Salient Object Detection: A Large Scale Evaluation

Ali Borji, Ming–Ming Cheng, Huaizu Jiang and Jia Li

Abstract—Detecting and segmenting salient objects in natural scenes, also known as salient object detection, has attracted a lot of focused research in computer vision and has resulted in many applications. However, while many such models exist, yet a deep understanding of achievements and issues is lacking. We aim to provide a comprehensive review and benchmark of the recent progress in this field. In the review part, we situate salient object detection among other closely related areas such as generic scene segmentation, active segmentation, object proposal generation, and saliency for fixation prediction. Covering 274 publications, we survey i) roots, key concepts, and tasks, ii) core techniques and main modeling trends, and iii) datasets and evaluation metrics in salient object detection. In the benchmark part, we extensively compare, quantitatively in terms of both accuracy and running time, 36 state-of-the-art models (24 salient object detection, 10 fixation prediction, 1 objectness, and 1 baseline) over 6 challenging datasets using three evaluation metrics. In retrospect, our evaluation shows a consistent rapid progress over the last few years. The top contenders in our benchmark significantly outperform the models identified as the best in the previous benchmark conducted just two years ago. We find that models designed specifically for salient object detection generally work better than models in closely related areas suggesting the right treatment of this problem. We analyse the influences of center-bias and scene complexity in model performance. Further, we identify hard cases for models hinting towards constructing more challenging large scale datasets. Finally, we propose solutions for tackling existing open problems such as evaluation metrics and dataset bias and suggest future research directions in salient object detection.

Index Terms—Salient object detection, salient region detection, saliency, explicit saliency, visual attention, regions of interest, objectness, segmentation, interestingness, importance, eye movements, scene understanding

1 Introduction

HUMANS are able to detect visually distinctive (so called salient) scene regions effortlessly and rapidly (pre-attentive stage). These filtered regions are then perceived and processed in finer detail for extraction of richer high-level information (attentive stage). This capability has long been studied by cognitive scientists and has recently attracted a lot of interest in computer vision community mainly because it helps find out the objects or regions that efficiently represent a scene and thus harness complex vision problems such as scene understanding.

One of the earliest saliency models, which generated the *first wave* of interest across multiple disciplines including cognitive psychology, neuroscience, and computer vision was proposed by Itti *et al.* [2] (see Fig. 1). This model was an implementation of earlier general computational frameworks and psychological theories of bottom-up attention based on center-surround mechanisms (e.g., *Feature Integration Theory (FIT)* by Treisman and Gelade [3], *Guided Search Model* by Wolfe *et al.* [4], and *Computational Attention Architecture* by Koch and Ullman [5]). In [2], Itti *et al.* showed examples where their model was able to detect spatial discontinuities in scenes. Subsequent behavioral (e.g., [6]) and computational studies (e.g., [7]) started to predict fixations with saliency maps to verify saliency models and to understand human visual attention. A *second wave* of interest (our main focus in this paper) appeared with works of Liu *et al.* [8], [9] and Achanta *et al.* [10] who defined saliency detection as a binary segmentation problem. These works themselves were inspired by some earlier models striving for detecting regions (e.g., Ma and Zhang [11], Liu and Gleicher [12], and Walther *et al.* [13]). Since then a plethora of saliency models have emerged that have faded the boundary between these two category of models. Further, it has been less clear where this new definition stands, as it shares many concepts with other established computer vision areas such as image segmentation algorithms (e.g., [14], [15]), category independent object proposals (e.g., [16], [17]), fixation prediction models (e.g. [7], [18]–[21]), and object detection methods (e.g., [22], [23]). One of our main goals here is to thoroughly review the literature, clarify less understood challenges, and offer a large scale evaluation of models.

In addition to the fast, bottom-up, involuntary, and stimulus-driven stage of attention (of main interest in computer vision community), there exists a slower, top-down, voluntary, and goal-driven stage of attention which is relatively less explored due to the complexity and variety of daily tasks and behaviors (see for example [24]–[26]). Further, subjective factors such as age, culture, and experience regulate attention. For example, a detective sees a crime scene differently than a policeman or a pedestrian.

Some related topics, closely or remotely, to visual saliency include: object importance [27]–[29], memorability [30], scene clutter [31], video interestingness [32], image interestingness [33]–[35], surprise [36], image quality assessment [37], [38], scene typicality [39], [40], aesthetic [35], and attributes [41].

An enormous amount of research has been undertaken

- *A. Borji is with the Department of Computer Science, University of Southern California, Los Angeles, CA, 90089. E-mail: borji@usc.edu*
- *M.M Cheng is with the Department of Engineering Science, University of Oxford, Parks Road, Oxford OX1 3PJ. E-mail: cmm.thu@gmail.com*
- *H. Jiang is with the Institute of Artificial Intelligence and Robotics, Xi'an Jiaotong University, China. E-mail: hzjiang@mail.xjtu.edu.cn*
- *J. Li is with State Key Laboratory of Virtual Reality Technology and Systems, School of Computer Science and Engineering, Beihang University. He is also with the International Research Institute for Multidisciplinary Science (IRIMS) at Beihang University, Beijing, China. E-mail: jiali@buaa.edu.cn*

- *An earlier version of this work has been published in ECCV 2012 [1].*
- *First two authors contributed equally.*
- *Manuscript received xx 2014.*

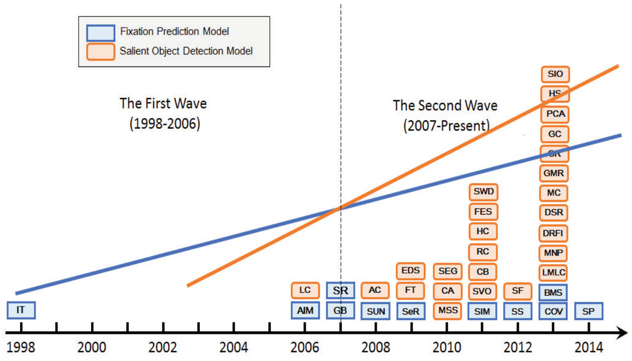


Fig. 1. A simplified chronicle of saliency modeling. Models in the first wave (1998-2007) were mainly dealing with fixation prediction while models in the seconds wave (2007-now) mainly addressed the detection and segmentation of the most salient objects. While both trends are still active research areas in computer vision and cognitive science, salient object detection has attracted more interest recently. Should be updated with the latest algorithms of Sect. 2

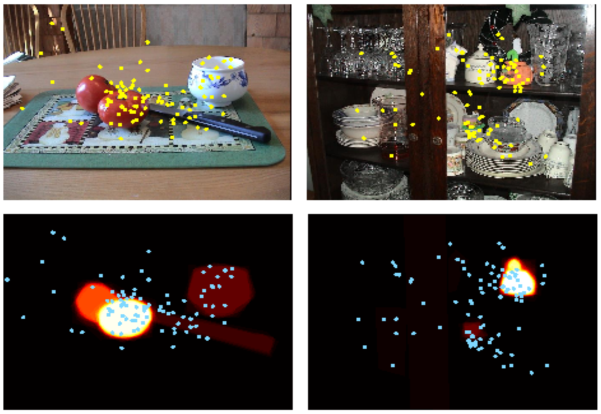


Fig. 2. Two sample images from the Bruce and Tsotsos dataset in Borji *et al.*'s experiment [46]. Left (right) column shows a case where humans are less (more) consistent in choosing the object that stands out the most in the scene. Dots represent 3-second free-viewing fixations.

to study attention through discovering where people look in an image (e.g., [2], [5], [42]). Previous research has shown that eyes are drawn to informative and salient areas in a scene (e.g., [2], [6], [7]). Fewer studies, however, have addressed what explicitly stands out in a scene, and therefore contribute more to perception, understanding, and representation of a scene. Elazary and Itti [43], analyzing LabelMe annotation data [44], showed that human observers tend to annotate more salient objects first. They hence concluded that salient objects are interesting. Masciocchi *et al.* [45] addressed decision processes by which humans choose points in a scene as the most interesting ones (by mouse clicking on 5 most interesting locations). Using a large observer population (>1000 in a web-based study), they found that interest selections are correlated with eye movements, and both types of data correlate with bottom-up saliency. Recently, Borji *et al.* [46] conducted two experiments in which they asked 70 observers to explicitly choose the most outstanding (i.e., salient) object in a scene. In the first experiment, observers viewed scenes with only two objects. In the second experiment, they asked observers to draw a polygon around the most salient object (see Fig. 2). These experiments showed that: 1) observers agree in their judgments, and 2) observers' judgments agree with saliency and eye movement maps. Similar results have been shown by Koehler *et al.* [47]. As in [45], [46], they asked observers to click on salient locations in natural scenes. While the most salient [46], important [27], or interesting [32], [43], [48] objects may tell us a lot about a scene, eventually it might be a subset of objects that can minimally describe a scene. This has been addressed in the past somewhat indirectly in contexts of saliency [49], language and attention [50], and phrasal recognition [41], [51].

1.1 What is Salient Object Detection about?

"Salient object detection" or "Salient object segmentation" is commonly interpreted in computer vision as a process that includes two stages: 1) detecting the most salient object and 2) segmenting the accurate boundary of that object. Rarely, however, models have explicitly distinguished these two stages (with few exceptions such as [52]–[54]). Following the traditional works by Itti *et al.* and Liu *et al.*, models adopted saliency concept to simultaneously perform two stages together. This is witnessed by the fact that these stages have not been separately evaluated and area-based scores have often been employed for model benchmarking (e.g., Precision-recall). The first stage does not necessarily being limited to one object. Majority of existing models, however, have attempted to segment the most salient object although their prediction maps can be used to find several objects in the scene. The second stage falls at the realm of classic segmentation problem in computer vision but has certain differences. For example, here accuracy is mainly determined/constrained by the most salient object. In the next subsection, we briefly review what people perceive as the most salient, interesting or outstanding in a scene from a cognitive perspective. In the rest of the paper, we then deal with the second stage (i.e., segmentation) in more detail. Eventually, in the Discussion section, we will get back to these points and propose new ways to address salient object detection.

In general, it is agreed that for good saliency detection, a model should meet at least the following three criteria: 1) Good detection. The probabilities of missing real salient regions and falsely marking background as salient regions should be low, 2) High resolution. Saliency maps should have high or full resolution to accurately locate salient objects and retain original image information as much as possible, and 3) Computational efficiency. As front-ends to other complex processes, these models should detect salient regions quickly.

1.2 Closely Related Research Areas

Here, we briefly explain similarities and differences of some closely related areas to salient object detection. Fig. 3 shows a illustration of models in these categories.

1.2.1 *Fixation prediction*

The emergence of salient object detection models was driven by the requirement of saliency-based applications (e.g., content-aware image resizing [49], [59], [60]) while fixation

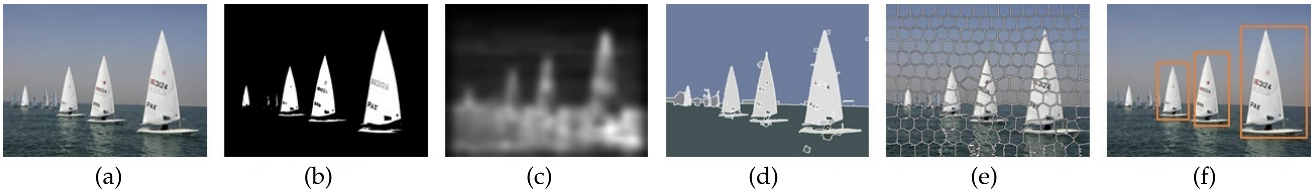


Fig. 3. Sample results produced by different models: (b) salient object detection [55], (c) fixation prediction [2], (d) image segmentation (regions with changing sizes) [56], (e) image segmentation (superpixels with comparable sizes) [57], and (f) object proposals (true positives) [58].

prediction models were constructed originally to understand human visual attention and eye movement prediction [2], [5]. Salient object detection and fixation prediction models have two fundamental differences. *First*, the former models aim to detect and segment the most salient object(s) as a whole by drawing pixel-accurate silhouettes, while the latter models only aim to predict points that people look at (free-viewing of natural scenes usually for 3-5 seconds). In theory/principle a model that works very well on one problem should not work well on the other. For example, salient object detection models segment the whole salient object and will generate a lot of false positives when evaluated with human fixations. On the contrary, fixation prediction models will miss a lot of points inside the salient object, leading to numerous false negatives when evaluated with salient object masks. *Second*, due to the existence of noise in eye tracking or observers' saccade landing (typically around 1 degree $\sim$ 30 pixels), highly accurate pixel-level saliency maps are less desired in fixation prediction. In fact, due to these noises, sometimes blurring/smoothing prediction maps increases the model performance [53], [61]. On the contrary, in salient object detection producing maps that can accurately distinguish object boundaries are highly desirable specially in applications. Note that a typical ground-truth fixation map includes several fixation dots, while a typical ground-truth salient object map usually contains several positive regions composed of many pixels.

In practice, models, whether they address salient object segmentation or fixation prediction, are applicable interchangeably as both entail generating similar saliency maps. For example, several researches have been thresholding their saliency maps, originally designed to predict fixations, to detect and segment salient proto-objects (e.g., [19], [62], [63]). Also different evaluation and benchmarks have been recently devoted for comparing models in these categories.

1.2.2 Image Segmentation

Image segmentation (including semantic scene labeling or semantic segmentation) has a long history in computer vision and is one of the very-well researched areas (e.g., [64]). The aim here is to assign each pixel a integer label that indicates which object or background it belongs to. In contrast, salient object detection models only care about the most salient object(s) and treat the segmentation task as a binary labeling problem (although they generate smooth maps which tell confidence in their decisions). They aim to tell whether a pixel belongs to the most salient object or not. In practice, it is possible to first segment the entire scene and then choose the object that is the most salient one. This approach, however, has not been followed in the past due to two possible reasons: 1) highly accurate general segmentation algorithms still do not exist and 2) such approach will be slow while detecting and segmenting salient objects should be fast since this process is often a front-end stage to more complex operations (i.e., salient object detection is often not the sole goal). To balance between these two challenges, recently salient detection models have taken advantage of superpixels (useful intermediate representation of a scene that extracts homogeneous regions with comparable sizes) which are not very accurate is segmenting objects (often over- or under- segment the scene) but are very fast to compute.

1.2.3 Object Proposal Generation

Object proposal generation models or objectness measures attempt to generate a small set (*e.g.*, a few hundreds or thousands) of object windows/regions, so that these windows/regions covers every objects in the input image, regardless the specific categories of those objects (*i.e.*, generic over categories) [58], [65]–[68]. When compared to traditional sliding window based object detection paradigm [23], [69], estimating object proposals in a pre-processing stage has three major advantages: 1) better accords with our human mental recognition behavior which quickly perceive objects before identifying them [70], [71], 2) greatly speed up the computation by reducing the search locations (e.g. from typically a few millions to less than a few thousands), especially when the number of object classes that need to be detected is high [72], and 3) also improves the detection accuracy by allowing the usage of stronger classifiers during testing [73].

Object proposal generation and salient object detection approaches are tightly linked. On one hand, the former approaches consider saliency as an useful cue for measuring objectness of a region [16], [65]. In other words, an object is more likely to be salient than a region on the background. This is based on a previous finding that image background is usually more structured and homogeneous (thus less salient) than objects [43], [46]. On the other hand, the latter approaches use objectness measures to assign higher saliency values to objects rather than the background (e.g., [74]).

1.3 Contributions and Organization of the Paper

In this paper we contribute in three ways: 1) In sections 2 to 4, we critically and exhaustively review the salient object detection literature as well as closely related topics such as segmentation, fixation prediction, and object proposal generation models. We also review common datasets, evaluation measures and issues pertaining to models, 2) In section 5, we conduct a large-scale benchmark of a total 36 models (24 salient object detection, 10 fixation prediction, and 1 objectness measure) over 6 challenging datasets using 3 evaluation measures. We compare models along several dimensions and also report their run time, and 3) Last but not least, we discuss in sections 6 and 7 the probable challenges and summarize learned lessons and highlight future directions to advance the field.

2 Survey of the State of the Art

Here, we review related works in three categories, including: 1) salient object detection models, 2) models in related areas such as fixation prediction, image segmentation and object proposal generation, and 3) applications of salient object detection. Note that many of these models are inherently correlated and in many cases a model can be interpreted from multiple perspectives. Thus our review will be mainly guided by the major "waves" in the chronicle of salient object detection models (as shown in Fig. 1). In what follows we use terms "salient object detection" and "salient region detection" interchangeably. Note that there is no sharp boundary among these models and in many cases prediction of a model can be used for several purposes. Our categorization here is mainly based on the author's use of the original model.

2.1 Salient Object Detection Models

In the past decades, hundreds of approaches have been proposed for detecting salient or interesting objects in images. Except for several studies on segmenting object-of-interest (e.g., [75]–[77]), most of these approaches aim to detect the salient subsets[1] from images first and then integrate them to segment the whole salient objects. Generally, these approaches share the following two major attributes:

(1) Block-based vs. Region-based analysis. In existing works, there exist mainly two kinds of visual subsets, including blocks and regions[2], that are used to detect salient objects. Blocks were usually adopted by many early approaches, while regions were increasingly popular with the development of superpixel algorithms.

(2) Intrinsic cues vs. Extrinsic cues. When detecting salient objects, a key step is to distinguish salient targets from distractors. Toward this end, some approaches propose to extract various cues only from the input image itself to pop-out targets and suppress distractors (*i.e.*, the intrinsic cues). However, other approaches argue that targets and distractors may share some common visual attributes and the intrinsic cues are often insufficient to distinguish them. Therefore, they incorporate extrinsic cues such as user annotations, depth map or statistical information of similar images to facilitate detecting salient objects in the input image.

According to the two attributes, in this review we divide most of existing salient object detection approaches into three major subgroups, including *block-based models with intrinsic cues*, *region-based model with intrinsic cues*, and *models with extrinsic cues*. In particular, there exist several approaches that can not be simply assigned to any of the three subgroups. For these approaches, we categorize them into a separate subgroup. In the rest part of this Section, we will introduce the representative models from all the four subgroups, while a detailed list of the representative models are illustrated in Fig. 7.

2.1.1 Block-based Models with Intrinsic Cues

In this section, we mainly review salient object detection models which utilized intrinsic cues extracted from blocks. Following the pioneer work of Itti *et al.* [2], salient object detection was widely defined as capturing the uniqueness, distinctiveness, or rarity of a scene, which was studied as pixel-wise center-surround contrast in early stages [10]–[12], [78]. Without any prior knowledge of the sizes of salient objects, multi-scale contrast was frequently adopted for robustness purpose. To that end, a L-layer Gaussian pyramid was first constructed in [8], [12]. Let $I^{(l)}$ be the image at the lth-level of the pyramid, the saliency score of pixel p was defined as:

$$s(p) = \sum_{l=1}^{L} \sum_{p' \in \mathcal{N}(p)} ||I^{(l)}(p) - I^{(l)}(p')||^2, \tag{1}$$

where $\mathcal{N}(p)$ was a 9×9 neighboring window centered at p.

In a later influential study, Achanta *et al.* [79] adopted a frequency-tuned approach to compute pixel-wise saliency maps before object segmentation. A full-resolution saliency map was efficiently computed by simply measuring the pixel-wise difference between the input image and its smoothed version. For a pixel p, its saliency value can be computed as:

$$s(p) = \|I_\mu - I_{\omega_{hc}}(p)\|^2, \tag{2}$$

where I_μ was the mean pixel value of the image (e.g., RGB/Lab features) and $I_{\omega_{hc}}$ was a Gaussian blurred version of the input image (using a 5×5 separable binomial kernel $\frac{1}{16}[1, 4, 6, 4, 1]$). Salient objects are then segmented by binarizing the pixel-wise saliency maps.

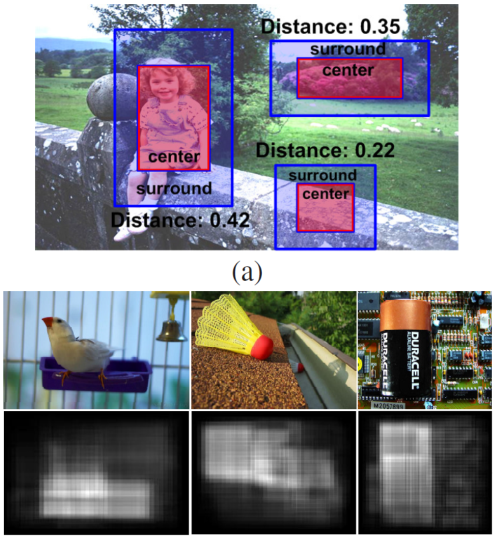
In [78], the input image was represented in a 2D space using polar transformation of its features. In this manner, pixels in each image region were mapped into the same 1D linear subspace. After that, the Generalized Principal Component Analysis (GPCA) [80] was used to estimate the linear subspaces without actually segmenting the image. Finally, the attentive (salient) regions were selected by measuring feature contrasts as well as geometric properties of regions.

In [81], Rosin proposed an extremely efficient approach for detecting salient objects. The whole approach was parameter-free and only required very simple pixel-wise operations such as edge detection, threshold decomposition and moment preserving binarization. The main objective of this approach was to provide a simple scheme to compute salience maps instead of achieving higher performance. Valenti *et al.* [82] proposed an isophote-based framework where the saliency map was estimated by linearly combining the saliency maps computed in terms of curvedness, color boosting, and isocenters clustering.

As intrinsic cues derived on pixel-level is often too poor to support object segmentation, even with multi-scale enhancement, the contrast analysis was extended to patch level. Following the center-surround mechanism suggested in [2],

[1] Visual subsets could be pixels, blocks, superpixels and regions. Blocks are rectangular patches uniformly sampled from the image (pixels are 1×1 blocks). Superpixel and region are perceptually homogenous image patches that are aligned with intensity edges. In the same image, superpixels often have comparable but different sizes, while the shapes and sizes of regions may change remarkably.

[2] In this review, the term "block" is used to represent pixels and blocks, while "superpixel" and "region" are used interchangeably.



(a)

(b)

Fig. 4. Illustration of center-surround contrast [8]. (a) center-surround contrast with different locations and sizes. (b) top row are input images and bottom row shows the saliency map based on center-surround contrast in Eq. 3. Image courtesy from [8].

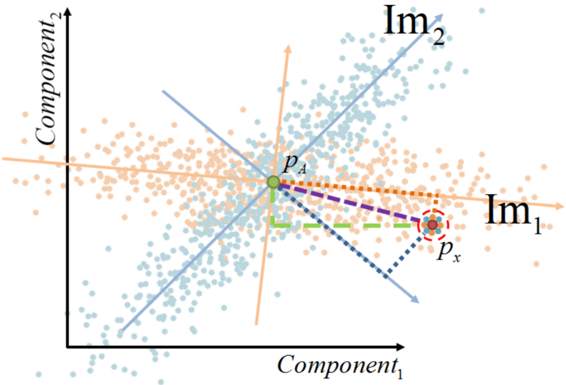


Fig. 5. Illustration of patch distinctiveness incorporating the distribution of patches. Im_1 and Im_2 represent two different images whose principal components are marked by the solid lines. Computing the length between p_x and the average patch p_A along the principal components of each image takes consideration of the distribution of patches of each image. The path for image Im_2 (dashed blue line) is longer than the path for image Im_1 (dashed orange line), correctly corresponding to the distinctness level of p_x in each image. Image courtesy from [85].

the contrast of a patch was always defined as its contrast with the surrounding patches [8], [10], [83], [84][3]. Given a rectangle $R(p)$ centered at pixel p and its surrounding strip $R_S(p)$ with the same area of $R(p)$, the uniqueness of p can be measured by the difference between $R(p)$ and $R_S(p)$, in terms of χ^2 distance of color histograms. In the implementation, the most distinct rectangle $R^*(p)$ and $R^*_S(p)$ were found by enumerating a large number of candidate aspect ratios and sizes for R and R_S until the largest center-surround difference was reached. As a consequence, the saliency score of a pixel p can be then computed as:

$$s(p) = \sum_{\{p'|p' \in R^*(p)\}} w_{pp'} \chi^2 \left(R^*(p'), R^*_S(p') \right), \tag{3}$$

where $w_{pp'}$ was a Gaussian weight between two pixels. See Fig. 4 for an illustration of such center-surround contrast.

Later in [83], such center-surround contrasts were computed in an information-theoretic way by using the Kullback-Leibler Divergence on difference features such as intensity, color and orientation. Li *et al.* [84] proposed to compute the center-surround contrast as a cost-sensitive max-margin classification problem. In particular, the center patch was thought of as a positive sample while the surrounding patches were all used as negative samples. The saliency of the center patch was determined by its inter-class separability from surroundings based the trained cost-sensitive Support Vector Machine (SVM).

Patch uniqueness was also defined as its global contrast with others [49]. A patch was considered to be salient if it was distinct from its K most similar patches, while their spatial distances were taken into account. Intuitively, a patch will be distinct from all the other pathces if it differs remarkably from the most similar pathes. In a recent work [85], Margolin *et al.* proposed to define the uniqueness of a patch by measuring its distance to the average patch based on the observation that indistinctive patches were mostly concentrated in the high-dimensional space while distinct patches were most scattered. To further incorporate the patch distribution on each single image, the uniqueness of a patch was measured by projecting its path to the average patch onto the principal components of the image. To this end, the saliency score of a patch p_x centered at pixel x was defined as

$$s(p_x) = \|\tilde{p}_x\|_1, \tag{4}$$

where $\tilde{p}_x$ was the coordinates of p_x in the PCA coordinated system. See Fig. 5 for an illustration.

Among the aforementioned approaches, a latent assumption, although unclarified, is that one block-based saliency map is accurate enough for segmenting salient objects. However, when benchmarking these models [86], we can see that block-based saliency maps generated by existing approaches are still far from perfect. Toward this end, Yu *et al.* [87] proposed to utilize the complementary characteristics of imperfect saliency maps generated by different contrast-based saliency models. As shown in Fig. 6, two complementary saliency maps were first generated for each image, including a sketch-like map and an envelope-like map. The sketch-like map can accurately locate parts of the most salient object (i.e., skeleton with high precision), while the envelope-like map can roughly cover the entire salient object (i.e., envelope with high recall). With these two maps, the reliable foreground and background regions can be detected from each image first to train a pixel classifier. By labeling all the other pixels with this classifier, the entire salient object can be detected as a whole.

To sum up, approaches in Sects. 2.1.1 aim to detect salient

[3] Though [8] is categorized as an extrinsic model in our review, the extraction of salient object features only involves intrinsic cues.

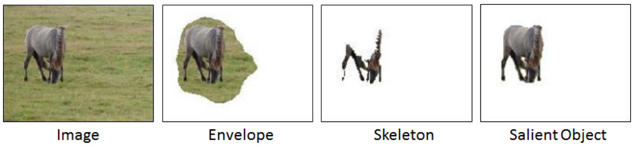


Fig. 6. Salient object can be detected by using the complementary characteristics of imperfect saliency maps in [87].

objects based on pixels or patches, where only intrinsic cues were utilized. These approaches usually suffer from two shortcomings: i) high-contrast edges usually stand out instead of the salient object, and ii) the boundary of the salient object is not preserved well (especially when using large blocks). To overcome these issues, more and more methods propose to compute the saliency map based on regions. This offers two advantages. On one hand, the number of regions are far less than the number of blocks, which implies the potential to develop highly efficient algorithms. On the other hand, more sophisticated features can be extracted from regions, leading to better performance. These region-based approaches will be discussed in the next subsection.

2.1.2 Region-based Models with Intrinsic Cues

Saliency models in the second subgroup adopt intrinsic cues extracted from image *regions* for estimating their saliency scores. Different from the block-based models, the region-based models often segment an input image into regions aligned with intensity edges first and then directly compute a regional saliency map. After that, the most salient superpixels are selected and combined to form salient objects.

As an early attempt of region-based salient object detection, Yu *et al.* [88] proposed a real-time clustering algorithm for fast image segmentation. Based on observations from background and salient regions, several rules were proposed to measure the background score of each region and a hierarchical set of attention regions were detected. In the hierarchy, salient objects were organized from the most salient to the least salient, which were used to provide adaptive image display.

Saliency, defined as uniqueness in terms of **global** regional contrast, was widely studied in existing approaches [55], [89], [96], [98]–[100], [104], [107], [110]. In [89], the input image was first fragmented into N regions $\{r_i\}_{i=1}^N$ and the saliency value of the region r_i can be measured as:

$$s(r_i) = \sum_{j=1}^{N} w_{ij} D_r(r_i, r_j). \tag{5}$$

where $D_r(r_i, r_j)$ captured the appearance contrast between two regions. Higher saliency scores will be assigned to regions which have large contrast to others. w_{ij} was a weight term between regions r_i and r_j, which can serve as spatial weighting purpose by giving farther regions less contributions to the saliency score than close ones. Sometimes, w_{ij} is also introduced to account for the irregular size of the region r_i if it comes from *e.g.*, the graph-based segmentation [111], mean-shift [112] algorithm, or clustering. Alternatively it can be uniformly set if compact superpixels were generated using such as SLIC [57] or Turbopixel [113] algorithms.

In [89], two regio-based salient object detection algorithms, HC and RC were introduced, where the regions comes from quantization (clustering) of color space of pixels and segmentation of the input image using [111]. Perazzi *et al.* [55] demonstrated if $D_r(r_i, r_j)$ was defined as Euclidean distance of colors between r_i and r_j, the global contrast can be efficiently computed using a Gaussian blurring kernel. In specific, Eq. 5 can be re-written as follows,

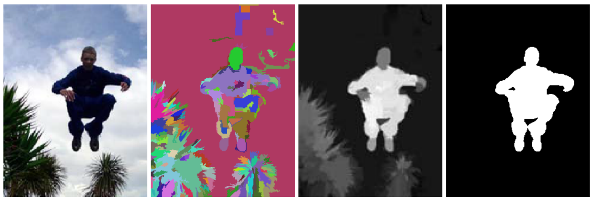


Fig. 8. From left to right, input image, superpixels, global regional contrast, and ground-truth annotation. Image courtesy from [89].

$$s(r_i) = \sum_{j=1}^{N} w_{ij} ||c_i - c_j||^2 \tag{6}$$

$$= c_i^2 \sum_{j=1}^{N} w_{ij} - 2c_i \sum_{j=1}^{N} c_j w_{ij} + \sum_{j=1}^{N} c_j^2 w_{ij}, \tag{7}$$

where c_i is the average color of r_i. The first term is a constant while the later two can be evaluated using a Gaussian blurring kernel on color c_j and squared color c_j^2. If w_{ij} is a Gaussian spatial weighting term, the complexity of Eq. 5 can be reduced from $O(N^2)$ to $O(N)$ of Eq. 7. In addition to color uniqueness, distinctiveness of complementary cues such as texture [99] and structure [98] were also considered for salient object detection. Margolin *et al.* [85] proposed to combine the regional uniqueness and patch distinctiveness to form the saliency map. See Fig. 8 for the illustration of global regional contrast.

Instead of simply maintaining a hard region index for each pixels, a soft abstraction is proposed in [100] based on histogram quantization using a global Gaussian Mixture Model (GMM) to generate a set of large scale perceptually homogeneous regions. By avoiding the hard decision boundaries of superpixels, such soft abstraction provided large spatial support and thus can uniformly highlight the entire salient object.

To further enhance the performance of region-based salient object detection, regional saliency scores were computed in multiple segmentations of the input image based on the **local** contrast in [91]. Specifically, the input image was segmented with the algorithm in [111] using K different groups of parameters. Denote r_i^n as the ith superpixel coming from the nth segmentation. Its saliency was computed as

$$s(r_i^n) = -w_i^n \log \left(1 - \sum_{r_j^n \in \mathcal{N}(r_i^n)} \alpha_{ij}^n D_r(r_i^n, r_j^n)\right). \tag{8}$$

Unlike global contrast, the uniqueness is only captured in a local range $\mathcal{N}(r_i^n)$, which was the set of neighbor regions of r_i^n. α_{ij}^n is the ratio between the area of r_j^n and total area of $\mathcal{N}(r_i^n)$, accounting for the influence of irregular regions. $D_r(r_i^n, r_j^n)$ is the color distance between regions r_i^n and r_j^n, computed as the χ^2 distance between the Lab color as well as the hue histograms of the two regions. As

#	Model	Pub	Year	Elements	Hypothesis		Integration (Optimization)	Code	Bench.
					Uniqueness	Prior			
1	**FG** [11]	MM	2003	PI	L	-	None	NA	
2	**RSA** [78]	MM	2005	PA	G	-	None	NA	
3	**RE** [12]	ICME	2006	mPI + RE	L	-	LI	NA	
4	**RU** [88]	TMM	2007	RE	-	P	LI	NA	
5	**AC** [10]	ICVS	2008	mPA	L	-	LI	NA	
6	**FT** [79]	CVPR	2009	PI	CS	-	None	C	✓
7	**ICC** [82]	ICCV	2009	PI	L	-	LI	NA	
8	**EDS** [81]	PR	2009	PI	ED	-	None	NA	
9	**CSM** [87]	MM	2010	PI + PA	L	SD	None	NA	
10	**RC** [89]	CVPR	2011	RE	G	-	None	C	✓
11	**HC** [89]	CVPR	2011	SRE	G	-	None	C	✓
12	**CC** [90]	ICCV	2011	mRE	-	CV		NA	
13	**CSD** [83]	ICCV	2011	mPA	CS	-	LI	NA	
14	**SVO** [74]	ICCV	2011	PA + RE	CS	O	EM	M + C	✓
15	**CB** [91]	BMVC	2011	mRE	L	CP	LI	M + C	✓
16	**SF** [55]	CVPR	2012	RE	G	SD	NL	C	✓
17	**ULR** [92]	CVPR	2012	RE	SPA	CP + COP	None	M + C	
18	**GS** [93]	ECCV	2012	PA/RE	-	B	None	NA	✓
19	**SIO** [94]	SPL	2013	RE	L	CP	None	NA	
20	**LMLC** [95]	TIP	2013	RE	CS	-	BA	M + C	✓
21	**HS** [96]	CVPR	2013	hRE	G	-	HI	EXE	✓
22	**GMR** [97]	CVPR	2013	RE	-	B	none	M	✓
23	**PISA** [98]	CVPR	2013	SRE	G	SD + CP	NL	NA	
24	**STD** [99]	CVPR	2013	SRE	G	-	None	NA	
25	**PCA** [85]	CVPR	2013	PA + PE	G	-	None	M+C	
26	**GU** [100]	ICCV	2013	SRE	G	-	None	C	✓
27	**GC** [100]	ICCV	2013	SRE	G	SD	AD	C	✓
28	**CHM** [84]	ICCV	2013	PA + mRE	CS + L	-	LI	M + C	✓
29	**DSR** [101]	ICCV	2013	mRE	-	B	BA	M + C	✓
30	**MC** [102]	ICCV	2013	RE	-	B	None	M + C	✓
31	**UFO** [103]	ICCV	2013	RE	G	F + O	NL	M + C	✓
32	**CIO** [104]	ICCV	2013	RE	G	O	None	NA	
33	**SLMR** [105]	BMVC	2013	RE	SPA	CON	None	NA	
34	**LSMD** [106]	AAAI	2013	RE	SPA	CP + COP	None	NA	
35	**SUB** [107]	CVPR	2013	RE	G	CP + COP + SD	None	NA	
36	**PDE** [108]	CVPR	2014	RE	-	CP + B + COP	None	NA	
37	**RBD** [109]	CVPR	2014	RE	-	CON	None	NA	

Fig. 7. Salient object detection models with intrinsic cues sorted by their publication year. For elements, {PI = pixel, PA = patch, RE = region, SRE = soft region}. For hypothesis {CP = center prior, G = global contrast, L = local contrast, ED = edge density, B = background prior, F = focusness prior, O = objectness prior, CV = Convexity prior, CS = center-surround contrast, COP = color prior, SD = spatial distribution, CON = connectivity prior, SPA = sparseness prior}. For hierarchy, {SL = single level, ML = multi-level, TS = two stage, EM = energy minization model}. For integration strategy, {LI = linear, NL = non-linear, AD = adaptive, HI = hierarchical}. For code, {M= Matlab, C= C/C++, NA = not available, EXE = executable}. We evaluate those models whose codes or executables are available.

the salient object usually lies near the center of the image, known as *center prior* for salient object detection, a Gaussian weight w_i^n was used to emphasize regions around the image center. Finally, saliency across multiple segmentations were combined to get pixel-wise saliency map.

$$s(p) = \frac{\sum_{n=1}^{K}\sum_{i=1}^{N(n)} s(r_i^n)c(p, r_i^n)\delta(p \in r_i^n)}{\sum_{n=1}^{K}\sum_{i=1}^{N(n)} c(p, r_i^n)\delta(p \in r_i^n)}, \tag{9}$$

where $N(n)$ was the number of regions in the nth segmentation. $c(p, r_i^n)$ captured the normalized color similarity of the region r_i^n and its contained pixel p. Similar idea of estimating regional saliency on multiple/hierarchical segmentations was also adopted in [96], [101] to increasing the reliability in detecting salient object.

Details are difficult to understand, too many symbols, can be replaced with a short description.

Li *et al.* [84] extended such pairwise local contrast by constructing a hypergraph to capture both internal consistency and external separation of regions for salient object detection. In specific, a hypergraph $\mathcal{G}^h = (\mathcal{V}^h, \mathcal{E}^h)$ was built, where $\mathcal{V}^h = \{v_i^h\}$ was the vertex set corresponding to the superpixels. $\mathcal{E}^h = \{e_j^h\}$ was the hyperedge set that was constructed by multi-scale clustering of superpixels. In this manner, a number of subsets of $\mathcal{V}^h$ were generated such that $\bigcup_{e^h \in \mathcal{E}^h} = \mathcal{V}^h$. As a consequence, the saliency score of vertex v_i^h was defined as

$$s(v_i^h) = \sum_{e^h \in \mathcal{E}^h} \Gamma(e^h)\delta(v_i^h \in e^h), \tag{10}$$

where $\Gamma(e^h)$ encoded the saliency on the hyperedge e^h. Intuitively, a salient hyperedge should be either enclosed by strong image edges or with small interesction with image boundaries. Therefore, $\Gamma(e^h)$ was defined as follows

$$\Gamma(e^h) = \omega_{e^h}\left(||I_g \circ M_g(e^h)||_1 - \rho(e^h)\right), \tag{11}$$

where ω_{e^h} was a scale-specific hyperedge weight (with larger weight on a large scale hyperedge). I_g was the thresholded gradient map of the input image I with Sobel operator. $M_g(e^h)$ was a binary mask indicating the pixels within a narrow band along the boundary of the hyperedge e^h. $\circ$ was the elementwise dot product operator, and $\rho(e^h)$ was a penalty factor that was equal to the number of the image boundary pixels shared by e^h.

Salient object, in terms of uniqueness, can also be defined as the sparse noises in a certain feature space, where the input image was represented as a low-rank matrix [92], [105], [106]. The basic assumption is that non-salient regions (background) can be explained by the low-rank matrix while

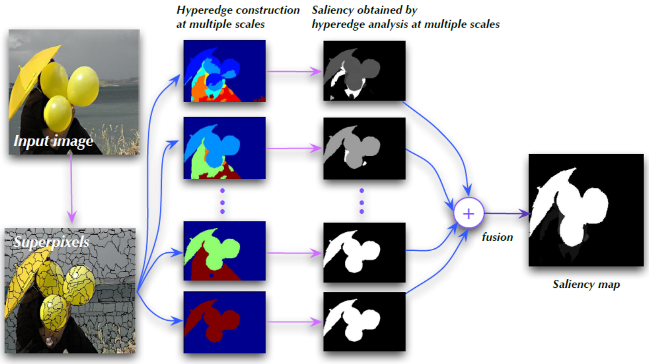

Hyperedge construction at multiple scales
Saliency obtained by hyperedge analysis at multiple scales
Input image
Superpixels
fusion
Saliency map

Fig. 9. An illustration of hypergraph modeling for salient object detection using multi-scale clustering of superpixels [84]. The first column shows an input image and its associated over-segmented image with a set of superpixels. The middle columns display the multi-scale hyperedges and their corresponding results of hyperedge saliency evaluation. The rightmost image shows the final saliency map generated by multi-scale hyperedge saliency fusion. Image courtesy from [84].

the salient regions were indicated by the sparse noises. Formally, each region r_i was described by a feature vector f_i. By stacking f_i together, we can get the feature matrix representation of the entier image $\mathbf{F} = [\mathbf{f}_1, \mathbf{f}_2, \cdots, \mathbf{f}_N] \in \mathcal{R}^{D \times N}$, where D is the dimension of the feature vector. $\mathbf{F}$ can be decomposed into two parts, the low-rank matrix $\mathbf{L} = [\mathbf{l}_1, \mathbf{l}_2, \cdots, \mathbf{l}_N] \in \mathcal{R}^{D \times N}$ and the sparse matrix $\mathbf{S} = [\mathbf{s}_1, \mathbf{s}_2, \cdots, \mathbf{s}_N] \in \mathcal{R}^{D \times N}$ by optimizing the following objective function WHAT SHOULD APPEAR UNDER MIN? L and S?

$$\begin{aligned} \min \quad & \|\mathbf{L}\|_* + \lambda \|\mathbf{S}\|_1, \\ s.t. \quad & \mathbf{F} = \mathbf{L} + \mathbf{S}, \end{aligned} \tag{12}$$

where λ is an coefficient to balance $\mathbf{L}$ and $\mathbf{S}$. $\mathbf{L}$ represented the background while $\mathbf{S}$ corresponded to the salient object. Thus the saliency of r_i can be defined as

$$s(r_i) = ||\mathbf{s}_i||_2, \quad \text{or} \quad s(r_i) = ||\mathbf{s}_i||_1. \tag{13}$$

Based on such general low-rank matrix recovery, Shen and Wu [92] proposed a unified approach to incorporate traditional low-level features with higher-level guidance, *e.g.*, central prior, face detection, and color prior, to detect salient objects based on a learned feature transformation[4]. Instead, Zou *et al.* [105] proposed to exploit bottom-up segmentation as a guidance cue of the low-rank matrix recovery for robustness purpose. Similar to [92], high-level priors were also adopted in [106], where a tree-structured sparsity-inducing norm regularization was introduced to hierarchically desribe the image structure with the aim to more uniformly highlight the entire salient object.

In addition to capturing the uniqueness, more and more priors were proposed for salient object detection as well. Spatial distribution [8] implies that the wider a color is distributed in the image, the less likelily a salient object contrains this color. In [8], [100], the pixels in the input image I was quantized by a Gaussian Mixture Model (GMM) $\{w_c, \mu_c, \Sigma_c\}_{c=1}^{C}$, where $\{w_c, \mu_c, \Sigma_c\}$ was the weight, mean color and the covariance matrix of the cth component. Each pixel p is assigned to a color component with the probability

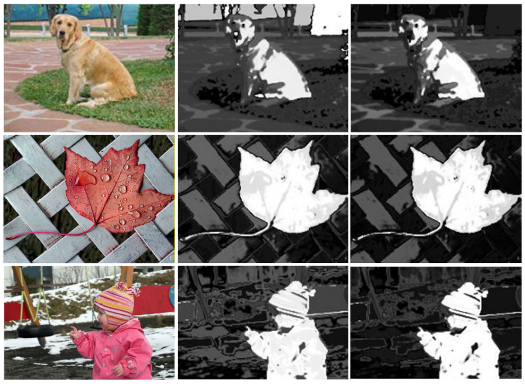


Fig. 10. From left to right, input image and spatial distribution with/without a center-weighted normalization term. Image courtesy from [8].

$$\mathbf{P}(c|I_p) = \frac{w_c \mathcal{N}(I_p|\mu_c, \Sigma_c)}{\sum_c w_c \mathcal{N}(I_p|\mu_c, \Sigma_c)}. \tag{14}$$

The horizontal spatial variance of the component c is then defined as

$$V_h(c) = \frac{1}{|P|_c} \sum_p \mathbf{P}(c|I_p) ||p_h - M_h(c)||^2, \tag{15}$$

$$M_h(c) = \frac{1}{|P|_c} \sum_p \mathbf{P}(c|I_p) p_h, \tag{16}$$

where p_h is the horizontal coordinate of the pixel p and $|P|_c = \sum_p \mathbf{P}(c|I_p)$. The vertical variance $V_v(c)$ was defined similarly. As a consequence, the spatial variance of each clustered component c is

$$V(c) = V_h(c) + V_v(c), \tag{17}$$

Finally, the saliency of each pixel p is defined as

$$s(p) = \sum_c \mathbf{P}(c|I_p)(1 - V(c)) \cdot (1 - D(c)), \tag{18}$$

where $D(c) = \sum_p \mathbf{P}(c|I_p) \cdot d_p$ is a center-weighted normalization term to balance the border cropping effect and d_p is the distance from pixel p to the image center.

The spatial distribution of superpixels can be efficiently evaluated in linear time using the Gaussian blurring kernel, in a similar way of computing the global uniqueness in Eq. 5. Such a spatial distribution prior was also considered in [98] evaluated in terms of both color and structure cues. See Fig. 10 for an illustration of regional spatial distribution.

Central prior assumes that a salient object is more likely to be put near the image center. In other words, the background is tend to be far away from the image center. To this end, the **backgroundness** prior was adopted for salient object detection [93], [97], [101], [102], assuming that a narrow border of the image is background region. With this pseudo-background B as reference, regional saliency can be computed as the contrast versus "background." In [93], a undirected weighted

[4] Though extrinsic ground-truth annotation were adopted to learn high-level priors and feature transformation, we put this model in intrinsic models to better organize the low-rank matrix recovery based approaches. Additionally, we tend to treat face and color priors as universal cues for salient object detection.

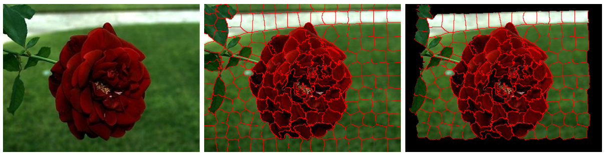


Fig. 11. Illustration of backgroundness prior. From left to right, input image, superpixels, and superpixels with background regions removed touching the image borde.

graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ was built, where the vertices were all regions (or patches) plus the pseudo-background B, *i.e.*, $\mathcal{V} = \{r_i\} \cup B$. There were two kinds of edges where the internal edges connected all adjacent regions and the boundary edges connected image border regions to the pseudo-background, $\mathcal{E} = \{(r_i, r_j)|r_i \text{ is adjacent to } r_j\} \cup \{(r_i, B)|r_i \text{ is on the image border}\}$. The geodesic distance between r_i and r_k was defined as the accumulated edge weights along the shortest path from r_i to r_k,

$$\begin{aligned} d_{geo}(r_i, r_k) = \min_{r_{i_1}=r_i, \cdots, r_{i_n}=r_k} \sum_{j=1}^{n-1} D_r(r_{i_j}, r_{i_{j+1}}). \\ s.t. \quad (r_{i_j}, r_{i_{j+1}}) \in \mathcal{E} \end{aligned} \tag{19}$$

The geodesic saliency of r_i was the geodesic distance from r_i to the pseudo-background node B on the graph $\mathcal{G}$,

$$s(r_i) = d_{geo}(r_i, B). \tag{20}$$

In [97], a two-stage saliency computation framework was proposed based on the manifold ranking on an undirected weighted graph as well, where the vertices were superpixels. In the first stage, the regional saliency scores were computed based on the relevances given each side of the pseudo-background queries. In the second stage, the saliency scores were refined based on the relevances given the foreground which was segmented using an adaptive threshold on the saliency scores obtained from the first stage.

In [101], saliency computation was formulated as the dense and sparse reconstruction errors w.r.t. the background templates. Similar to [92], the input image was described by a feature matrix $\mathbf{F}$. The background templates, defined as the superpixels touching the image border, was described as $\mathbf{B} = [\mathbf{b}_1, \mathbf{b}_2, \cdots, \mathbf{b}_M] \in \mathcal{R}^{D \times M}$, where M is the number of border superpixels. Given the Principal Component Analysis (PCA) basis $\mathbf{U_B} = [\mathbf{u}_1, \mathbf{u}_2, \cdots, \mathbf{u}_{D'}]$ of the background template $\mathbf{B}$, which were eigenvectors of the normalized covariance matrix of $\mathbf{B}$ corresponded to the largest D' eigenvalues, the *dense* reconstruction error of region r_i is computed as

$$\varepsilon_i^d = ||\mathbf{f}_i - (\mathbf{U_B}\beta_i + \bar{\mathbf{f}})||^2, \tag{21}$$

$$\beta_i = \mathbf{U_B}^T(\mathbf{f}_i - \bar{\mathbf{f}}), \tag{22}$$

where $\bar{\mathbf{f}}$ was the mean feature of $\mathbf{F}$. For the sparse reconstruction error, each region r_i was encoded with sparse representation based on the background templates $\mathbf{B}$ by

$$\alpha_i = \arg\min_{\alpha_i} ||\mathbf{f}_i - \mathbf{B}\alpha_i||^2 + \gamma||\alpha_i||_1, \tag{23}$$

and the reconstruction error was

$$\varepsilon_i^s = ||\mathbf{f}_i - \mathbf{B}\alpha_i||^2. \tag{24}$$

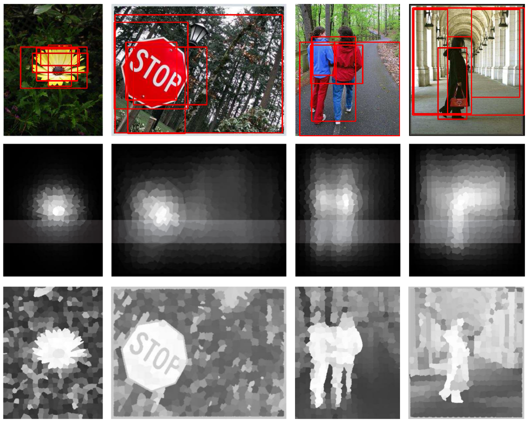


Fig. 12. AIllustration of regional dieverse density score. From top to bottom: input image, pixel-wise objectness prior, and regional dieverse density scores. Image courtsey from [104].

These two types of reconstruction erros will be computed and propagated to pixels on multiple segmentations, which will be fused to form the final saliency map by Bayesian inference.

Jiang *et al.* [102] formulated the saliency detection via absorbing Markov chain where the transient and absorbing nodes were superpixels around image center and border, respectively. The saliency of each superpixel was computed as the absorbed time for the transient node to the absorbing nodes of the Markov chain.

Beyond these approaches, the generic **objectness** prior[5] was also exploited to facilitate the detection of salient objects by leveraging the object proposal generation model [16]. Chang *et al.* [74] presented a computational framework by fusing the objectness and regional saliency into a graphical model. These two terms were jointly estimated by iteratively minimizing the energy function which encoded the mutual interaction between them. During the optimization, objectness can help to improve the estimation of regional saliency, and vice versa. In [103], regional objectness was defined as the average objectness values of its contained pixels. Jia and Han [104] computed the regional saliency by comparing it to the "soft" foreground and background according to the objectness prior. With a set of bounding box object candidates $\{(B_1, b_1), (B_2, b_2), \cdots, (B_J, b_J)\}$ where B_j is the bounding box and b_j is its confidence score [16], pixel-wise objectness score can be defined as

$$o(p) = \left[\sum_{j=1}^{J} b_j^2 \delta(p \in B_j) \exp(-\theta d(p, B_j))\right]^{1/2}, \tag{25}$$

where $d(p, B_j)$ was the normalized distance between the pixel p and B_j. The regional objectness score $o(r_i)$ of r_i is the average objectness scores of its contained pixels. Regional diverse density was then computed as

$$\begin{aligned} &DD(r_i) \\ &= \sum_{j=1}^{N} D_r(r_i, r_j)o(r_j) + (1 - D_r(r_i, r_j))(1 - o(r_j)). \end{aligned} \tag{26}$$

If a region was very distinct from the potential background

[5] Although it was learned from training data, we also tend to treat is as a universal intrinsic cue for salient object detection.

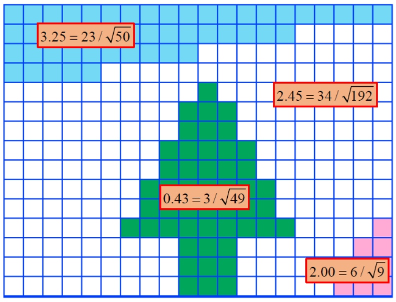


Fig. 13. An illustrative example of the boundary connectivity [109]. The synthetic image consists of four regions with their boundary connectivity values overlaid. The boundary connectivity is large for background regions and small for object regions.

or very similar to the potential foreground, it will be assigned a higher saliency score. Finally, a fully-connected Gaussian Markov Random Field (GMRF) between each pair of regions was constructed to enforce the consistency between salient regions. Fig. 12 provides an illustration of regional diverse density scores.

Salient object detection relying on the pseudo-background assumption sometimes may fail especially when the object touched the image border. Zhu *et al.* [109] further proposed the **boundary connectivity** prior. As shown in Fig. 13, regions from salient object are much less connected to image border than the ones from the background. THESE THREE EQUATIONS CANNOT BE SELF-EXPLAINED HERE. Boundary connectivity of a region r_i was defined as follows:

$$BC(r_i) = \frac{L_{bnd}(r_i)}{\sqrt{Area(r_i)}}, \tag{27}$$

where $Area(r_i)$ is the spanning area of r_i, defined as

$$Area(r_i) = \sum_{j=1}^{N} \exp(-\frac{d_{geo}(r_i, r_j)^2}{2\sigma_{clr}^2}) = \sum_{j=1}^{N} S(r_i, r_j). \tag{28}$$

$L_{bnd}(r_i)$ was the length along the image boundary, defined as

$$L_{bnd}(r_i) = \sum_{j=1}^{N} S(r_i, r_j)\delta(p_j \in B). \tag{29}$$

Such a boundary connectivity score was then integrated into a quadratic objective function to get the final optimized saliency map. It is worth pointing out that similar ideas of boundary connectivity prior were also investigated in [105] as *segmentation prior* and as *surroundness* score in [114].

Focusness prior, reflecting the fact that a salient object is often photographed in focus to attract more attention, was investigated in recent works [103], [115]. Jiang *et al.* [103] defined the focusness from the degree of focal blur. By modeling such de-focus blur as the convolution of a shape image with a point spread function, approximated by a Gaussian kernel, the pixel-level focusness was casted as estimating the standard deviation of the Gaussian kernel by scale space analysis. Regional focusness score can be computed by propagating the focusness and/or sharpness at the boundary and interior edge pixels. The saliency score was defined as a non-linear combination of uniqueness, objectness, and focusness.

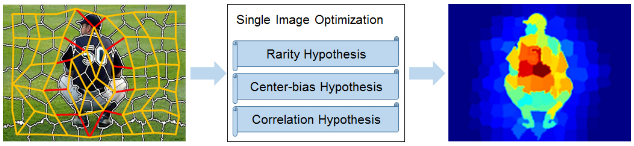


Fig. 14. Regional saliency scores are directly optimized to meet various hypotheses in [94]

Li *et al.* [94] proposed to optimize the saliency values of all superpixels in an image to simultaneously meet several saliency hypotheses on visual rarity, center-bias and mutual correlation. As shown in Fig. 14, the saliency values of all superpixels were treated as the optimization objective on each single image. Given the average Lab colors of all the N superpixels $\{\mathbf{v}_i\}_{i=1}^N$, the correlation w_{ij} between the ith and the jth superpixels can be computed as a kind of visual similarity:

$$w_{ij} = \exp\left(-\frac{\|\mathbf{v}_i - \mathbf{v}_j\|_1}{3}\right). \tag{30}$$

Based on such correlations, the saliency values $\{s_i\}_{i=1}^N$ can be optimized by solving:

$$\begin{aligned} \min_{\{s_i\}_{i=1}^N} & \sum_{i=1}^{N} s_i \sum_{j\neq i}^{N} w_{ij} + \lambda_c \sum_{i=1}^{N} s_i e^{d_i/d_D} \\ & + \lambda_r \sum_{i=1}^{N}\sum_{j\neq i}^{N} (s_i - s_j)^2 w_{ij} e^{-d_{ij}/d_D} \\ s.t. \;\; & 0 \le s_i \le 1, \forall i, \\ & \sum_{i=1}^{N} s_i = 1. \end{aligned} \tag{31}$$

where d_D is half the image diagonal length. d_{ij} and d_i are the distances from the ith superpixel to the jth superpixel and image center, respectively. In the optimization, the saliency value of each superpixel was optimized by quadratic programming when considering the influences of all the other superpixels. Finally, two adaptive thresholds were used to select the most reliable foreground and background according to the saliency values of superpixels. By classifying other superpixels whose saliency values fell between the two thresholds according to their similarity with foreground and background, the whole salient object was segmented as a whole. Similarly, Zhu *et al.* [109] also adopted such optimization-based framework to integrate multiple foreground/background cues as well as the smoothness terms to automatically infer the optimal saliency values.

Performance of salient object detection based on regions might be affected by the segmentation parameters. In addition to other approaches based on multi-scale regions [84], [91], [96], single-scale potential salient regions were extracted by solving the facility location problem in [107]. An input image was first represented as an undirected graph on superpixels, where a much smaller set of candidate region centers were then generated on these superpixels through agglomerative clustering. On this set, a submodular objective function was built to maximize the similarity, which incorporated both low-level cues as well as high-level priors such as face prior, center prior and color prior, between potential salient region centers and all superpixels assigned to that center. Moreover, a penalty term was also added to the optimization objective so as to penalize the number of selected potential salient region centers and to avoid over-segmentation. By applying a greedy algorithm, the objective function can be iteratively optimized to group superpixels into regions, whose saliency was further measured through global regional contrast and spatial distribution.

In [95], a Bayesian framework was proposed for salient object detection, which was formulated as estimating the posterior probability of being foreground at each pixel p given the input image I. Denote $s(p)$ as s_p for short, the posterior probability was computed as

$$\begin{aligned}&\mathbf{P}(s_p = 1|I_p)\\&= \frac{\mathbf{P}(s_p = 1)\mathbf{P}(I_p|s_p = 1)}{\mathbf{P}(s_p = 1)\mathbf{P}(I_p|s_p = 1) + \mathbf{P}(s_p = 0)\mathbf{P}(I_p|s_p = 0)}\end{aligned} \tag{32}$$

where $\mathbf{P}(s_p = 0|I_p) = 1 - \mathbf{P}(s_p = 1|I_p)$. To estimate the saliency prior, a convex hull H was first estimated around the detected interest points. To leverage the regional information, superpixels were grouped into larger regions based on a Laplacian sparse subspace clustering method. Suppose the pixel p belongs to the region r after grouping, the saliency prior $\mathbf{P}(s_p = 1)$ was defined as

$$\mathbf{P}(s_p = 1) = \frac{|r \cap H|}{|r|}, \text{ where } p \in r. \tag{33}$$

The convex hull H, which divided the image I into the inner region R_I and outside region R_O, provided a coarse estimation of foreground as well as background and can be adopted for likelihood computation. With the color representation $[l(p), a(p), b(p)]$ for each pixel p in the CIELab color space, color histograms for R_I and R_O were constructed on each channel. Assuming each channel was independent, the likelihood at pixel p can be computed as

$$\mathbf{P}(I_p|s_p = 1) = \prod_{f\in\{l,a,b\}} \frac{|\{q|q \in R_I, f(q) = f(p)\}|}{|\{q|q \in R_I\}|}, \tag{34}$$

$$\mathbf{P}(I_p|s_p = 0) = \prod_{f\in\{l,a,b\}} \frac{|\{q|q \in R_O, f(q) = f(p)\}|}{|\{q|q \in R_O\}|}. \tag{35}$$

Liu *et al.* [108] adopted an optimization-based framework for detecting salient objects. Similar to [95], a convex hull was roughly estimated from Harris corners and contour points to bipartite an image into pure background and potential foreground. Then saliency seeds were learned from the image, while a guidance map was learned from background regions as well as human prior knowledge. Under the assistance of these cues, a general Linear Elliptic System with Dirichlet boundary was introduced to model the diffusions from seeds to other regions to generate a saliency map.

Among all the models reviewed in this subsection, there are mainly three types of regions adopted for saliency computation. Irregular regions with varying sizes can be generated using the graph-based segmentation algorithm [111], mean-shift algorithm [112], or clustering (quantization). On the other hand, with recent progress on superpixels algorithm, compact regions with comparable sizes were also popular choices using the SLIC algorithm [57], Turbopixel algorithm [113], etc. The main difference of these two types of regions is that whether the influence of region size should be taken into account. Furthermore, soft regions were also considered for saliency analysis, where every pixel maintained a likelihood belonging to each of all the regions instead of only a hard region label (*e.g.*, fitted using the GMM). To further enhance robustness of segmentation, regions can be generated based on multiple segmentations or in a hierarchcial way. Generally, single-scale segmentation is faster while multi-scale segmentation can improve the overall performance.

To measure the saliency of regions, uniqueness, usually in the form of regional contrast, is still the most frequently used feature. In addition, more and more complementary priors for the regional saliency were investigated as well to improve the overall performance, such as backgroundness, objectness, focusness and boundary connectivity. Compared with the block-based saliency models, the extension of these priors is also a main advantage of the region-based saliency models. Furthermore, regions provide more sophisticated cues (*e.g.*, color histogram) to better capture the salient object of a scene in contrast to pixels and patches. Another benefit to define saliency upon region is related to the efficiency. Since the number of regions in an image is far less than the number of pixels, computing saliency at region level can significantly reduce the computational cost while producing full-resolution saliency maps.

Notice that the approaches discussed in this subsection only utilize intrinsic cues. In the next subsection, we will review how to incorporate extrinsic cues to facilitate the detection of salient objects.

2.1.3 Models with Extrinsic Cues

Models in the third subgroup adopt the *extrinsic cues* to assist the detection of salient objects in images and videos. In addition to the visual cues observed from the single input image, the extrinsic cues can be derived from the ground-truth annotations of the training images, similar images, the video sequence, a set of input images containing the common salient objects, depth maps, or light field images. In this section, we will review these models according to the types of extrinsic cues. Fig. 15 lists all the models with extrinsic cues, where each method is highlighted with several pre-defined attributes.

Supervised salient object detection: while machine learning approaches have been widely studied in other areas of computer vision and state-of-the-art performance are achieved, *e.g.*, in the fields of object recognition and image classification, it is somehow surprising that few research interests are attracted to this direction for salient object detection. All of the current works focus on the supervised scenario, *i.e.*, learning a salient object detector given a set of training samples with ground-truth annotations which aims to separate the salient elements from the background elements.

The first step to apply a trained classifier to salient object detection is feature extraction. Each element (*e.g.*, a pixel or a region) in the input image will be represented by a feature vector $\mathbf{x} \in \mathcal{R}^d$, where d is the feature dimension. Such a feature vector is then mapped to a saliency score $y \in \mathcal{R}^+$ based on the learned linear or non-linear mapping function $f : \mathcal{R}^d \to \mathcal{R}^+$.

One can assume the mapping function f is linear, *i.e.*, $y = \mathbf{w}^{\mathrm{T}}\mathbf{x}$, where $\mathbf{w}$ denotes the combination weights of all components in the feature vector. Liu *et al.* [8] proposed to learn the weights with the Conditional Random Field (CRF) model trained on the rectangular annotations of the salient objects. Recently, the large-margin framework was adopted

#	Model	Pub	Year	Cues	Elements	Hypothesis		Integration	Classifier	Code	Bench.
						Uniqueness	Prior	(Optimization)			
1	**LTD** [8]	CVPR	2007	GT	mPI + PA + SRE	L + CS	SD	LI	CRF	NA	
2	**OID** [116]	ECCV	2010	GT	mPI + PA + SRE	L + CS	SD	LI	mixtureSVM	NA	
3	**LGCR** [117]	BMVC	2010	GT	RE	-	P	None	BDT	NA	
4	**DRFI** [118]	CVPR	2013	GT	RE	L	B + P	LI	RF	M + C	✓
5	**LOS** [119]	CVPR	2014	GT	RE	L + G	PRA + B + SD + CP	None	SVM	NA	
6	**HDCT** [120]	CVPR	2014	GT	RE	L + G	SD + P + HD	None	BDT + LS	M	

#	Model	Pub	Year	Cues	Elements	Hypothesis		Integration	Annotation	Code	Bench.
						Uniqueness	Prior	(Optimization)			
7	**VSIT** [121]	ICCV	2009	SI	PI	-	SS	None	yes	NA	
8	**FIEC** [122]	CVPR	2011	SI	PI + PA	AM	-	LI	no	NA	
9	**SA** [123]	CVPR	2013	SI	PI	-	CMP	CRF	yes	NA	
10	**LBI** [124]	CVPR	2013	SI	PA	SP	-	None	no	M + C	

#	Model	Pub	Year	Cues	Elements	Hypothesis		Integration	Type	Code	Bench.
						Uniqueness	Prior	(Optimization)			
11	**LC** [125]	MM	2006	TC	PI + PA	L	-	LI	online	NA	
12	**VA** [126]	ICPR	2008	TC	mPI + PA + SRE	L	CS + SD + MCO	CRF	offline	NA	
13	**SEG** [127]	ECCV	2010	TC	PA + PI	CS	MCO	CRF	offline	M + C	✓
14	**RDC** [128]	CSVT	2013	TC	RE	G	-	None	offline	NA	

#	Model	Pub	Year	Cues	Elements	Hypothesis		Integration	ImgNum	Code	Bench.
						Uniqueness	Prior	(Optimization)			
15	**CSIP** [129]	TIP	2011	SCO	mRE	-	RS	LI	two	M + C	
16	**CO** [130]	CVPR	2011	SCO	PI + PA	G	RP	None	multiple	NA	
17	**CBCO** [131]	TIP	2013	SCO	SRE	G	SD + RP	NL	multiple	NA	
18	**GRS** [132]	Vis.Comp.	2014	SCO	RE	G	GCOP	None	multiple	NA	

#	Model	Pub	Year	Cues	Elements	Hypothesis		Integration	Source	Code	Bench.
						Uniqueness	Prior	(Optimization)			
19	**LS** [133]	CVPR	2012	DP	RE	G	KA	NL	stereo images	NA	
20	**DRM** [134]	BMVC	2013	DP	RE	G	-	SVM	Kinect	NA	
21	**SDLF** [115]	CVPR	2014	LF	mRE	G	F + B + O	NL	Lytro camera	NA	

Fig. 15. Salient object detection models with extrinsic cues grouped by their cues adopted. For cues, {GT = ground-truth annotation, SI = similar images, TC = temporal cues, SCO = saliency co-occurrence, DP = depth, and LF = light field}. For saliency hypothesis, {P = generic properties, PRA = pre-attention cues, HD = discriminativity in high-dimensional feature space, SS = saliency similarity, AM = annomaly, CMP = complement of saliency cues, SP = sampling probability, MCT = motion contrast, MCO = motion coherence, RP = repetiveness, RS = region similarity, GCOP = global color prior, and KA = knowledge assistance.}. Others, {CRF = conditional random field, SVM = support vector machine, BDT = boosted decision tree, RF = random forest, and LS = least-square solver.}.

to learn the weights $\mathbf{w}$ in [119]. Specifically, denote the training samples as $\{(\mathbf{x}_i^{(k)}, y_i^{(k)})\}$, where $\mathbf{x}_i^{(k)}$ is the saliency feature of the i-th superpixel of the k-th training image and $y_i^{(k)}$ is its saliency score. The optimal weights can be learned as:

$$\mathbf{w}^* = \arg\min_{\mathbf{w}} \frac{1}{2}\alpha||\mathbf{w}||^2 + \sum_k \sum_{\{ij|y_i^{(k)}=1, y_i^{(k)}=0\}} \max\left(0, 1 - (y_i^{(k)}\mathbf{w}^{\mathrm{T}}\mathbf{x}_i^{(k)} - y_j^{(k)}\mathbf{w}^{\mathrm{T}}\mathbf{x}_j^{(k)})\right), \tag{36}$$

where α is a manually set parameter.

Due to the highly non-linear essence of the saliency mechanism, the linear mapping might not perfectly capture the characteristics of saliency. To this end, such a linear integration was extended in [116], where a mixture of linear Support Vector Machines (SVM) was adopted to partition the feature space into a set of sub-regions that were linearly separable using a divide-and-conquer strategy. In each region, a linear SVM, the mixture weights, and the combination parameters of the saliency features were learned for better saliency estimation. Alternatively, other non-linear classifiers were also utilized, the boosted decision trees (BDT) [117], [120] and the random forest (RF) [118].

Generally speaking, supervised approaches allow richer representations for the elements compared with the heuristic methods. In the pioneer work of the supervised salient object detection, Liu *et al.* [8] proposed a set of features including the local multi-scale contrast, regional center-surround histogram distance, and global color spatial distribution. Similar to the models with only intrinsic cues, the region-based representation for salient object detection becomes increasingly popular as more sophisticated descriptors can be extracted in region level. Mehrani and Veksler [117] demonstrated promising results by considering standard regional generic properties, *e.g.*, color and shape, which are widely used in other applications like image classification. Jiang *et al.* [118] proposed a regional saliency descriptor including the regional local contrast, regional backgroundness, and regional generic properties. In [119], [120], each region was described by a set of features such as local and global contrast, backgroundness, spatial distribution, and the center position. The pre-attentive features were also considered in [119].

Usually, the richer representations result in feature vectors with higher dimensions, *e.g.*, $d = 93$ in [118] and $d = 75$ in [120]. With the availability of a large collections of training samples, the learned classifier is capable of automatically integrating such richer features and pick up the most discriminative ones. Therefore, better performance can be achieved compared with the heuristic methods.

Salient object detection with similar images: with the availability of increasingly larger amount of visual content on the web, salient object detection by leveraging the visually similar images to the input image was studied in recent years. Generally, given the input image I, K similar images $\mathcal{C}_I = \{I_k\}_{k=1}^K \cup I$ are first retrieved from a large collection of images $\mathcal{C}$. The salient object detection on the input I can be assisted by examining these similar images.

In existing studies, there often exist various kinds of assumptions on these similar images. For example, some approaches assumed that the saliency annotation of $\mathcal{C}$ was available. Specifically, Marchesotti *et al.* [121] proposed to describe each indexed image I_k by a pair of descriptors $(\mathbf{f}^+_{I_k}, \mathbf{f}^-_{I_k})$, where $\mathbf{f}^+_{I_k}$ and $\mathbf{f}^-_{I_k}$ denoted the feature descriptors (fisher vector) of the salient and non-salient regions according to the saliency annotations, respectively. To compute the saliency map, the input image is represented as a set of patches $\{a_i\}_{i=1}^P$ and each patch a_i is described by a fisher vector $\mathbf{f}_i$. The saliency of the region r, defined as the neighbors of a_i, is then computed as: sometimes r is used to represent region, while R is used in some other cases. I just changed R to r in this equation. Perhaps we should check the consistency of symbols throughout the paper.

$$s(r) = ||\mathbf{f}_r - \mathbf{f}_{BG}||_1 - ||\mathbf{f}_R - \mathbf{f}_{FG}||_1, \tag{37}$$

where $\mathbf{f}_r = \sum_{a_i \in r} \mathbf{f}_i, \mathbf{f}_{FG} = \sum_{k=1}^K \mathbf{f}^+_{I_k}, \mathbf{f}_{BG} = \sum_{k=1}^K \mathbf{f}^-_{I_k}$. And $||\cdot||_1$ denotes the L_1 norm. Finally, the saliency of the region r was propagated to its contained pixels,

$$s(p) = \frac{\sum_r w_r \cdot s(r)}{\sum_r w_r}, \tag{38}$$

where w_r is Gaussian weight measuring the spatial distance of the pixel p to the geometrical center of the region r.

Alternatively, based on the observation that different features contributed differently to the saliency analysis on each individual image, Mai *et al.* [123] proposed to learn the image specific rather than the universal weights to fuse the saliency maps that are computed on different feature channels. To this end, the retrieved similar images along with their saliency annotations are adopted to learn the fusion weights of the input image in a CRF framework to account for the dependence of aggregation on individual images[6].

Similar image retrieval works well only on the large scale image collections. Saliency annotation is time consuming, tedious, and even intractable on such collections, however. To this end, some methods tried to leverage the *unannotated* similar images. With the web-scale image collection $\mathcal{C}$, Wang *et al.* [122] proposed a simple yet effective saliency estimation algorithm. The pixel-wise saliency map is computed as:

$$s(p) = \sum_{k=1}^K ||I(p) - \tilde{I}_k(p)||_2, \tag{39}$$

where $\tilde{I}_k$ is the geometrically warpped version of I_k with the reference I and $||\cdot||_2$ is the L_2 norm. The main insight is that similar images offer good approximations to the background regions while the salient regions might not be well approximated.

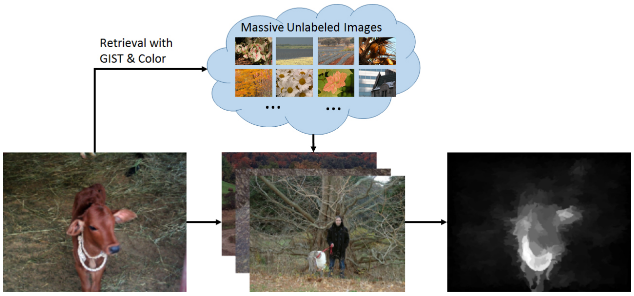


Fig. 16. Images with similar color and spatial layouts can serve as extrinsic cues for salient object detection in [124]

Siva *et al.* [124] proposed a probabilistic formulation for saliency computation as a sampling problem. A patch a is considered to be salient if it has the low probability of being sampled from the images $\mathcal{C}_I$ which are similar to the input image I. As shown in Fig. 16, such similar images can be retrieved from a massive unlabeled image database by using color and GIST features [135]. The saliency of a is then defined as:

$$s(a) = 1 - p_a, \tag{40}$$

where p_a is a number proportional to the probability of sampling patch a from $\mathcal{C}_I$.

It was assumed that the probability of sampling a patch a from an image $J \in \mathcal{C}_I$ can be formulated by uniformly selecting a patch b in J and then perturbing it by some noise in an informative feature space. We can get:

$$\begin{aligned} p_a &\propto \mathbf{P}(A = a|\mathcal{C}_I) \\ &= \int_{\mathcal{C}_I} \mathbf{P}(A = a|J) \ \mathrm{d}J \\ &= \int_{\mathcal{C}_I} \int_J \mathbf{P}(A = a|b)\mathbf{P}(b|J) \ \mathrm{d}b \ \mathrm{d}J \\ &\propto \int_{\mathcal{C}_I} \int_J \mathbf{P}(A = a|b) \ \mathrm{d}b \ \mathrm{d}J \end{aligned} \tag{41}$$

Assuming the noise is uniform and Gaussian over the sapce of image patches, p_a can be further decomposed as:

$$p_a \propto \int_{\mathcal{C}_I} \int_J \exp(-\frac{d(a,b)^2}{\sigma^2}) \ \mathrm{d}b \ \mathrm{d}J \tag{42}$$

By noting that the Gaussian distribution is *short-tailed*, such integration can be approximated for efficiency purpose:

$$\begin{aligned} p_a \approx & \sum_{b \in \mathcal{N}_m(a, \mathcal{C}_I \backslash \{I\})} \exp(-\frac{d(a,b)^2}{\sigma^2}) + \\ & \sum_{b \in \mathcal{N}_m(a, \{I\})} \exp(-\frac{d_I(a,b)^2}{\sigma^2}) \end{aligned} \tag{43}$$

where $\mathcal{N}_m(a, \mathcal{C}_I \backslash \{I\})$ are the m approximated nearest neighbors of patch a, which are taken from all images in $\mathcal{C}_I$ except I.

Co-salient object detection: instead of concentrating on computing saliency on a single image, co-salient object detection algorithms (or namely, co-saliency detection) focus on discovering the *common* salient objects shared by multiple images $\{I^i\}_{i=1}^M$. That is, such objects can be the same object with different view points or the objects of the same category sharing similar visual appearances. Note that the

[6] We will further discuss more technical details about [123] in Sect. 2.1.4.

key characteristic of co-salient object detection algorithms is that their input is *a set* of images, while classical salient object detection models only need *one* image.

Co-saliency detection is closely related to the concept of image co-segmentation that aims to segment similar objects from multiple images [136], [137]. As stated in [131], three major differences exist between co-saliency and co-segmentation. First, co-saliency detection algorithms only focus on detecting the common salient objects while the similar but non-salient background might be also segmented out in co-segmentation approaches [138], [139]. Second, some co-segmentation methods, *e.g.*, [137], need user inputs to guide the segmentation process under ambiguous situations. Third, salient object detection always serves as a pre-processing step, and thus more efficient algorithms are also preferred than the co-segmentation algorithms, especially, on a large number of images.

Too many technical details for the algorithms in co-saliency!! Perhaps we can introduce one representative algorithm with equations and leave the others described with only text.

Li and Ngan [129] proposed a co-saliency detection method from an image pair $\{I^1, I^2\}$ that contain some objects in common. The co-saliency was defined as the inter-image correspondence, *i.e.*, low saliency values should be given to the dissimilar regions. According to this principle, the co-saliency of the element (*e.g.*, pixel or region) $e_1 \in I^1$ is defined as:

$$s_m^1(e_1) = \max_{e_2 \in I^2} sim\left(e_1, e_2\right). \tag{44}$$

To compute $sim(e_1, e_2)$, which measures the similarity between two elements e_1 and e_2, a co-multilayer graph is constructed by dividing I^1 and I^2 into spatial pyramid representations by using hierarchical image segmentation. Each node corresponds to a region that was described by color and texture descriptors. $sim(e_1, e_2)$ is then computed using the normalized single-pair SimRank algorithm.

Traditional salient object detection methods on a single image always consider the *distinctiveness* property to compute saliency. In [130], Chang *et al.* proposed to compute the co-saliency by exploiting the additional *repeatedness* property across multiple images. The co-saliency for the pixel $p \in I^i$ is defined as:

$$s_m^i(p) = s^i(p) w^i(p), \tag{45}$$

where $s^i(p)$ is the output from any single-image saliency detection algorithm and $w^i(p)$ measures the likelihood of repeatedness of p over $\{I^k\}_{k \neq i}$.

Fu *et al.* [131] proposed a cluster-based co-saliency detection algorithm by exploiting the global contrast, spatial distribution, and corresponding cues over multiple images. K clusters $\{C^k\}_{k=1}^K$ were first obtained from the input images $\{I^i\}_{i=1}^M$, where each cluster C^k is associated with a cluster center μ^k. Furthermore, pixels in the image I^j are denoted as $\{p_i^j\}_{i=1}^{N_j}$, where each pixel p_i^j is associated with its normalized location z_i^j. The global contrast for the cluster C^k is defined as:

$$w_c(C^k) = \sum\nolimits_{i=1, i \neq k}^{K} \frac{n^i}{N} ||\mu^k - \mu^i||_2, \tag{46}$$

where n^i represented the pixel number of the cluster C^i and N was the total pixel number of all images. The spatial distribution is defined as:

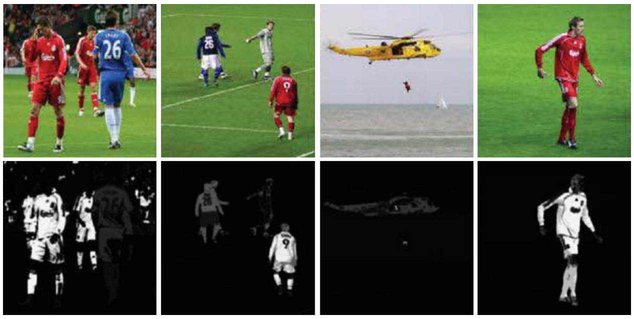


Fig. 17. The co-saliency detection approaches propose to pop-out the salient object shared by multiple images (saliency maps are produced by in [131])

$$w_s(C^k) = \frac{1}{n^k} \sum_{j=1}^{M} \sum_{i=1}^{N_j} G\left(||z_i^j - o^j||_2, \sigma^2\right) \delta(p_i^j \in C^k), \tag{47}$$

where o^j is the center of the image I^i. $G(\cdot, \sigma^2)$ is the zero-mean Gaussian function with σ^2 being the normalized radius of images. To obtain the saliency with the corresponding cue, a M-bin histogram $\hat{\mathbf{q}}^k = \{\hat{q}_j^k\}_{j=1}^M$ was adopted to describe the distribution of the cluster C^k in M images:

$$\hat{q}_j^k = \frac{1}{n^k} \sum\nolimits_{i=1}^{N_j} \delta(p_i^j \in C^k), j = 1, \cdots, M. \tag{48}$$

The corresponding cue is then defined as:

$$w_d(C^k) = \frac{1}{1 + \text{var}(\hat{\mathbf{q}}^k)}, \tag{49}$$

where $\text{var}(\hat{\mathbf{q}}^k)$ measures the variance of the histogram $\hat{\mathbf{q}}^k$. Finally, the co-saliency of each cluster is obtained as:

$$s_m(C^k) = w_c(C^k) w_s(C^k) w_d(C^k). \tag{50}$$

Salient object detection on videos: in addition to the spatial information in a single image, video sequence provides the temporal cue, *e.g.*, motion to facilitate salient object detection. Zhai and Shah [125] first estimated the keypoint correspondences between two consecutive frames. Denote $\mathbf{p}_i$ as the i-th keypoint and $\mathbf{p}'_i$ its corresponding keypoint in the consecutive frame, its saliency is defined as:

$$s_t(\mathbf{p}_i) = \sum\nolimits_{j=1}^{n} DistT(\mathbf{p}_i, \mathbf{p}_j), \tag{51}$$

where n is the number of correspondences. $DistT(\mathbf{p}_i, \mathbf{p}_j)$ captures the motion contrast between $\mathbf{p}_i$ and $\mathbf{p}_j$. For simplicity, the motion model was assumed to be homography. Specifically, M homographies $\{\mathbf{H}_m\}_{m=1}^M$ were estimated by RANSAC, which model different motion segments in the scene. For each homography $\mathbf{H}_m$, a set of points $\mathcal{L}_m = \{\mathbf{p}_1^m, \cdots, \mathbf{p}_{n_m}^m\}$ were considered as its inliers, with n_m being the number of inliers for $\mathbf{H}_m$. The motion contrast can be defined as:

$$DistT(\mathbf{p}_i, \mathbf{p}_j) = \epsilon(\mathbf{p}_i, \mathbf{H}_m), \tag{52}$$

where $\mathbf{p}_j \in \mathcal{L}_m$. $\epsilon(\mathbf{p}_i, \mathbf{H}_m) = |\mathbf{p}'_i - \hat{\mathbf{p}}'_i|$ measures the projection error of $\mathbf{p}_i$ given $\mathbf{H}_m$, where $\hat{\mathbf{p}}'_i$ is the projected keypoint of $\mathbf{p}_i$. Finally, the saliency for the target keypoint

$\mathbf{p}$ is defined as:

$$s_t(\mathbf{p}) = \sum_{j=1}^{M} \alpha_j \epsilon(\mathbf{p}, \mathbf{H}_j). \tag{53}$$

$\alpha_j \in [0, 1]$ is the normalized spanning area of $\mathbf{H}_j$ introduced to suppress the background regions which have larger areas but less keypoints. The average saliency value of all the inliers of each homography was then assigned to its corresponding spanning region to get the final saliency map.

Liu *et al.* [126] extended their original spatial saliency features [8] to the motion field resulting from the optical flow algorithm. With the colorized motion field as the input image, the local multi-scale contrast, regional center-surround distance, and global spatial distribution are computed and finally integrated in a linear way. Rahtu *et al.* [127] integrated the spatial saliency into the energy minimization framework, where the temporal coherence constraint is considered.

Li *et al.* [128] extended the regional contrast-based saliency to the spatiotemporal domain. Given the over-segmentation of the frames of the video sequence, spatial and temporal region matchings between each two consecutive frames in a frame were estimated based on their color, texture, and motion features in a interactive manner on an undirected un-weighted matching graph. The regional saliency was determined by computing its local contrast to the surrounding regions not only in the present frame but also in the temporal domain.

Salient object detection with depth: human beings live in real 3D environments, where stereoscopic contents provide additional depth cues for understanding the surroundings and play an important role in visual attention. This is further validated by Lang *et al.* [140] through experimental analysis of the importance of depth cues for eye fixation prediction. Recently, researchers started to study how to exploit the depth cues for salient object detection [133], [134], which might be captured from either the stereo images indirectly or the depth camera (*e.g.*, Kinect) directly.

The most straightforward extension is to adopt the widely used hypotheses introduced in sections 2.1.1 and 2.1.2 to the depth channel, *e.g.*, the global contrast on the depth map [133], [134]. Furthermore, Niu *et al.* [133] demonstrated how to leverage the domain knowledge in stereoscopic photography to compute the saliency map. The input image is first segmented into regions $\{r_i\}$. In practice, the content of interest is often given small or zero disparities to minimize the vergence-accommodation conflict. Therefore, the first type of regional saliency based on the disparity was defined as:

$$s_{d,1}(r_i) = \begin{cases} \frac{d_{max} - \bar{d}_i}{d_{max}} & \text{if } \bar{d}_i \geq 0 \\ \frac{d_{min} - \bar{d}_i}{d_{min}} & \text{if } \bar{d}_i < 0 \end{cases} \tag{54}$$

where d_{max} and d_{min} are the maximal and minimal disparities. $\bar{d}_i$ is the average disparity in region r_i.

Additionally, objects with negative disparities were perceived popping out from the scene and attracted more attention. The second type of regional stereo saliency is then defined as:

$$s_{d,2}(r_i) = \frac{d_{max} - \bar{d}_i}{d_{max} - d_{min}}. \tag{55}$$

The stereo saliency can be computed by linearly combining these two types of saliency:

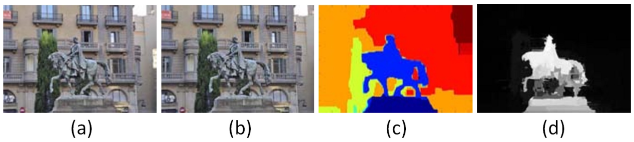


Fig. 18. Depth (or disparity) map serves as an additional extrinsic cue for detecting salient objects in 3D scenes [133]. (a) left image, (b) right image, (c) disparity map, and (d) stereo saliency map.

$$s_d(r_i) = (1 - \lambda)s_{d,1}(r_i) + \lambda s_{d,2}(r_i) \tag{56}$$

$$\lambda = \gamma + \frac{n_{d_p<0}}{n}(1 - \gamma) \tag{57}$$

where $n_{d_p<0}$ is the number of pixels with negative disparities and n is the total number of pixels. γ is a parameter with default value of 0.5.

Salient object detection on light field: recently the light field for salient object detection was proposed in [115]. A light field, which can be captured using the specifically designed camera, *e.g.*, Lytro, can be essentially viewed as an array of images captured by a grid of cameras towards the scene. The light field data offer two benefits for salient object detection. On one hand, the light field allows synthesizing a stack of images focusing at different depths. On the other hand, a light field provides an approximation to scene depth and occlusions.

With these additional information, Li *et al.* [115] first utilized the focusness and objectness priors to robustly choose the background and select the foreground candidates. Specifically, the layer with the estimated background likelihood score was used to estimate the background regions. Regions, coming from Mean-shift algorithm, with the high foreground likelihood score were chosen as salient object candidates. Finally, the estimated background and foreground were utilized to compute the contrast-based saliency map on the all-focus image.

2.1.4 Other Models

In previous sections, we mainly reviewed models which focus on computing saliency map for the input image, which might be useful for some applications like image retargeting. There exist some algorithms which aim at directly segmenting or localizing salient objects with bounding boxes, or whose main research effort is not on the saliency map computation.

Localization models: Feng *et al.* [141] defined saliency for each sliding window as its composition cost using the remaining parts of the image. The local maxima, which were found among all the sliding windows in a brute-force manner, were believed to be corresponding to salient objects. The input image is first segmented into a set of superpixels. Given an image window W, inside segments $\{s_i^n\}$ are those superpixels whose areas within the window are larger than areas outside. Outside segments are those whose outside areas are greater than areas inside. The saliency for W is defined as the composition cost of inside segments $\{s_i^n\}$ using outside segments $\{s_o^n\}$. Specifically, for two segments (regions) s_1 and s_2, their composition cost is defined as:

$$c(s_1, s_2) = [1 - d_s(s_1, s_2)] \cdot d_a(s_1, s_2) + d_s(s_1, s_2) \cdot d_a^{max}, \tag{58}$$

where $d_a(s_1, s_2)$ is the intersection distance between the Lab color histograms of s_1 and s_2. Their spatial distance $d_s(s_1, s_2)$ is Hausdorff distance normalized by the longer image side length. d_a^{max} is the largest appearance distance within the image. The composition cost between $\{s_i^n\}$ and $\{s_o^n\}$ can be computed in a greedy manner, leveraging the fast pre-computation and incremental updating.

Unlike previous approaches, which usually assume that at least one salient object exist in the input image, Wang *et al.* [142] studied the problem of detecting the *existence* and the location of salient objects on thumbnail images as some background images contain no salient objects at all. Each image is described by a set of saliency features extracted on multiple channels. The existence of salient objects is formulated as a binary classification problem. For localization, a regression function was learned using Random Forest on training samples to directly output the position of the salient object.

Segmentation models: segmenting salient objects is closely related to the figure-ground problem, which is essentially a binary classification problem by separating the salient object from the background. Lu *et al.* [90] exploited the convexity (concavity) prior for salient object segmentation, which assumed that the region on the convex side of a curved boundary tends to belong to the foreground. Based on this assumption, concave arcs were first found on the contours of superpixels. For a concave arc, its convexity context was defined as the windows which is tightly around the arc. An undirected weight graph was then built over the superpixels with concave arcs where the weights between vertices were determined by the summation of concavity context on different scales in the hierarchical segmentation of the image. Finally the Normalized Cut algorithm [143] was performed to separate the salient object from the background.

In order to more effectively leverage the contextual cues, Wang *et al.* [144] proposed to integrate an auto-context classifier [145] into an iterative energy minimization framework to automatically segment the salient object. The auto-context model is a multi-layer Boosting classifier on each pixel and its surroundings to predict if it is associated with the target concept. The subsequent layer is built on the classification of the previous layer. Hence through the layered learning process, the spatial context is automatically incorporated for more accurate segmentation of the salient object.

Aggregation models: to leverage the output saliency maps of existing salient object detection algorithms, some models focus on aggregating them more effectively. Given M saliency maps $\{S_i\}_{i=1}^M$ which may come from different salient object detection models or are computed on hierarchical segmentations of the input image, aggregation models try to integrate them to form a more accurate saliency map to facilitate the detection of salient objects.

Denote $S_i(p)$ as the saliency value at pixel p of the i-th saliency map. In [1], Borji *et al.* proposed a standard saliency aggregation method as follows:

$$S(p) = P(y_p = 1|\mathbf{x}(p)) \propto \frac{1}{Z} \sum\nolimits_{i=1}^{m} \zeta(S_i(p)) \tag{59}$$

where $\mathbf{x}(p) = (S_1(p), S_2(p), \ldots, S_m(p))$ is the saliency scores for pixel p and $y_p = 1$ indicates p is labeled as salient. $\zeta(\cdot)$ is a real-valued function which can take the following forms:

$$\zeta_1(x) = x; \ \zeta_2(x) = \exp(x); \ \zeta_3(x) = -\frac{1}{\log(x)}. \tag{60}$$

Inspired by the aggregation model in [1], Mai *et al.* [123] proposed two aggregation solutions. The first solution also adopted the pixel-wise aggregation:

$$P(y_p = 1|\mathbf{x}(p); \lambda) = \sigma\left(\sum_{i=1}^{m} \lambda_i S_i(p) + \lambda_{m+1}\right) \tag{61}$$

where $\lambda = \{\lambda_i | i = 1 \ldots m{+}1\}$ is the set of model parameters and $\sigma(z) = 1/(1 + \exp(-z))$. However, they proposed that one potential problem of such direct aggregation is its ignorance of the interaction between neighboring pixels. Inspired by [9], they proposed the second solution by using the CRF in aggregating saliency analysis results from multiple methods to capture the relation between neighboring pixels. The parameters of the CRF aggregation model were optimized on the training data and the saliency aggregation result for each pixel was the posterior probability of being labeled as salient with the trained CRF.

In [146], Tian *et al.* extended the segmentation framework in [87] by learning the complementary saliency maps for salient object segmentation. Given the saliency maps computed by eight existing saliency methods, two aggregation functions, including an additive function and a multiplicative function, were learned to generate two kinds of saliency maps, one with high recall (i.e., the "envelope" map highlighting a large area containing the objects) and one with high precision (i.e., the "sketch" map highlighting small areas inside each salient object). Given such envelope map and sketch map that encoded the foreground/background priors in a given image, the foreground (salient) objects were then segmented by using graph-cuts for image pixel labeling.

Alternatively, Yan *et al.* [96] integrate the saliency maps computed on multi-layer segmentation of the image into a hierarchical tree-structure graphical model, where each node corresponds to a region in every layer. Specifically, for a node corresponding to region i in layer $\mathcal{L}^l$, its saliency variable is denoted as s_i^l. Set $\mathcal{S}$ contains all of them. Saliency aggreation is to optimize the following energy function:

$$E(\mathcal{S}) = \sum_l \sum_i E_D(s_i^l) + \sum_l \sum_{i, R_i^l \subseteq R_j^{l+1}} E_S(s_i^l, s_j^{l+1}). \tag{62}$$

The energy contains two parts. Data term $E_D(s_i^l)$ is to gather separate saliency confidence, defined as:

$$E_D(s_i^l) = \beta^l ||s_i^l - \bar{s}_i^l||_2^2, \tag{63}$$

where β^l controls the layer confidence and $\bar{s}_i^l$ is the initial saliency value. The hierarchy smoothness term $E_S(s_i^l, s_j^{l+1})$ enforces consistency between corresponding regions in two consecutive layers, defined as:

$$E_S(s_i^l, s_j^{l+1}) = \lambda^l ||s_i^l - s_j^{l+1}||_2^2, \tag{64}$$

where λ^l controls the strength of consistency between layers. Thanks to the tree structure, such an energy function can efficiently minimize using the belief propagation. In fact, solving the three layer hierarchical model is equivalent to applying a weighted average to all single-layer saliency

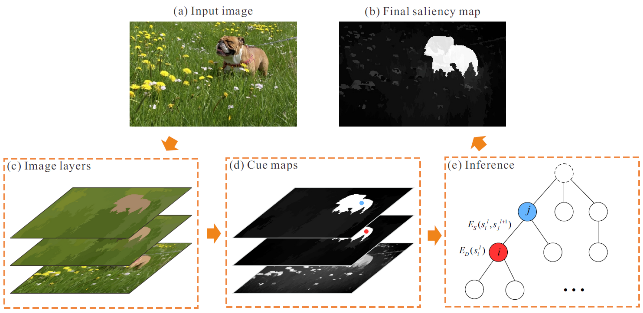


Fig. 19. Hierarchical salient object detection model by Yan et al. Image courtesy from [96].

maps. Different from naive multi-layer fusion, this hierarchical inference algorithm can select optimal weights for each region instead of global weighting.

Active models: Inspired by the interactive segmentation models (see subsection 2.2.2) and to tackle the bias of salient object datasets, a new trend has emerged recently by explicitly decoupling two stages of saliency detection mentioned in subsection 1.1: detecting the most salient object and segmenting it. Instead, some studies propose to perform active segmentation which utilize the advantage of both fixation prediction and segmentation models. As a representative model of active segmentation, Mishra *et al.* [52] combined multiple cues (e.g., color, intensity, texture, stereo and/or motion) to predict fixations. As a result, the "optimal" closed contour for salient object around the fixation point was segmented in polar space. Li *et al.* [53] proposed a salient object segmentation model containing two components: a *segmenter* that proposes candidate regions, and a *selector* that gives each region a saliency score with fixation models. Similarly, Borji [54] proposed to first roughly locate the salient object at the peak of the fixation map (or its estimation using a fixation prediction model) and then segment the object using superpixels. The last two algorithms used annotations to determine the upper-bound of the segmentation performance, proposed datasets with multiple objects in scenes, and provided new insight to the inherent connections of fixation prediction and salient object segmentation.

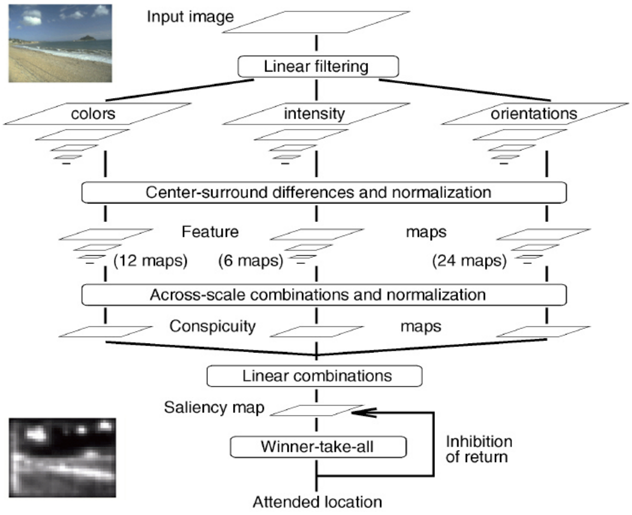


Fig. 20. Itti's model [2] and its output for a sample image.

2.2 Models in Closely Related areas

2.2.1 Fixation Prediction Models

Reviewing all fixation prediction models goes beyond the scope of this paper (See [47], [86], [147], [148] for reviews and benchmarks of these models). Here we give pointers to the most important trends and works in this domain. Inclusion of these models here is to measure their performance versus salient object detection models.

Over the years, a large body of fixation prediction models have been proposed, most of which base themselves on low-level image features like color, intensity and orientation. Such models basically analyze visual uniqueness, unpredictability, rarity, or surprise of a region, and is often attributed to variations in image attributes like color, gradient, edges, and boundaries (e.g., the most famous model proposed by Itti *et al.* [2]; see Fig. 20). As opposed to salient object detection models, these models often produce higher saliency values near edges instead of uniformly highlighting salient objects which is not good as some studies have claimed that people look at the center of objects [149].

Fixation prediction models have been used to predict where people look during free-viewing of static natural scenes (e.g., [18], [20], [21], [150] or dynamic scenes/vidoes (e.g., [151]–[153]) by employing motion, flicker, optical flow (e.g., [154], [155]), or spatiotemporal interest points learned from image regions at fixated locations (e.g., [156] used for action recognition). It is believed that at early stages of free viewing (first few hundred milliseconds), mainly image-based conspicuities guide attention and later on, high-level factors (e.g., actions and events in scenes) direct eye movements [6], [157]. These high-level factors may not necessarily translate to bottom-up saliency (e.g., based on color, intensity, or orientation). For instance, a human's head may not stand out from the rest of the scene but may attract attention. Thus, combining high-level concepts and low-level features have been used to scale up current models and reach the human performance.

Some top-down factors in free-viewing are already known although active investigation still continues to discover and explain more semantic factors and reduce the semantic gap between models and humans. For instance, Einhäuser *et al.* [158] proposed that objects are better predictors of fixations than bottom-up saliency. Although we have shown that this study has some shortcomings [159], there is evidence that object might be important in guiding attention and fixations [149], [160]–[162]. Cerf *et al.* [163] showed that faces and text attract human gaze. Li *et al.* [164] proposed that human fixation can be better predicted by incorporating the prior knowledge learned from millions of unlabeled images. Subramanian *et al.* [165], by recording eye fixations over a large affective image dataset, observed that fixations are directed toward emotional and action stimuli and duration of fixations are longer on such stimuli. Similarly, Judd *et al.* [18], by plotting image regions at the top salient locations of the human saliency map (made of fixations), observed that humans, faces, cars, text, and animals attract human gaze probably because they convey more information in a scene. Castelhano [166] and Borji et al. [161] have shown that gaze direction guides eye movements in free-viewing of natural scenes. Alongside, some personal characteristics

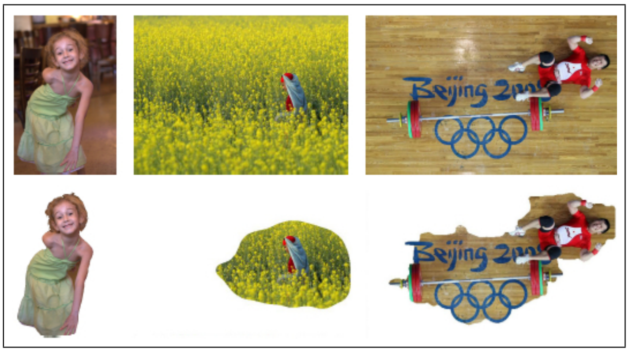


Fig. 21. Saliency concept according to CA model [49].

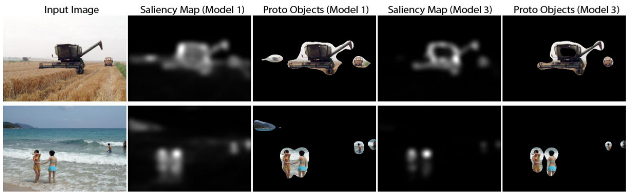


Fig. 22. Salient proto objects from COV model [63].

such as experience, age, gender, and culture change the way humans look at images (e.g., [167], [168]). Some models are can handle both fixation prediction and salient object detection (e.g., BMS [114]). Some authors have thresholded their saliency maps (e.g., top 20% of activation) to detect proto-objects (e.g., CA [63] and COV [63] models; see Figs. 21 and 22, respectively).

Fig. 23 provides a list of fixation prediction models considered in this study. All of these models are based on pure low-level mechanisms and have shown to be very efficient in previous fixation prediction benchmarks [147], [148].

2.2.2 Image Segmentation Models

Segmentation is a fundamental problem studied in computer vision and usually adopted as a pre-process step to image analysis. Without any prior knowledge of the content of the scene, the task of segmentation is to partition an image into perceptually coherent regions. Many algorithms have been proposed in last several decades. Typical approaches are graph-based, where pixels of the image are connected to form a weighted graph. This graph is partitioned into components by minimizing a cost function of the inter and/or intra components, e.g., Minimum Cut [172], Normalized Cut [143], and Ratio Cut [173]. A representative graph-based image segmentation algorithm is [111], which adopts a local optimization criteria and conducts a bottom-up strategy to heuristically aggregate data points into more compact clusters. In addition to graph-based approaches, there also exist other methods. Among these, Mean-Shift [56] is a non-parametric clustering algorithm, which iteratively seeks the mode in the feature space by finding the local maxima of a density function. Pixels that converge to the same mode belong to the same region. Recently, hierarchical segmentations are generated based on the gPb edge detector [14] by converting the ultrametric contour map (UCM) into a hierarchical region tree using oriented watershed transformation (OWT).

#	Model	Ref.	Pub	Year	Code	Bench.
1	**IT**	Itti *et al.* [2]	PAMI	1998	M	✓
2	**AIM**	Bruce & Tsotsos [7]	JOV	2006	M	✓
3	**GB**	Harel *et al.* [169]	NIPS	2007	M + C	✓
4	**SR**	Hou & Zhang [19]	CVPR	2007	M	✓
5	**SUN**	Zhang *et al.* [170]	JOV	2008	M	✓
6	**SeR**	Seo & Milanfar [62]	JOV	2009	M	✓
7	**SIM**	Murray *et al.* [171]	CVPR	2011	M	✓
8	**SS**	Hou *et al.* [61]	PAMI	2012	M	✓
9	**COV**	Erdem & Erdem [63]	JOV	2013	M	✓
10	**BMS**	Zhang *et al.* [114]	ICCV	2013	M + C	✓

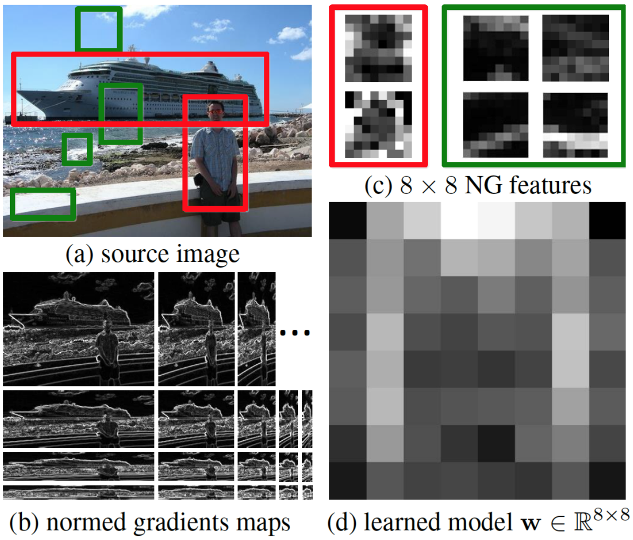
Fig. 23. Fixation prediction models. JOV = Journal of Vision.

Since visual cues utilized in these algorithms, such as intensity, color, and texture, are usually low-level, none of the current segmentation algorithms can produce reliable partitioning of a natural image. Therefore, parameters of these algorithms are tuned to generate an over-segmentation of the image, forming a set of *superpixels*. The local boundaries of the image will be well preserved by the superpixels, which gives the chance of other algorithms to group these superpixels into larger valid regions by considering more powerful cues. To more efficiently generate reliable superpixels, several algorithms are proposed recently. Quick-Shift [174], as a variant of the Mean-Shift algorithm, simply moves each point in the feature space to the nearest neighbor to seek the mode where an increment of the density can be achieved. SLIC (Simple Linear Iterative Clustering) algorithm [57] generates superpixels efficiently by clustering the pixels in terms of color and spatial distance using the k-means algorithm. Turbo pixels [113], a geometric-flow based algorithm, iteratively refines the boundaries between regions starting from the initial seeds. In addition to the superpixels, high-level tasks can also benefit from the multiple segmentations by segmenting the image with different parameters or increasingly grouping superpixels. For example, various levels of spatial support resulting from multiple segmentations are adopted for surface layout prediction [175] and object detection [176].

Interactive segmentation algorithms are also developed in recent years, where the goal is to separate the foreground object from the background with the help of users. Some scratches [177] or a bounding box [178] are usually drawn on the image to mark the candidate foreground and background regions. These provided marks will serve as foreground and background seeds to refine the rest of the image, maybe along with the further interaction of the user. To make the cutted foreground object look more natural, alpha matting algorithm [179] is usually adopted to post-process the foreground. Mortensen and Barrett [9] have proposed a boundary-based interactive segmentation method that requires the user to control the mouse along the boundary of the object and place several marks. They then used Dijkstra's shortest path algorithm to finish the segmentation of the object. Another example is the active contour method [10], which is able to capture salient image contour. In this method, an initial contour is placed near the boundary of the object of interest and the contour is evolved to catch the object boundary. A recent computer-assisted annotation system has been proposed by Maire *et al.*, for recovering hierarchical scene structure [180].

2.2.3 Object Proposal Generation Models

Recently, some researchers have started to concentrate on generating a small set of category-independent object proposals, in terms of either bounding boxes or segments (regions), each of which contain an object rather than other



(a) source image

(b) normed gradients maps

(c) 8×8 NG features

(d) learned model $\mathbf{w} \in \mathbb{R}^{8\times 8}$

Fig. 24. BING objectness measure by Cheng *et al.* [58]. Although object (red) and non-object (green) windows present huge variation in the image space (a), in proper scales and aspect ratios where they correspond to a small fixed size (b), their corresponding normed gradients, i.e. a NG feature (c), share strong correlation. We learn a single 64D linear model (d) for selecting object proposals based on their NG features.

stuff without any category-specific information. Compared with the huge amount of sliding windows (*e.g.*, $1,000,000$) or regions of various scales, a relatively small set of (*e.g.*, 1000) object proposals can be generated with these models in the pre-processing stage, which have high degrees of *objectness* or *object plausibility*. More computational budget with high-level, category-specific prior knowledge can then be put on the later stage to get efficient object detections.

Object proposal generation models fall in between general object detection and salient object detection literatures. On one hand, generic object detection models always exploit category-specific information to train an object detector. The object proposal models usually serve as a pre-processing step to accelerate the object detection. On the other hand, though saliency can be used as a visual cue to determine the objectness, a generic object is not necessarily to be visually salient. It is worth noting that the salient object localization algorithms we introduced in Sec. 2.1.4 also aim to *directly* predict a bounding box around the salient object while the goal of object proposal generation methods is to generate a set of candidate proposals to speed up further category-specific object detection. Moreover, salient object localization approaches tend to fail on complicated images where many objects exist and they are usually not dominant.

In the pioneer work [16] of objectness, Alexe *et al.* argue that some generic characteristics, such as boundary closure, high local contrast, and sometimes visual saliency, are shared by objects of any categories. Based on these generic properties of objects, several cues including multiscale visual saliency, color contrast, edge density, and superpixel straddling are integrated in a Bayesian framework to determine how likely for a bounding box to contain an object of any class. Rahu *et al.* [181] later proposed to discriminatively compute the objectness score of each proposal as a linear combination of the features including the superpixel straddling, boundary edge distribution, and window symmetry, where the combining weights are learned based on a structured output ranking objective function. Selective search [182] took advantage of multiple candidate segmentations to generate bounding box proposals. Starting from over-segmentation of the input image, adjacent regions were gradually merged to form hierarchical segmentations. Object proposals are defined as tight bounding boxes surrounding the regions in the hierarchy. Ristin *et al.* [183] demonstrated that local context prior can be utilized to generate object proposals. Very few but large image patches are randomly sampled initially. These local patches together provide contextual information to estimate the prior distribution of object locations. In [184], the input image is modeled as an undirected graph where the vertices correspond to superpixels and the weights capture the probability that two neighboring superpixels belong to the same object. Object proposals can be efficiently generated as the bounding boxes surrounding the random partial spanning trees of the graph. In a recent work [58] (see Fig. 24), Cheng *et al.* demonstrated simple binarized normed gradients (BING) feature of a bounding box which is resized to a fixed size (*e.g.*, 8×8) even along with simple linear SVMs are surprisingly powerful to rank the sliding windows of the image. Additionally, it is highly computationally cheap (more than $1000\times$ faster than most widely used alternatives). Deep convolutional neural networks (DNNs) have achieved state-of-the-art performance on a number of image recognition benchmarks. Recently, the Deep Neural Network (DNN) is adopted to generate bounding box object proposals [185], where proposal generation is defined as a regression problem to the coordinates and objectness scores of output bounding boxes. Compared with pervious approaches with hand-crafted features, such a deep model is fully learned. In [186], it is observed that the number of contours wholly enclosed by a bounding box is indicative of the likelihood of the box containing an object. To this end, the objectness score of each bounding box is defined as the number of edges that exist in the box minus those that are members of contours that intersect the box's boundary.

Concurrently, some other researchers focus on ranking the segments (regions) of the input image. In the pioneer work [187], a framework is presented to output a set of object hypotheses, represented as figure-ground segmentations of the input image, based on a mid-level feature vector. Initial hypotheses are first automatically extracted by solving a sequence of constrained parametric min-cut problems (CPMC) on the pixel grids. A Random Forest regressor is learned to rank these hypotheses in the descendant degrees of objectness, where each of them is jointly described by a feature descriptor, including the graph partition properties, region properties, and Gestalt properties. Similarly, Endres and Hoiem [17] proposed to produce diverse category-independent object region proposals by labeling the superpixels as foreground or background on the CRF framework based on a seed region and learned affinity function between regions. Such region proposals are then learned to rank in the structured learning framework based on appearance features that generalize well across all object categories. The local selective search strategy of [182] was further augmented by the global search in [188]. On each hierarchy resulting from grouping adjacent regions, all the regions were classified as either foreground or background by minimizing the energy function with different foreground and background hypotheses and parameters of the energy function to form object proposals. In a recent work [189], multi-scale regions coming from hierarchical segmentations of the image will be grouped to generate object region proposals by efficiently exploiting their combinatorial grouping space (*i.e.*, singletons, pairs, triplets, and 4-tuples of regions). Kim and Grauman [190] introduced a global category-independent shape prior for object proposal generation based on the observation that shapes were often shared between objects of different categories. Associated exemplar shapes from a given database were projected to the test image, which with high overlapping will be further grouped to form shape priors of the novel input image. By performing a series of figure-ground segmentations using graph-cuts based on each group of the shape priors, a set of object proposals were generated. Similarly, a large collections of example object regions were also maintained in [191]. Objectness score of each segment of the input image can be computed based on its nearest neighbors in the database by combining segment properties, mutual consistency across the nearest exemplar regions, and the prior probability of each exemplar region.

#	Model	Pub	Year	Output	Code	Bench.
1	**OBJ** [16]	CVPR	2010	BB	M+C	✓
2	**CPMC** [187]	CVPR	2010	SG	M	
3	**CIOP** [17]	ECCV	2010	SG	M+C	
4	**LOC** [181]	ICCV	2011	BB	C	
5	**SELECT** [182]	ICCV	2011	BB	M	
6	**SS** [190]	ECCV	2012	SG	M+C	
7	**LCP** [183]	ACCV	2012	BB	NA	
8	**PRIME** [184]	ICCV	2013	BB	M+C	
9	**BING** [58]	CVPR	2014	BB	C	
10	**GLS** [188]	CVPR	2014	SG	-	
11	**DNN** [185]	CVPR	2014	BB	-	
12	**MCG** [189]	CVPR	2014	SG	M+C	
13	**EB** [186]	ECCV	2014	BB	-	
14	**GOP** []	ECCV	2014	SG	-	
14	**DDO** [191]	PAMI	2014	SG	NA	
15	**VOP** [192]	CVPRW	2012	SG	NA	
16	**ODSA** [193]	ICRA	2013	SG	M+C	

Fig. 25. Object proposal generation models. For the proposal form, {BB = bounding boxes, SG = segments}.

Those models discussed above are all working on a single input image. Object proposals can also be extended to leverage the spatial-temporal cues [192] and multi-modal data [193]. In [192], object proposals [17] are independently extracted at each frame. These proposals are then linked over frames into spatial-temporal object hypotheses. These hypotheses are used as higher order potentials of an energy function to label the superpixels of each frame. By varying the parameters of the energy function, multiple segmentations as well as a large pool of video object proposals are generated. Karpathy *et al.* [193] generate object proposals from 3D meshes of indoor environments. The scene is first decomposed into a set of candidate mesh segments. Five intrinsic shape measures, including compactness, symmetry, smoothness, and local and global convexity, along with a shape recurrence measure of each segment are then extracted to jointly compute its objectness score.

Fig. 25 provides a list of reviewed object proposal generation models. With more and more research effort put on this direction, it is hard to tell which kind of proposal output is more advantageous. Regular bounding boxes may allow efficient feature extraction, *e.g.*, the binary features in [58]. On the other hand, region proposals are more natural for object representation.

2.3 Applications of Salient Object Detection

The value of salient object detection models lies on their applications in many fields in computer vision, graphics, and robotics. They have been utilized for several applications such as object detection and recognition [201]–[208], image and video compression [209], [210], video summarization [211]–[214], photo collage/media retargeting/cropping/thumb-nailing [194], [215], [216], image quality assessment [217]–[220], image segmentation [221]–[225], content-based image retrieval and image collection browsing [198], [226]–[228], image editing and manipulating [195], [197], [199], [200], visual tracking [229]–[236], object discovery [193], [237], and human-robot interaction [238]–[241]. Fig. 26 shows some of these example applications.

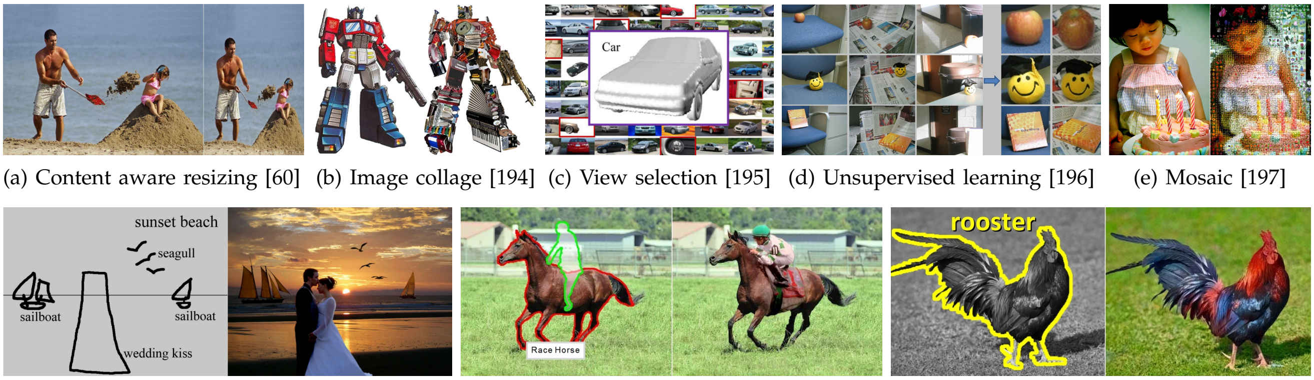
3 Datasets and Evaluation Metrics

3.1 Datasets for Salient Object Detection

As more models have been proposed in the literature, more datasets have been introduced to further challenge saliency detection models. Early attempts aimed to collect images with salient objects being annotated with bounding boxes (e.g., **MSRA-A** and **MSRA-B** [8]), while later efforts annotated such salient objects with pixel-wise binary masks (e.g., **ASD** [79] and **DUT-OMRON** [242]). Typically, images, which can be annotated with accurate masks, contain only limited objects (usually one) and simple/clear background regions. On the contrary, recent attempts have been made to collect datasets with multiple objects and complex/cluttered background (e.g., [46], [53], [54]). As we mentioned in the Introduction Section, a much more sophisticated mechanism is required to determine which is the most salient object when several candidate objects are presented in the same scene. For example, Borji [54] and Li *et al.* [53] used the peak of human fixation map to determine which object is the most salient one (i.e., the one that humans look at the most; see section 1.2).

In this Section, we review the most influential datasets in the field of salient object detection. We have listed 21 salient object datasets in Fig. 27, which contains 20 datasets with still images and only 1 dataset for evaluating salient object detection models in video. This fact implies that more video datasets are needed in the literature[7]. Note that all images or video frames in these datasets are annotated with binary masks or rectangles, while some image datasets also provide the fixation data for each image collected during free-viewing conditions. When annotating these datasets, subjects are asked to label the only salient object in each scene (e.g., [8]) or annotate the most salient one among several candidates (e.g., [46])

[7] Some spatiotemporal salient object detection models proposed to use the surveillance video with foreground targets annotated by rectangles for quantitative evaluation



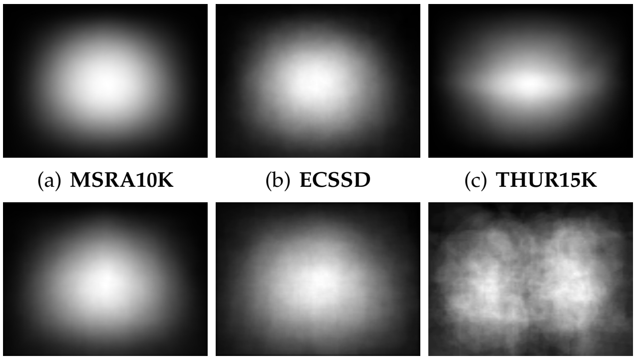
(a) Content aware resizing [60] (b) Image collage [194] (c) View selection [195] (d) Unsupervised learning [196] (e) Mosaic [197]

(f) Image montage [198] (g) Object manipulation [199] (h) Semantic colorization [200]

Fig. 26. Sample applications of salient object detection. Images are reproduced from corresponding references.

#	Dataset	Reference	Year	Images	Objects	Annotation	Resolution	Annotators	Eye data	Bench.
									Still image datasets (Spatial)	
1	**MSRA-A**	[8], [243]	2007	20K	~1	Bounding Box	400 × 300	3	-	
2	**MSRA-B**	[8], [243]	2007	5K	~1	Bounding Box	400 × 300	9	-	
3	**SED1**	[1], [244]	2007	100	1	Pixel-wise	~300 × 225	3	-	
4	**SED2**	[1], [244]	2007	100	2	Pixel-wise	~300 × 225	3	-	✓
5	**ASD**	[8], [79]	2009	1000	~1	Pixel-wise	400 × 300	1	-	
6	**SOD**	[14], [245]	2010	300	~3	Pixel-wise	481 × 321	7	-	
7	**iCoSeg**	[137]	2010	643	~1	Pixel-wise	~500 × 400	1	-	
8	**MSRA5K**	[8], [91]	2011	5K	~1	Pixel-wise	400 × 300	1	-	
9	**Infrared**	[246], [247]	2011	900	~5	Pixel-wise	1024 × 768	2	15	
10	**ImgSal**	[230]	2013	235	~ 2	Pixel-wise	640 × 480	19	50	
11	**CSSD**	[96]	2013	200	~1	Pixel-wise	~400 × 300	1	-	✓
12	**ECSSD**	[96], [248]	2013	10K	~1	Pixel-wise	~400 × 300	1	-	
13	**MSRA10K**	[8], [249]	2013	10K	~1	Pixel-wise	400 × 300	1	-	✓
14	**THUR15K**	[8], [249]	2013	15K	~1	Pixel-wise	400 × 300	1	-	✓
15	**DUT-OMRON**	[242]	2013	5,172	~5	Bounding Box	400 × 400	5	5	✓
16	**Bruce-A**	[7], [46]	2013	120	~4	Pixel-wise	681 × 511	70	20	
17	**Judd-A**	[54], [250]	2014	900	~5	Pixel-wise	1024 × 768	2	15	✓
18	**PASCAL-S**	[53]	2014	850	~5	Pixel-wise	variable	12	8	
19	**UCSB**	[47]	2014	700	~5	Point-wise Clicks	405 × 405	100	8	
20	**OSIE**	[160]	2014	700	~5	Pixel-wise	800 × 600	1	15	
									Video datasets (Spatio-temporal)	
21	**RSD**	[251]	2009	62,356	variable	Bounding Box	variable	23	-	

Fig. 27. Overview of popular salient object datasets (sorted based on their publication year). **MSRA10K** is a superset of **MSRA5K** and **ASD**. Top: image datasets, Bottom: video datasets.



(a) **MSRA10K** (b) **ECSSD** (c) **THUR15K**

(d) **DUT-OMRON** (e) **Judd-A** (f) **SED2**

Fig. 28. Average annotation map (AAM) of six datasets used in our benchamark.

3.1.1 MSRA and Its Descendants

- **MSRA-A & MSRA-B** [8][8]: This is the first "large-scale" dataset in the literature for quantitative evaluation of salient object detection models. It contains two parts: **MSRA-A** consisting of 20,840 images and **MSRA-B** containing 5,000 highly unambiguous images selected from **MSRA-A**. These images cover a large variety of scenarios such as flowers, fruits, animals, indoor and outdoor scenes. In the annotation, each image was resized to have a maximum side length of 400 pixels and the salient object(s) was manually annotated by rectangles (3 subjects for **MSRA-A** and 9 subjects for **MSRA-B**). Since the bounding box is inaccurate and often fail to reveal the accurate object boundaries, this dataset can be best used for salient object localization other than pixel-wise model evaluation. Moreover, a major shortcoming of this dataset is that most images only contain one object that is highly biased to image

[8] http://research.microsoft.com/en-us/um/people/jiansun/

centers (see Fig. 27 and Fig. 28). Consequently, some other datasets which are built upon the images from this dataset also suffer from this drawback.

- **ASD** [79][9]: This dataset is the most popular dataset in the literature (a.k.a., **MSRA1K**). It contains 1000 images from the **MSRA-B** dataset, while a binary pixel-wise object mask is provided for each image. When selecting images from the **MSRA-B** dataset, one standard is the minimum ambiguity on salient objects. Therefore, images in this dataset often have only one salient object and clean background, leading to extremely high performances when using simple algorithms. Actually, the performances of recent approaches seem to reach an "upper bound" on the **ASD** dataset, which poses a pressing demand on larger datasets with more complex testing images (e.g., multiple objects with cluttered background).
- **MSRA5K** [118][10]: In a recent work, Jiang et al., fully annotated the 5,000 images from the **MSRA-B** dataset with pixel-wise salient object masks.
- **MSRA10K** [252][11]: This dataset contains 10,000 images sampled from both **MSRA-A** and **MSRA-B** datasets with annotations for all pixels. Such a large-scale benchmark makes it very challenging and also suitable for more comprehensive model evaluation as well as performance analysis. A model benchmark on this dataset can be found in [252].

3.1.2 Other Datasets

- **SOD** [245][12]: This dataset is a collection of salient object boundaries based on Berkeley Segmentation Dataset (BSD) [15]. In the annotation process, seven subjects were asked to choose one or multiple salient objects in each image. For each object mask from each subject, a consistency score is computed from the labeling results of the other six subjects. Following [93], a binary salient object mask in each image was finally obtained by removing all labeled objects whose consistency scores are smaller than a threshold (set to 0.7 empirically) and combining the masks of objects with the highest inter-subject consistency.
- **iCoSeg** [137][13]: This dataset is originally introduced for the co-segmentation of foreground objects from a group of related images. It contains totally 643 images in 38 groups. Each image has pixel-wise annotation that may cover one or multiple salient objects. It is used for evaluating the salient object detection models in [118].
- **SED** [244][14]: This dataset consists of two parts. The first part, denoted as the "single-object database" (**SED1**), consists of 100 images with only one salient object in each image (i.e., similar to the **ASD** dataset). The second part, denoted as "two-objects database" (**SED2**), contains another 100 images with exactly two salient objects in each image. In the quantitative studies conducted in [1], Borji *et al.* demonstrated that saliency object detection models usually perform significantly worse on **SED2** than on the simple datasets (e.g., **ASD**). In their recent work [46], Borji *et al.* further asked 70 observers to select the most salient objects among the two objects in each image of **SED2**. The selection results can be found at: http://ilab.usc.edu/borji/Exp1Data.zip.

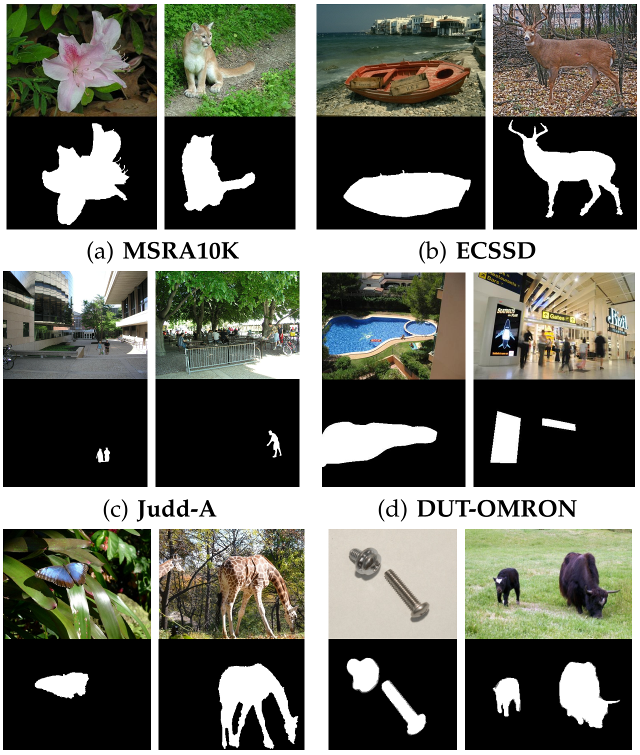

(a) **MSRA10K**
(b) **ECSSD**
(c) **Judd-A**
(d) **DUT-OMRON**
(e) **THUR15K**
(f) **SED2**

Fig. 29. Sample images from benchmarked salient object datasets along with their pixel-level annotations.

- **CSSD & ECSSD** [96][15]: Since images in the **ASD** dataset often contain only one object and simple background, a new dataset, denoted as Complex Scene Saliency Dataset (**CSSD**), was proposed in [96]. This dataset contains 200 images with diversified patterns in both foreground and background. Furthermore, the Extended Complex Scene Saliency Dataset (**ECSSD**) extends **CSSD** and has 1000 semantically meaningful but structurally complex images. Images in these two datasets are acquired from the **BSD** dataset [15], PASCAL VOC [253] and Internet, while the binary masks for salient objects are produced by 5 subjects.
- **ImgSal** [230][16]: Since the problems of salient object detection and fixation prediction are tightly correlated with each other, it would be valuable to construct an image dataset with both binary masks and human fixations. Toward this end, Li *et al.* introduced a dataset in [230] with these two kinds of information. In particular, they divided the 235 images from this dataset into 6 categories, including:
  1) images with large salient regions,
  2) images with intermediate salient regions,
  3) images with small salient regions,
  4) images with cluttered backgrounds,
  5) images with repeating distractors, and

[9] http://ivrgwww.epfl.ch/supplementary_material/RK_CVPR09/
[10] http://www.jianghz.com/projects/saliency_drfi/
[11] http://mmcheng.net/salobj/
[12] http://elderlab.yorku.ca/∼vida/SOD/
[13] http://chenlab.ece.cornell.edu/projects/touch-coseg/
[14] http://www.wisdom.weizmann.ac.il/∼vision/Seg_Evaluation_DB/
[15] http://www.cse.cuhk.edu.hk/leojia/projects/hsaliency/
[16] http://www.cim.mcgill.ca/∼lijian/

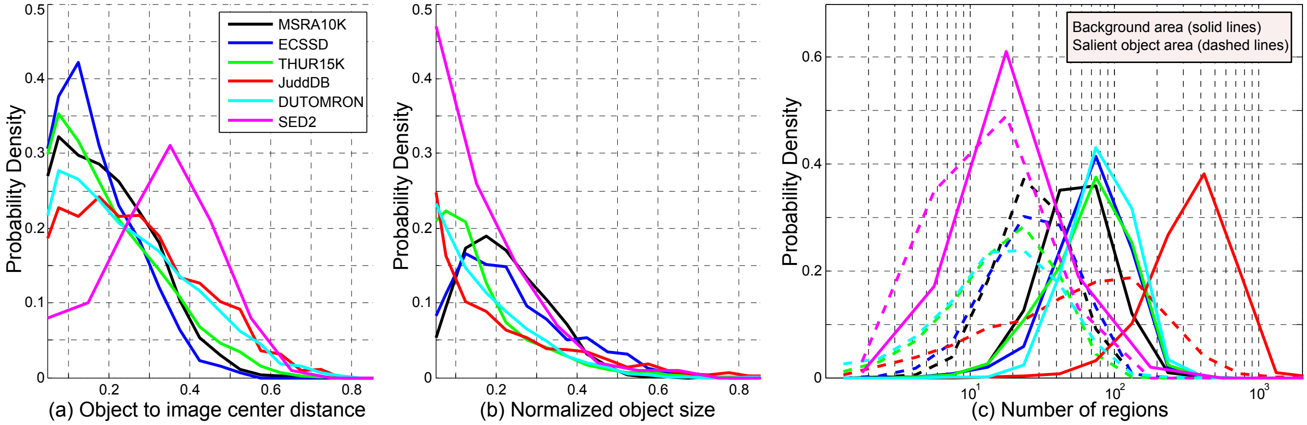


Fig. 30. Statistics of the benchmark dataset we used. a) Distribution of normalized object distance from image center, b) Distribution of normalized object, and c) Distribution of number of superpixels on salient objects and image background.

6) images with both large and small salient regions.

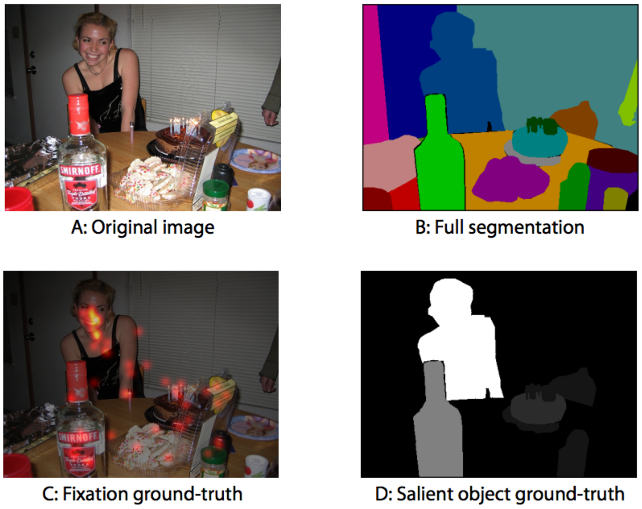
For these images, 19 subjects are asked to choose the salient objects and a pixel will become salient if it has been selected more than 50% subjects. One advantage of this dataset is that it provides rich information for each image, such as fixation data, object masks and category information. However, one major drawback is that it contains only 235 images and the limited number of scenes may lead to over-fitting when using the learning-based algorithms.

- **THUR15K** [254][17]: This dataset, containing a set of categorized images, is originally introduced for evaluating sketch-based image retrieval algorithms. Around 3,000 images are crawled from Flickr© for each of the 5 keywords, including "butterfly", "coffee mug", "dog jump", "giraffe" and "plane." Totally, around 15,000 images are collected. For each image, if there exists an object that is perfectly matched with the query keyword, such an object will be manually annotated with pixel-wise mask. Note that only the salient objects that are almost fully visible get labeled since partially occluded objects are often less useful for shape matching. As a consequence, some images have no salient region in the **THUR15K** dataset.
- **DUT-OMRON** [97][18]: The **DUT-OMRON** dataset also aims to overcome the drawbacks of the **ASD** dataset (i.e., limited objects and simple background). It contains 5,168 high quality images manually selected from more than 140,000 natural images. These images have one or more salient objects and relatively complex background. In the annotation process, each image is resized to have a maximum side length of 400 pixels, while both bounding boxes and pixel-wise object masks are provided for each image. In addition, the fixation data are also recorded using an eye tracking device. These three kinds of user data makes this dataset suitable for simultaneously evaluating salient object localization and detection models as well as fixation prediction models, which could provide an feasible way to explore the latent connections between these three research fields. One possible shortcoming of this dataset is that images have been presented for two seconds to free-viewing observers back to back which can cause high-center bias (Fig. 28) and memory confounds (previous image may influence eye movements over the next).
- **Bruce-A & Judd-A** [46], [54][19]: These datasets were created by Borji et al., mainly for checking generality of salient object detection models over complex scenes with several objects and high background clutter. These two datasets (**Bruce** also known as **Toronto** [7] and **Judd** also known as **MIT** [18]) have been frequently used for fixation prediction. Note that these datasets are center-biased (in terms of fixations) and also salient objects often fall at the image center (as shown in Fig. 28). This means that eye movement center-bias in this dataset is due to photographer bias which is tendency of photographers to frame salient, interesting, or important objects at the image center [157].
- **PASCAL-S** [53][20]: In a similar effort to [46], this dataset contains annotations of the most salient objects over complex scenes taken from PASCAL VOC [253] dataset, which consists of 20 object categories labeled over a large collection of photos. Each photo has been annotated by one subject and checked by another subject. One drawback of this dataset is that objects beyond the chosen 20 categories are not annotated. Contrary to **Bruce-A** and **Judd-A**, Li *et al.* took annotations of PASCAL dataset and recorded eye movements of observers on them (as shown in Fig. 31).

In addition to salient object datasets listed in Fig. 27, there exist some other datasets where objects, instead of "salient" objects, are manually annotated (e.g., [44], [255]). Since objects in these datasets often have well annotated binary masks, a feasible way to turn them into salient object datasets is to acquire the saliency scores from multiple subjects on the object level. Similar to [46], by summarizing these subjective saliency score on objects, the most salient objects can be easily inferred.

Beyond the object dataset, there also exist a number of fixation benchmarks that are publicly available. With simple post-processing, these fixation datasets can be also turned into salient object datasets. Intuitively, the object around the

[17] http://mmcheng.net/gsal/
[18] http://ice.dlut.edu.cn/lu/dut-omron/homepage.htm
[19] http://ilab.usc.edu/borji/
[20] http://yinli.cvpr.net/


A: Original image

B: Full segmentation

C: Fixation ground-truth

D: Salient object ground-truth

Fig. 31. An illustration of PASCAL-S dataset and the model by Li *et al.* [53]. This dataset provides both eye fixation (C) and salient object (D) mask. The labeling of salient objects is based on the full segmentation (B). A notable difference between PASCAL-S and its predecessors is that each image in PASCAL-S is labeled by multiple labelers without restrictions on the number of salient objects.

peak of the fixation density map can be selected as the most salient one (as in [54], **UCSB** [47], [160] and **OSIE** [160]). Note that acquiring the fixation data could also become a necessary prerequisite step to annotate the most salient object in a scene with complex content.

To sum up, there exist a large number of image benchmarks in the literature whose settings such as the image number, object number per image, annotation form and resolution are changing remarkably. To demonstrate the properties of these benchmarks, we show some representative samples (i.e., images and annotations) for each benchmark in Fig. 29. Furthermore, we illustrate in Fig. 30 the statistical information on some of the most popular benchmarks.

In Fig. 30(a), we show the normalized distances from salient objects to the corresponding image centers. We can see that the salient objects in **ECSSD** have the shortest distance to image centers, while the salient objects in **SED2** have the longest distances. This is reasonable since images in **SED2** usually have two objects aligned around opposite image borders. Moreover, we can see that the spatial distribution of salient objects in **Judd-A** have a larger variety than other datasets, indicating that this dataset have smaller positional bias (e.g., center-bias of salient objects and border-bias of background regions).

In Fig. 30(b), we demonstrate the average object sizes of these benchmarks, while the size of each object is normalized by the size of the corresponding image. We can see that **MSRA10K** and **RCCSD** have larger objects while **SED2** have smaller objects. In particular, we can see that some benchmarks contain a limited number of images with large foreground objects. By jointly considering the center-bias property of these benchmarks, it becomes very easy to achieve a extremely high precision on these images.

In Fig. 30(c), we aim to show the complexity of these benchmarks. Toward this end, we apply the graph-based superpixel segmenation algorithm [111] to see how many superpixels (i.e., homogenous regions) can be obtained on average from the salient objects and background regions of each image, respectively. In this manner, we can use this measure to reflect how challenging a benchmark is since massive superpixels often indicate complex foreground objects and cluttered background. From Fig. 30(c), we can see that **Judd-A** is the most challenging benchmark since it has an average number of 493 superpixels from the background of each image. On the contrary, **SED2** contains fewer number of superpixels in foreground and background regions, indicating that images in this benchmark often contain uniform regions and are easy to process.

What is a representative suitable dataset for salient object evaluation? We believe that for model benchmarking, it is important to assess models over *large number of scenes* as well as *complex scenes with multiple objects*. So far such a dataset satisfying two conditions does not exist mainly because manual annotation of multiple objects (as in [254]) or tracking human fixations (as in [53]) is very time consuming. In particular, when processing images with multiple foreground and complex background, the subjective annotations from various users may diversify from each other (as in [46]). Toward this end, existing studies often adopt several datasets with different properties for model evaluation, which is also the solution we adopt in this study.

3.2 Evaluation Measures

Here, we explain three universally-agreed, standard, and easy-to-understand measures for evaluating the salient object detection model. The first two evaluation metrics are based on the overlapping area between subjective annotation and saliency prediction, including the precision-recall (PR) and the receiver operating characteristics (ROC). From these two metrics, we also report the F-Measure, which jointly considers recall and precision, and AUC, which is the area under the ROC curve. Moreover, we also introduce the third measure which directly compute the mean absolute error (MAE) between the estimated saliency map and ground-truth annotation. For the sake of simplification, we use S to represent the predicted saliency map normalized to $[0, 255]$ and G be the ground-truth binary mask of salient objects. For a binary mask, we use $|\cdot|$ to represent the number of non-zero entries in the mask.

Precision-recall (PR). For a saliency map S, we can convert it to a binary mask M and compute $Precision$ and $Recall$ by comparing M with ground-truth G:

$$Precision = \frac{|M \cap G|}{|M|}, \quad Recall = \frac{|M \cap G|}{|G|} \tag{65}$$

From this definition, we can see that the binarization of S is the key step in the evaluation. Usually, there are three popular ways to perform the binarization. In the first solution, Achanta *et al.* [79] proposed the image-dependent adaptive threshold for binarizing S, which is computed as twice the mean saliency of S:

$$T_a = \frac{2}{W \times H} \sum_{x=1}^{W} \sum_{y=1}^{H} S(x, y), \tag{66}$$

where W and H are the width and the height of the saliency map S, respectively.

The second way to bipartite S is to use a fixated threshold which changes from 0 to 255. On each threshold, a pair of ($Precision$, $Recall$) scores are computed, which are finally combined to form a precision-recall (PR) curve to describe the model performance at different situations.

The third way to perform the binarization is to use the GrabCut-like algorithm (e.g., in [89]). In this solution, the PR curve is first computed and the threshold that leads to 95% recall is further selected. With this threshold, the initial binary mask is generated, which can be used to initialize the iterative GrabCut segmentation [178]. After several iterations, the binary mask can be gradually refined, which will be used to re-compute the precision-recall value.

F-measure. Usually, neither $Precision$ nor $Recall$ can comprehensively evaluate the quality of a saliency map. To this end, the F-measure is proposed as a weighted harmonic mean of $Precision$ and $Recall$ with a non-negative weight β^2:

$$F_\beta = \frac{(1+\beta^2)Precision \times Recall}{\beta^2 Precision + Recall}. \qquad (67)$$

As suggested by many salient object detection works (*e.g.*, [79]), β^2 is set to 0.3 to raise more importance to the $Precision$ value. The reason for weighting precision more than recall is that recall rate is not as important as precision (see also [9]). For instance, 100% recall can be easily achieved by setting the whole region to foreground.

According to the changing ways for saliency map binarization, there exist two ways to compute F-Measure. When the adaptive threshold or GrabCut algorithm is used for the binarization, we can generate a single F_β for each image and the final F-Measure is computed as the average F_β. If a unique PR curve is generated on all the testing images, we compute a F_β for each precision-recall pair and report the average. As defined in (67), F-Measure is a weighted harmonic mean of precision and recall, thus share the same value bounds as precision and recall values, *i.e.* [0, 1].

Receiver operating characteristics (ROC) curve. In addition to the $Precision$, $Recall$ and F_β, we can also report the false positive rate (FPR) and true positive rate (TPR) when binarizing the saliency map with a set of fixed thresholds:

$$TPR = \frac{|M \cap G|}{|G|}, \quad FPR = \frac{|M \cap G|}{|M \cap G| + |\bar{M} \cap \bar{G}|} \qquad (68)$$

where $\bar{M}$ and $\bar{G}$ denote the opposite of the binary mask M and ground-truth G, respectively. The ROC curve is the plot of TPR versus FPR by testing all possible thresholds.

Arear under ROC curve (AUC). While ROC is a two-dimensional representation of a model's performance, the AUC distils this information into a single scalar. As the name implies, it is calculated as the area under the ROC curve. A perfect model will score an AUC of 1, while random guessing will score an AUC around of 0.5.

Mean absolute error (MAE). The overlap-based evaluation measures introduced above do not consider the true negative saliency assignments, *i.e.*, the pixels correctly marked as non-salient. This favors methods that successfully assign high saliency to salient pixels but fail to detect non-salient regions over methods that successfully detect non-salient pixels but make mistakes in determining the salient ones [55], [100]. Moreover, in some application scenarios [256] the quality of the weighted, continuous saliency maps may be of higher importance than the binary masks. For a more comprehensive comparison we therefore also evaluate the mean absolute error (MAE) between the continuous saliency map S and the binary ground-truth G, both normalized in the range [0, 1]. The MAE score is defined as:

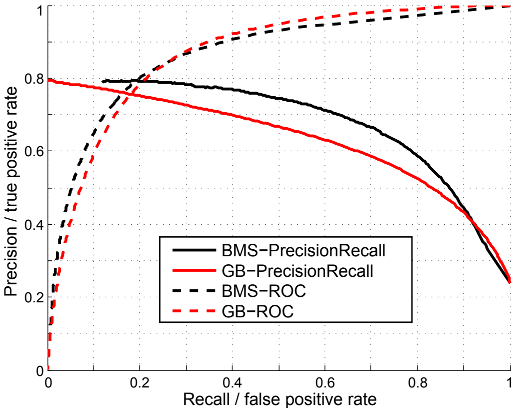


Fig. 32. Precision-recall and ROC curves for BMS [114] and GB [169] models over the **ECSSD** dataset.

$$MAE = \frac{1}{W \times H} \sum_{x=1}^{W} \sum_{y=1}^{H} |S(x,y) - G(x,y)|. \qquad (69)$$

Note that these scores sometimes do not agree with each other. For example, Fig. 32 shows a comparison of two models over **ECSSD** dataset using PR and ROC metrics. While there is not a big difference in ROC curves (thus about the same AUC), one model clearly scores better using the PR curve (thus having higher F-measure). Such disparity between the ROC and PR measures has been extensively studied in [257]. For evaluating the problem of salient object detection, where the number of negative examples (none salient object pixels) is typically much bigger than the number of positive examples (salient object pixels), PR curves are more informative than ROC curves, which can present an over optimistic view of an algorithm's performance [257]. Thus we mainly based our conclusions on the PR curves scores (F-Measure scores), and also reporting others scores for comprehensive comparisons as well as facilitating specific application requirements. It's worth mentioning that active research is on going to figure out the better ways of measuring salient object detection and segmentation models (see for example [258]).

4 Salient Object Detection Benchmark

In the field of salient object detection, it still lacks public datasets for the comprehensive evaluation of the models with extrinsic cues (e.g., depth map and co-segmentation). Therefore, in this study we focus on evaluating the models whose input is a single image, which is the main research direction in the literature.

4.1 Models for Comparison

In the comparison, 36 models are incorporated in total, including 24 salient object detection models, 10 fixation prediction models, one objectness proposal model, and one baseline model (as listed in Fig. 7, Fig. 15, Fig. 23, and Fig.

25). The reason to include models from different categories is to perform across-category comparison and study whether models specifically designed for salient object detection show actually advantage over fixation prediction and objectness models. This is particularly important since these models usually have different goals and generate visually distinctive maps (see Sec. 2). We also include a baseline model in the comparison, denoted as Average Annotation Map (AAM). An AAM is simply the average of ground-truth annotations collected from all images on each dataset. Since the AAM usually has larger activation around the image center (center bias; see Fig. 28), by using this map we can study the effect of center bias in model comparison.

The source codes (or executables) of these models are publicly accessible for evaluation. These models are written in Matlab, or C/C++, or mixture of them. We have created a unified repository for sharing code and data where you can run models with a single click or you can add new models for benchmarking purpose[21].

4.2 Datasets for Comparison

As shown in Sec. 3, different image datasets have different statistical attributes which may influence the performance of specific models. Therefore, it is necessary to test models over several datasets for fair comparisons and a good model should demonstrate impressive performance over almost all the testing datasets. Therefore, we choose the datasets based on the following criteria: 1) being widely-used, 2) containing a large number of images, 3) having different biases such as number of salient objects, and center-bias, and 4) potential to be the standard benchmarks for future.

Following these criteria, six datasets are chosen for model comparisons, including **MSRA10K**, **ECSSD**, **THUR15K**, **Judd-A**, **DUT-OMRON** and **SED2**. Among these datasets, **MSRA10K** is chosen because it contains the largest number of images that cover the images of **ASD** and **MSRA5K**. **THUR15K** and **DUT-OMRON** are also chosen for containing a large number of images, while **ECSSD** is chosen as it contains many semantically meaningful but structurally complex natural images. The reason to include **Judd-A** is to assess performance of models over scenes with multiple objects and high background clutter. Finally, we also evaluate models on **SED2** to check whether salient object detection models can perform well on images containing multiple objects (two objects for each image of **SED2**). Fig. 29 shows an overview of the representative images selected from these six datasets, while Fig. 28 also shows the AAMs generated on these six datasets so as to illustrate their different bias.

4.3 Quantitative Comparison of Models

To quantitatively measure the performance of various models on the six datasets, we adopt five evaluation metrics. In Fig. 33 and Fig. 34, we show the PR curves and ROC curves of each model on each of the six datasets. Moreover, Fig. 36 and Fig. 37 show the performance of these models in terms of AUC and MAE, respectively. Furthermore, Fig. 35 shows the mean F-measure scores of all the models (the first column of each dataset). To facilitate the reading, we highlight the best three models in each table as red, green and blue colors, with decreasing performances.

From the results obtained so far, we find that the following models performs among the best when using different evaluation metrics:

- With the PR and ROC curves, DRFI outperforms all other models on six datasets with large margin. Besides, DSR and MC achieve close performance and both perform slightly better than other models.
- with the mean F-measure, DRFI consistently wins over all datasets while MC ranks the second best over four datasets. DSR ranks the second on **THUR15K** and the third on **DUT-OMRON**.
- With the AUC, DRFI again ranks the best over all six datasets, while DSR reaches the second places over five datasets and MC ranks the third on two datasets. It is worth pointing out that all the models perform above the chance level (AUC = 0.5) on the six benchmark datasets.
- With the MAE, model rankings are diverse from either the mean F-measure or the AUC. DRFI and DSR again rank on the top, but both of them are out of the top three models on the **Judd-A** dataset. The third rank here belongs to FES and RC models. MC, which performs well in terms of mean F-measure and AUC, is not included in the top three models of any datasets.

In particular, we further compare the performance of fixation prediction models on the six datasets. Among fixation prediction models, COV and BMS perform the best while SR and SUN act the worst. Surprisingly, COV and BMS even outperform several salient object detection models in terms of mean F-measure, AUC and MAE, implying that they are suitable for both fixation prediction and proto-object detection. Beyond these two outliers, the cross-category comparisons show that the fixation prediction and objectness proposal models on average perform worse than salient object detection models. For example, most of the top three models are specifically designed for salient object detection. Additionally, Fig. 38 shows the distribution of mean F-measure, AUC and MAE scores of all salient object detection models versus all fixation prediction models over all benchmark datasets. We see a sharp separation of models especially for the mean F-measure, where most of the top models are salient object detection models. This result is consistent with the conclusion of [1] where they showed fixation prediction model perform lower than salient object detection models. Though stemming from the branch of fixation prediction, research in salient object detection demonstrates its unique properties and has truly added to what traditional saliency models already offer.

Most salient object detection algorithms outperform the baseline AAM. For instance, among the 24 salient object detection models, AAM only outperforms 2 models over **MSRA10K**, 8 models over **ECSSD**, 4 models on **THUR15K**, 11 models on **Judd-A**, 4 on **DUT-OMRON** in term of the mean F-measure. Interestingly, AAM does not outperform any model over the **SED2** dataset which means that indeed there is less location-bias (center-bias) in this dataset and salient object detection models can detect off-center objects. Notice that AAM performs the lowest on **SED2** compared to other datasets. In particular, we would like to point out

[21] Our code, data, and results are available in our online benchmark website: http://mmcheng.net/salobjbenchmark/

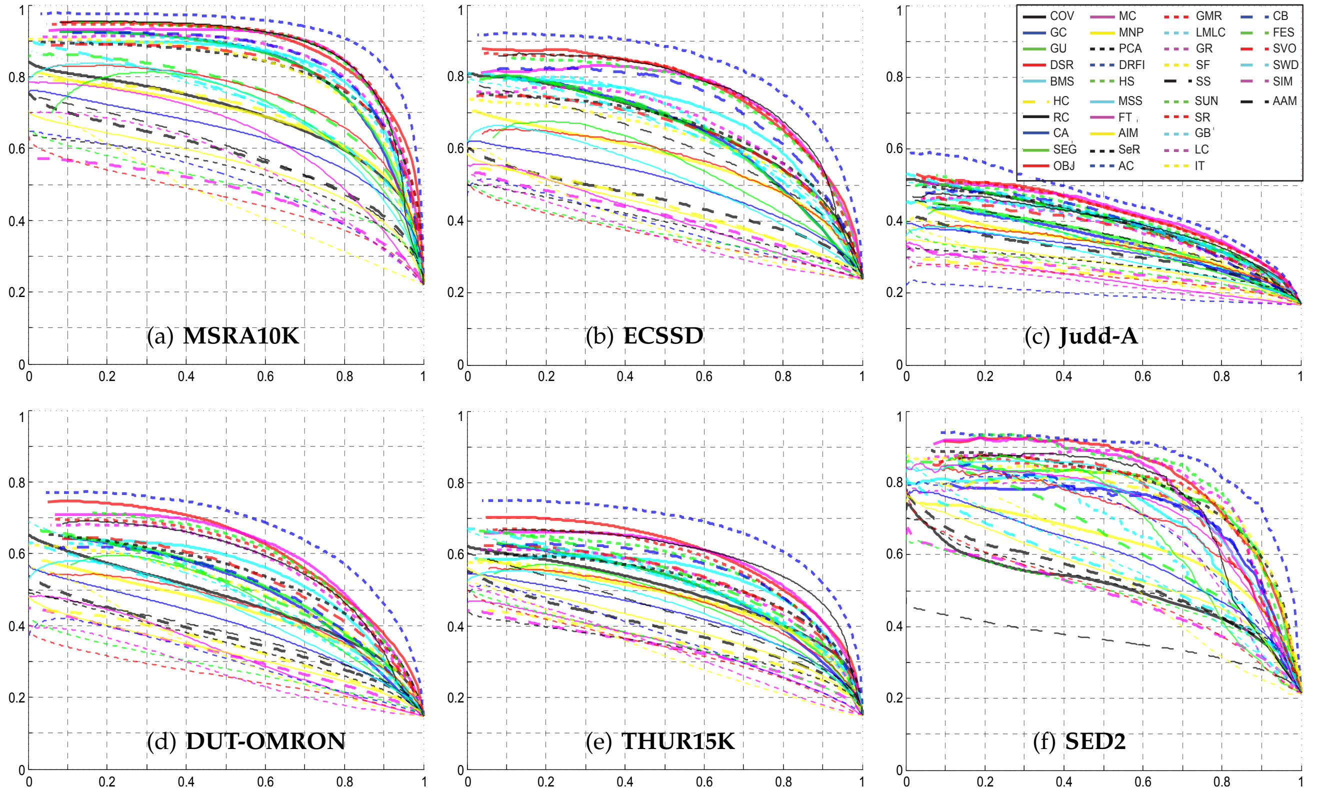


Fig. 33. Precision (vertical axis) and recall (horizontal axis) curves of saliency methods on 6 popular benchmark datasets.

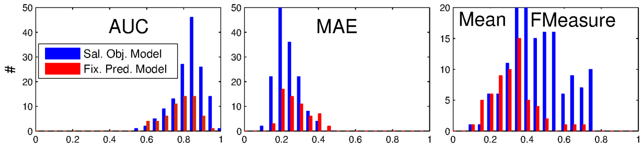


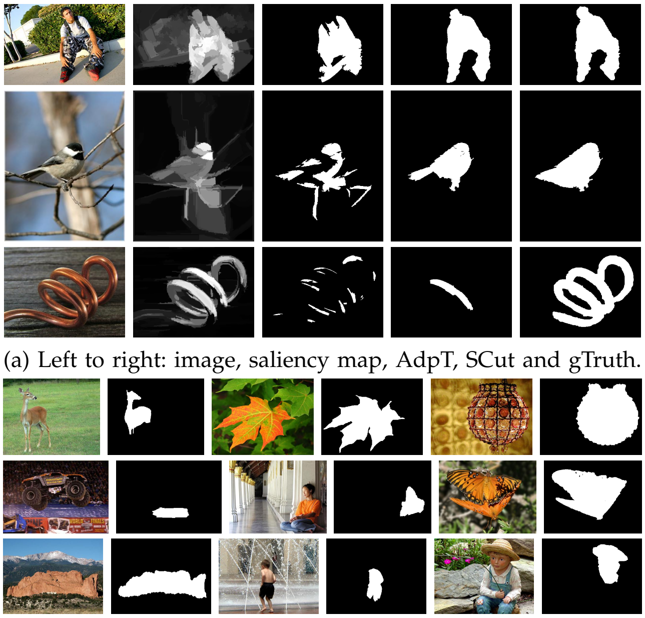
Fig. 38. Histogram of AUC, MAE, and Mean F-meaure scores for salient object detection models (blue) versus fixation pediction models (red) collapsed over all six datasets.

that it does not necessarily mean that models below AAM are not good, as taking advantage of the location prior may further enhance their performance (*e.g.*, LC and FT).

On average, over all evaluation metrics, all these models achieve lower performance on **Judd-A**, **DUT-OMRON** and **THUR15K**, indicating that these datasets are more challenging. The low model performance on **Judd-A** may arise from both less center-bias and smaller objects in images, while the noisy annotation of **DUT-OMRON** dataset might be a reason for low model performance. By investigating some images from these two datasets on which these models perform unsatisfactory, we find that there are several objects in each image that can be potentially the most salient one. This makes the ground-truth of these images somehow subjective and lead to certain labeling ambiguity, although the most salient object in **Judd-A** has been objectively defined to be the most looked-at one measured by eye movement data.

4.4 Performance Analysis

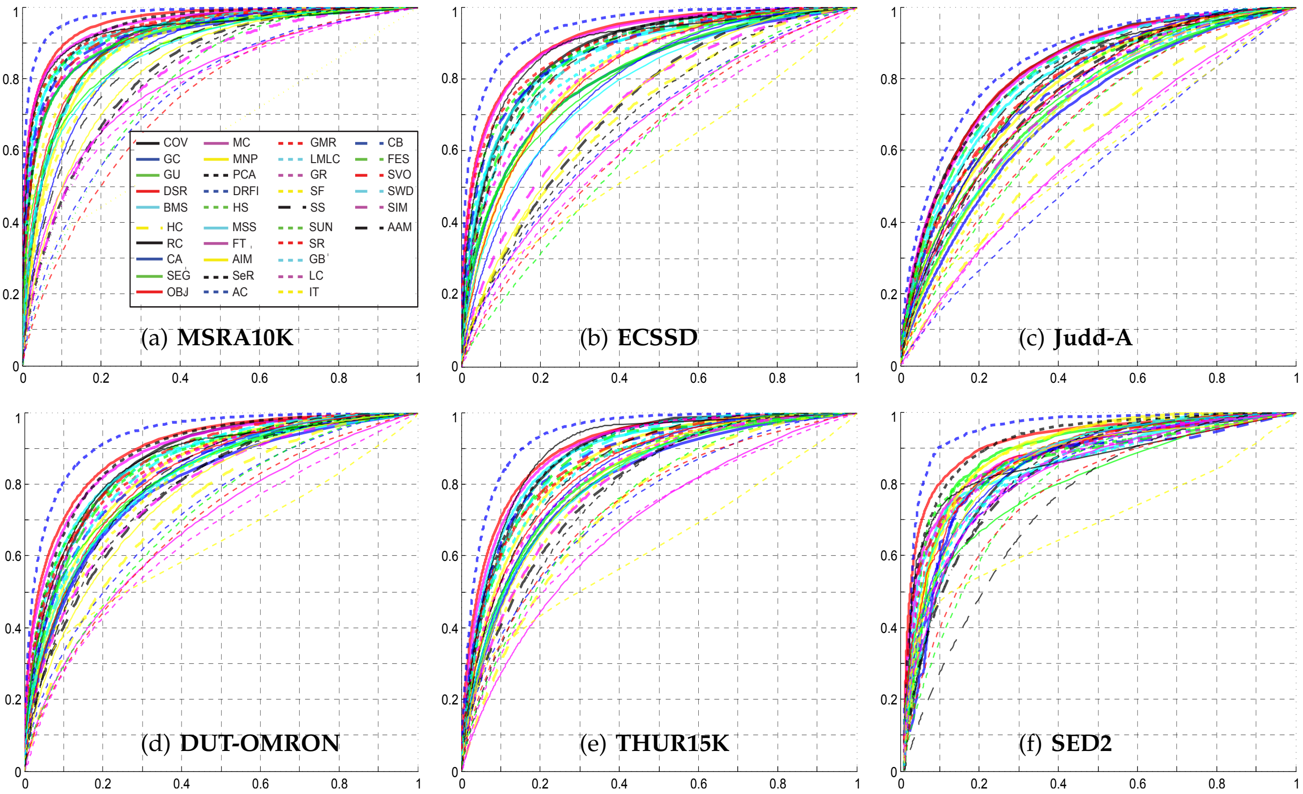
4.4.1 Analysis of Object Segmentation Methods


(a) Left to right: image, saliency map, AdpT, SCut and gTruth.

(b) DRFI model output fed to the SCut algorithm.

Fig. 39. Samples of salient object segmentation results.

In many computer vision and computer graphics applications, segmenting regions of interest is of great practical importance [194]–[196], [198], [200], [228], [259]. The simplest way of segmenting a salient object is to binarize the saliency map using a fixed threshold, as we discussed in the previous section. Such a fixed threshold might be hard to choose, however. In this section, we extensively evaluate two additional most commonly used salient object segmentation methods, including adaptive threshold [79],



(a) **MSRA10K** (b) **ECSSD** (c) **Judd-A** (d) **DUT-OMRON** (e) **THUR15K** (f) **SED2**

Fig. 34. ROC curves of models on 6 benchmarks. False and true positive rates are shown in x and y axes, respectively.

and SaliencyCut [89]. The mean F-measure scores when using different segmentation methods are shown in Fig. 35, while each segmentation algorithm was feeded with saliency maps produced by all the 36 models.

Except for the datasets **Judd-A** and **SED2**, the best segmentation results are all achieved via SaliencyCut method combined with a sophistical salient object detection model (DRFI, DSR, or MC). This indicates that the segmentation of salient objects can benefit from enforcing the label consistency by using graph-based segmentation and global appearance statistics. Note that the default SaliencyCut algorithm [89] only output the most dominate salient object, making the model performance on **SED2** and **Judd-A** less optimal since images in these two datasets often contain multiple salient objects.

Actually, the label consistency of nearby pixels is already widely recognized in the field of image segmentation. In Fig. 39(a), we demonstrated the segmentation results obtained by further enforcing label consistency among nearby and similar pixels. If a model can detect the majority of the salient object pixels, enforcing such label consistency often helps to recover the "missing" pixels.

However, when applying such consistency hypothesis on challenging scenarios such as complex object topology, spindle components, and similar appearance to the background, failures may occur. As shown in the last row of Fig. 39(a), due to the complex topology of the salient object, the hypothesis of label consistency may not always hold in each local region. In addition, the appearances various parts of an object may become very distinct due to the existence of shading and reflection, which makes the segmentation of the whole object very challenging. In Fig. 39(b), we have shown more results on images with various complexity when combining the best model, DRFI, with the best segmentation algorithm, SaliencyCut.

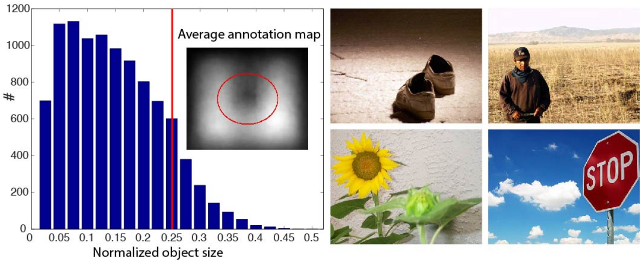


Fig. 41. Left: Histogram of object center over all images, threshold (red line = 0.247), and annotation map over 1000 less center-biased images in **MSRA10K** dataset. Right: Four less center-biased images. The overlaid circle illustrates the center-bias threshold.

4.4.2 Analysis of Center-bias

Center-bias is a severe problem in designing and evaluating fixation prediction and salient object detection models. Usually, the center-bias, caused by the photographer's bias for placing targets around the center of the view, may lead to unfair comparisons by favoring models that emphasize regions around image centers. To address this bias in model comparison, some approaches propose to adding a Gaussian center prior to all models before comparison. However, this solution is insufficient for fair comparisons since many salient object detection models already contain different levels of center-bias.

To remove the influence of such center-bias in model comparison, we choose 1000 images with no/less center bias

Model	MSRA10K				ECSSD				THUR15K				Judd-A				DUT-OMRON				SED2			
	Max	Mean	AdpT	SCut	Max	Mean	AdpT	SCut	Max	Mean	AdpT	SCut	Max	Mean	AdpT	SCut	Max	Mean	AdpT	SCut	Max	Mean	AdpT	SCut
GC	0.794	0.729	0.777	0.780	0.641	0.577	0.612	0.593	0.533	0.488	0.517	0.497	0.384	0.357	0.321	0.342	0.535	0.489	0.528	0.506	0.729	0.664	0.730	0.616
GU	0.793	0.728	0.770	0.801	0.638	0.573	0.599	0.622	0.534	0.488	0.510	0.523	0.392	0.364	0.327	0.370	0.539	0.485	0.515	0.536	0.759	0.700	0.741	0.654
DSR	0.835	0.765	0.824	0.833	0.737	0.645	0.717	0.703	0.611	0.551	0.604	0.597	0.454	0.402	0.421	0.410	0.626	0.560	0.614	0.593	0.794	0.722	0.821	0.632
MC	0.847	0.733	0.824	0.855	0.742	0.611	0.704	0.745	0.610	0.515	0.603	0.600	0.460	0.392	0.420	0.434	0.627	0.529	0.603	0.615	0.779	0.669	0.803	0.630
MNP	0.668	0.512	0.724	0.822	0.568	0.433	0.555	0.709	0.495	0.390	0.523	0.603	0.367	0.303	0.337	0.405	0.467	0.370	0.486	0.576	0.621	0.509	0.778	0.765
PCA	0.782	0.566	0.782	0.845	0.646	0.465	0.627	0.720	0.544	0.413	0.558	0.601	0.432	0.330	0.404	0.368	0.554	0.429	0.554	0.624	0.754	0.576	0.796	0.701
DRFI	0.881	0.761	0.838	0.905	0.787	0.667	0.733	0.801	0.670	0.573	0.607	0.674	0.475	0.387	0.419	0.447	0.665	0.555	0.605	0.669	0.831	0.741	0.839	0.702
HS	0.845	0.752	0.800	0.870	0.731	0.623	0.659	0.769	0.585	0.516	0.549	0.602	0.442	0.384	0.358	0.428	0.616	0.520	0.565	0.616	0.811	0.739	0.776	0.713
GMR	0.847	0.766	0.825	0.839	0.740	0.643	0.712	0.736	0.597	0.533	0.594	0.579	0.454	0.403	0.409	0.432	0.610	0.537	0.591	0.591	0.773	0.711	0.789	0.643
LMLC	0.801	0.714	0.773	0.857	0.659	0.562	0.601	0.724	0.540	0.460	0.519	0.588	0.375	0.340	0.307	0.379	0.521	0.417	0.493	0.552	0.653	0.537	0.712	0.674
GR	0.816	0.666	0.770	0.830	0.664	0.518	0.583	0.677	0.551	0.444	0.509	0.546	0.418	0.343	0.338	0.378	0.599	0.471	0.540	0.580	0.798	0.686	0.753	0.639
SF	0.779	0.556	0.759	0.573	0.619	0.414	0.576	0.378	0.500	0.365	0.495	0.342	0.373	0.244	0.319	0.219	0.519	0.401	0.512	0.377	0.764	0.583	0.794	0.509
CB	0.815	0.686	0.775	0.857	0.717	0.571	0.656	0.761	0.581	0.474	0.556	0.615	0.444	0.360	0.375	0.435	0.542	0.433	0.534	0.593	0.730	0.626	0.704	0.657
FES	0.717	0.471	0.753	0.534	0.645	0.421	0.655	0.467	0.547	0.390	0.575	0.426	0.424	0.304	0.411	0.333	0.520	0.370	0.555	0.380	0.617	0.355	0.785	0.174
SVO	0.789	0.540	0.585	0.863	0.639	0.443	0.357	0.737	0.554	0.382	0.441	0.609	0.414	0.314	0.279	0.419	0.557	0.361	0.407	0.609	0.744	0.516	0.667	0.746
SWD	0.689	0.478	0.705	0.871	0.624	0.443	0.549	0.781	0.528	0.371	0.560	0.649	0.434	0.296	0.386	0.454	0.478	0.338	0.506	0.613	0.548	0.411	0.714	0.737
HC	0.677	0.597	0.663	0.740	0.460	0.401	0.441	0.499	0.386	0.347	0.401	0.436	0.286	0.260	0.257	0.280	0.382	0.331	0.380	0.435	0.736	0.648	0.759	0.646
RC	0.844	0.729	0.820	0.875	0.741	0.637	0.701	0.776	0.610	0.530	0.586	0.639	0.431	0.360	0.370	0.425	0.599	0.506	0.578	0.621	0.774	0.699	0.807	0.649
CA	0.621	0.455	0.679	0.748	0.515	0.378	0.494	0.625	0.458	0.346	0.494	0.557	0.353	0.280	0.330	0.394	0.435	0.328	0.458	0.532	0.591	0.429	0.737	0.565
SEG	0.697	0.491	0.585	0.812	0.568	0.441	0.408	0.715	0.500	0.371	0.425	0.580	0.376	0.302	0.268	0.393	0.516	0.378	0.450	0.562	0.704	0.502	0.640	0.669
MSS	0.696	0.389	0.711	0.362	0.530	0.290	0.536	0.203	0.478	0.280	0.490	0.200	0.341	0.191	0.324	0.089	0.476	0.288	0.490	0.193	0.743	0.457	0.783	0.298
FT	0.635	0.420	0.628	0.472	0.434	0.286	0.431	0.257	0.386	0.268	0.400	0.238	0.278	0.187	0.250	0.132	0.381	0.268	0.388	0.259	0.715	0.523	0.734	0.436
AC	0.520	0.171	0.566	0.014	0.411	0.191	0.410	0.038	0.410	0.219	0.431	0.068	0.227	0.115	0.199	0.049	0.354	0.188	0.383	0.040	0.684	0.375	0.729	0.140
LC	0.569	0.413	0.589	0.432	0.390	0.274	0.396	0.219	0.386	0.291	0.408	0.289	0.264	0.199	0.246	0.156	0.327	0.251	0.353	0.243	0.683	0.527	0.752	0.486
OBJ	0.718	0.535	0.681	0.840	0.574	0.434	0.456	0.698	0.498	0.373	0.482	0.593	0.368	0.295	0.282	0.413	0.481	0.353	0.445	0.578	0.685	0.523	0.723	0.731
COV	0.667	0.363	0.755	0.394	0.641	0.371	0.677	0.413	0.510	0.333	0.587	0.398	0.429	0.291	0.427	0.315	0.486	0.334	0.579	0.373	0.518	0.312	0.724	0.212
BMS	0.805	0.693	0.798	0.822	0.683	0.582	0.659	0.690	0.568	0.485	0.578	0.594	0.434	0.363	0.404	0.416	0.573	0.503	0.576	0.580	0.713	0.625	0.760	0.627
SS	0.572	0.388	0.642	0.675	0.467	0.331	0.441	0.574	0.415	0.304	0.482	0.523	0.344	0.263	0.321	0.397	0.396	0.295	0.443	0.502	0.533	0.385	0.696	0.641
SIM	0.498	0.365	0.585	0.794	0.433	0.332	0.391	0.672	0.372	0.269	0.429	0.568	0.295	0.231	0.292	0.384	0.358	0.258	0.402	0.539	0.498	0.373	0.685	0.725
AIM	0.555	0.390	0.575	0.750	0.449	0.303	0.357	0.571	0.427	0.289	0.461	0.559	0.317	0.218	0.260	0.360	0.361	0.246	0.377	0.495	0.541	0.409	0.718	0.693
SeR	0.542	0.436	0.607	0.755	0.419	0.346	0.391	0.596	0.374	0.305	0.419	0.536	0.316	0.271	0.285	0.388	0.385	0.296	0.411	0.532	0.521	0.430	0.714	0.702
SUN	0.505	0.337	0.596	0.670	0.388	0.270	0.376	0.478	0.387	0.261	0.432	0.486	0.303	0.194	0.291	0.285	0.321	0.227	0.360	0.445	0.504	0.357	0.661	0.613
SR	0.473	0.165	0.569	0.001	0.381	0.143	0.385	0.001	0.374	0.162	0.457	0.002	0.279	0.117	0.270	0.001	0.298	0.137	0.363	0.000	0.504	0.211	0.700	0.002
GB	0.688	0.470	0.737	0.837	0.624	0.438	0.613	0.765	0.526	0.374	0.571	0.650	0.419	0.304	0.396	0.455	0.507	0.361	0.548	0.638	0.571	0.420	0.746	0.695
IT	0.471	0.351	0.586	0.158	0.407	0.177	0.414	0.003	0.373	0.195	0.437	0.005	0.297	0.146	0.283	0.000	0.378	0.216	0.449	0.005	0.579	0.289	0.697	0.008
AVG	0.580	0.475	0.692	0.779	0.597	0.484	0.627	0.756	0.458	0.358	0.569	0.620	0.392	0.313	0.367	0.411	0.406	0.325	0.514	0.534	0.388	0.305	0.524	0.640

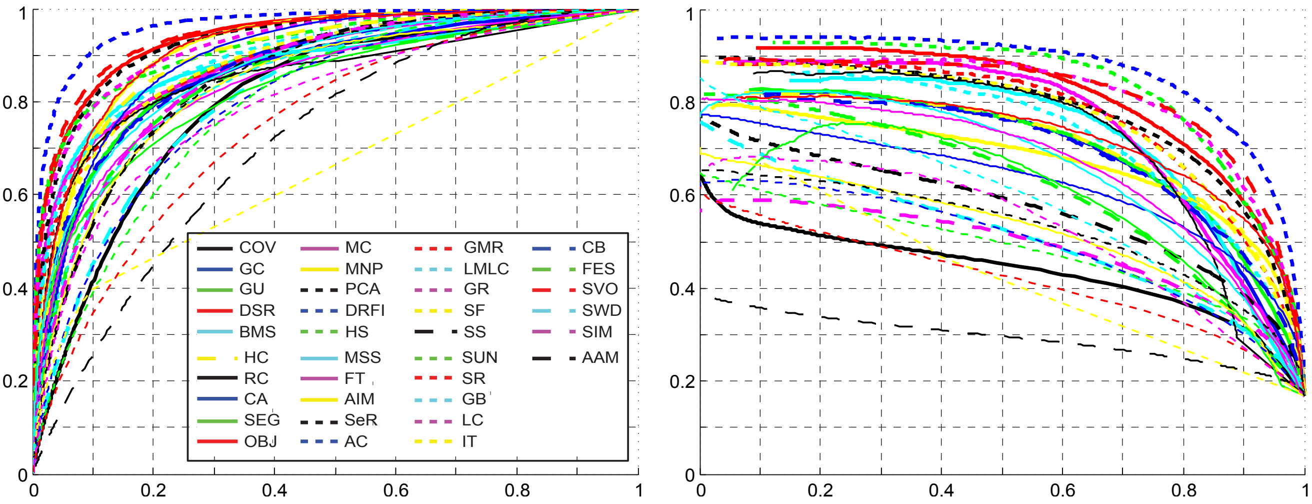
Fig. 35. Average F_β (larger is better) over images, using best F_β of varying thresholds, adaptive threshold, and SaliencyCut.

Model	MSRA10K	ECSSD	THUR15K	Judd-A	DUT-OMRON	SED2
GC	0.912	0.805	0.803	0.702	0.796	0.846
GU	0.916	0.808	0.816	0.718	0.810	0.877
DSR	0.959	0.914	0.902	0.826	0.899	0.915
MC	0.951	0.910	0.895	0.823	0.887	0.877
MNP	0.895	0.820	0.854	0.768	0.835	0.888
PCA	0.941	0.876	0.885	0.804	0.887	0.911
DRFI	0.978	0.944	0.938	0.851	0.933	0.944
HS	0.933	0.883	0.853	0.775	0.860	0.858
GMR	0.944	0.889	0.856	0.781	0.853	0.862
LMLC	0.936	0.849	0.853	0.724	0.817	0.826
GR	0.925	0.831	0.829	0.747	0.846	0.854
SF	0.905	0.817	0.799	0.711	0.803	0.871
CB	0.927	0.875	0.870	0.760	0.831	0.839
FES	0.898	0.860	0.867	0.805	0.848	0.838
SVO	0.930	0.857	0.865	0.784	0.866	0.875
SWD	0.901	0.857	0.873	0.812	0.843	0.845
HC	0.867	0.704	0.735	0.626	0.733	0.880
RC	0.936	0.892	0.896	0.775	0.859	0.852
CA	0.872	0.784	0.830	0.774	0.815	0.853
SEG	0.882	0.808	0.818	0.747	0.825	0.796
MSS	0.875	0.779	0.813	0.726	0.817	0.871
FT	0.790	0.661	0.684	0.593	0.682	0.820
AC	0.756	0.668	0.740	0.548	0.721	0.831
LC	0.771	0.627	0.696	0.586	0.654	0.827
OBJ	0.907	0.818	0.839	0.750	0.822	0.870
COV	0.904	0.879	0.883	0.826	0.864	0.833
BMS	0.929	0.865	0.879	0.788	0.856	0.852
SS	0.823	0.725	0.792	0.754	0.784	0.826
SIM	0.808	0.734	0.797	0.727	0.783	0.833
AIM	0.833	0.730	0.814	0.719	0.768	0.846
SeR	0.813	0.695	0.778	0.746	0.786	0.835
SUN	0.778	0.623	0.746	0.674	0.708	0.789
SR	0.736	0.633	0.741	0.676	0.688	0.769
GB	0.902	0.865	0.882	0.815	0.857	0.839
IT	0.640	0.577	0.623	0.586	0.636	0.682
AVG	0.857	0.863	0.849	0.797	0.814	0.736

Fig. 36. AUC: area under ROC curve (larger is better).

Model	MSRA10K	ECSSD	THUR15K	Judd-A	DUT-OMRON	SED2
GC	0.139	0.214	0.192	0.258	0.197	0.185
GU	0.142	0.224	0.201	0.270	0.214	0.167
DSR	0.121	0.173	0.142	0.196	0.139	0.140
MC	0.145	0.204	0.184	0.231	0.186	0.182
MNP	0.229	0.307	0.255	0.286	0.272	0.215
PCA	0.185	0.248	0.198	0.181	0.206	0.200
DRFI	0.118	0.166	0.150	0.213	0.155	0.130
HS	0.149	0.228	0.218	0.282	0.227	0.157
GMR	0.126	0.189	0.181	0.243	0.189	0.163
LMLC	0.163	0.260	0.246	0.303	0.277	0.269
GR	0.198	0.285	0.256	0.311	0.259	0.189
SF	0.175	0.230	0.184	0.218	0.183	0.180
CB	0.178	0.241	0.227	0.287	0.257	0.195
FES	0.185	0.215	0.155	0.184	0.156	0.196
SVO	0.331	0.404	0.382	0.422	0.409	0.348
SWD	0.267	0.318	0.288	0.292	0.310	0.296
HC	0.215	0.331	0.291	0.348	0.310	0.193
RC	0.137	0.187	0.168	0.270	0.189	0.148
CA	0.237	0.310	0.248	0.282	0.254	0.229
SEG	0.298	0.342	0.336	0.354	0.337	0.312
MSS	0.203	0.245	0.178	0.204	0.177	0.192
FT	0.235	0.291	0.241	0.267	0.250	0.206
AC	0.227	0.265	0.186	0.239	0.190	0.206
LC	0.233	0.296	0.229	0.277	0.246	0.204
OBJ	0.262	0.337	0.306	0.359	0.323	0.269
COV	0.197	0.217	0.155	0.182	0.156	0.210
BMS	0.151	0.216	0.181	0.233	0.175	0.184
SS	0.266	0.344	0.267	0.301	0.277	0.266
SIM	0.388	0.433	0.414	0.412	0.429	0.384
AIM	0.286	0.339	0.298	0.331	0.322	0.262
SeR	0.310	0.404	0.345	0.379	0.352	0.290
SUN	0.306	0.396	0.310	0.319	0.349	0.307
SR	0.232	0.266	0.175	0.200	0.181	0.220
GB	0.222	0.263	0.229	0.261	0.240	0.242
IT	0.213	0.273	0.199	0.200	0.198	0.245
AVG	0.260	0.276	0.248	0.343	0.288	0.405

Fig. 37. MAE: Mean Absolute Error (smaller is better).



Method	GC	GU	DSR	MC	MNP	PCA	DRFI	HS	GMR	LMLC	GR	SF	CB	FES	SVO	SWD	HC	RC
F_β	0.697	0.699	0.776	0.764	0.661	0.750	0.831	0.815	0.754	0.720	0.791	0.747	0.693	0.621	0.792	0.521	0.700	0.744
AUC	0.860	0.871	0.938	0.888	0.912	0.928	0.964	0.918	0.886	0.896	0.925	0.885	0.872	0.839	0.942	0.813	0.898	0.855
MAE	0.164	0.169	0.117	0.171	0.188	0.162	0.127	0.150	0.148	0.201	0.183	0.150	0.207	0.160	0.325	0.291	0.176	0.177

Method	CA	SEG	MSS	FT	AC	LC	OBJ	**COV**	**BMS**	**SS**	**SIM**	**AIM**	**SeR**	**SUN**	**SR**	**GB**	**IT**	AAM
F_β	0.620	0.629	0.666	0.671	0.521	0.569	0.708	0.463	0.739	0.571	0.515	0.540	0.546	0.498	0.444	0.590	0.460	0.328
AUC	0.896	0.828	0.868	0.843	0.800	0.797	0.915	0.805	0.879	0.852	0.858	0.836	0.849	0.795	0.750	0.850	0.655	0.716
MAE	0.199	0.300	0.167	0.183	0.177	0.192	0.243	0.176	0.146	0.225	0.363	0.265	0.273	0.276	0.184	0.208	0.165	0.406

Fig. 40. Results of center-bias analysis over 1000 less center-biased images chosen from the **MSRA10K** dataset. Top: ROC and PR curves, Bottom: Mean F-measure, AUC, and MAE scores for all models.

from **MSRA10K** (i.e., off-center images). First, the distance from the centroid of the salient object to the image center is computed for each image in **MSRA10K**. Those images whose such distances are larger than a threshold are then chosen. Fig. 41 shows some images with no/less center-bias as well as an illustration of the threshold used in choosing images. The average annotation of less center-biased images shows two peaks at the left and right of the image, which is suitable for testing the performance of salient object detection models on off-center images.

On these 1000 images, we re-evaluate all the 36 models using the same metrics as in the quantitative experiment. As shown in Fig. 40, DRFI and DSR again perform the best. Overall, most models perform worse when tested on off-center images (*e.g.*, the AUC score of MC declines from 0.951 to 0.888), while some models gain an increase in their performances. For example, the AUC score of SVO increases from 0.930 to 0.942 and it gets the second place. Some models, *e.g.*, HS, gain a lower performance but higher ranks (e.g., HS takes the second place in terms of mean F-measure) than on **MSRA10K**. DRFI still wins over other models here with a large margin and its performance only slightly changes on **MSRA10K** and these 1000 off-center images (the differences are 0.05, 0.05, and 0.009 on mean F-measure, AUC, and MAE, respectively). This means that this model is not taking much advantage of center-bias. In the contrast, CB uses a lot of location prior and that is why its performance drops heavily when applied to these images (the differences are 0.122, 0.122, and 0.029 on mean F-measure, AUC, and MAE, respectively).

Additionally, it can be observed from Fig. 28(f) that there is less center bias over the **SED2** dataset. We can therefore study the center bias on it. Similarly, DRFI and DSR outperforms other models in terms of mean F-measure, AUC, and MAE scores, indicating they are more robust to the location variations of salient objects. HS again ranks second according to the mean F-measure score.

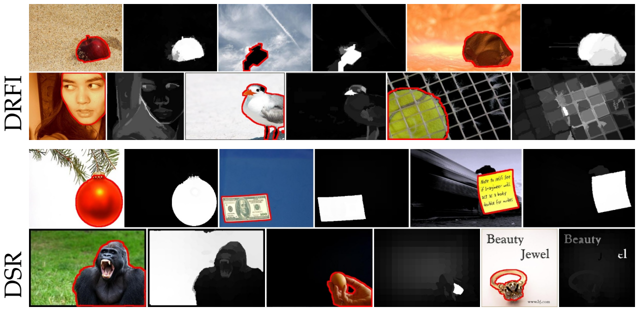


Fig. 42. Top and Bottom rows for each model illustrae Best and worse cases in UnCentered images.

Overall, all the models perform well above the chance level over either the 1000 off-center images or **SED2**. It is also worth noticing that the AAM model performs significantly worse on these two datasets, as well as **Judd-A**, which validates the effectiveness of the experiments on studying center bias on them.

4.4.3 Analysis of the Existence of Salient Object

Almost all of existing salient object detection models assume that there is at least one salient object in the input image. This impractical assumption might lead to less optimal performance on “background images”, which do not contain any dominated salient objects, as studied in [142]. To verify the effectiveness of models on background images,

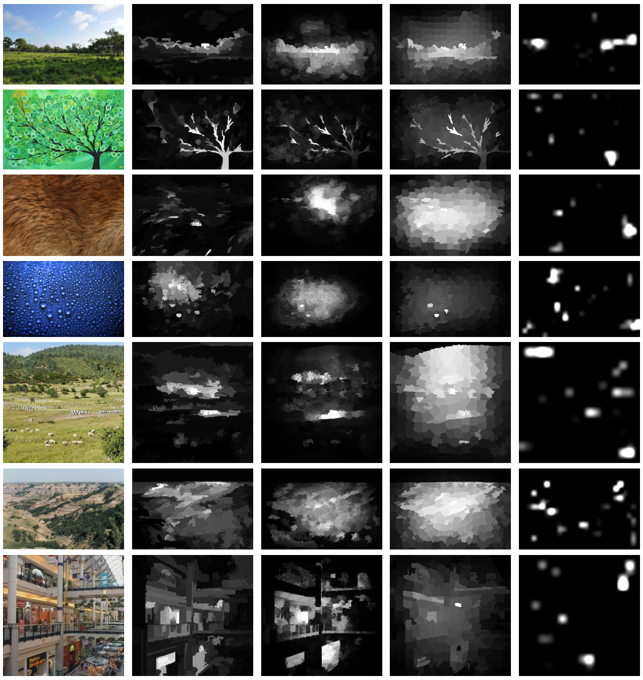


Fig. 43. Sample background-only images and prediction maps of DRFI, DSR, MC, and IT models.

we collected 800 images from the web and evaluated compared models on them. As shown in Fig. 43, there exist no dominated salient objects in these background images. On such images full of textures or cluttered background, a good model should generate a dark saliency map, *i.e.*, without any activation as there are no salient objects. For quantitative evaluation, we only report the MAE score of each model, which is basically the sum of non-zero elements of the output saliency map. Note that it is not feasible to calculate PR and ROC curves here since the ground-truth positive labeling here is empty. Also notice that ground-truth of eye fixations do exist on such background images.

In Fig. 43, we also show some representative background images and the saliency maps generated by the top three salient object detection models (on the other six datasets) and a classical fixation prediction model. Fig. 44 reports MAE scores of all compared 35 models, except the baseline AAM. We can see that the top salient object detection models on the other six datasets, such as DRFI, DSR, and MC, do not perform well and often generate activations on the background images even though only regular textures exist (the third and fourth rows of Fig. 43). This is reasonable as they always assume there exist salient objects in the input image and will try their best to find some ones. On the other hand, they can be distracted by the clutter in the background since high contrast always exist on the cluttered region. Most of existing salient object detection models compute saliency based on contrast values. In this manner, the cluttered background regions are more likely to be considered as salient.

From Fig. 44, we can see that fixation prediction models, COV and IT, perform the best on background images in terms of the MAE scores. Compared with the saliency maps with dense salient regions produced by salient object models, the output of fixation prediction often generate sparse activations on the scene (see Fig. 43 for the output of IT). The sum of non-zero elements of such sparse saliency maps are smaller and thus the performance of COV and IT are better.

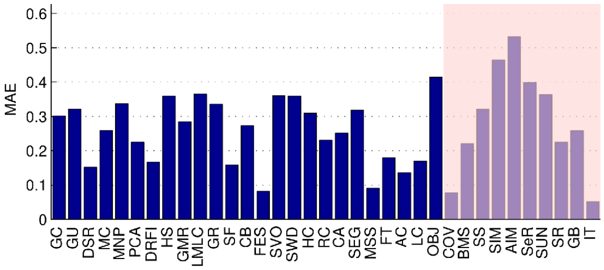


Fig. 44. MAE over background images with no salient objects. Shaded area belongs to fixation prediction models.

4.4.4 Analysis of the Worst and Best Cases for Top Models

To understand what are the real challenges for existing salient object detection models, we illustrate the best and worst cases for the top models over all the six datasets (models are sorted by mean F-measure). Due to the limited spaces, we only show the results of DRFI and MC in Fig. 45, while more illustrations can be found on our website.

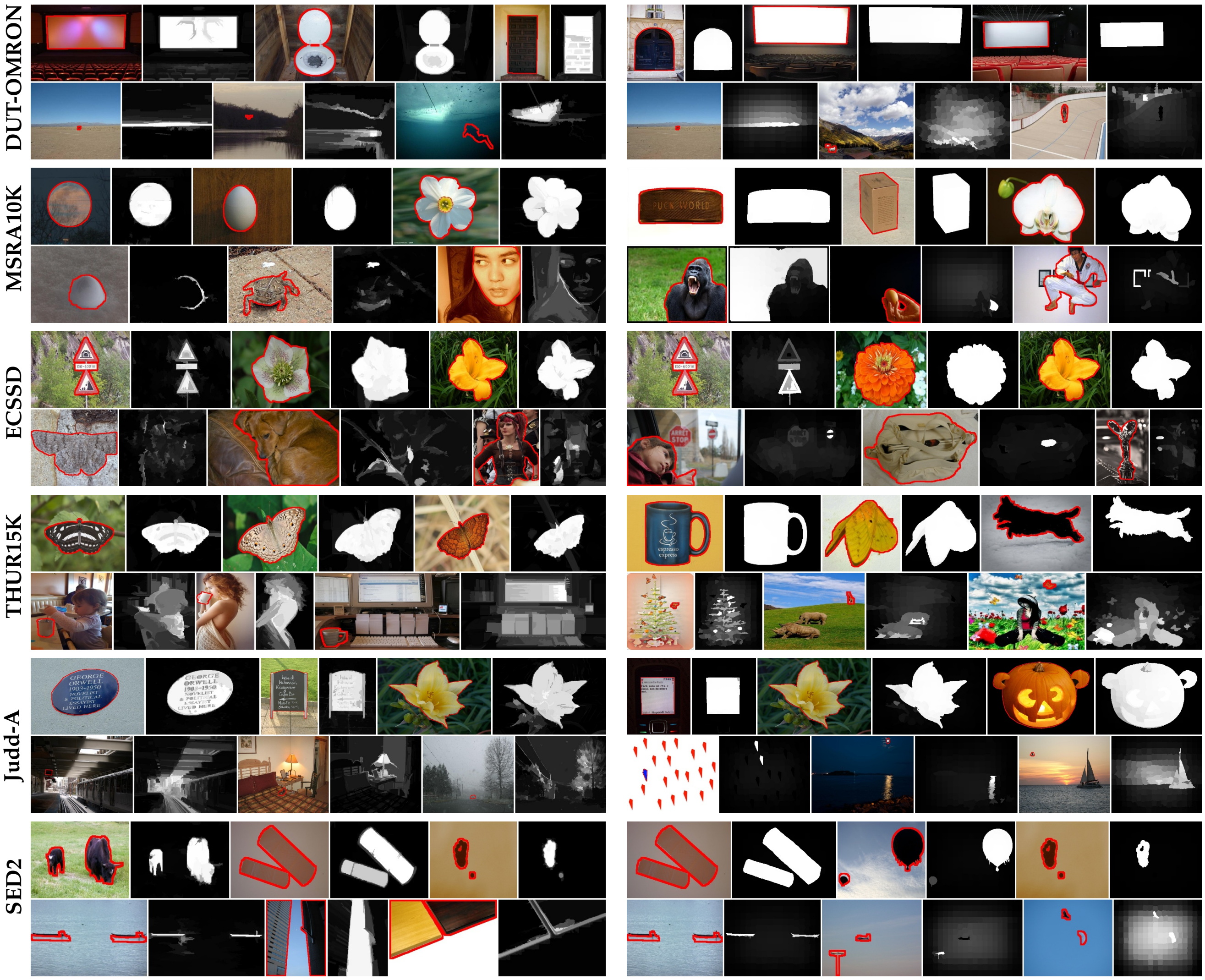
From Fig. 45, we notice that models share the same easy and difficult cases. Both DRFI and MC perform substantially well on the cases where a dominated salient object exists with relatively clean background. However, these models may fail in the following cases:

- Complex scenario: a complex scene with cluttered background may prevent the successful detection of salient foreground.
- Sematic saliency: the salient object is *semantically* salient as a whole while the object or certain parts it contains may be visually similar to the background (*e.g.*, DRFI fails on the facial image of **MSRA10K**)
- Object size: It is challenging for both DRFI and MC detecting small objects (e.g., the bad cases on **DUT-OMRON** and **Judd-A** datasets), while some models may fail on detecting extremely large objects.
- Assumption failure: MC relies on the assumption that the image border areas are background. Thus it may fail on the scenes where the salient object touches the image border, *e.g.*, the gorilla image in the fourth row of the right column of Fig. 45.

4.4.5 Runtime Analysis

As the front-end of many saliency-driven applications, the computational complexity of these saliency detection models is also a major concern. In Fig. 46, we show the runtime of all the compared models over the 10,000 images from **MSRA10K**. Most of these images have the resolution 400 × 300 and the testing platform contains an Intel Xeon E5645 2.40GHz CPU with 8 GB RAM.

As shown in Fig. 46, LC is the fastest (about 0.009 seconds per image), followed by HC and GU. The best model in our benchmark (DRFI) needs about 13.7 seconds to process one image. On one hand, the code of this model is written with a combination of Matlab and C, while the fastest three models are implemented purely in C. On the other hand,



(a) DRFI (b) MC

Fig. 45. Best (1st rows for each model on a dataset) and worst (2nd rows) cases of DRFI (left column) and MC (right column) models. Ground-truth object(s) is denoted by a red contour.

Method	GC	GU	DSR	MC	MNP	PCA	DRFI	HS	GMR	LMLC	GR	SF	CB	FES	SVO	SWD	HC
Time(s)	0.037	0.03	10.2	0.195	21	4.34	13.7	0.528	0.149	140	1.35	0.202	2.24	0.0960	56.5	0.19	0.017
Code	C	C	M+C	M+C	M+C	M+C	M+C	C	C	M+C	M+C	C	M+C	M+C	M+C	M+C	C

Method	RC	CA	SEG	MSS	FT	AC	LC	OBJ	**COV**	**BMS**	**SS**	**SIM**	**AIM**	**SeR**	**SUN**	**SR**	**GB**	**IT**
Time(s)	0.136	49.0	10.9	0.076	0.072	0.129	0.009	3.01	25.4	0.575	0.053	1.11	8.66	1.31	3.56	0.04	0.735	0.3
Code	C	M+C	M	C	C	M	C	M+C	M	M+C	M	M	M	M	M	M	M+C	M

Fig. 46. Average time for computing a saliency map, tested on **MSRA10K** databset (typical resolution of 400×300) using a desktop machine with Intel Xeon E5645 2.40GHz CPU and 8 GB RAM (M= Matlab, C= C/C++). Three fastet models are marked with red, blue, and green colors, respectively. See also the project page for a latest updates of new models: http://mmcheng.net/salobjbenchmark/. Fixation prediction models are shown in bold face.

since more sophisticated cues are adopted, DRFI performs best and runs slowest.

4.4.6 Performance Summarization

From the experimental results obtained so far, we can thus conclude the overall performance of each model. To facilitate the conclusion, we first conduct a qualitative experiment to demonstrate the output maps of all models on an image with relatively complex background. As shown in Fig. 47, dark blue areas are less salient while dark red areas are more salient. Compared with other models, top models like DRFI and DSR suppress most of the background well while almost successfully detect the whole salient object. They thus give higher precision scores and less false positives. Some models which include a center-bias component also results in appealing maps (*e.g.*, CB). Interestingly, region-based approaches, *e.g.*, RC, HS, DRFI, BMR, CB, and DSR always preserve the object boundary well compared with the block-based models.

We can also clearly see the distinctness of different categories of models. The saliency maps of salient object models will try to highlight the whole salient object and suppress

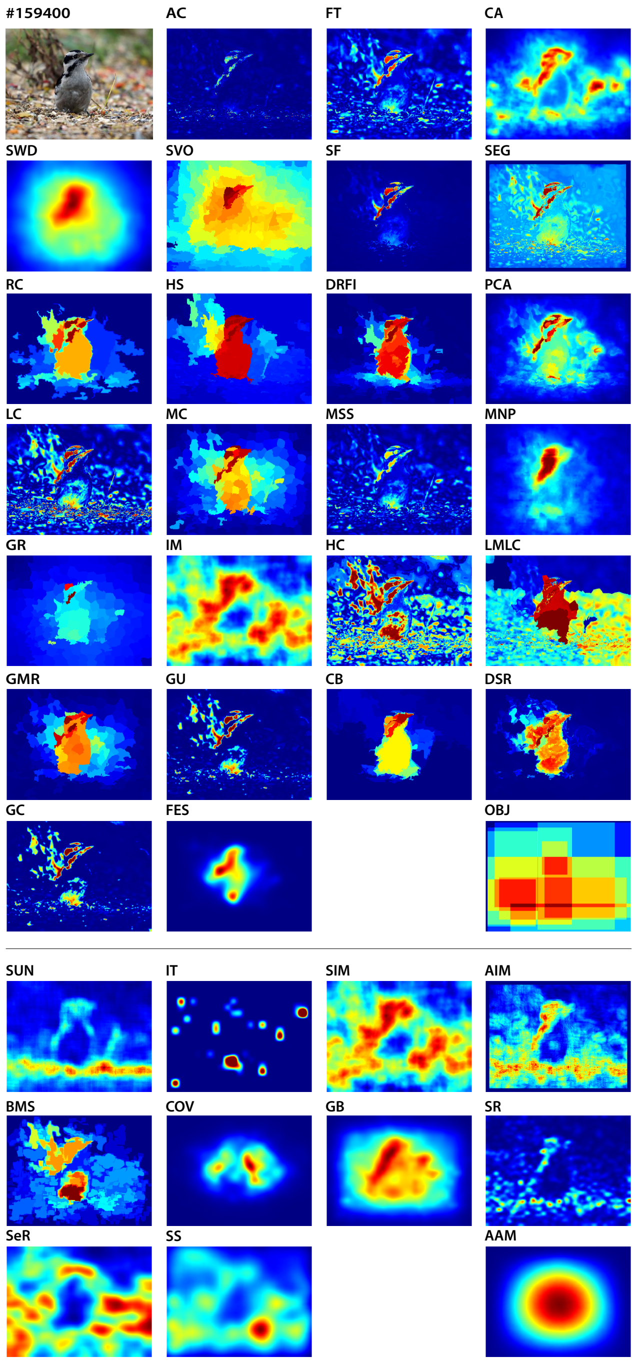


Fig. 47. Sample prediction maps of models. Top: salient object detection and objencess models, Bottom: fixation prediction models and the average annotation map.

the background. On the contrary, fixation prediction models often produce blob-like and sparse saliency maps corresponding to the fixation areas of human beings on the scene. The objectness map is a rough indication of the salient object. The output of the latter two types of models might not suit to segment the whole salient object well.

Finally, in Fig. 48 we summarize the rankings of models. These raking scores are based on average ranking over all 6 dataset, except that the 7^{th} row shows F-Measure results using best fixed thresholding over none-center biased dataset (*i.e.*, Fig. 40). The first 4 rows shows ranking of F-Measures scores, using best fixed thresholding, mean F-Measure of all threshold, adaptive thresholding [79], and SaliencyCut [252] respectively. The $5^{th}, 6^{th}, 8^{th}$ rows shows average ranking using area under ROC curve (AUC), mean absolute error (MAE) and timing used respectively. Best three models are highlighted with red, green, and blue colors in order. Based on such overall rankings (see Fig. 47 for example results using different methods), we can conclude that DRFI, DSR, MC, GMR, HS and RC are the top six models for addressing the problem of salient object detection.

5 Discussions

5.1 Design Choices

In the past decades, hundreds of methods for salient object detection have been proposed and a large number of design choices have be explored. By comparing the detailed method summarization (see Fig. 7 & 15) and evaluation results (see Fig. 34-37), our extensive evaluations does suggest some clear messages about commonly used design choices, which are valuable for the design of future algorithms:

5.1.1 Block-based vs. Region-based

From the chronological ordered method summarization Fig. 1, we observed a consistent evolution from block-based analysis to region-based analysis. Behind this evolution is the significant performance advantage of region level analysis, which we believe comes from three major reasons. First, the number of regions is typically much smaller than pixels or blocks, making the computation of high order feature or relations computationally feasible (e.g. all pairs correlations). Second, decomposing an image into perceptually homogeneous element helps to abstract out unnecessary details and is important for high quality saliency detection [55], [100]. Third, the region itself contains some important cues which could be missing at pixel/patch level, such as shapes, aspect ratios, perimeter [118]. For instance, the region-based method, RC [89], achieves 90% segmentation precision in the most widely used benchmark [79], and outperforms previously best reported results (75% segmentation precision in the pixel-based method FT [79]) by a large margin. This suggests that region based analysis tends to be preferred over pixel-level analysis when designing future salient object detection algorithms.

5.1.2 Intrinsic vs. Extrinsic

The effectiveness of intrinsic cues (see Sec. 2.1 for definition) has been validated in the past, indicated by the fact that there are 4 purely intrinsic cues based methods (MC, DSR, GMR, and RC) among the top 5 methods (see Sec. 4.3). Among various intrinsic cues, it's worth noticing that all the five leading methods explicitly explore the background prior. While some people prefer to call this background prior (or center prior) of salient object, others might prefer to call it center-bias of the dataset (see more discussions in Sec. 5.3). There is also a consistent trend of moving from local cues to global cues, possibly because the later tends to assign similar saliency values across similar image regions rather than highlighting only the boundary regions.

Despite the great success of purely intrinsic cues based methods, the DRFI method, which takes into account both intrinsic and extrinsic cues, consistently achieves better performance on all benchmarks than other methods. This

Method	GC	GU	DSR	MC	MNP	PCA	DRFI	HS	GMR	LMLC	GR	SF	CB	FES	SVO	SWD	HC	RC
Best	13	12	3	2	23	10	1	5	4	14	7	16	8	15	11	17	26	6
Mean	8	7	1	5	15	13	2	4	3	12	11	17	10	22	14	21	19	6
AdpT	14	15	2	3	17	7	1	10	4	21	13	16	12	9	33	18	25	5
SCut	25	21	14	8	10	9	1	4	12	13	19	29	7	28	6	2	26	5
AUC	24	19	2	3	15	4	1	8	6	18	16	22	13	14	7	12	27	5
MAE	10	12	1	6	24	13	2	14	4	23	22	9	18	5	35	29	26	3
NCB	16	15	5	6	20	8	1	2	7	12	4	9	17	22	3	30	14	10
Time	4	3	28	14	31	26	30	17	12	35	22	15	23	9	34	13	2	11

Method	CA	SEG	MSS	FT	AC	LC	OBJ	**COV**	**BMS**	**SS**	**SIM**	**AIM**	**SeR**	**SUN**	**SR**	**GB**	**IT**	AAM
Best	24	19	22	28	30	32	21	20	9	27	35	29	31	34	36	18	33	25
Mean	24	16	30	29	34	27	18	26	9	28	32	31	25	33	36	20	35	23
AdpT	22	32	20	28	31	29	23	8	6	24	30	34	27	35	36	11	26	19
SCut	23	16	33	31	34	32	11	30	17	24	18	22	20	27	36	3	35	15
AUC	20	25	23	33	31	35	17	9	10	28	29	26	30	32	34	11	36	21
MAE	25	33	11	21	16	20	31	8	7	28	36	30	34	32	15	19	17	27
NCB	23	21	19	18	29	26	13	33	11	25	31	28	27	32	35	24	34	36
Time	33	29	8	7	10	1	24	32	18	6	20	27	21	25	5	19	16	-

Fig. 48. Summary rankings of models under different evaluation metric. Fixation prediction models are shown in bold face.

results suggest that targets and distractors may share some common visual attributes and the intrinsic cues might be insufficient to distinguish them.

Compared to intrinsic cues, the usage of extrinsic cues such as salient object training data, similar images and saliency co-occurrence is still less explored. How to efficiently use these cues in different application scenarios remains an open question.

5.1.3 Heuristic vs. Learning From Data

Like the early years of other areas [260], [261], most existing salient object detection studies still focused on creating effective features and using heuristic model to detect salient object [55], [79], [89], [97]. To date, there are various of features have been shown to be helpful to salient object detection (see see Fig. 7 & 15), including local contrast, global contrast, edge density, background prior, focusses, objectness, convexity, spatial distribution, spareness, etc. It is becoming more and more challenging to design heuristic models which is able to fully explore the potentials of these rich features. The leading method in our benchmark (*i.e.* DRFI), discriminatively train a regression model to predict region saliency according to a 86 dimensional feature vector (including 26D local contrast, 26D background prior and 34D region properties). The impressive performance of this learning-based method suggested that data driven method is the right way to go for salient object detection.

The simplicity and training-free properties of many successful salient object detection models, has been an attracting advantage for their popularity in many application areas (see Sec. 2.3). By eliminating the requirement for training, third party applications could directly use those heuristic salient object detection method without preparing expensive training data. An emerging question is: for salient object detection, will the data-driven idea be conflict with the easy useability? Unlike other classical computer vision problems, e.g. generic object detection, classification, the dada-driven approach in salient object detection seems to have surprisingly good generalization ability. Despite the huge characteristics different among datasets we have evaluated (see Sec. 3), the DRFI approach was only trained on a small subset of **MSRA5K**, and it still consistently outperforms other methods when we test it directly on all other dataset. This encouraging results suggested that we might be able to further explore data driven based salient object detection without losing the simplicity and useability in the application side.

5.2 Salient Object Detection, Fixation Prediction, and Objectness Proposals

The attentive visual search mechanisms have been studied in different background and problem focuses: including salient object detection [1], [9], [89], [118], fixation prediction [2], [19], [61], and objectness proposals [58], [65]–[68]. Salient object detection models usually aim to detect only the most salient objects in a scene and segment the whole extent of those objects. Fixation prediction models typically try to predict where human looks, *i.e.*, a small set of fixation points. Both of these two types of methods output a single saliency map, where higher value in this map indicates the corresponding image pixel have more chance to belong to salient objects or be fixated. Both recall and precision are important for a high-quality saliency map. While both these two areas have made great progress in the last few decades and enabled many practical applications (Sec. 2.3), generating a single saliency map to indicate the locations of all objects in an image is still challenging and even impossible (*e.g.* for images with multiple objects occluding each other). Objectness proposal generation models typically aim at predicting a small set (typically a few hundreds or thousands) of candidate object bounding boxes or region proposals (often overlapped with each other). High recall at a small set of proposals is a major objective.

According to object-based attention theory [13], [262], [263], human brain groups similar pixels into proto-objects and the saliency of proto-objects is estimated and incorporated together. Strictly speaking, attentional focus and gaze shifts do not always coincide: attentional focus can be directed to new target without accompanied eye-movements [264], [265]. However, a strong correlation between fixations and salient objects exists and the definition of a salient object is highly consistent among human subjects [53], [266]. Objectness and salient object detection are highly correlated as well, saliency estimation is even explicitly used as a cue for objectness detection methods [16], [65]. Recent study [263] suggested that the unit of attention depends on the task, the field of view, and the observer's intention [267]. Attention might adopt a spatial-based behavior within complex extended objects, be object-based on the global scale, and be directed to any well-formed perceptually distinguishable surface, depending on which of these factors will dominate [268].

For the task of salient object detection, our comprehensive benchmark results suggested that fixation prediction models and objectness proposal methods on average perform worse than salient object detection models (Sec. 4.3). This is mainly due to the disparity among the design goals: fixation prediction models usually try to predict sparse pixels that may be fixated; salient object detection models try to segment the dense pixels that form the entire salient object; many objectness proposal methods even do not produce saliency maps. Having different priors of sparsity significantly limit models of one type to have good performance on the other task, and results in different ground-truth representation and preferred evaluation criteria [53], [252], [257]. The fact that state-of-the-art fixation prediction models often assign high saliency values at or near high-contrast edges prevents them get high performance on salient object detection tasks where segmenting the salient object as a whole is the goal [1], [89], [263], [269].

5.3 Dataset Bias

Datasets play as one of the most important reasons for the rapid progress in saliency detection researches. On one hand, they supply large scale training data and enable comparing performance of competing algorithms. On the other hand, each dataset is a specific/small sampling of the original huge/unlimited problem/application domain, and can easily contains certain degree of bias. Our large scale benchmark results suggest similar conclusions as in [1], [53]: *models performance drops significantly when migrating from current datasets, often with a single object, to more cluttered scenes with many objects.* To date, there seems to be a unanimous agreement on the presence of bias (i.e. skewness) in underlying structure of datasets.

Consequently, there are studies to address the effect of bias in visual datasets. For instance, Torralba & Efros identify three biases in computer vision datasets, namely: *selection bias, capture bias* and *negative set bias* [270]. Selection bias is caused by preference of a particular kind of image during data gathering. It results in qualitatively similar images in a dataset. This is evidenced by the strong color contrast (see [53], [252]) in most frequently used salient object benchmark dataset [79]. Thus two practices in dataset construction are proffered: i) *having independent image selection and annotation process* [53], and ii) *crossing the most salient object first and then segmenting it.* Negative set bias is the consequence of the lack of rich and unbiased negative set, *i.e.* one should avoid being focused on a particular image of interest and datasets should model the whole world. Negative set bias may affect the ground-truth by incorporating annotator's personal preference to some object types. Thus, having variety of images is motivated in such datasets. Capture bias conveys the effect of image composition on the dataset. The most popular kind of such a bias is the tendency of composing objects in the central region of the image, *i.e.* center bias. The existence of bias in a dataset makes the generic quantitative evaluation of models difficult and sometimes even misleading. For instance, a toy saliency model, which consists of a Gaussian blob at the center of image, often scores higher than many fixation prediction models [18], [147], [157].

Bias is often closely related with application tasks and sometimes could be deliberately utilized as a prior in a specific task in order to improve the performance of an algorithm. For instance, due to aesthetics reasons (*e.g.* rule of thirds), photographers tend to frame the salient object near center of the an image [157], [271]. From an application point of view, most images we are dealing with are intentionally captured by humans with salient object away from image borders (except for images from surveillance camera and driving recorder). In our large scale benchmark (see Sec. 4), all top performance algorithms use the location prior cues. Notice that although performance scores (F-measure and AUC) might drop significantly from easier dataset to more difficult dataset, the ranking of top scoring models are quite consistent.

5.4 Promising Future Directions

From the experimental results and discussions listed above, here we propose several promising research directions for constructing more effective models and benchmarks.

5.4.1 Beyond Working with Single Image

Most benchmarks and saliency models discussed in this study are about still images. Unfortunately, the motion information, which is one of the most important visual cues in human vision system, is ignored. As concluded in [272], motion strength and direction are two important preattentive features that are capable to capture our visual attention. Without these two features, existing image saliency models, which only consider the spatial features, often fail when applied on video frames. Therefore, it is necessary to study how to extend existing models to the spatiotemporal domain so as to address the problem of salient object detection in the real-world scenario.

Before launching the study of salient object detection in video, three challenges should be addressed: *public benchmark, neurobiological evidences* and *model efficiency*. As we proposed in the Benchmark Section, there is only one video benchmarks in the literature, while the videos are selected from very limited scenarios (e.g., cartoons and news). For these videos, only bounding boxes are provided for the key frame to roughly localize the salient object. Without a promising benchmark, it is infeasible to obtain fair comparisons of video saliency models (recall the booming of image saliency models after the publication of **MSRA-B** and **ASD**). Moreover, motion information is only heuristically used in existing studies due to the lack of neurobiological evidences on what constitutes salient motion. For example, some studies proposed to use motion consistency (e.g., [125], [211], [273]) as a saliency measure, while [36] argued that the inter-frame variation acts as surprise to attract human attention. However, these models often encounter scenarios that such consistency and surprise hypotheses fail to work. Actually, the direct correlation between motion and saliency is still unclear in neuroscience. Furthermore, one important aspect that video saliency detection differs from image saliency detection is the requirement on efficiency. Although off-line video saliency analysis is also acceptable in very limited scenarios, we need real-time saliency analysis to process live video streams in most cases. This poses a high standard that most existing image saliency models fail to meet: some models even need around one minute to process one 400×300 image (e.g., DRFI: 13.7s, CA: 49.0s, SVO: 56.5s). To sum up, these three challenges, especially the benchmark challenge, should be addressed first before the booming of video-based salient object detection models.

5.4.2 Gauging Progress and Model Comparison

Despite its popularity, this field still lacks a coherent and unified comparison framework which we addressed this shortcoming and discussed evaluation scores and datasets. One area of research would be designing scores for tackling dataset biases and evaluation of saliency segmentation maps with respect to ground-truth annotation similar to [258].

To steadily monitor and organize progress in this filed, we have launched a new website [274] for fair model evaluation and comparison, where authors could contribute by sharing their software, datasets, and scores. This has been already started in the object recognition community (PASCAL VOC challenge) and text information retrieval community (TREC datasets) as well as face recognition (e.g., FERET).

5.4.3 Other Directions

Traditionally, saliency models have added feature channels such as face, car, animal, human body and object center bias to better predict fixations. Explicit addition of these channels using object detectors to salient object detection models may improve the overall performance. Another future direction, similar to models mentioned in Sec. 2.1.4, will be combining several different saliency models to achieve higher performance. Since these models are based on different mechanisms, it is likely that their combination may increase accuracy.

Here we directly compared models against annotation data, one might also consider comparing models based on their accuracy in specific applications, for instance in image thumbnailing or object detection. Majority of existing models try to correctly segment the salient object region (often evaluated using ROC and F-measure). It would be interesting to evaluate models in terms of their accuracy in preserving boundary of objects similar to general segmentation algorithms (e.g., [14]). Currently, salient object detection algorithms assume that at least one salient object exists in each image. Some images, however, may not contain salient objects at all (see Fig. 43 for some examples). The performance of algorithms on such background images need to be investigated further.

An emerging trend is active salient object segmentation (see Sec. 2.1.4). Here the idea is to separate the detection of the most salient object in a scene from its segmentation. This trend can help tackle the center bias in datasets. Existing models have convoluted these two stages and have been very successful on the biased benchmarks. However, they may fail on complex scenes when there are multiple objects in a complex background. Further capitalizing on this idea can greatly help extend the applicability and performance of models to unconstrained conditions.

Some other remaining questions include: How many (salient) objects are necessarily to represent a scene? Will map smoothing change the scores and model ranking? How is salient object detection different form other fields? What is the best way to tackle center bias (due to photographer bias) in model evaluation? A collaborative engagement with other related fields such as visual attention, computer graphics, scene labeling and categorization, general segmentation, and object recognition can help to answer these questions and situate the field better.

6 Summary and Conclusion

In this paper, we exhaustively reviewed salient object detection models and closely related areas to it. We also benchmark a large number of salient object detection models over several datasets and discussed existing challenges.

The numerous works on salient object detection call for a methodological approach for evaluating results. We reviewed a large body of work in saliency modeling and discussed their pros and cons. We categorized the related works in four divisions including: *classic salient object/region detection and segmentation*, *fixation prediction*, *category-independent object measures or object proposals*, and *active segmentation* approaches. The last may not necessary be a saliency model but when the seed point is the most salient image location, then it becomes a salient object detection approach.

Detecting and segmenting salient objects are very useful for scene understanding. Objects in an image will automatically catch more attention than background stuff, such as grass, trees and sky. Therefore, if in the first place, we can detect all generic objects then we can perform detailed reasoning and scene understanding at the next stage. Compared to traditional special-purpose object detectors, salient object detection models are general, typically fast (which allows processing a large number of images with low cost), often without need for training or annotation.

Future works should focus on better situating salient object detection models among other related areas such as fixation prediction, objectness measures, and general segmentation algorithms. In particular, connections between salient object detection and fixation prediction models can help enhance performance of both types of models. In this regard, datasets that offer both salient object judgements of humans and eye movements are highly desirable. Conducting behavioral studies to understand how humans perceive and prioritize objects in scenes and how this concept is related to language, scene description, attributes, etc. can offer invaluable insights. Further, it will be rewarding to focus more on evaluating and comparing salient object models to gauge future progress. Tackling dataset biases such as center bias and selection bias and moving toward more challenging images is important. In such scenarios, two components i) detecting salient objects in a scene and ii) segmenting the extent of the objects efficiently are important (i.e., decoupling the two steps). In this regard, it is important to design challenging datasets which helps us move forward in this direction. Finally, the main remaining questions are: "what we want from salient object detection models and what is the best salient object detection algorithm?".

Although salient object detection and segmentation methods have made great strides in recent years, a very robust salient object that is able to get high quality results for nearly every image is still lacking. Even for humans, what is the most salient object in the image is sometimes a quite ambiguous question. To this end, a general suggestion:

> *Don't ask what segments can do for you, ask what you can do for the segments*[22]. — Jitendra Malik

[22] http://www.cs.berkeley.edu/~malik/student-tree-2010.pdf

is particularly important to build robust applications. For instance, when dealing with noisy Internet images, although salient detection and segmentation methods cannot guarantee robust performance on individual images, their efficiency and simplicity makes it possible to automatically process a large number of images, which can then be further filtered for reliability and accuracy, thus enabling many applications running robustly [132], [194], [195], [198], [200], [252] and even supports unsupervised learning [196].

Acknowledgments

Authors would like to thank anonymous reviewers for their helpful comments on the paper. Ali Borji was supported by Defense Advanced Research Projects Agency (government contract no. HR0011-10-C-0034), the National Science Foundation (CRCNS grant number BCS-0827764), the General Motors Corporation, and the Army Research Office (grant number W911NF-08-1-0360). Jia Li was supported in part by grants from the Chinese National Natural Science Foundation under contract No. 61370113.

References

[1] A. Borji, D. N. Sihite, and L. Itti, "Salient object detection: A benchmark," in *ECCV*, 2012, pp. 414–429.

[2] L. Itti, C. Koch, and E. Niebur, "A model of saliency-based visual attention for rapid scene analysis," *IEEE TPAMI*, 1998.

[3] A. M. Treisman and G. Gelade, "A feature-integration theory of attention," *Cognitive Psychology*, pp. 97–136, 1980.

[4] J. M. Wolfe, K. R. Cave, and S. L. Franzel, "Guided search: an alternative to the feature integration model for visual search." *Journal of Experimental Psychology: Human perception and performance*, vol. 15, no. 3, p. 419, 1989.

[5] C. Koch and S. Ullman, "Shifts in selective visual attention: towards the underlying neural circuitry," in *Matters of Intelligence*. Springer, 1987, pp. 115–141.

[6] D. Parkhurst, K. Law, and E. Niebur, "Modeling the role of salience in the allocation of overt visual attention," *Vision Research*, vol. 42, no. 1, pp. 107–123, 2002.

[7] N. D. Bruce and J. K. Tsotsos, "Saliency based on information maximization," pp. 155–162, 2005.

[8] T. Liu, J. Sun, N. Zheng, X. Tang, and H.-Y. Shum, "Learning to detect a salient object," in *CVPR*, 2007, pp. 1–8.

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

[10] R. Achanta, F. Estrada, P. Wils, and S. Süsstrunk, "Salient region detection and segmentation," in *Computer Vision Systems*. Springer, 2008, vol. 5008, pp. 66–75.

[11] Y.-F. Ma and H.-J. Zhang, "Contrast-based image attention analysis by using fuzzy growing," in *ACM Multimedia*, 2003.

[12] F. Liu and M. Gleicher, "Region enhanced scale-invariant saliency detection," in *ICME*, 2006, pp. 1477–1480.

[13] D. Walther and C. Koch, "Modeling attention to salient proto-objects," *Neural Networks*, vol. 19, no. 9, pp. 1395–1407, 2006.

[14] P. Arbelaez, M. Maire, C. Fowlkes, and J. Malik, "Contour detection and hierarchical image segmentation," *IEEE TPAMI*, vol. 33, no. 5, pp. 898–916, 2011.

[15] D. R. Martin, C. C. Fowlkes, and J. Malik, "Learning to detect natural image boundaries using local brightness, color, and texture cues," *IEEE TPAMI*, vol. 26, no. 5, pp. 530–549, 2004.

[16] B. Alexe, T. Deselaers, and V. Ferrari, "What is an object?" in *CVPR*, 2010, pp. 73–80.

[17] I. Endres and D. Hoiem, "Category independent object proposals," in *ECCV*. Springer, 2010, vol. 6315, pp. 575–588.

[18] T. Judd, K. Ehinger, F. Durand, and A. Torralba, "Learning to predict where humans look," in *ICCV*, 2009, pp. 2106–2113.

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

[20] A. Borji and L. Itti, "Exploiting local and global patch rarities for saliency detection," in *CVPR*, 2012, pp. 478–485.

[21] A. Borji, "Boosting bottom-up and top-down visual features for saliency estimation," in *CVPR*, 2012, pp. 438–445.

[22] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," in *CVPR*, vol. 1, 2001, pp. I–511.

[23] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan, "Object detection with discriminatively trained part-based models," *IEEE TPAMI*, pp. 1627–1645, 2010.

[24] M. Hayhoe and D. Ballard, "Eye movements in natural behavior," *Trends in cognitive sciences*, pp. 188–194, 2005.

[25] V. Navalpakkam and L. Itti, "Modeling the influence of task on attention," *Vision Research*, vol. 45, no. 2, pp. 205–231, 2005.

[26] A. Borji, D. Sihite, and L. Itti, "What/where to look next? modeling top-down visual attention in complex interactive environments," *IEEE TSMC*, vol. 44, no. 5, pp. 523–538, 2014.

[27] M. Spain and P. Perona, "Measuring and predicting object importance," *IJCV*, vol. 91, no. 1, pp. 59–76, 2011.

[28] A. C. Berg, T. L. Berg, H. Daume, J. Dodge, A. Goyal, X. Han, A. Mensch, M. Mitchell, A. Sood, K. Stratos *et al.*, "Understanding and predicting importance in images," in *CVPR*, 2012, pp. 3562–3569.

[29] B. M't Hart, H. C. Schmidt, C. Roth, and W. Einhäuser, "Fixations on objects in natural scenes: dissociating importance from salience," *Frontiers in psychology*, vol. 4, 2013.

[30] P. Isola, J. Xiao, A. Torralba, and A. Oliva, "What makes an image memorable?" in *CVPR*, 2011, pp. 145–152.

[31] R. Rosenholtz, Y. Li, and L. Nakano, "Measuring visual clutter," *Journal of Vision*, vol. 7, no. 2, 2007.

[32] Y.-G. Jiang, Y. Wang, R. Feng, X. Xue, Y. Zheng, and H. Yang, "Understanding and predicting interestingness of videos," *AAAI*, 2013.

[33] H. Katti, K. Y. Bin, T. S. Chua, and M. Kankanhalli, "Pre-attentive discrimination of interestingness in images," in *IEEE ICME*, 2008, pp. 1433–1436.

[34] M. Gygli, H. Grabner, H. Riemenschneider, F. Nater, and L. Van Gool, "The interestingness of images," *ICCV*, 2013.

[35] S. Dhar, V. Ordonez, and T. L. Berg, "High level describable attributes for predicting aesthetics and interestingness," in *CVPR*, 2011, pp. 1657–1664.

[36] L. Itti and P. Baldi, "Bayesian surprise attracts human attention," in *NIPS*, 2005, pp. 547–554.

[37] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE TIP*, vol. 13, no. 4, pp. 600–612, 2004.

[38] Z. Wang, A. C. Bovik, and L. Lu, "Why is image quality assessment so difficult?" in *IEEE ICASSP*, vol. 4, 2002.

[39] J. Vogel and B. Schiele, "A semantic typicality measure for natural scene categorization," in *Pattern Recognition*. Springer, 2004, vol. 3175, pp. 195–203.

[40] K. A. Ehinger, J. Xiao, A. Torralba, and A. Oliva, "Estimating scene typicality from human ratings and image features," in *Proceedings of the 33rd Annual Cognitive Science Conference*. Cognitive Science Society, Inc., 2011.

[41] A. Farhadi, I. Endres, D. Hoiem, and D. Forsyth, "Describing objects by their attributes," in *CVPR*, 2009, pp. 1778–1785.

[42] N. H. Mackworth and A. J. Morandi, "The gaze selects informative details within pictures," *Perception & Psychophysics*, vol. 2, no. 11, pp. 547–552, 1967.

[43] L. Elazary and L. Itti, "Interesting objects are visually salient," *Journal of Vision*, vol. 8, no. 3, pp. 3, 1–15, 2008.

[44] B. C. Russell, A. Torralba, K. P. Murphy, and W. T. Freeman, "Labelme: a database and web-based tool for image annotation," *International journal of computer vision*, vol. 77, no. 1-3, pp. 157–173, 2008.

[45] C. Masciocchi, S. Mihalas, D. Parkhurst, and E. Niebur, "Everyone knows what is interesting: Salient locations which should be fixated." *Journal of Vision*, vol. 9, pp. 1–22, 2009.

[46] A. Borji, D. N. Sihite, and L. Itti, "What stands out in a scene? a study of human explicit saliency judgment," *Vision Research*, vol. 91, no. 0, pp. 62–77, 2013.

[47] K. Koehler, F. Guo, S. Zhang, and M. P. Eckstein, "What do saliency models predict?" *Journal of Vision*, vol. 14, no. 3, p. 14, 2014.

[48] C. M. Masciocchi, S. Mihalas, D. Parkhurst, and E. Niebur, "Everyone knows what is interesting: Salient locations which should be fixated," *Journal of Vision*, vol. 9, no. 11, 2009.

[49] S. Goferman, L. Zelnik-Manor, and A. Tal, "Context-aware saliency detection," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 34, no. 10, pp. 1915–1926, 2012.

[50] L. Itti and M. A. Arbib, "Attention and the minimal subscene," *Action to language via the mirror neuron system*, pp. 289–346, 2006.

[51] G. Kulkarni, V. Premraj, S. Dhar, S. Li, Y. Choi, A. C. Berg, and T. L. Berg, "Baby talk: Understanding and generating simple image descriptions," in *CVPR*, 2011, pp. 1601–1608.

[52] A. K. Mishra, Y. Aloimonos, L. F. Cheong, and A. Kassim, "Active visual segmentation," *IEEE TPAMI*, vol. 34, no. 4, pp. 639–653, 2012.

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

[54] A. Borji, "What is a salient object? a dataset and a baseline model for salient object detection," in *IEEE Transactions on Image Processing*, 2014, p. submitted.

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

[56] D. Comaniciu and P. Meer, "Mean shift: A robust approach toward feature space analysis," *IEEE TPAMI*, vol. 24, no. 5, pp. 603–619, 2002.

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

[58] M.-M. Cheng, Z. Zhang, W.-Y. Lin, and P. H. S. Torr, "BING: Binarized normed gradients for objectness estimation at 300fps," in *CVPR*, vol. 2, 2014, p. 4.

[59] S. Avidan and A. Shamir, "Seam carving for content-aware image resizing," in *ACM SIGGRAPH, 2007*. New York, NY, USA: ACM, 2007.

[60] G.-X. Zhang, M.-M. Cheng, S.-M. Hu, and R. R. Martin, "A shape-preserving approach to image resizing," *Computer Graphics Forum*, vol. 28, no. 7, pp. 1897–1906, 2009.

[61] X. Hou, J. Harel, and C. Koch, "Image signature: Highlighting sparse salient regions," *IEEE TPAMI*, vol. 34, no. 1, pp. 194–201, 2012.

[62] H. J. Seo and P. Milanfar, "Static and space-time visual saliency detection by self-resemblance," *Journal of Vision*, vol. 9, no. 12, pp. 15, 1–27, 2009.

[63] E. Erdem and A. Erdem, "Visual saliency estimation by nonlinearly integrating features using region covariances," *Journal of Vision*, vol. 13, no. 4, pp. 11, 1–20, 2013.

[64] H.-D. Cheng, X. Jiang, Y. Sun, and J. Wang, "Color image segmentation: advances and prospects," *Pattern Recognition*, vol. 34, no. 12, pp. 2259–2281, 2001.

[65] B. Alexe, T. Deselaers, and V. Ferrari, "Measuring the objectness of image windows," *IEEE TPAMI*, vol. 34, no. 11, 2012.

[66] I. Endres and D. Hoiem, "Category-independent object proposals with diverse ranking," *IEEE TPAMI*, to appear.

[67] J. Uijlings, K. van de Sande, T. Gevers, and A. Smeulders, "Selective search for object recognition," *IJCV*, vol. 104, no. 2, pp. 154–171, 2013.

[68] Z. Zhang, J. Warrell, and P. H. Torr, "Proposal generation for object detection using cascaded ranking svms," in *CVPR*, 2011, pp. 1497–1504.

[69] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *CVPR*, vol. 1, 2005, pp. 886–893.

[70] H. Teuber, "Physiological psychology," *Annual Review of Psychology*, vol. 6, no. 1, pp. 267–296, 1955.

[71] R. Desimone and J. Duncan, "Neural mechanisms of selective visual attention," *Annual review of neuroscience*, vol. 18, no. 1, pp. 193–222, 1995.

[72] T. Dean, M. A. Ruzon, M. Segal, J. Shlens, S. Vijayanarasimhan, and J. Yagnik, "Fast, accurate detection of 100,000 object classes on a single machine," in *CVPR*, 2013.

[73] X. Wang, M. Yang, S. Zhu, and Y. Lin, "Regionlets for generic object detection," *ICCV*, 2013.

[74] K.-Y. Chang, T.-L. Liu, H.-T. Chen, and S.-H. Lai, "Fusing generic objectness and visual saliency for salient object detection," in *ICCV*, 2011, pp. 914–921.

[75] G. Hua, Z. Liu, Z. Zhang, and Y. Wu, "Iterative local-global energy minimization for automatic extraction of objects of interest," *IEEE TPAMI*, vol. 28, no. 10, pp. 1701–1706, 2006.

[76] B. C. Ko and J.-Y. Nam, "Automatic object-of-interest segmentation from natural images," in *International Conference on Pattern Recognition (ICPR)*, vol. 4, 2006, pp. 45–48.

[77] M. Allili and D. Ziou, "Object of interest segmentation and tracking by using feature selection and active contours," in *CVPR*, 2007, pp. 1–8.

[78] Y. Hu, D. Rajan, and L.-T. Chia, "Robust subspace analysis for detecting visual attention regions in images," in *ACM Multimedia*, ser. MULTIMEDIA '05. New York, NY, USA: ACM, 2005, pp. 716–724.

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

[80] R. Vidal, Y. Ma, and S. Sastry, "Generalized principal component analysis (gpca)," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 27, no. 12, pp. 1945–1959, 2005.

[81] P. L. Rosin, "A simple method for detecting salient regions," *Pattern Recognition*, vol. 42, no. 11, pp. 2363–2371, 2009.

[82] R. Valenti, N. Sebe, and T. Gevers, "Image saliency by isocentric curvedness and color," in *ICCV*, 2009, pp. 2185–2192.

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

[84] X. Li, Y. Li, C. Shen, A. R. Dick, and A. van den Hengel, "Contextual hypergraph modeling for salient object detection," in *ICCV*, 2013, pp. 3328–3335.

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

[86] A. Borji and L. Itti, "State-of-the-art in visual attention modeling," *IEEE TPAMI*, vol. 35, no. 1, pp. 185–207, 2013.

[87] H. Yu, J. Li, Y. Tian, and T. Huang, "Automatic interesting object extraction from images using complementary saliency maps," in *Proceedings of the International Conference on Multimedia*, ser. MULTIMEDIA '10. New York, NY, USA: ACM, 2010, pp. 891–894.

[88] Z. Yu and H.-S. Wong, "A rule based technique for extraction of visual attention regions based on real-time clustering," *IEEE Transactions on Multimedia*, vol. 9, no. 4, pp. 766–784, 2007.

[89] M.-M. Cheng, G.-X. Zhang, N. Mitra, X. Huang, and S.-M. Hu, "Global contrast based salient region detection," in *CVPR*, 2014.

[90] Y. Lu, W. Zhang, H. Lu, and X. Xue, "Salient object detection using concavity context," in *ICCV*, 2011, pp. 233–240.

[91] H. Jiang, J. Wang, Z. Yuan, T. Liu, and N. Zheng, "Automatic salient object segmentation based on context and shape prior," in *British Machine Vision Conference (BMVC)*. BMVA Press, 2011, pp. 110.1–110.12.

[92] X. Shen and Y. Wu, "A unified approach to salient object detection via low rank matrix recovery," in *CVPR*, 2012, pp. 853–860.

[93] Y. Wei, F. Wen, W. Zhu, and J. Sun, "Geodesic saliency using background priors," in *ECCV*. Springer, 2012, vol. 7574, pp. 29–42.

[94] J. Li, Y. Tian, L. Duan, and T. Huang, "Estimating visual saliency through single image optimization," *IEEE Signal Processing Letters*, vol. 20, no. 9, pp. 845–848, 2013.

[95] Y. Xie, H. Lu, and M.-H. Yang, "Bayesian saliency via low and mid level cues," *IEEE TIP*, vol. 22, no. 5, pp. 1689–1698, 2013.

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

[97] 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.

[98] K. Shi, K. Wang, J. Lu, and L. Lin, "Pisa: Pixelwise image saliency by aggregating complementary appearance contrast measures with spatial priors," in *CVPR*, 2013, pp. 2115–2122.

[99] C. Scharfenberger, A. Wong, K. Fergani, J. S. Zelek, and D. A. Clausi, "Statistical textural distinctiveness for salient region detection in natural images," in *CVPR*, 2013, pp. 979–986.

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

[101] 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.

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

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

[104] Y. Jia and M. Han, "Category-independent object-level saliency detection," in *ICCV*, 2013.

[105] W. Zou, K. Kpalma, Z. Liu, J. Ronsin *et al.*, "Segmentation driven low-rank matrix recovery for saliency detection," in *British Machine Vision Conference (BMVC)*, 2013, pp. 1–13.

[106] H. Peng, B. Li, R. Ji, W. Hu, W. Xiong, and C. Lang, "Salient object detection via low-rank and structured sparse matrix decomposition," in *Twenty-Seventh AAAI Conference on Artificial Intelligence*, 2013.

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

[108] R. Liu, J. Cao, G. Zhong, Z. Lin, S. Shan, and Z. Su, "Adaptive partial differential equation learning for visual saliency detection," in *CVPR*, 2014.

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

[110] H. Liu, S. Jiang, Q. Huang, C. Xu, and W. Gao, "Region-based visual attention analysis with its application in image browsing on small displays," in *Proceedings of the 15th Annual ACM International Conference on Multimedia*, ser. MULTIMEDIA '07. New York, NY, USA: ACM, 2007, pp. 305–308.

[111] P. F. Felzenszwalb and D. P. Huttenlocher, "Efficient graph-based image segmentation," *IJCV*, vol. 59, no. 2, pp. 167–181, Sep. 2004.

[112] D. Comaniciu and P. Meer, "Mean shift: A robust approach toward feature space analysis," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 24, no. 5, pp. 603–619, 2002.

[113] A. Levinshtein, A. Stere, K. N. Kutulakos, D. J. Fleet, S. J. Dickinson, and K. Siddiqi, "Turbopixels: Fast superpixels using geometric flows," *IEEE TPAMI*, vol. 31, no. 12, pp. 2290–2297, 2009.

[114] J. Zhang and S. Sclaroff, "Saliency detection: A boolean map approach," in *ICCV*, 2013, pp. 153–160.

[115] N. Li, J. Ye, Y. Ji, H. Ling, and J. Yu, "Saliency detection on light fields," in *CVPR*, 2014.

[116] P. Khuwuthyakorn, A. Robles-Kelly, and J. Zhou, "Object of interest detection by saliency learning," in *ECCV*. Springer, 2010, vol. 6312, pp. 636–649.

[117] P. Mehrani and O. Veksler, "Saliency segmentation based on learning and graph cut refinement." in *British Machine Vision Conference (BMVC)*, 2010, pp. 1–12.

[118] H. Jiang, J. Wang, Z. Yuan, Y. Wu, N. Zheng, and S. Li, "Salient object detection: A discriminative regional feature integration approach," in *IEEE CVPR*, 2013, pp. 2083–2090.

[119] S. Lu, V. Mahadevan, and N. Vasconcelos, "Learning optimal seeds for diffusion-based salient object detection," in *CVPR*, 2014.

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

[121] L. Marchesotti, C. Cifarelli, and G. Csurka, "A framework for visual saliency detection with applications to image thumbnailing," in *ICCV*, 2009, pp. 2232–2239.

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

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

[124] P. Siva, C. Russell, T. Xiang, and L. Agapito, "Looking beyond the image: Unsupervised learning for object saliency and detection," in *CVPR*, 2013, pp. 3238–3245.

[125] Y. Zhai and M. Shah, "Visual attention detection in video sequences using spatiotemporal cues," in *Proceedings of the 14th Annual ACM International Conference on Multimedia*, ser. MULTIMEDIA '06. New York, NY, USA: ACM, 2006, pp. 815–824.

[126] T. Liu, N. Zheng, W. Ding, and Z. Yuan, "Video attention: Learning to detect a salient object sequence," in *International Conference on Pattern Recognition (ICPR)*, 2008, pp. 1–4.

[127] E. Rahtu, J. Kannala, M. Salo, and J. Heikkilä, "Segmenting salient objects from images and videos," in *ECCV*. Springer, 2010, vol. 6315, pp. 366–379.

[128] S. Bin, Y. Li, L. Ma, W. Wu, and Z. Xie, "Temporally coherent video saliency using regional dynamic contrast," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 23, no. 12, pp. 2067–2076, 2013.

[129] H. Li and K. N. Ngan, "A co-saliency model of image pairs," *IEEE TIP*, vol. 20, no. 12, pp. 3365–3375, 2011.

[130] K.-Y. Chang, T.-L. Liu, and S.-H. Lai, "From co-saliency to co-segmentation: An efficient and fully unsupervised energy minimization model," in *CVPR*, 2011, pp. 2129–2136.

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

[132] M.-M. Cheng, N. J. Mitra, X. Huang, and S.-M. Hu, "Salientshape: group saliency in image collections," *The Visual Computer*, vol. 30, no. 4, pp. 443–453, 2014.

[133] Y. Niu, Y. Geng, X. Li, and F. Liu, "Leveraging stereopsis for saliency analysis," in *CVPR*, 2012, pp. 454–461.

[134] K. Desingh, K. M. Krishna, D. Rajan, and C. Jawahar, "Depth really matters: Improving visual salient region detection with depth," in *British Machine Vision Conference (BMVC)*, 2013.

[135] A. Oliva and A. Torralba, "Modeling the shape of the scene: A holistic representation of the spatial envelope," *IJCV*, no. 3, pp. 145–175, 2001.

[136] C. Rother, T. P. Minka, A. Blake, and V. Kolmogorov, "Cosegmentation of image pairs by histogram matching - incorporating a global constraint into mrfs," in *CVPR*, 2006, pp. 993–1000.

[137] D. Batra, A. Kowdle, D. Parikh, J. Luo, and T. Chen, "icoseg: Interactive co-segmentation with intelligent scribble guidance," in *CVPR*, 2010, pp. 3169–3176.

[138] L. Mukherjee, V. Singh, and J. Peng, "Scale invariant cosegmentation for image groups," in *CVPR*, 2011, pp. 1881–1888.

[139] G. Kim, E. P. Xing, F.-F. Li, and T. Kanade, "Distributed cosegmentation via submodular optimization on anisotropic diffusion," in *ICCV*, 2011, pp. 169–176.

[140] C. Lang, T. V. Nguyen, H. Katti, K. Yadati, M. S. Kankanhalli, and S. Yan, "Depth matters: Influence of depth cues on visual saliency," in *ECCV*, 2012, pp. 101–115.

[141] J. Feng, Y. Wei, L. Tao, C. Zhang, and J. Sun, "Salient object detection by composition," in *ICCV*, 2011, pp. 1028–1035.

[142] P. Wang, J. Wang, G. Zeng, J. Feng, H. Zha, and S. Li, "Salient object detection for searched web images via global saliency," in *CVPR*, 2012, pp. 3194–3201.

[143] J. Shi and J. Malik, "Normalized cuts and image segmentation," *IEEE TPAMI*, vol. 22, no. 8, pp. 888–905, 2000.

[144] L. Wang, J. Xue, N. Zheng, and G. Hua, "Automatic salient object extraction with contextual cue," in *ICCV*, 2011, pp. 105–112.

[145] Z. Tu and X. Bai, "Auto-context and its application to high-level vision tasks and 3d brain image segmentation," *IEEE TPAMI*, vol. 32, no. 10, pp. 1744–1757, 2010.

[146] Y. Tian, J. Li, S. Yu, and T. Huang, "Learning complementary saliency priors for foreground object segmentation in complex scenes," 2014.

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

[148] A. Borji, H. R. Tavakoli, D. N. Sihite, and L. Itti, "Analysis of scores, datasets, and models in visual saliency prediction," in *ICCV*, 2013, pp. 921–928.

[149] A. Nuthmann and J. M. Henderson, "Object-based attentional selection in scene viewing," *Journal of Vision*, vol. 10, no. 8, 2010.

[150] S. Barthelmé, H. Trukenbrod, R. Engbert, and F. Wichmann, "Modeling fixation locations using spatial point processes," *Journal of Vision*, vol. 13, no. 12, p. 1, 2013.

[151] M. Dorr, T. Martinetz, K. R. Gegenfurtner, and E. Barth, "Variability of eye movements when viewing dynamic natural scenes," *Journal of Vision*, vol. 10, no. 10, p. 28, 2010.

[152] S. Marat, T. Ho Phuoc, L. Granjon, N. Guyader, D. Pellerin, and A. Guérin-Dugué, "Modelling spatio-temporal saliency to predict gaze direction for short videos," *IJCV*, vol. 82, no. 3, pp. 231–243, May 2009.

[153] E. Vig, M. Dorr, T. Martinetz, and E. Barth, "Intrinsic dimensionality predicts the saliency of natural dynamic scenes," *IEEE TPAMI*, vol. 34, no. 6, pp. 1080–1091, 2012.

[154] L. Itti, "Quantitative modelling of perceptual salience at human eye position," *Visual cognition*, vol. 14, no. 4-8, pp. 959–984, 2006.

[155] L. Itti, N. Dhavale, and F. Pighin, "Realistic avatar eye and head animation using a neurobiological model of visual attention," in *Optical Science and Technology, SPIE's 48th Annual Meeting*, vol. 5200. International Society for Optics and Photonics, 2004, pp. 64–78.

[156] W. Kienzle, M. O. Franz, B. Schölkopf, and F. A. Wichmann, "Center-surround patterns emerge as optimal predictors for human saccade targets," *Journal of Vision*, vol. 9, no. 5, p. 7, 2009.

[157] B. W. Tatler, "The central fixation bias in scene viewing: Selecting an optimal viewing position independently of motor biases and image feature distributions," *Journal of Vision*, vol. 7, no. 14, p. 4, 2007.

[158] W. Einhâuser, M. Spain, and P. Perona, "Objects predict fixations better than early saliency," *Journal of Vision*, vol. 8, no. 14, pp. 18, 1–26, 2008.

[159] A. Borji, D. N. Sihite, and L. Itti, "Objects do not predict fixations better than early saliency: A re-analysis of einhäuser et al.'s data," *Journal of Vision*, vol. 13, no. 10, p. 18, 2013.

[160] J. Xu, M. Jiang, S. Wang, M. S. Kankanhalli, and Q. Zhao, "Predicting human gaze beyond pixels," *Journal of Vision*, vol. 14, no. 1, pp. 1–20, 2014.

[161] A. Borji, D. Parks, and L. Itti, "Complementary effects of gaze direction and early saliency in guiding fixations during free-viewing," *Journal of Vision*, 2014.

[162] H.-C. Wang and M. Pomplun, "The attraction of visual attention to texts in real-world scenes," *Journal of Vision*, vol. 12, no. 6, p. 26, 2012.

[163] M. Cerf, J. Harel, W. Einhäuser, and C. Koch, "Predicting human gaze using low-level saliency combined with face detection." in *NIPS*, 2007, pp. 241–248.

[164] J. Li, Y. Tian, and T. Huang, "Visual saliency with statistical priors," *IJCV*, vol. 107, no. 3, pp. 239–253, 2014.

[165] R. Subramanian, D. Shankar, N. Sebe, and D. Melcher, "Emotion modulates eye movement patterns and subse-quent memory for the gist and details of movie scenes," *Journal of Vision*, 2014.

[166] M. S. Castelhano, M. Wieth, and J. M. Henderson, "I see what you see: Eye movements in real-world scenes are affected by perceived direction of gaze," in *Attention in cognitive systems. Theories and systems from an interdisciplinary viewpoint*. Springer, 2007, pp. 251–262.

[167] J. Shen and L. Itti, "Top-down influences on visual attention during listening are modulated by observer sex," *Vision Research*, vol. 65, pp. 62–76, Jul 2012, front cover of journal, July 2012.

[168] H. F. Chua, J. E. Boland, and R. E. Nisbett, "Cultural variation in eye movements during scene perception," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 102, no. 35, pp. 12 629–12 633, 2005.

[169] J. Harel, C. Koch, and P. Perona, "Graph-based visual saliency," in *NIPS*, 2007, pp. 545–552.

[170] L. Zhang, M. H. Tong, T. K. Marks, H. Shan, and G. W. Cottrell, "Sun: A bayesian framework for saliency using natural statistics," *Journal of Vision*, vol. 8, no. 7, pp. 32, 1–20, 2008.

[171] N. Murray, M. Vanrell, X. Otazu, and C. A. Parraga, "Saliency estimation using a non-parametric low-level vision model," in *CVPR*, 2011, pp. 433–440.

[172] Z. Wu and R. M. Leahy, "An optimal graph theoretic approach to data clustering: Theory and its application to image segmentation," *IEEE TPAMI*, vol. 15, no. 11, pp. 1101–1113, 1993.

[173] S. Wang and J. M. Siskind, "Image segmentation with ratio cut," *IEEE TPAMI*, vol. 25, no. 6, pp. 675–690, 2003.

[174] A. Vedaldi and S. Soatto, "Quick shift and kernel methods for mode seeking," in *ECCV*, vol. 5305, 2008, pp. 705–718.

[175] D. Hoiem, A. A. Efros, and M. Hebert, "Recovering surface layout from an image," *IJCV*, vol. 75, no. 1, pp. 151–172, 2007.

[176] T. Malisiewicz and A. A. Efros, "Improving spatial support for objects via multiple segmentations," in *British Machine Vision Conference (BMVC)*, 2007, pp. 1–10.

[177] Y. Li, J. Sun, C.-K. Tang, and H.-Y. Shum, "Lazy snapping," *ACM Transactions on Graphics*, vol. 23, no. 3, pp. 303–308, 2004.

[178] C. Rother, V. Kolmogorov, and A. Blake, ""grabcut": interactive foreground extraction using iterated graph cuts," *ACM Transactions on Graphics*, vol. 23, no. 3, pp. 309–314, 2004.

[179] J. Wang and M. F. Cohen, "Image and video matting: A survey," *Foundations and Trends in Computer Graphics and Vision*, vol. 3, no. 2, pp. 97–175, 2007.

[180] M. Maire, S. X. Yu, and P. Perona, "Hierarchical scene annotation," in *British Machine Vision Conference (BMVC)*, 2013, pp. 84.1–84.11.

[181] E. Rahtu, J. Kannala, and M. Blaschko, "Learning a category independent object detection cascade," in *ICCV*, 2011, pp. 1052–1059.

[182] K. E. A. van de Sande, J. R. R. Uijlings, T. Gevers, and A. W. M. Smeulders, "Segmentation as selective search for object recognition," in *ICCV*, 2011, pp. 1879–1886.

[183] M. Ristin, J. Gall, and L. J. V. Gool, "Local context priors for object proposal generation," in *ACCV*, 2012, pp. 57–70.

[184] S. Manen, M. Guillaumin, and L. J. V. Gool, "Prime object proposals with randomized prim's algorithm," in *ICCV*, 2013, pp. 2536–2543.

[185] D. Erhan, C. Szegedy, A. Toshev, and D. Anguelov, "Scalable object detection using deep neural networks," in *CVPR*, 2014.

[186] C. L. Zitnick and P. Dollar, "Edge boxes: Locating object proposals from edges," in *ECCV*, 2014.

[187] J. Carreira and C. Sminchisescu, "Constrained parametric min-cuts for automatic object segmentation," in *CVPR*, 2010, pp. 3241–3248.

[188] P. Rantalankila and E. R. Juho Kannala, "Generating object segmentation proposals using global and local search," in *CVPR*, 2014.

[189] P. Arbelaez, J. Pont-Tuset, J. Barron, F. Marqués, and J. Malik, "Multiscale combinatorial grouping," in *CVPR*, 2014.

[190] J. Kim and K. Grauman, "Shape sharing for object segmentation," in *ECCV (7)*, 2012, pp. 444–458.

[191] H. Kang, A. Efros, T. Kanade, and M. Hebert, "Data-driven objectness," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 99, no. PrePrints, p. 1, 2014.

[192] G. Sharir and T. Tuytelaars, "Video object proposals." in *CVPR Workshops*. IEEE, 2012, pp. 9–14.

[193] A. Karpathy, S. Miller, and L. Fei-Fei, "Object discovery in 3d scenes via shape analysis," in *ICRA*, 2013, pp. 2088–2095.

[194] H. Huang, L. Zhang, and H.-C. Zhang, "Arcimboldo-like collage using internet images," *ACM Transactions on Graphics*, vol. 30, no. 6, p. 155, 2011.

[195] H. Liu, L. Zhang, and H. Huang, "Web-image driven best views of 3d shapes," *The Visual Computer*, vol. 28, no. 3, pp. 279–287, 2012.

[196] J.-Y. Zhu, J. Wu, Y. Wei, E. Chang, and Z. Tu, "Unsupervised object class discovery via saliency-guided multiple class learning," in *IEEE CVPR*. IEEE, 2012, pp. 3218–3225.

[197] R. Margolin, L. Zelnik-Manor, and A. Tal, "Saliency for image manipulation," *The Visual Computer*, pp. 1–12, 2013.

[198] T. Chen, M.-M. Cheng, P. Tan, A. Shamir, and S.-M. Hu, "Sketch2photo: internet image montage," *ACM Transactions on Graphics*, vol. 28, no. 5, p. 124, 2009.

[199] C. Goldberg, T. Chen, F.-L. Zhang, A. Shamir, and S.-M. Hu, "Data-driven object manipulation in images," *Computer Graphics Forum*, vol. 31, no. 2pt1, pp. 265–274, 2012.

[200] A. Y.-S. Chia, S. Zhuo, R. K. Gupta, Y.-W. Tai, S.-Y. Cho, P. Tan, and S. Lin, "Semantic colorization with internet images," *ACM Transactions on Graphics*, vol. 30, no. 6, p. 156, 2011.

[201] U. Rutishauser, D. Walther, C. Koch, and P. Perona, "Is bottom-up attention useful for object recognition?" in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2004*, vol. 2, 2004, pp. II–37.

[202] C. Kanan and G. Cottrell, "Robust classification of objects, faces, and flowers using natural image statistics," in *CVPR*, 2010, pp. 2472–2479.

[203] F. Moosmann, D. Larlus, and F. Jurie, "Learning saliency maps for object categorization," in *ECCV International Workshop on The Representation and Use of Prior Knowledge in Vision*, 2006.

[204] A. Borji, M. N. Ahmadabadi, and B. N. Araabi, "Cost-sensitive learning of top-down modulation for attentional control," *Machine Vision and Applications*, vol. 22, no. 1, pp. 61–76, 2011.

[205] A. Borji and L. Itti, "Scene classification with a sparse set of salient regions," in *IEEE International Conference on Robotics and Automation (ICRA)*, 2011, pp. 1902–1908.

[206] H. Shen, S. Li, C. Zhu, H. Chang, and J. Zhang, "Moving object detection in aerial video based on spatiotemporal saliency," *Chinese Journal of Aeronautics*, vol. 26, no. 5, pp. 1211–1217, 2013.

[207] Z. Ren, S. Gao, L.-T. Chia, and I. Tsang, "Region-based saliency detection and its application in object recognition," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. PP, no. 99, pp. 1–1, 2013.

[208] M. Rudinac and P. P. Jonker, "Saliency based method for object localization," in *Proceedings of the 16th annual conference of the advanced school for computing and imaging*, 2010.

[209] C. Guo and L. Zhang, "A novel multiresolution spatiotemporal saliency detection model and its applications in image and video compression," *IEEE TIP*, vol. 19, no. 1, pp. 185–198, 2010.

[210] L. Itti, "Automatic foveation for video compression using a neurobiological model of visual attention," *IEEE TIP*, vol. 13, no. 10, pp. 1304–1318, 2004.

[211] Y.-F. Ma, X.-S. Hua, L. Lu, and H.-J. Zhang, "A generic framework of user attention model and its application in video summarization," *IEEE Transactions on Multimedia*, vol. 7, no. 5, pp. 907–919, 2005.

[212] P. Bodesheim, "Spectral clustering of rois for object discovery," in *DAGM-Symposium*, ser. Lecture Notes in Computer Science, R. Mester and M. Felsberg, Eds., vol. 6835, 2011, pp. 450–455.

[213] Y. J. Lee, J. Ghosh, and K. Grauman, "Discovering important people and objects for egocentric video summarization," in *CVPR*, 2012, pp. 1346–1353.

[214] Q.-G. Ji, Z.-D. Fang, Z.-H. Xie, and Z.-M. Lu, "Video abstraction based on the visual attention model and online clustering," *Signal Processing: Image Communication*, vol. 28, no. 3, pp. 241 – 253, 2012.

[215] S. Goferman, A. Tal, and L. Zelnik-Manor, "Puzzle-like collage," *Computer Graphics Forum*, vol. 29, no. 2, pp. 459–468, 2010.

[216] J. Wang, L. Quan, J. Sun, X. Tang, and H.-Y. Shum, "Picture collage," in *CVPR*, vol. 1, 2006, pp. 347–354.

[217] A. Ninassi, O. Le Meur, P. Le Callet, and D. Barbba, "Does where you gaze on an image affect your perception of quality? applying visual attention to image quality metric," in *IEEE ICIP*, vol. 2, 2007, pp. II–169.

[218] H. Liu and I. Heynderickx, "Studying the added value of visual attention in objective image quality metrics based on eye movement data," in *Image Processing (ICIP), 2009 16th IEEE International Conference on*. IEEE, 2009, pp. 3097–3100.

[219] K. B. Raja and M. Pedersen, "Artifact detection in gamut mapped images using saliency," in *Colour and Visual Computing Symposium (CVCS), 2013*, 2013, pp. 1–6.

[220] A. Li, X. She, and Q. Sun, "Color image quality assessment combining saliency and fsim," in *Fifth International Conference on Digital Image Processing*, vol. 8878. International Society for Optics and Photonics, 2013, pp. 88 780I–88 780I.

[221] M. Donoser, M. Urschler, M. Hirzer, and H. Bischof, "Saliency driven total variation segmentation," in *ICCV*, 2009, pp. 817–824.

[222] Q. Li, Y. Zhou, and J. Yang, "Saliency based image segmentation," in *International Conference on Multimedia Technology (ICMT)*, 2011, pp. 5068–5071.

[223] C. Qin, G. Zhang, Y. Zhou, W. Tao, and Z. Cao, "Integration of the saliency-based seed extraction and random walks for image segmentation," *Neurocomputing*, vol. 129, no. 0, pp. 378 – 391, 2013.

[224] L. Duan, J. Gu, Z. Yang, J. Miao, W. Ma, and C. Wu, "Bio-inspired visual attention model and saliency guided object segmentation," in *Genetic and Evolutionary Computing*. Springer, 2014, pp. 291–298.

[225] M. Johnson-Roberson, J. Bohg, M. Bjorkman, and D. Kragic, "Attention-based active 3d point cloud segmentation," in *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2010, pp. 1165–1170.

[226] S. Feng, D. Xu, and X. Yang, "Attention-driven salient edge (s) and region (s) extraction with application to cbir," *Signal Processing*, vol. 90, no. 1, pp. 1–15, 2010.

[227] J. Sun, J. Xie, J. Liu, and T. Sikora, "Image adaptation and dynamic browsing based on two-layer saliency combination," *IEEE Transactions on Broadcasting*, vol. 59, no. 4, pp. 602–613, 2013.

[228] L. Li, S. Jiang, Z. Zha, Z. Wu, and Q. Huang, "Partial-duplicate image retrieval via saliency-guided visually matching," *IEEE MultiMedia*, vol. 20, no. 3, pp. 13–23, 2013.

[229] S. Stalder, H. Grabner, and L. Van Gool, "Dynamic objectness for adaptive tracking," in *Computer Vision–ACCV 2012*. Springer, 2013, pp. 43–56.

[230] J. Li, M. Levine, X. An, X. Xu, and H. He, "Visual saliency based on scale-space analysis in the frequency domain," *IEEE TPAMI*, vol. 35, no. 4, pp. 996–1010, 2013.

[231] G. M. García, D. A. Klein, J. Stückler, S. Frintrop, and A. B. Cremers, "Adaptive multi-cue 3d tracking of arbitrary objects," in *Pattern Recognition*. Springer, 2012, pp. 357–366.

[232] A. Borji, S. Frintrop, D. N. Sihite, and L. Itti, "Adaptive object tracking by learning background context," in *CVPR*, 2012, pp. 23–30.

[233] D. A. Klein, D. Schulz, S. Frintrop, and A. B. Cremers, "Adaptive real-time video-tracking for arbitrary objects," in *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2010, pp. 772–777.

[234] S. Frintrop and M. Kessel, "Most salient region tracking," in *IEEE International Conference on Robotics and Automation, 2009*, 2009, pp. 1869–1874.

[235] D. Sidibé, D. Fofi, F. Mériaudeau *et al.*, "Using visual saliency for object tracking with particle filters," *Eusipco 2010*, pp. 1776–1780, 2010.

[236] G. Zhang, Z. Yuan, N. Zheng, X. Sheng, and T. Liu, "Visual saliency based object tracking," in *Computer Vision–ACCV 2009*. Springer, 2010, vol. 5995, pp. 193–203.

[237] S. Frintrop, G. M. Garcıa, and A. B. Cremers, "A cognitive approach for object discovery," in *International Conference on Pattern Recognition (ICPR)*, 2014.

[238] D. Meger, P.-E. Forssén, K. Lai, S. Helmer, S. McCann, T. Southey, M. Baumann, J. J. Little, and D. G. Lowe, "Curious george: An attentive semantic robot," *Robotics and Autonomous Systems*, vol. 56, no. 6, pp. 503–511, 2008.

[239] J. Shin, R. Triebel, R. Y. Siegwart, R. Y. Siegwart, and R. Y. Siegwart, *Unsupervised 3d object discovery and categorization for mobile robots*. Eidgenössische Technische Hochschule Zürich, Autonomous System Lab, 2011.

[240] Y. Sugano, Y. Matsushita, and Y. Sato, "Calibration-free gaze sensing using saliency maps," in *CVPR*, 2010, pp. 2667–2674.

[241] Y. Cai, Y. Yang, A. G. Hauptmann, and H. D. Wactlar, "A cognitive assistive system for monitoring the use of home medical devices," in *ACM international workshop on Multimedia indexing and information retrieval for healthcare*. ACM, 2013, pp. 59–66.

[242] Z. Yang, D. Li, J. Wang, and X. Li, "Saliency detection based on manifold learning," in *Eighth International Symposium on Multispectral Image Processing and Pattern Recognition*, vol. 8919. International Society for Optics and Photonics, 2013, pp. 891 906–891 906.

[243] MSRA, "http://research.microsoft.com/en-us/um/people/jiansun/."

[244] S. Alpert, M. Galun, R. Basri, and A. Brandt, "Image segmentation by probabilistic bottom-up aggregation and cue integration," in *CVPR*, 2007, pp. 1–8.

[245] V. Movahedi and J. H. Elder, "Design and perceptual validation of performance measures for salient object segmentation," in *IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2010, pp. 49–56.

[246] M. Brown and S. Susstrunk, "Multi-spectral sift for scene category recognition," in *CVPR*, 2011, pp. 177–184.

[247] Q. Wang, P. Yan, Y. Yuan, and X. Li, "Multi-spectral saliency detection," *Pattern Recognition Letters*, vol. 34, no. 1, pp. 34–41, 2013.

[248] ECSSD, "http://www.cse.cuhk.edu.hk/leojia /projects/hsaliency/."

[249] THUR15000, "http://mmcheng.net/gsal/."

[250] Judd, "http://ilab.usc.edu/borji."

[251] J. Li, Y. Tian, T. Huang, and W. Gao, "A dataset and evaluation methodology for visual saliency in video," in *IEEE ICME*, 2009, pp. 442–445.

[252] M.-M. Cheng, N. J. Mitra, X. Huang, P. H. S. Torr, and S.-M. Hu, "Global contrast based salient region detection," *IEEE TPAMI*, 2014.

[253] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The pascal visual object classes (voc) challenge," *IJCV*, vol. 88, no. 2, pp. 303–338, Jun. 2010.

[254] M.-M. Cheng, N. J. Mitra, X. Huang, and S.-M. Hu, "Salientshape: group saliency in image collections," *The Visual Computer*, vol. 30, no. 4, pp. 443–453, 2014.

[255] M. R. Greene, "Statistics of high-level scene context," *Frontiers in psychology*, vol. 4, 2013.

[256] S. Avidan and A. Shamir, "Seam carving for content-aware image resizing," *ACM Transactions on Graphics*, vol. 26, no. 3, p. 10, 2007.

[257] J. Davis and M. Goadrich, "The relationship between precision-recall and roc curves," in *ICML*, 2006, pp. 233–240.

[258] R. Margolin, L. Zelnik-Manor, and A. Tal, "How to evaluate foreground maps?" in *CVPR*, 2014.

[259] J. He, J. Feng, X. Liu, T. Cheng, T.-H. Lin, H. Chung, and S.-F. Chang, "Mobile product search with bag of hash bits and boundary reranking," in *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*. IEEE, 2012, pp. 3005–3012.

[260] A. W. Smeulders, M. Worring, S. Santini, A. Gupta, and R. Jain, "Content-based image retrieval at the end of the early years," *IEEE TPAMI*, vol. 22, no. 12, pp. 1349–1380, 2000.

[261] K. Mikolajczyk and C. Schmid, "A performance evaluation of local descriptors," *IEEE PAMI*, vol. 27, no. 10, pp. 1615–1630, 2005.

[262] P. R. Roelfsema, V. A. Lamme, and H. Spekreijse, "Object-based attention in the primary visual cortex of the macaque monkey," *Nature*, vol. 395, no. 6700, pp. 376–381, 1998.

[263] V. Yanulevskaya, J. Uijlings, J.-M. Geusebroek, N. Sebe, and A. Smeulders, "A proto-object-based computational model for visual saliency," *Journal of Vision*, vol. 13, no. 13, p. 27, 2013.

[264] T. S. Horowitz, E. M. Fine, D. E. Fencsik, S. Yurgenson, and J. M. Wolfe, "Fixational eye movements are not an index of covert attention," *Psychological Science*, vol. 18, no. 4, pp. 356–363, 2007.

[265] T. A. Kelley, J. T. Serences, B. Giesbrecht, and S. Yantis, "Cortical mechanisms for shifting and holding visuospatial attention," *Cerebral Cortex*, vol. 18, no. 1, pp. 114–125, 2008.

[266] M. I. Posner, "Orienting of attention," *Quarterly journal of experimental psychology*, vol. 32, no. 1, pp. 3–25, 1980.

[267] B. J. Scholl, "Objects and attention: The state of the art," *Cognition*, vol. 80, no. 1, pp. 1–46, 2001.

[268] W. Einhäuser, U. Rutishauser, and C. Koch, "Task-demands can immediately reverse the effects of sensory-driven saliency in complex visual stimuli," *Journal of Vision*, vol. 8, no. 2, p. 2, 2008.

[269] A. F. Russell, S. Mihalaş, R. von der Heydt, E. Niebur, and R. Etienne-Cummings, "A model of proto-object based saliency," *Vision Research*, vol. 94, pp. 1–15, 2014.

[270] A. Torralba and A. A. Efros, "Unbiased look at dataset bias," in *IEEE CVPR*, 2011, pp. 1521–1528.

[271] R. Datta, D. Joshi, J. Li, and J. Z. Wang, "Studying aesthetics in photographic images using a computational approach," in *ECCV*, 2006, pp. 288–301.

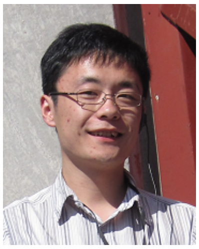
[272] J. M. Wolfe, "Guidance of visual search by preattentive information," in *Neurobiology of Attention*, 1st ed., L. Itti, G. Rees, and J. Tsotsos, Eds. Amsterdam: Elsevier Press, 2005, pp. 101–104.

[273] S. Li and M.-C. Lee, "Efficient spatiotemporal-attention-driven shot matching," in *Proceedings of the 15th Annual ACM International Conference on Multimedia*, ser. MULTIMEDIA '07. New York, NY, USA: ACM, 2007, pp. 178–187.

[274] A. Borji, M.-M. Cheng, H. Jiang, and J. Li, "Salient object detection benchmark," in *http://mmcheng.net/salobjbenchmark/*.

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. He then spent a year at University of Bonn, Germany as a research assistant. He is currently a postdoctoral scholar at iLab, University of Southern California, Los Angeles, CA. His research interests include: bottom-up and top-down visual attention, visual search and active learning, object and scene recognition, machine learning, cognitive robotics, cognitive and computational neurosciences and biologically plausible vision models.

Ming-Ming Cheng received his PhD degree from Tsinghua University in 2012. He is currently a research fellow in Oxford University, working with Prof. Philip Torr. His research interests includes computer graphics, computer vision, image processing, and image retrieval. He has received the Google PhD fellowship award, the IBM PhD fellowship award, and the new PhD Researcher Award from Chinese Ministry of Education.



Huaizu Jiang is currently working as a research assistant at Institute of Artificial Intelligence and Robotics, Xi'an Jiaotong University. Before that, 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. Specifically, his research interests include object detection, large-scale visual recognition, and (3D) scene understanding.

Jia Li received his B.E. degree from Tsinghua University in 2005 and Ph.D. degree from the Chinese Academy of Sciences in 2011. During 2011 and 2013, he used to serve as a research fellow and visiting assistant professor in Nanyang Technological University, Singapore. He is currently an associate professor at Beihang University, Beijing, China. His research interests include visual attention/saliency modeling, multimedia analysis, and vision from Big Data.