [56], face detection and recognition [57], texture recognition [58], fine-grained recognition [59], and multi-view 3D shape recognition [60]. In this paper, we also investigate CNN features for both image-level salient object existence and region-level saliency features.

Image-level salient object existence features. We adopt the learned representation [21] of the input image, which is pre-trained on ImageNet for multi-class image classification. It contains 5 convolutional layers and 2 fully connected layers. We use the output of the $fc7$ layer. The feature dimension is reduced via Principal Component Analysis (PCA), leading to a 1885-dimensional $\Phi^{ext}(I)$ feature vector, where 95% energy (variance) is kept.

Region-level saliency features. For regional saliency, we investigate the global context (GC) features proposed in [18]. We briefly introduce it below and refer readers to [18] for more technical details. To extract the GC feature of each superpixel, the input image is first padded with mean pixel values around the center of the superpixel, where the mean pixel values come from the training images of CNN pre-training. The padded image is then warped to 227×227 pixels and fed into the CNN. By doing so, the GC feature of each region contains the position-aware information of the entire image. Moreover, regional saliency labels can be used to fine tune the pre-trained CNN [18]. We use the fine-tuned Clarifai CNN model [61], which contains 5 convolutional layers and two fully connected layers. The output of the $fc7$ layer is used as the representation of each superpixel. After PCA, we get a 39-dimensional saliency feature vector $\Phi_i^{sal}(I)$ for the superpixel r_i, keeping 95% energy.

4 Learning and Inference

In this section, we describe how to learn our model parameters $\mathbf{w}$ from training samples (Sec. 1) and how to infer both the salient object existence label y and regional saliency labels $\mathbf{s}$ given a test image (Sec. 2).

4.1 Large Margin Learning

Given a set of training samples $\{(I_m, y_m, \mathbf{s}_m)\}_{m=1}^M$, we find the optimal model parameters by minimizing the following regularized empirical risk [62],

$$\min_{\mathbf{w}} \; L(\mathbf{w}) = \frac{\lambda}{2}\|\mathbf{w}\|^2 + \frac{1}{M}\sum_{m=1}^{M} R_m(\mathbf{w}), \tag{5}$$

where λ controls the trade off between the regularization term and the loss term. $R_m(\mathbf{w})$ is a hinge loss function defined as

$$\begin{aligned} R_m(\mathbf{w}) = \max_{y,\mathbf{s}} \; & (\langle \mathbf{w}, \Psi(I_m, y, \mathbf{s})\rangle + \Delta(y_m, y, \mathbf{s}_m, \mathbf{s})) \\ & - \langle \mathbf{w}, \Psi(I_m, y_m, \mathbf{s}_m)\rangle, \end{aligned} \tag{6}$$

where the loss function $\Delta(y_m, y, \mathbf{s}_m, \mathbf{s})$ is defined as follows

$$\Delta(y_m, y, \mathbf{s}_m, \mathbf{s}) = \delta(y_m \neq y) + \alpha(\mathbf{s}_m, \mathbf{s}). \tag{7}$$

The first term is the 0/1 loss widely used for multi-class classification. In addition, we introduce the second term to constrain the salient object segmentation. We measure the misclassifications of regions by counting the number of incorrectly classified superpixels, weighted by their normalized areas. The second loss term can be written as

$$\alpha(\mathbf{s}_m, \mathbf{s}) = \frac{1}{Z}\sum_{l=1}^{N} \beta_l \delta(s_{m,l} \neq s_l), \tag{8}$$

where β_l is the area of the region r_l. $Z = \sum_{l=1}^N \beta_l$ is a normalization term to ensure $\alpha(\mathbf{s}_m, \mathbf{s}) \in [0, 1]$.

Eq. 5 can be efficiently minimized using the bundle optimization method [62], which iteratively builds an increasingly accurate piecewise quadratic approximation of the objective function $L(\mathbf{w})$ based on its sub-gradient $\partial L(\mathbf{w})$. Let

$$\begin{aligned} \mathbf{s}_y^* &= \arg\max_{\mathbf{s}} \left(\langle \mathbf{w}, \Psi(I_m, y, \mathbf{s})\rangle + \Delta(y_m, y, \mathbf{s}_m, \mathbf{s})\right), \\ y^* &= \arg\max_{y \in \mathcal{Y}} \left(\langle \mathbf{w}, \Psi(I_m, y, \mathbf{s})\rangle + \Delta(y_m, y, \mathbf{s}_m, \mathbf{s}_y^*)\right), \end{aligned} \tag{9}$$

The sub-gradient $\partial L(\mathbf{w})$ can then be computed as

$$\partial L(\mathbf{w}) = \lambda \mathbf{w} + \Psi(I_m, y^*, \mathbf{s}_{y^*}^*) - \Psi(I_m, y_m, \mathbf{s}_m).$$

Given the sub-gradient $\partial L(\mathbf{w})$, the optimal model parameters can then be learned by minimizing Eq. 5 using [62].

4.2 Inference

Given a test image I, we maximize Eq. 3 to jointly predict its salient object existence label y^* and regional saliency labels $\mathbf{s}^*$ as follows,

$$(y^*, \mathbf{s}^*) = \arg\max_{y \in \mathcal{Y}, \mathbf{s}} \langle \mathbf{w}, \Psi(I, y, \mathbf{s})\rangle. \tag{10}$$

Since $y \in \mathcal{Y} = \{-1, 1\}$, we can iterate over all its possible values. Given any $y \in \mathcal{Y}$, we utilize the max-flow algorithm [24]

to optimize the Eq. 3 to get the optimal regional saliency labels.

During training, we have to solve the loss-augmented energy function in Eq. 9. Luckily, we can incorporate the loss of regional saliency labels into the unary term of Eq. 3. Therefore, we can again utilize the max-flow algorithm [24] for efficient inference.

To output a saliency map, we diffuse the latent segmentation result of the salient object using the quadratic energy function [49] as follows,

$$\mathbf{z} = \frac{\gamma(\mathbf{I} + \gamma\mathbf{L})^{-1}\mathbf{I}(\mathbf{s} + 1)}{2}, \tag{11}$$

where $\mathbf{z} = [z_i]_{i=1}^{N}$. $z_i \in [0, 1]$ is the saliency value of the superpixel r_i. $\mathbf{I}$ is the identity matrix. $\mathbf{V} = [v_{ij}]$ and $\mathbf{D} = diag\{d_{11}, \cdots, d_{NN}\}$ is the degree matrix, where $d_{ii} = \sum_j v_{ij}$. $\mathbf{L} = \mathbf{D} - \mathbf{V}$ is the Laplacian matrix.

5 Experimental Results

5.1 Setup

To the best of our knowledge, the only publicly available background images dataset in the literature is the thumbnail background image dataset [23]. Images in this dataset, however, are of low resolution (130×130). Furthermore, since we are interested in images with common sizes (*e.g.*, 400×300), this dataset is not suitable for our scenarios. To this end, we collected 6182 background images from the SUN dataset [63], describable texture dataset [64], Flickr, and Bing image search engines. We randomly sample 5000 background images to train our model and leave other 1182 images for testing. Additionally, we randomly sample 5000 images from the MSRA10K dataset [33] for training and 1237 images for testing. In total, we have 10000 images for training and 2419 for testing.

We also test our proposed approach (SSVM) on background images introduced by [54]. It contains 1631 images gathered from different benchmark datasets. Since some images are also sampled from the SUN dataset [63], we exclude these overlapping images from testing, resulting in 688 background images. We denote this dataset as Salient Object Subitizing Background (SOSB).

For the salient object detection task, we evaluate our proposed approach on MSRA-B [15] and ECSSD [39] datasets with pixel-wise annotations. MSRA-B contains 5000 images with variations including natural scenes, animals, indoor scenes, etc. There are 1000 semantically salient but structurally complex images in ECSSD, making it very challenging for models.

methods	superv.	task	pub. & year	deep models
SVO [44]	unspvd.	detection	ICCV 2011	✗
CA [29]	unspvd.	detection	CVPR 2010	✗
CB [36]	unspvd.	detection	BMVC 2011	✗
RC [33]	unspvd.	detection	PAMI 2015	✗
SF [14]	unspvd.	detection	CVPR 2012	✗
LRK [65]	unspvd.	detection	CVPR 2012	✗
HS [39]	unspvd.	detection	CVPR 2013	✗
GMR [41]	unspvd.	detection	CVPR 2013	✗
PCA [32]	unspvd.	detection	CVPR 2013	✗
MC [42]	unspvd.	detection	IJCV 2017	✗
DSR [17]	unspvd.	detection	ICCV 2013	✗
RBD [16]	unspvd.	detection	CVPR 2014	✗
DRFI [15]	spvd.	detection	CVPR 2013	✗
HDCT [51]	spvd.	detection	CVPR 2014	✗
MCDL [18]	spvd.	detection	CVPR 2015	✓
GS [23]	spvd.	existence localization	CVPR 2012	✗
SOS [54]	spvd.	existence counting	CVPR 2015	✗
SSVM	joint + spvd.	detection existence		✓

Table 1: Summaries of different salient object detection algorithms based on supervision type (superv.) and tasks that each method can solve. Unspvd./spvd. denote unsupervised/supervised, respectively.

We compare our approaches with 15 state-of-the-art salient object detection models, including 12 unsupervised methods and 3 supervised models, which are summarized in Tab. 1. Following the benchmark [26], for quantitative comparisons, we binarize a saliency map with a fixed threshold ranging from 0 to 255. At each threshold, we compute Precision and Recall scores. We can then plot a Precision-Recall (PR) curve. To obtain a scalar metric, we report the average precision (AP) score defined as the area under the PR curve. We also report F_θ scores defined as $F_\theta = \frac{(1+\theta^2)\times Precision\times Recall}{\theta^2\times Precision+Recall}$. We set $\theta^2 = 0.3$ following the previous works. Additionally, we report the Mean Absolute Error (MAE) scores between saliency maps and the ground-truth binary masks.

5.2 Empirical Analysis of Our Approach

Here, we empirically analyze our proposed approach on four benchmark datasets. In particular, we quantitatively study

	Test Set	MSRA-B	ECSSD	SOSB
linear SVM [54]	98.51	98.10	95.60	85.71
rbf SVM	98.64	98.38	**95.90**	85.96
[23]	90.64	89.26	72.50	75.23
SSVM (joint)	**99.22**	**98.66**	94.40	**88.36**

Table 2: Salient object existence prediction accuracy (%) of different approaches on benchmark datasets. Best results are highlighted with bold fonts.

	Test Set	MSRA-B	ECSSD	SOSB
w/o smoothness	3.14	**8.05**	16.33	15.36
w/o existence	1.69	15.30	6.41	12.44
SSVM (joint)	**1.55**	15.28	**6.39**	**8.97**

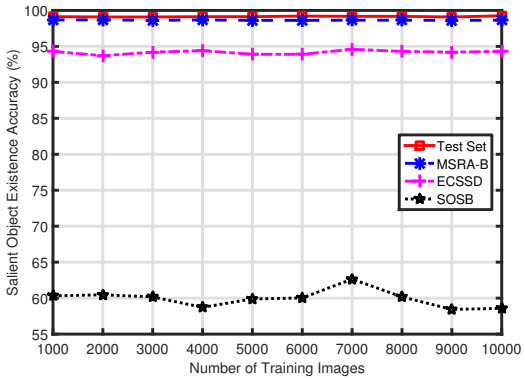
Table 3: Segmentation errors (%) comparison on benchmark datasets. Best results are highlighted with bold fonts.

the performance of both salient object segmentation error and salient object existence prediction accuracy. For each experiment, we randomly select a set of training samples to train the one-class SVM. Such a procedure is repeated 10 times and the average performance is reported.

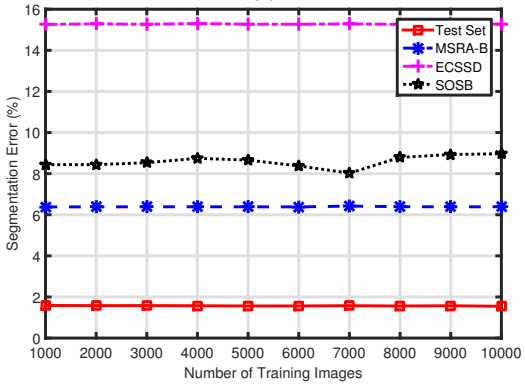
Effectiveness of Joint Learning. To validate the effectiveness of our joint learning approach, *i.e.*, whether salient object segmentation and existence prediction benefit from each other in a single integrated formulation, we compare our model with several baseline approaches.

On one hand, we investigate salient object existence prediction. Without object segmentation, Eq. 3 becomes a standard primal linear SVM formulation. As can be seen from Tab. 2, the joint formulation outperforms the baseline on three benchmark datasets. The improvement on the test set and MSRA-B dataset is valuable considering the high performance of the baseline model. On the most challenging SOSB dataset, the improvement is significant, resulting in an accuracy improvement of around 13%. The reason why the baseline performs better on ECSSD dataset might be due to the dataset bias since we run *cross-dataset* evaluation.

On the other hand, we study salient object detection in terms of segmentation error. We introduce two baselines. The first one, like most existing supervised models, predicts regional saliency labels independently without considering their spatial correlations. The second one takes spatial correlation (*i.e.*, the smoothness term in Eq. 3) of adjacent superpixel labels into account. Similar to our joint formulation, we train segmentation parameters for salient object images



(a)



(b)

Fig. 2: Empirical analysis of our approach on benchmark datasets: (a) accuracy of salient object existence prediction and (b) segmentation error of salient object detection versus different number of training images (M in Eq. 5).

and background images, respectively. During testing, we first predict salient object existence label based on the linear SVM and then choose corresponding segmentation parameters. As we can observe in Tab. 3, salient object segmentation indeed benefits from salient object existence in our joint formulation. The reason why the first baseline model performs the best on MSRA-B dataset might be that the regional saliency CNN model is fine-tuned on the MSRA10K dataset [33], which has a large overlap with MSRA-B (3,000 out of 10,000 images).

Number of Training Images. As can be seen from Fig. 2, both salient object existence prediction accuracy and salient object segmentation error are quite steady w.r.t. the number of training images. One possible reason is that CNN features, which are learned from a huge amount of training data, are so powerful that a good decision boundary could be learned from a few training samples.

	AP		F_θ		MAE			
	MSRA-B	ECSSD	MSRA-B	ECSSD	MSRA-B	ECSSD	Test Set	SOSB
linear SVM + SVO	0.691	0.560	0.580	0.526	0.348	0.412	0.166	0.275
linear SVM + CA	0.559	0.471	0.524	0.466	0.247	0.338	0.118	0.160
linear SVM + CB	0.717	0.591	0.676	0.588	0.188	0.280	0.089	0.168
linear SVM + RC	0.734	0.611	0.730	0.625	0.136	0.234	0.071	0.113
linear SVM + SF	0.661	0.573	0.612	0.432	0.172	0.273	0.089	**0.097**
linear SVM + LRK	0.743	0.590	0.636	0.544	0.214	0.303	0.117	0.167
linear SVM + HS	0.693	0.591	0.733	0.625	0.159	0.266	0.074	0.203
linear SVM + GMR	0.780	0.638	0.745	0.615	0.128	0.236	0.066	0.144
linear SVM + PCA	0.731	0.569	0.634	0.528	0.188	0.288	0.093	0.102
linear SVM + MC	0.777	0.630	0.735	0.610	0.144	0.250	0.077	0.144
linear SVM + DSR	0.763	**0.642**	0.740	0.605	0.119	0.226	0.063	**0.098**
linear SVM + RBD	**0.796**	0.633	0.749	0.602	**0.112**	**0.225**	**0.057**	0.154
linear SVM + **DRFI**	**0.816**	**0.686**	**0.759**	**0.658**	0.129	0.230	0.064	0.171
linear SVM + **HDCT**	0.777	0.616	0.714	0.592	0.148	0.249	0.075	0.107
linear SVM + **MCDL**	0.556	0.540	**0.852**	**0.731**	**0.057**	**0.172**	**0.026**	0.107
SSVM	**0.883**	**0.701**	**0.816**	**0.689**	**0.107**	**0.214**	**0.051**	**0.101**

Table 4: AP, F_θ and MAE scores compared with state-of-the-art approaches on different benchmark datasets, where supervised approaches are marked with bold fonts. The best three scores are highlighted with red, green, and blue fonts, respectively.

5.3 Salient Object Existence Prediction

Here, we quantitatively study our proposed approach for the salient object existence prediction task. Compared with the state-of-the-art approach [23], our approach has two advantages, more powerful features and incorporation of regional saliency information. Though a non-linear classifier (Random Forest) is utilized in [23], as we can see from Tab. 2, our approach has significantly higher prediction accuracy on all datasets. Moreover, compared with [23], our approach is able to jointly address salient object existence and detection problems.

5.4 Salient Object Detection

In this section, we compare our SSVM approach with other state-of-the-art salient object detection approaches. Our SSVM approach is designed to address the limitations of conventional approaches, where they impractically assume that at least one salient object exists in the input image. For a fair comparisons, we introduce a two-stage scheme, as in [23]. Specifically, we first predict the existence label of salient objects using the linear SVM introduced in Sec. 3. If there are no salient objects, we output an all-black saliency map. Otherwise, we generate saliency maps using different approaches.

Since ground-truth annotations of background images are all-black images, only MAE scores are feasible to report on the test set and SOSB dataset. See Tab. 4 and Fig. 3 for quantitative comparisons. We can see in Fig. 3 that PR curves of our approach are higher than others in most cases. Although the PR curves of the MCDL are close to ours, there are no precision values for MCDL when the recall scores are small. To this end, SSVM outperforms other unsupervised and supervised approaches on both MSRA-B and ECSSD datasets in terms of AP scores. Specifically, SSVM performs better than the second best method by 8.2% (DRFI) on MSRA-B and by 2.2% (DRFI) on ECSSD. While the F_θ scores are not as superior as the AP scores, SSVM is ranked as the second best on both MSRA-B and ECSSD datasets, slightly worse than MCDL. In terms of MAE scores, SSVM performs the second best on three datasets and performs the third best on the most challenging SOSB dataset. The reason why SSVM performs inferior to MCDL might be because our approach can not always produce all-black saliency maps for background images as other methods[2)]. Moreover, MCDL considers multi-context (including global and local context of a region).

In Fig. 5, we provide qualitative comparisons of our approach and other top performing approaches. As can be seen, our SSVM approach can produce appealing saliency maps on images where salient objects touch the image border. On background images, our SSVM approach generates near all-black saliency maps, denoting no existence of salient objects.

[2)] Recall that we produce an all-black saliency map if the linear SVM recognizes an input as a background image.

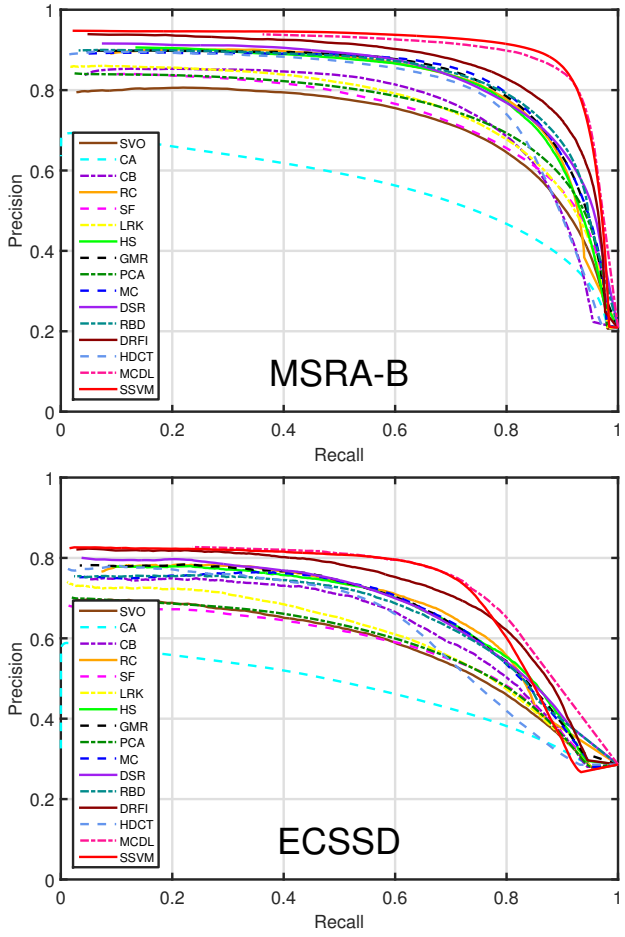


Fig. 3: Precision-Recall curves of different approaches on MSRA-B and ECSSD benchmark datasets.

On a PC equipped with an Intel i7 CPU (3.4GHz) and 32GB RAM, it takes about 2 hours to train our approach using MATLAB code. During testing, our approach takes around 3s to extract features (GPU is not required), and 0.02s for joint inference of the existence label and regional saliency labels. Notice that feature computation dominate the computational time of our method, which could potentially be improved by highly efficient recent deep saliency models [53].

5.5 Limitations and Future Work

Sometimes our approach makes incorrect classifications between salient object images and background images. See Fig. 4 for some failure cases. In the top row, the color of the lady's face is similar to the background, making it hard for a classifier to make a correct prediction. In the bottom row, the classifier is distracted by the artistic content (*e.g.*, high contrast of edges) of the image, resulting in an incorrect classification.

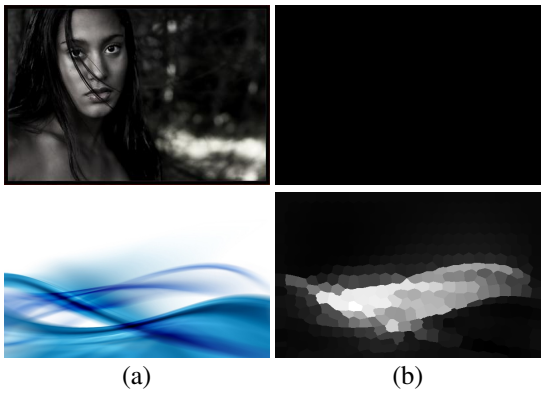


Fig. 4: Failure cases of our SSVM approach. Top row is a salient object image that is incorrectly recognized as a background image. Bottom row is a background image misclassified as a salient object image. From left to right: (a) input images, (b) saliency maps produced by SSVM.

For future work, we plan to investigate end-to-end parameters learning of not only the joint salient object detection and existence prediction objective function, but also the image-level and region-level CNNs. By fine tuning CNNs, better results can be achieved. For instance, it can be expected by a fine-tuned image-level CNN to correctly classify the background image in Fig. 4. Another direction that is worth exploring is the *weakly supervised* setting: learning the model purely from image-level annotations. This would greatly reduce the burden of annotating salient objects.

6 Conclusion

In this paper, we propose a supervised learning approach for joint salient object detection and existence prediction. Our approach is aware of salient object existence and thus produces less false positives on background images that contain no salient objects at all. Moreover, our formulation integrates both image-level and region-level classifications, where two tasks are dependent and benefit from each other. Experimental results validate the effectiveness of our approach.

For potential applications, jointly detecting salient objects and predict their existence is crucial for processing vast amount of heterogenous internet images [12], getting the most dominate regions while suppress false alarms. Thus, our system could potentially be used for image composition [6, 66], big data analysis [67], semantic segmentation [68], and internet based colorization [69].

Acknowledgements This research was sponsored by NSFC (NO. 61572264, 61620106008) and CAST young talents plan.

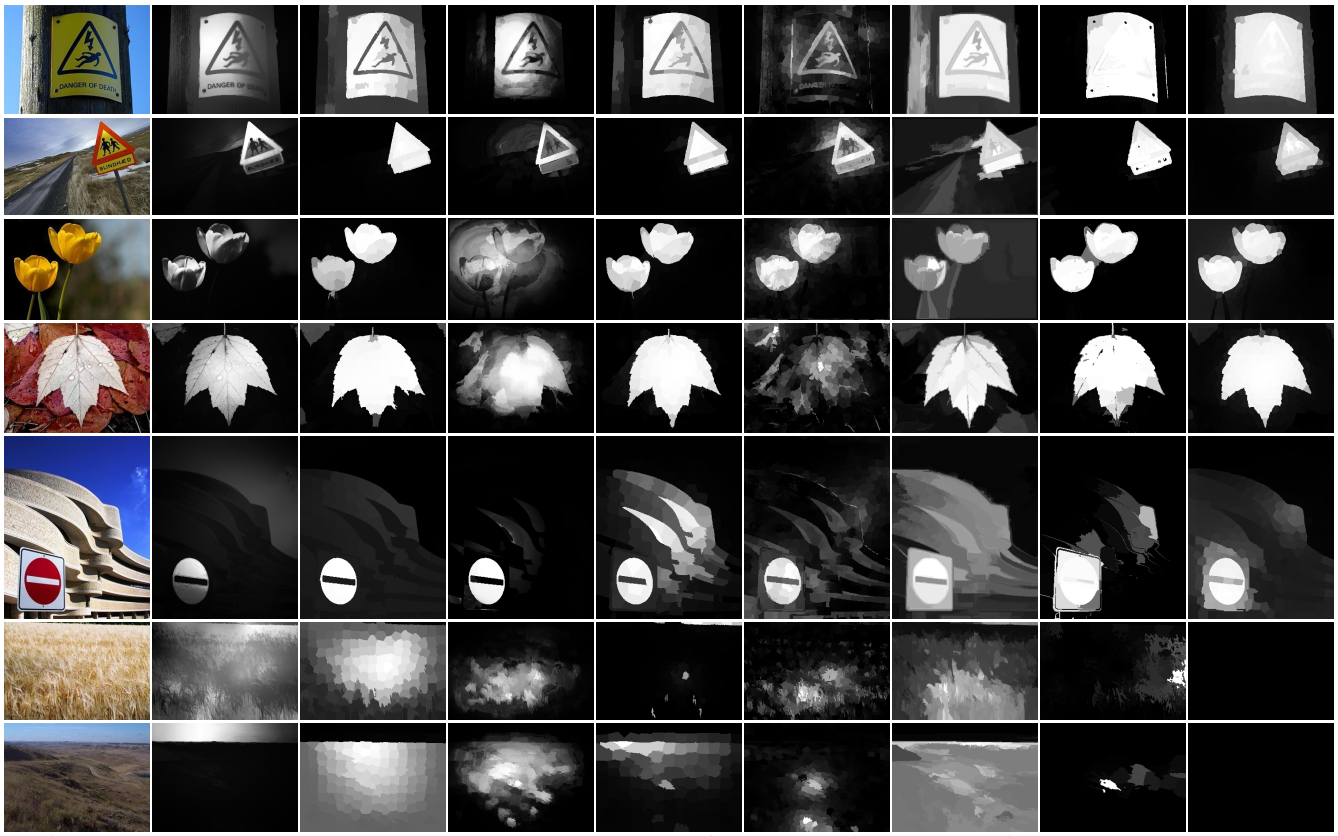


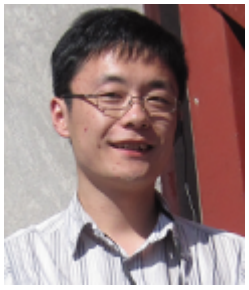
Fig. 5: Qualitative comparisons of saliency maps produced by different approaches. From left to right: input images, saliency maps of state-of-the-art approaches, and saliency maps of our proposed approach SSVM.

References

1. L. Itti, C. Koch, and E. Niebur. A model of saliency-based visual attention for rapid scene analysis. *IEEE TPAMI*, 1998.
2. A. Borji and L. Itti. State-of-the-art in visual attention modeling. *IEEE TPAMI*, 35(1):185–207, 2013.
3. A. Borji, D.N. Sihite, and L. Itti. Quantitative analysis of human-model agreement in visual saliency modeling: A comparative study. *IEEE TIP*, 22(1):55–69, 2013.
4. Tie Liu, Zejian Yuan, Jian Sun, Jingdong Wang, Nanning Zheng, Xiaoou Tang, and Heung-Yeung Shum. Learning to detect a salient object. *IEEE TPAMI*, 33(2):353–367, 2011.
5. Guo-Xin Zhang, Ming-Ming Cheng, Shi-Min Hu, and Ralph R. Martin. A shape-preserving approach to image resizing. *Computer Graphics Forum*, 28(7):1897–1906, 2009.
6. Tao Chen, Ming-Ming Cheng, Ping Tan, Ariel Shamir, and Shi-Min Hu. Sketch2photo: internet image montage. *ACM TOG*, 2009.
7. Tao Chen, Ping Tan, Li-Qian Ma, Ming-Ming Cheng, Ariel Shamir, and Shi-Min Hu. Poseshop: Human image database construction and personalized content synthesis. *IEEE TVCG*, (5), 2013.
8. Ming-Ming Cheng, Niloy J. Mitra, Xiaolei Huang, and Shi-Min Hu. Salientshape: group saliency in image collections. *The Visual Computer*, 30(4):443–453, 2014.
9. Jingdong Wang, Long Quan, Jian Sun, Xiaoou Tang, and Heung-Yeung Shum. Picture collage. In *CVPR*, pages 347–354, 2006.
10. Ashwan Abdulmunem, Yu-Kun Lai, and Xianfang Sun. Saliency guided local and global descriptors for effective action recognition. *Computational Visual Media*, 2(1):97–106, 2016.
11. Jianguang Zhang, Yahong Han, and Jianmin Jiang. Tucker decomposition-based tensor learning for human action recognition. *Multimedia Systems*, 22(3):343–353, 2016.
12. Shi-Min Hu, Tao Chen, Kun Xu, Ming-Ming Cheng, and Ralph R Martin. Internet visual media processing: a survey with graphics and vision applications. *The Visual Computer*, pages 1–13, 2013.
13. Ming-Ming Cheng, Qi-Bin Hou, Song-Hai Zhang, and Paul L Rosin. Intelligent visual media processing: When graphics meets vision. *JCST*, 2017.
14. F. Perazzi, P. Krahenbuhl, Y. Pritch, and A. Hornung. Saliency filters: Contrast based filtering for salient region detection. In *CVPR*, 2012.
15. Jingdong Wang, Huaizu Jiang, Zejian Yuan, Ming-Ming Cheng, Xiaowei Hu, and Nanning Zheng. Salient object detection: A discriminative regional feature integration approach. *IJCV*, 2017.
16. Wangjiang Zhu, Shuang Liang, Yichen Wei, and Jian Sun. Saliency optimization from robust background detection. In *CVPR*, 2014.

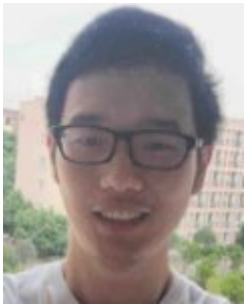
17. Xiaohui Li, Huchuan Lu, Lihe Zhang, Xiang Ruan, and Ming-Hsuan Yang. Saliency detection via dense and sparse reconstruction. In *ICCV*, 2013.
18. Rui Zhao, Wanli Ouyang, Hongsheng Li, and Xiaogang Wang. Saliency detection by multi-context deep learning. In *CVPR*, 2015.
19. Guanbin Li and Yizhou Yu. Visual saliency based on multiscale deep features. In *CVPR*, pages 5455–5463, 2015.
20. Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.
21. Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. Imagenet classification with deep convolutional neural networks. In *NIPS*, pages 1106–1114, 2012.
22. Ali Borji. What is a salient object? a dataset and a baseline model for salient object detection. In *IEEE TIP*. 2014.
23. Peng Wang, Jingdong Wang, Gang Zeng, Jie Feng, Hongbin Zha, and Shipeng Li. Salient object detection for searched web images via global saliency. In *CVPR*, pages 3194–3201, 2012.
24. Yuri Boykov and Vladimir Kolmogorov. An experimental comparison of min-cut/max-flow algorithms for energy minimization in vision. *IEEE TPAMI*, 26(9):1124–1137, 2004.
25. Ali Borji, Ming-Ming Cheng, Huaizu Jiang, and Jia Li. Salient object detection: A survey. *arXiv preprint arXiv:1411.5878*, 2014.
26. Ali Borji, Ming-Ming Cheng, Huaizu Jiang, and Jia Li. Salient object detection: A benchmark. *IEEE TIP*, 24(12):5706–5722, 2015.
27. Junwei Han, Nian Liu, and Dingwen Zhang. Visual saliency detection and applications: A survey. *Frontiers of Computer Science*, 2017.
28. R. Achanta, S. Hemami, F. Estrada, and S. Süsstrunk. Frequency-tuned salient region detection. In *CVPR*, 2009.
29. Stas Goferman, Lihi Zelnik-Manor, and Ayellet Tal. Context-aware saliency detection. *IEEE TPAMI*, 34(10), 2012.
30. Yonghong Tian, Jia Li, Shui Yu, and Tiejun Huang. Learning complementary saliency priors for foreground object segmentation in complex scenes. *IJCV*, 2015.
31. Shu Fang, Jia Li, Yonghong Tian, Tiejun Huang, and Xiaowu Chen. Learning discriminative subspaces on random contrasts for image saliency analysis. *IEEE TNNLS*, 2016.
32. Ran Margolin, Ayellet Tal, and Lihi Zelnik-Manor. What makes a patch distinct? In *CVPR*, pages 1139–1146, 2013.
33. Ming-Ming Cheng, Niloy J. Mitra, Xiaolei Huang, Philip H. S. Torr, and Shi-Min Hu. Global contrast based salient region detection. *IEEE TPAMI*, 37(3):569–582, 2015.
34. A. Borji and L. Itti. Exploiting local and global patch rarities for saliency detection. In *CVPR*, pages 478–485, 2012.
35. Wei Qi, Ming-Ming Cheng, Ali Borji, Huchuan Lu, and Lian-Fa Bai. Saliencyrank: Two-stage manifold ranking for salient object detection. *Computational Visual Media*, 1(4):309–320, 2015.
36. Huaizu Jiang, Jingdong Wang, Zejian Yuan, Tie Liu, and Nanning Zheng. Automatic salient object segmentation based on context and shape prior. In *BMVC*, 2011.
37. Pedro F. Felzenszwalb and Daniel P. Huttenlocher. Efficient graph-based image segmentation. *IJCV*, pages 167–181, 2004.
38. Ming-Ming Cheng, Yun Liu, Qibin Hou, Jiawang Bian, Philip Torr, Shi-Min Hu, and Zhuowen Tu. HFS: Hierarchical feature selection for efficient image segmentation. In *ECCV*, 2016.
39. Qiong Yan, Li Xu, Jianping Shi, and Jiaya Jia. Hierarchical saliency detection. In *CVPR*, pages 1155–1162. CVPR, 2013.
40. Yichen Wei, Fang Wen, Wangjiang Zhu, and Jian Sun. Geodesic saliency using background priors. In *ECCV*, pages 29–42. 2012.
41. Chuan Yang, Lihe Zhang, Huchuan Lu, Xiang Ruan, and Ming-Hsuan Yang. Saliency detection via graph-based manifold ranking. In *CVPR*, 2013.
42. Bowen Jiang, Lihe Zhang, Huchuan Lu, Chuan Yang, and Ming-Hsuan Yang. Saliency detection via absorbing markov chain. In *ICCV*, 2013.
43. Jianming Zhang and Stan Sclaroff. Saliency detection: A boolean map approach. In *ICCV*, pages 153–160, 2013.
44. Kai-Yueh Chang, Tyng-Luh Liu, Hwann-Tzong Chen, and Shang-Hong Lai. Fusing generic objectness and visual saliency for salient object detection. In *ICCV*, pages 914–921, 2011.
45. Peng Jiang, Haibin Ling, Jingyi Yu, and Jingliang Peng. Salient region detection by ufo: Uniqueness, focusness and objectness. In *ICCV*, 2013.
46. Yangqing Jia and Mei Han. Category-independent object-level saliency detection. In *ICCV*, 2013.
47. Ming-Ming Cheng, Jonathan Warrell, Wen-Yan Lin, Shuai Zheng, Vibhav Vineet, and Nigel Crook. Efficient salient region detection with soft image abstraction. In *ICCV*, pages 1529–1536, 2013.
48. Long Mai, Yuzhen Niu, and Feng Liu. Saliency aggregation: A data-driven approach. In *CVPR*, pages 1131–1138, 2013.
49. Song Lu, Vijay Mahadevan, and Nuno Vasconcelos. Learning optimal seeds for diffusion-based salient object detection. In *CVPR*, 2014.
50. Paria Mehrani and Olga Veksler. Saliency segmentation based on learning and graph cut refinement. In *BMVC*, pages 1–12, 2010.
51. Jiwhan Kim, Dongyoon Han, Yu-Wing Tai, and Junmo Kim. Salient region detection via high-dimensional color transform. In *CVPR*, 2014.
52. Pattaraporn Khuwuthyakorn, Antonio Robles-Kelly, and Jun Zhou. Object of interest detection by saliency learning. In *ECCV*. 2010.
53. Qibin Hou, Ming-Ming Cheng, Xiao-Wei Hu, Ali Borji, Zhuowen Tu, and Philip Torr. Deeply supervised salient object detection with short connections. In *IEEE CVPR*, 2017.
54. Jianming Zhang, Shugao Ma, Mehrnoosh Sameki, Stan Sclaroff, Margrit Betke, Zhe Lin, Xiaohui Shen, Brian Price, and Radom??r M??ch. Salient object subitizing. In *CVPR*, 2015.
55. Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *IEEE CVPR*, 2009.
56. Ross B. Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *CVPR*, pages 580–587, 2014.
57. Shuo Yang, Ping Luo, Chen Change Loy, and Xiaoou Tang. From facial parts responses to face detection: A deep learning approach. *ICCV*, 2015.
58. Mircea Cimpoi, Subhransu Maji, and Andrea Vedaldi. Deep filter banks for texture recognition and segmentation. In *CVPR*, 2015.

59. Tsung-Yu Lin, Aruni RoyChowdhury, and Subhransu Maji. Bilinear CNN models for fine-grained visual recognition. *ICCV*, 2015.
60. Hang Su, Subhransu Maji, Evangelos Kalogerakis, and Erik G. Learned-Miller. Multi-view convolutional neural networks for 3d shape recognition. *ICCV*, 2015.
61. Matthew D. Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. In *ECCV*, pages 818–833, 2014.
62. Trinh Minh Tri Do and Thierry Artières. Regularized bundle methods for convex and non-convex risks. *JMLR*, 13:3539–3583, 2012.
63. Jianxiong Xiao, James Hays, Krista A. Ehinger, Aude Oliva, and Antonio Torralba. SUN database: Large-scale scene recognition from abbey to zoo. In *CVPR*, pages 3485–3492, 2010.
64. Mircea Cimpoi, Subhransu Maji, Iasonas Kokkinos, Sammy Mohamed, and Andrea Vedaldi. Describing textures in the wild. In *CVPR*, pages 3606–3613, 2014.
65. Xiaohui Shen and Ying Wu. A unified approach to salient object detection via low rank matrix recovery. In *CVPR*, 2012.
66. Hua Huang, Lei Zhang, and Hong-Chao Zhang. Arcimboldo-like collage using internet images. *ACM TOG*, 30(6):155, 2011.
67. Hong Liu, Lei Zhang, and Hua Huang. Web-image driven best views of 3d shapes. *The Visual Computer*, 2012.
68. Yunchao Wei, Xiaodan Liang, Yunpeng Chen, Xiaohui Shen, Ming-Ming Cheng, Jiashi Feng, Yao Zhao, and Shuicheng Yan. Stc: A simple to complex framework for weakly-supervised semantic segmentation. *IEEE TPAMI*, 2016.
69. Alex Yong-Sang Chia, Shaojie Zhuo, Raj Kumar Gupta, Yu-Wing Tai, Siu-Yeung Cho, Ping Tan, and Stephen Lin. Semantic colorization with internet images. *ACM TOG*, 30(6):156, 2011.



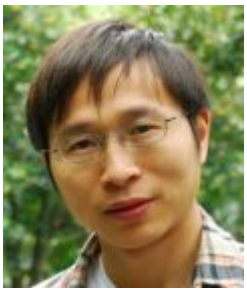
Huaizu Jiang is currently a PhD student in College of Information and Computer Sciences, University of Massachusetts, Amherst. He received his BS and MS degrees from Xi'an Jiaotong University, China, in 2005 and 2009, respectively. He is interested in how to teach an intelligent machine to understand the visual scene like a human.

Ming-Ming Cheng received his PhD degree from Tsinghua University in 2012. Then, he did 2 years research fellow, with Prof. Philip Torr in Oxford. He is now an associate professor at Nankai University. His research interests includes computer graphics, computer vision, and image processing.



Shi-Jie Li received his BS degree from University of Electronic Science and Technology of China, Chengdu, China 2016. He is now a master student in department of computer science, Nankai University, working with Prof. Ming-Ming Cheng.

Ali Borji received his BS and MS degrees in computer engineering from Petroleum University of Technology, Tehran, Iran, 2001 and Shiraz University, Shiraz, Iran, 2004, respectively. He did his Ph.D. in cognitive neurosciences at Institute for Studies in Fundamental Sciences (IPM) in Tehran, Iran, 2009 and spent four years as a postdoctoral scholar at iLab, University of Southern California from 2010 to 2014. He is currently an assistant professor at University of Central Floria, Orlando. His research interests include visual attention, active learning, object and scene recognition, and cognitive and computational neurosciences.



Jingdong Wang received the B.Eng. and M.Eng. degrees in automation from the Department of Automation, Tsinghua University, Beijing, China, in 2001 and 2004, respectively, and the Ph.D. degree in computer science from the Department of Computer Science and Engineering, The Hong Kong University of Science and Technology, Hong Kong, in 2007. He is currently a Lead Researcher with the Internet Media Group, Microsoft Research, Beijing, China. His current research interests include computer vision, machine learning, and multimedia. He has served as an Area Chair in CVPR 2017, ECCV 2016, ACMMM 2015, and ICME 2015, a Track Chair in ICME 2012. He is an Editorial Board Member of the IEEE TRANSACTIONS ON MULTIMEDIA and the International Journal of Multimedia Tools and Applications and an Associate Editor of the International Journal of Neurocomputing. He has shipped 10+ technologies to Microsoft products, including XiaoIce Chatbot, Microsoft cognitive service, and Bing search.