Salient Object Detection: A Survey

Article *in* Computational Visual Media · November 2014

DOI: 10.1007/s41095-019-0149-9 · Source: arXiv

4 authors:

Ali Borji

177 PUBLICATIONS **11,208** CITATIONS

SEE PROFILE

Ming-Ming Cheng
Nankai University

225 PUBLICATIONS **17,368** CITATIONS

SEE PROFILE

Huaizu Jiang
University of Massachusetts Amherst

12 PUBLICATIONS **2,994** CITATIONS

SEE PROFILE

Jia Li
Beihang University (BUAA)

99 PUBLICATIONS **3,010** CITATIONS

SEE PROFILE

Salient Object Detection: A Survey

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, a deep understanding of achievements and issues is lacking. We aim to provide a comprehensive review of the recent progress in this field. We situate salient object detection among other closely related areas such as generic scene segmentation, object proposal generation, and saliency for fixation prediction. Covering 256 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. We also discuss open problems such as evaluation metrics and dataset bias in model performance and suggest future research directions.

Index Terms—Salient object detection, salient region detection, saliency, explicit saliency, visual attention, regions of interest, objectness, object proposal, 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 details 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 the computer vision community mainly because it helps find 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, is proposed by Itti *et al.* [1] (see Fig. 1). This model is 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 [2], *Guided Search Model* by Wolfe *et al.* [3], and *Computational Attention Architecture* by Koch and Ullman [4]). In [1], Itti *et al.* show examples where their model is able to detect spatial discontinuities in scenes. Subsequent behavioral (e.g., [5]) and computational studies (*e.g.*, [6]) start 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) appears with works of Liu *et al.* [7], [8] and Achanta *et al.* [9] who define saliency detection as a binary segmentation problem. These works themselves are inspired by some earlier models striving for detecting regions (*e.g.*, Ma and Zhang [10], Liu and Gleicher [11], and Walther *et al.* [12]). Since then a plethora of saliency models have emerged that have blurred 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.*, [13], [14]), category independent object proposal generation approaches (*e.g.*, [15], [16]), fixation prediction models (*e.g.* [6], [17]–[20]), and object detection methods (*e.g.*, [21], [22]). One of our main goals here is to thoroughly review the literature, clarify less understood challenges, and offer learned lessons from existing works.

In addition to the fast, bottom-up, involuntary, and stimulus-driven stage of attention which is 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 [23]–[25]). 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 [26]–[28], memorability [29], scene clutter [30], video interestingness [31]–[34], surprise [35], image quality assessment [36], [37], scene typicality [38], [39], aesthetic [33], and attributes [40].

An enormous amount of research has been undertaken to study attention through discovering where people look in an image (*e.g.*, [1], [4], [41]). Previous research has shown that eyes are drawn to informative and salient areas in a scene (e.g., [1], [5], [6]). 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 [42], analyzing LabelMe annotation data [43], demonstrate that human observers tend to annotate more salient objects first. They hence conclude that salient objects are interesting. Masciocchi *et al.* [44] address 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 find that interest selections are correlated with eye movements, and both types of data correlate with bottom-up saliency. Recently, Borji *et al.* [45] conduct two experi-

- *A. Borji is with the Computer Science Department, University of Wisconsin, Milwaukee, WI 53211. E-mail: borji@uwm.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*
- *First two authors contributed equally.*

- *Manuscript received xx 2014.*

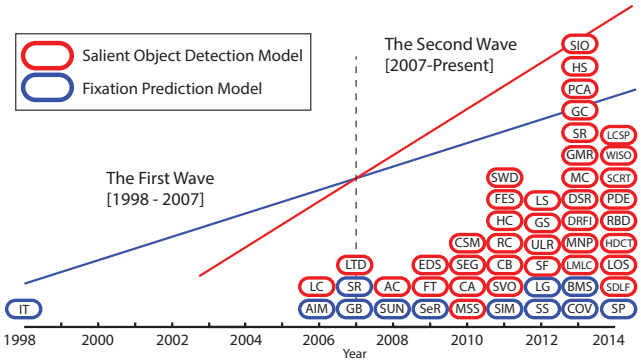


Fig. 1. A simplified chronicle of saliency modeling. Models in the first wave (1998-2007) are mainly dealing with fixation prediction while models in the seconds wave (2007-now) mainly addressed the detection and segmentation of the most salient objects. While both trends are still active research areas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

ments in which they asked 70 observers to explicitly choose the most outstanding (*i.e.*, salient) object in a scene. In the first experiment, observers view scenes with only two objects. In the second experiment they ask observers to draw a polygon around the most salient object (see Fig. 2). These experiments reveal 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.* [46]. As in [44], [45], they ask observers to click on salient locations in natural scenes. While the most salient [45], important [26], or interesting [34], [42], [47] 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 [48], language and attention [49], and phrasal recognition [40], [50].

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, have models explicitly distinguished these two stages (with few exceptions such as [51]–[53]). Following the traditional works by Itti *et al.* [1] and Liu *et al.* [7], models adopt 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 need to be limited to one object. The majority of existing models 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 in the realm of classic segmentation problems in computer vision but has certain differences. For example, here accuracy is mainly determined by the most salient object. In the next section, we briefly review what people perceive as the most salient, interesting, or outstanding in a scene from a cognitive perspective. Finally, in the discussion section, we will get back to these points and propose new ways to address salient object detection.

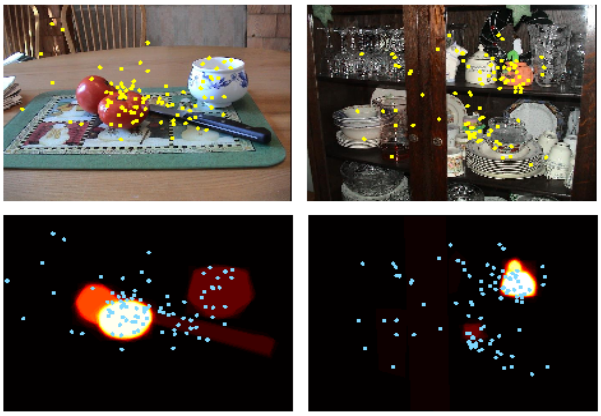


Fig. 2. Two sample images from the Bruce and Tsotsos dataset in Borji *et al.*'s experiment [45]. Left (right) column shows a case where humans are less (more) consistent in choosing the object that stands out the most in the scene. Dots represent 3-second free-viewing fixations.

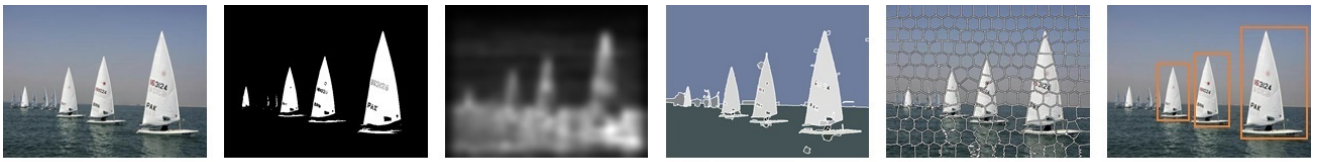
In general, it is agreed that for good saliency detection a model should meet at least the following three criteria: 1) good detection: the probabilities of missing real salient regions and falsely marking background as salient regions should be low; 2) high resolution: saliency maps should have high or full resolution to accurately locate salient objects and retain original image information; 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 an illustration of models in these categories.

1.2.1 Fixation prediction

The emergence of salient object detection models is driven by the requirement of saliency-based applications (e.g., content-aware image resizing [48], [58], [59]), while fixation prediction models are constructed originally to understand human visual attention and eye movement prediction [1], [4]. 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, 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 [52], [60]. On the contrary, salient object detection producing maps that can accurately distinguish object boundaries are highly desirable, especially in



(a) (b) (c) (d) (e) (f)

Fig. 3. Sample results produced by different models: (b) salient object detection [54], (c) fixation prediction [1], (d) image segmentation (regions with various sizes) [55], (e) image segmentation (superpixels with comparable sizes) [56], and (f) object proposals (true positives) [57].

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., [18], [61], [62]). 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) is one of the very-well researched areas in computer vision (*e.g.*, [63]). Their aim is to assign each pixel a label indicating the 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 similar to the classic figure-ground problem in segmentation literature (although they generate smooth maps with confidence levels at each point). The aim is to tell whether a pixel belongs to the most salient object. 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 pre-processing 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) which are not very accurate in 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 regions, so that these regions cover every objects in the input image, regardless of the specific categories of those objects (*i.e.*, generic over categories) [57], [64]–[67]. When compared to traditional sliding window based object detection paradigm [22], [68], estimating object proposals in a pre-processing stage has three major advantages: 1) better accords with our human mental recognition behavior which quickly perceives objects before identifying them [69], [70]; 2) greatly speeds 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 [71], and 3) also improves the detection accuracy by allowing the usage of stronger classifiers during testing [72].

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 [15], [64]. In other words, an object is more likely to be salient than a region on the background. This is based on a finding that image background is usually more structured and homogeneous (thus less salient) than objects [42], [45]. On the other hand, the latter approaches use objectness measures to assign higher saliency values to objects rather than the background (*e.g.*, [73]).

1.3 Organization of the Paper

In Sec. 2, 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. In Sec. 3, we review common datasets, evaluation measures and issues pertaining to models, and, finally, in Sec. 4 and 5, we discuss the probable challenges and summarize learned lessons and highlight future directions to advance the field.

2 Survey of the State of the Art

In this section, we review related works in 3 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 (as shown in Fig. 1). Here, we use terms "salient object detection" and "salient region detection" interchangeably. Note that there is no sharp boundary among these models and often model predictions 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, a lot of approaches have been proposed for detecting salient or interesting objects in images. Except for several studies on segmenting object-of-interest (*e.g.*, [74]–[76]), most of these approaches aim to identify the salient subsets[1] from images first (*i.e.*, compute a saliency

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 homogeneous 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.

map) 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 are mainly two kinds of visual subsets, including blocks and regions[2], that are used to detect salient objects. Blocks are usually adopted by many early approaches, while regions are 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 image.

Since most of the existing models aim to produce saliency maps, in this review we divide most of existing salient object detection approaches into three major subgroups according to such two attributes, including *block-based models with intrinsic cues*, *region-based model with intrinsic cues*, and *models with extrinsic cues*. In these groups, we focus on reviewing how to compute saliency maps. Salient object segmentation can be achieved by binarizing a saliency map with a fixed or adaptive threshold. Meanwhile, some other researchers concentrate on more sophisticated segmentation or localization algorithms. There also exist several approaches that can not be simply assigned to any of previous three subgroups. We categorize them as *other models* and review them in a separate subgroup. In the rest part of this section, we review models from all the four subgroups and provide algorithm details of the representative ones. A complete list of the reviewed models are illustrated in Fig. 4, Fig. 5, and Fig. 6.

2.1.1 Block-based Models with Intrinsic Cues

In this subsection, we mainly review salient object detection models which utilize intrinsic cues extracted from blocks. Following the seminal work of Itti *et al.* [1], salient object detection is widely defined as capturing the uniqueness, distinctiveness, or rarity of a scene.

In early stages [9]–[11], uniqueness is widely studied as pixel-wise center-surround contrast. Hu *et al.* [77] represent the input image in a 2D space using polar transformation of its features so that each region in the images can be mapped into a 1D linear subspace. After that, the Generalized Principal Component Analysis (GPCA) [78] is used to estimate the linear subspaces without actually segmenting the image. Finally, the attentive (salient) regions are selected by measuring feature contrasts as well as geometric properties of regions. Rosin [79] proposes an efficient approach for detecting salient objects. The whole approach is parameter-free and requires only very simple pixel-wise operations such as edge detection, threshold decomposition and moment preserving binarization. Valenti *et al.* [80] propose an isophote-based framework where the saliency map is estimated by linearly combining the saliency maps computed in terms of curvedness, color boosting, and isocenters clustering.

In an influential study, Achanta *et al.* [81] adopt a frequency-tuned approach to compute full resolution saliency maps by simply measuring the pixel-wise color difference between the smoothed image pixels and average color of the image. The saliency of pixel x is computed as:

$$s(x) = \|I_\mu - I_{\omega_{hc}}(x)\|^2, \tag{1}$$

where I_μ is the mean pixel value of the image (*e.g.*, RGB/Lab features) and $I_{\omega_{hc}}$ is a Gaussian blurred version of the input image (*e.g.*, using a 5×5 kernel).

Without any prior knowledge of the sizes of salient objects, multi-scale contrast is frequently adopted for robustness purpose [7], [11]. To that end, a L-layer Gaussian pyramid is first constructed in [7], [11]. Let $I^{(l)}$ be the image at the lth-level of the pyramid, the saliency score of pixel x is defined as:

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

where $\mathcal{N}(x)$ is a 9×9 neighboring window centered at x. Even with such multi-scale enhancement, intrinsic cues derived on pixel-level are often too poor to support object segmentation. The contrast analysis is thus extended to the patch level. Following the center-surround mechanism suggested in [1], the contrast of a patch is always defined as its contrast with the surrounding patches [7], [9], [82], [83][3]. Given a rectangle $R(x)$ centered at pixel x and its surrounding strip $R_S(x)$ with the same area of $R(x)$, the uniqueness of x can be measured by the difference between $R(x)$ and $R_S(x)$ (*e.g.*, the χ^2 distance of color histograms). In the implementation, the most distinct rectangle $R^*(x)$ and $R_S^*(x)$ are found by enumerating a large number of candidate aspect ratios and sizes for R and R_S until the largest center-surround difference is reached. As a consequence, the saliency score of a pixel x can be computed as:

$$s(x) = \sum_{\{x' | x' \in R^*(x)\}} w_{xx'} \chi^2 \left(R^*(x'), R_S^*(x')\right), \tag{3}$$

where $w_{xx'}$ is a Gaussian weight between two pixels, which has small value if they are far away.

Later in [82], such center-surround contrasts are computed in an information-theoretic way by using the Kullback-Leibler Divergence on difference features such as intensity, color and orientation.

Li *et al.* [83] study the center-surround contrast as a cost-sensitive max-margin classification problem. In particular, the center patch is thought of as a positive sample while the surrounding patches are all used as negative samples. The saliency of the center patch is determined by its separability from surroundings based on the trained cost-sensitive Support Vector Machine (SVM).

Patch uniqueness is also defined as its global contrast with others [48]. Intuitively, a patch is considered to be salient if it is remarkably distinct from its most similar patches, while their spatial distances are taken into account.

In a recent work [84], Margolin *et al.* propose to define the uniqueness of a patch by measuring its distance to the average patch based on the observation that distinct

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

3. Though [7] is categorized as an extrinsic model in our review, the extraction of salient object features only involves intrinsic cues.

patches are more scattered than non-distinct ones in the high-dimensional space. To further incorporate the patch distributions on each single image, the uniqueness of a patch is 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 is defined as

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

where $\tilde{p}_x$ is the coordinates of p_x in the PCA coordinated system.

To sum up, approaches in Sec. 2.1.1 aim to detect salient objects based on pixels or patches where only intrinsic cues are 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 main advantages. First, the number of regions is far less than the number of blocks, which implies the potential to develop highly efficient algorithms. Second, 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* to estimate their saliency scores. Different from the block-based models, region-based models often segment an input image into regions aligned with intensity edges first and then compute a regional saliency map.

As an early attempt, in [11], the regional saliency score is defined as the average saliency scores of its contained pixels, defined in terms of multi-scale contrast. Yu *et al.* [85] propose a set of rules to determine the background scores of each region based on observations from background and salient regions.

Saliency, defined as uniqueness in terms of **global regional contrast**, is widely studied in existing approaches [86]–[90]. In [86], a region-based saliency algorithm is introduced by measuring the global contrast between the target region with respect to all other regions in the image. Generally, the input image is first fragmented into N regions $\{r_i\}_{i=1}^N$. 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 $D_r(r_i, r_j)$ captures the appearance contrast between two regions. Higher saliency scores will be assigned to regions with large global contrast. w_{ij} is 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 depending on the method used for segmentation (*e.g.*, the graph-based segmentation [91], mean-shift [55] algorithm, or clustering). It is often uniformly set if compact superpixels are generated using such as SLIC [56] or Turbopixel [92] algorithms.

Perazzi *et al.* [54] demonstrate that if $D_r(r_i, r_j)$ is 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:

$$s(r_i) = \sum_{j=1}^{N} w_{ij} ||\mathbf{c}_i - \mathbf{c}_j||^2 \tag{6}$$

$$= \mathbf{c}_i^2 \sum_{j=1}^{N} w_{ij} - 2\mathbf{c}_i \sum_{j=1}^{N} \mathbf{c}_j w_{ij} + \sum_{j=1}^{N} \mathbf{c}_j^2 w_{ij}, \tag{7}$$

where $\mathbf{c}_i$ 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 the color $\mathbf{c}_j$ and squared color $\mathbf{c}_j^2$. If w_{ij} is a Gaussian spatial weighting term, the complexity of evaluating (6) for all elements can be reduced from $O(N^2)$ to $O(N)$ by evaluating (7) using efficient filtering based techniques [93].

In addition to color uniqueness, distinctiveness of complementary cues such as texture [88] and structure [94] are also considered for salient object detection. Margolin *et al.* [84] propose to combine the regional uniqueness and patch distinctiveness to form the saliency map.

Instead of maintaining a hard region index for each pixel, a soft abstraction is proposed in [89] to generate a set of large scale perceptually homogeneous regions using histogram quantization and a global Gaussian Mixture Model (GMM). By avoiding the hard decision boundaries of superpixels, such soft abstraction provides large spatial support and can more uniformly highlight the salient object.

In [97], Jiang *et al.* propose a **multi-scale local region contrast** based approach, which calculates saliency values across multiple segmentations for robustness purpose and combines these regional saliency values to get a pixel-wise saliency map. Specifically, the input image is segmented with the algorithm in [91] using N_s different groups of parameters. Denote r_i^n as the ith superpixel coming from the nth segmentation. Its saliency is 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 is the set of neighbor regions of r_i^n. α_{ij}^n is the normalized ratio between the area of r_j^n and total area of $\mathcal{N}(r_i^n)$, accounting for the influence of irregular regions. As 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 is used to emphasize regions around the image center. Finally, regional saliency scores across multiple segmentations are combined to get the pixel-wise saliency map.

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

where $N(n)$ is the number of regions in the nth segmentation. $c(x, r_i^n)$ is defined as the normalized color similarity of the region r_i^n and its contained pixel x. A similar idea of estimating regional saliency on multiple/hierarchical segmentations is adopted in [87], [102] to increase the detection reliability.

Li *et al.* [83] extend the pairwise local contrast by building a hypergraph, constructed by non-parametric multi-scale clustering of superpixels, to capture both internal consistency and external separation of regions. The salient

#	Model	Pub	Year	Elements	Hypothesis		Aggregation (Optimization)	Code
					Uniqueness	Prior		
1	**FG** [10]	MM	2003	PI	L	-	-	NA
2	**RSA** [77]	MM	2005	PA	G	-	-	NA
3	**RE** [11]	ICME	2006	mPI + RE	L	-	LN	NA
4	**RU** [85]	TMM	2007	RE	-	P	LN	NA
5	**AC** [9]	ICVS	2008	mPA	L	-	LN	NA
6	**FT** [81]	CVPR	2009	PI	CS	-	-	C
7	**ICC** [80]	ICCV	2009	PI	L	-	LN	NA
8	**EDS** [79]	PR	2009	PI	-	ED	-	NA
9	**CSM** [95]	MM	2010	PI + PA	L	SD	-	NA
10	**RC** [86]	CVPR	2011	RE	G	-	-	C
11	**HC** [86]	CVPR	2011	RE	G	-	-	C
12	**CC** [96]	ICCV	2011	mRE	-	CV	-	NA
13	**CSD** [82]	ICCV	2011	mPA	CS	-	LN	NA
14	**SVO** [73]	ICCV	2011	PA + RE	CS	O	EM	M + C
15	**CB** [97]	BMVC	2011	mRE	L	CP	LN	M + C
16	**SF** [54]	CVPR	2012	RE	G	SD	NL	C
17	**ULR** [98]	CVPR	2012	RE	SPS	CP + CLP	-	M + C
18	**GS** [99]	ECCV	2012	PA/RE	-	B	-	NA
19	**LMLC** [100]	TIP	2013	RE	CS	-	BA	M + C
20	**HS** [87]	CVPR	2013	hRE	G	-	HI	EXE
21	**GMR** [101]	CVPR	2013	RE	-	B	-	M
22	**PISA** [94]	CVPR	2013	RE	G	SD + CP	NL	NA
23	**STD** [88]	CVPR	2013	RE	G	-	-	NA
24	**PCA** [84]	CVPR	2013	PA + PE	G	-	NL	M+C
25	**GU** [89]	ICCV	2013	RE	G	-	-	C
26	**GC** [89]	ICCV	2013	RE	G	SD	AD	C
27	**CHM** [83]	ICCV	2013	PA + mRE	CS + L	-	LN	M + C
28	**DSR** [102]	ICCV	2013	mRE	-	B	BA	M + C
29	**MC** [103]	ICCV	2013	RE	-	B	-	M + C
30	**UFO** [104]	ICCV	2013	RE	G	F + O	NL	M + C
31	**CIO** [105]	ICCV	2013	RE	G	O	GMRF	NA
32	**SLMR** [106]	BMVC	2013	RE	SPS	BC	-	NA
33	**LSMD** [107]	AAAI	2013	RE	SPS	CP + CLP	-	NA
34	**SUB** [90]	CVPR	2013	RE	G	CP + CLP + SD	-	NA
35	**PDE** [108]	CVPR	2014	RE	-	CP + B + CLP	-	NA
36	**RBD** [109]	CVPR	2014	RE	-	BC	LS	M

Fig. 4. Salient object detection models with intrinsic cues (sorted by year). Element, {PI = pixel, PA = patch, RE = region}, where prefixes m and h indicate multi-scale and hierarchical versions, respectively. 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, CLP = color prior, SD = spatial distribution, BC = boundary connectivity prior, SPS = sparse noises}. Aggregation/optimization, {LN = linear, NL = non-linear, AD = adaptive, HI = hierarchical, BA = Bayesian, GMRF = Gaussian MRF, EM = energy minimization, and LS = least-square solver.}. Code, {M= Matlab, C= C/C++, NA = not available, EXE = executable}.

object detection is casted as finding salient vertices and hyperedges in the hypergraph.

Salient objects, in terms of uniqueness, can also be defined as the **sparse noises** in a certain feature space where the input image is represented as a low-rank matrix [98], [106], [107]. The basic assumption is that non-salient regions (background) can be explained by the low-rank matrix while the salient regions are indicated by the sparse noises. Formally, each region r_i is described by a feature vector $\mathbf{f}_i$. By stacking $\mathbf{f}_i$ together, we can get the feature matrix representation of the entier image $\mathbf{F} = [\mathbf{f}_1, \mathbf{f}_2, \cdots, \mathbf{f}_N] \in \mathbb{R}^{D\times N}$, where D is the dimension of the feature vector. $\mathbf{F}$ can be decomposed into two parts, the low-rank matrix $\mathbf{L} = [\mathbf{l}_1, \mathbf{l}_2, \cdots, \mathbf{l}_N] \in \mathbb{R}^{D\times N}$ and the sparse matrix $\mathbf{S} = [\mathbf{s}_1, \mathbf{s}_2, \cdots, \mathbf{s}_N] \in \mathbb{R}^{D\times N}$ by optimizing the following objective function

$$\min_{\mathbf{L},\mathbf{S}} \quad \|\mathbf{L}\|_* + \lambda\|\mathbf{S}\|_1, \quad s.t. \quad \mathbf{F} = \mathbf{L} + \mathbf{S}, \tag{10}$$

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

$$s(r_i) = ||\mathbf{s}_i||_2, \quad \text{or} \quad s(r_i) = ||\mathbf{s}_i||_1. \tag{11}$$

Based on such a general low-rank matrix recovery framework, Shen and Wu [98] propose a unified approach to incorporate traditional low-level features with higher-level guidance, *e.g.*, **center prior**, **face prior**, and **color prior**, to detect salient objects based on a learned feature transformation[4]. Instead, Zou *et al.* [106] propose to exploit bottom-up segmentation as a guidance cue of the low-rank matrix recovery for robustness purpose. Similar to [98], high-level priors are also adopted in [107], where a tree-structured sparsity-inducing norm regularization is introduced to hierarchically describe the image structure with the aim to more uniformly highlight the entire salient object.

In addition to capturing the uniqueness, more and more priors are proposed for salient object detection as well. **Spatial distribution prior** [7] implies that the wider a color is distributed in the image, the less likely a salient object contains this color. In [7], [89], pixels in the input image I are quantized by a GMM $\{w_c, \mu_c, \mathbf{\Sigma}_c\}_{c=1}^{C}$, where $\{w_c, \mu_c, \mathbf{\Sigma}_c\}$ are the weight, mean color and the covariance matrix of the cth component. Each pixel x is assigned to a color component with the probability

$$\mathbf{P}(c|I_x) = \frac{w_c\mathcal{N}(I_x|\mu_c, \mathbf{\Sigma}_c)}{\sum_c w_c\mathcal{N}(I_x|\mu_c, \mathbf{\Sigma}_c)}. \tag{12}$$

The horizontal spatial variance of the component c is:

$$V_h(c) = \frac{1}{|P|_c}\sum_x \mathbf{P}(c|I_x)||x_h - M_h(c)||^2, \tag{13}$$

$$M_h(c) = \frac{1}{|P|_c}\sum_x \mathbf{P}(c|I_x)x_h, \tag{14}$$

4. Though extrinsic ground-truth annotations are adopted to learn high-level priors and the 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 intrinsic cues for salient object detection.

where x_h is the x-coordinate of the pixel x and $|P|_c = \sum_x \mathbf{P}(c|I_x)$. The vertical variance $V_v(c)$ is defined similarly. Finally, the saliency of each pixel p is defined as:

$$s(x) = \sum_c \mathbf{P}(c|I_x)(1 - V(c)) \cdot (1 - D(c)), \tag{15}$$

where $V(c)$ is the spatial variance of the component c, defined as $V(c) = V_h(c) + V_v(c)$. $D(c) = \sum_x \mathbf{P}(c|I_x) \cdot d_x$ is a center-weighted normalization term to balance the border cropping effect and d_x is the distance from pixel x to the image center. The spatial distribution of superpixels can be efficiently evaluated in linear time using the Gaussian blurring kernel as well, in a similar way of computing the global regional contrast in Eq. 5. Such a spatial distribution prior is also considered in [94] evaluated in terms of both color and structure cues.

Center prior assumes that a salient object is more likely to be put near the image center. In other words, the background tends to be far away from the image center. To this end, the **backgroundness prior** is adopted for salient object detection [99], [101]–[103], assuming that a narrow border of the image is the background region, *i.e.*, the pseudo-background. With this pseudo-background B as a reference, regional saliency can be computed as the contrast versus "background". In [99], a undirected weighted graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ is built, where the vertices are the regions (or patches) plus the pseudo-background B, *i.e.*, $\mathcal{V} = \{r_i\} \cup B$. There are two kinds of edges where the internal edges connect all adjacent regions and the boundary edges connect 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 is defined as the accumulated edge weights along the shortest path from r_i to r_k:

$$\begin{aligned} d_{geo}(r_i, r_k) = & \min_{r_{i_1}=r_i,\cdots,r_{i_n}=r_k} \sum_{j=1}^{n-1} D_r(r_{i_j}, r_{i_{j+1}}). \\ & s.t. \quad (r_{i_j}, r_{i_{j+1}}) \in \mathcal{E} \end{aligned} \tag{16}$$

The saliency score of r_i is defined as its geodesic distance to the pseudo-background node B on the graph $\mathcal{G}$,

$$s(r_i) = d_{geo}(r_i, B). \tag{17}$$

In [101], a two-stage saliency computation framework is proposed based on the manifold ranking on an undirected weighted graph. In the first stage, the regional saliency scores are computed based on the relevances given to each side of the pseudo-background queries. In the second stage, the saliency scores are refined based on the relevances given to the initial foreground which is segmented using an adaptive threshold on the saliency scores obtained from the first stage.

In [102], saliency computation is formulated as the dense and sparse reconstruction errors w.r.t. the pseudo-background. The dense reconstruction error of each region is computed based on the Principal Component Analysis (PCA) basis of the background templates, while the sparse reconstruction error is defined as the residual based on the sparse representation of the background templates. These two types of reconstruction errors are propagated to pixels on multiple segmentations, which will be fused to form the final saliency map by Bayesian inference.

Jiang *et al.* [103] propose to formulate the saliency detection via absorbing Markov Chain where the transient and absorbing nodes are superpixels around the image center and border, respectively. The saliency of each superpixel is 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] is also exploited to facilitate the detection of salient objects by leveraging the object proposal generation model [15]. Chang *et al.* [73] present a computational framework by fusing the objectness and regional saliency into a graphical model. These two terms are jointly estimated by iteratively minimizing the energy function that encodes their mutual interactions. In [104], regional objectness is defined as the average objectness values of its contained pixels, which will be incorporated for regional saliency computation.

Jia and Han [105] compute the saliency of each region by comparing it to the "soft" foreground and background according to the objectness prior. Regional saliency of r_i, called the diverse density score, is computed as:

$$DD(r_i) = \sum_{j=1}^{N} D_r(r_i, r_j)o(r_j) + (1 - D_r(r_i, r_j))(1 - o(r_j)) \tag{18}$$

where $o(r_i)$ is the objectness score of region r_i, computed as spatially weighted average objectness scores of its contained pixels. A higher saliency score will be assigned to a region if it is distinct from the potential background or similar to the potential foreground.

Salient object detection relying on the pseudo-background assumption sometimes may fail, especially when the object touches the image border. To this end, **boundary connectivity prior** is explored in [86], [109]. Intuitively, salient objects are much less connected to the image border than the ones in the background. Thus, the boundary connectivity score of a region could be estimated according to the ratio between its length along the image border and the spanning area of this region [109], which are computed based on its geodesic distances to the pseudo-background and other regions, respectively. Such a boundary connectivity score is 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 are also investigated in [106] as *segmentation prior* and as *surroundness* in [110].

The **focusness prior**, reflecting the fact that a salient object is often photographed in focus to attract more attention, has been investigated in recent works [104], [111]. Jiang *et al.* [104] define the focusness from the degree of focal blur. By modeling such a de-focus blur as the convolution of a sharp image with a point spread function, approximated by a Gaussian kernel, the pixel-level focusness is 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 is finally derived from the non-linear combination of uniqueness (global contrast), objectness, and focusness scores.

Performance of salient object detection based on regions might be affected by the segmentation parameters. In addi-

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

tion to other approaches based on multi-scale regions [83], [87], [97], single-scale potential salient regions are extracted by solving the facility location problem in [90]. An input image is first represented as an undirected graph on superpixels, where a much smaller set of candidate region centers are then generated through agglomerative clustering. On this set, a submodular objective function is built to maximize the similarity. By applying a greedy algorithm, the objective function can be iteratively optimized to group superpixels into regions whose saliency values are further measured via the regional global contrast and spatial distribution.

The Bayesian framework is studied for saliency computation [100], [112], which is formulated as estimating the posterior probability of being foreground at each pixel x given the input image I. Denote $s(x)$ as s_x for short, the posterior probability in [100] is computed as

$$\begin{aligned}&\mathbf{P}(s_x = 1|I_x)\\ &= \frac{\mathbf{P}(s_x = 1)\mathbf{P}(I_x|s_x = 1)}{\mathbf{P}(s_x = 1)\mathbf{P}(I_x|s_x = 1) + \mathbf{P}(s_x = 0)\mathbf{P}(I_x|s_x = 0)}\end{aligned} \tag{19}$$

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

$$\mathbf{P}(s_x = 1) = \frac{|r \cap H|}{|r|}, \text{ where } x \in r. \tag{20}$$

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

$$\mathbf{P}(I_x|s_x = 1) = \prod_{v\in\{l,a,b\}} \frac{|\{x'|x' \in R_I, v(x') = v(x)\}|}{|\{x'|x' \in R_I\}|}, \tag{21}$$

$$\mathbf{P}(I_x|s_x = 0) = \prod_{v\in\{l,a,b\}} \frac{|\{x'|x' \in R_O, v(x') = v(x)\}|}{|\{x'|x' \in R_O\}|}. \tag{22}$$

Liu *et al.* [108] adopt an optimization-based framework for detecting salient objects. Similar to [100], a convex hull is roughly estimated to bipartite an image into pure background and potential foreground. Then saliency seeds are learned from the image, while a guidance map is learned from background regions as well as human prior knowledge. Under the assistance of these cues, a general Linear Elliptic System with Dirichlet boundary is 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 [91], mean-shift algorithm [55], or clustering (quantization). On the other hand, with recent progress on superpixels algorithms, compact regions with comparable sizes are also popular choices using the SLIC algorithm [56], Turbopixel algorithm [92], etc. The main difference between these two types of regions is whether the influence of region size should be taken into account. Furthermore, soft regions are also considered for saliency analysis, where every pixel maintains a probability belonging to each of all the regions (components) instead of only a hard region label (*e.g.*, fitted by a GMM). To further enhance robustness of segmentation, regions can be generated based on multiple segmentations or in a hierarchical 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 global/local regional contrast, is still the most frequently used feature. In addition, more and more complementary priors for the regional saliency are investigated 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 of defining 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. 5 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 salient object detection. All of existing learning-based 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.

Each element (*e.g.*, a pixel or a region) in the input image will be represented by a feature vector $\mathbf{f} \in \mathbb{R}^D$, where D is the feature dimension. Such a feature vector is then mapped to a saliency score $s \in \mathbb{R}^+$ based on the learned linear or non-linear mapping function $f : \mathbb{R}^D \to \mathbb{R}^+$.

One can assume the mapping function f is linear, *i.e.*, $s = \mathbf{w}^{\mathrm{T}}\mathbf{f}$, where $\mathbf{w}$ denotes the combination weights of all components in the feature vector. Liu *et al.* [7] propose to learn the weights with the Conditional Random Field (CRF)

#	Model	Pub	Year	Cues	Elements	Hypothesis		Aggregation (Optimization)	GT Form	Code
						Uniqueness	Prior			
1	**LTD** [7]	CVPR	2007	GT	mPI + PA + RE	L + CS	SD	CRF	BB	NA
2	**OID** [113]	ECCV	2010	GT	mPI + PA + RE	L + CS	SD	mixtureSVM	BB	NA
3	**LGCR** [114]	BMVC	2010	GT	RE	-	P	BDT	BM	NA
4	**DRFI** [115]	CVPR	2013	GT	mRE	L	B + P	RF	BM	M + C
5	**LOS** [116]	CVPR	2014	GT	RE	L + G	PRA + B + SD + CP	SVM	BM	NA
6	**HDCT** [117]	CVPR	2014	GT	RE	L + G	SD + P + HD	BDT + LS	BM	M

#	Model	Pub	Year	Cues	Elements	Hypothesis		Aggregation (Optimization)	GT Necessity	Code
						Uniqueness	Prior			
7	**VSIT** [118]	ICCV	2009	SI	PA	-	SS	-	yes	NA
8	**FIEC** [119]	CVPR	2011	SI	PI + PA	L	-	LN	no	NA
9	**SA** [120]	CVPR	2013	SI	PI	-	CMP	CRF	yes	NA
10	**LBI** [121]	CVPR	2013	SI	PA	SP	-	-	no	M + C

#	Model	Pub	Year	Cues	Elements	Hypothesis		Aggregation (Optimization)	Type	Code
						Uniqueness	Prior			
11	**LC** [122]	MM	2006	TC	PI + PA	L	-	LN	online	NA
12	**VA** [123]	ICPR	2008	TC	mPI + PA + RE	L	CS + SD + MCO	CRF	offline	NA
13	**SEG** [112]	ECCV	2010	TC	PA + PI	CS	MCO	CRF	offline	M + C
14	**RDC** [124]	CSVT	2013	TC	RE	L	-	-	offline	NA

#	Model	Pub	Year	Cues	Elements	Hypothesis		Aggregation (Optimization)	Image Number	Code
						Uniqueness	Prior			
15	**CSIP** [125]	TIP	2011	SCO	mRE	-	RS	LN	two	M + C
16	**CO** [126]	CVPR	2011	SCO	PI + PA	G	RP	-	multiple	NA
17	**CBCO** [127]	TIP	2013	SCO	RE	G	SD + C	NL	multiple	NA

#	Model	Pub	Year	Cues	Elements	Hypothesis		Aggregation (Optimization)	Source	Code
						Uniqueness	Prior			
18	**LS** [128]	CVPR	2012	DP	RE	G	DK	NL	stereo images	NA
19	**DRM** [129]	BMVC	2013	DP	RE	G	-	SVM	Kinect	NA
20	**SDLF** [111]	CVPR	2014	LF	mRE	G	F + B + O	NL	Lytro camera	NA

Fig. 5. Salient object detection models with extrinsic cues grouped by their adopted cues. 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, CMP = complement of saliency cues, SP = sampling probability, MCO = motion coherence, RP = repeatedness, RS = region similarity, C = corresponding, and DK = domain knowledge.}. Others, {CRF = conditional random field, SVM = support vector machine, BDT = boosted decision tree, and RF = random forest.}.

model trained on the rectangular annotations of the salient objects. In a recent work [116], the large-margin framework is adopted to learn the weights $\mathbf{w}$.

Due to the highly non-linear essence of the saliency mechanism, however, the linear mapping might not perfectly capture the characteristics of saliency. To this end, such a linear integration is extended in [113], where a mixture of linear Support Vector Machines (SVM) is adopted to partition the feature space into a set of sub-regions that are linearly separable using a divide-and-conquer strategy. In each region, a linear SVM, its mixture weights, and the combination parameters of the saliency features are learned for better saliency estimation. Alternatively, other non-linear classifiers are also utilized; the boosted decision trees (BDT) [114], [117] and the random forest (RF) [115].

Generally speaking, supervised approaches allow richer representations for the elements compared with the heuristic methods. In the seminal work of the supervised salient object detection, Liu *et al.* [7] propose 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 [114] demonstrate 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.* [115] propose a regional saliency descriptor including the regional local contrast, regional backgroundness, and regional generic properties. In [116], [117], each region is described by a set of features such as local and global contrast, backgroundness, spatial distribution, and the center prior. The pre-attentive features are also considered in [116].

Usually, the richer representations result in feature vectors with higher dimensions, *e.g.*, $D = 93$ in [115] and $D = 75$ in [117]. With the availability of a large collections of training samples, the learned classifier is capable of automatically integrating such richer features and picking up the most discriminative ones. Therefore, better performance can be expected 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 has been studied in recent years. Generally, given the input image I, K similar images $\mathcal{C}_I = \{I_k\}_{k=1}^K$ 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 some studies, it is assumed that saliency annotations of $\mathcal{C}$ is available. Specifically, Marchesotti *et al.* [118] propose 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 $\mathbf{f}_{I_k}^-$ denote 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 $\{p_x\}_{x=1}^{P}$ and each patch p_x is described by a fisher vector $\mathbf{f}_x$. For robustness, the saliency score is computed for a neighborhood $\mathcal{N}_x$ of p_x,

$$s(\mathcal{N}_x) = ||\mathbf{f}_{\mathcal{N}_x} - \mathbf{f}_{BG}||_1 - ||\mathbf{f}_{\mathcal{N}_x} - \mathbf{f}_{FG}||_1, \quad (23)$$

where $\mathbf{f}_{\mathcal{N}_x} = \sum_{p_x \in \mathcal{N}_x} \mathbf{f}_x, \mathbf{f}_{FG} = \sum_{k=1}^{K} \mathbf{f}_{I_k}^{+}, \mathbf{f}_{BG} = \sum_{k=1}^{K} \mathbf{f}_{I_k}^{-}$. Finally, the saliency of $\mathcal{N}_x$ is propagated to its contained pixels,

$$s(x) = \sum_{\mathcal{N}_x} w_{\mathcal{N}_x} \cdot s(\mathcal{N}_x), \quad (24)$$

where $w_{\mathcal{N}_x}$ is a normalized Gaussian weight measuring the spatial distance of the pixel x to the geometrical center of the neighborhood region $\mathcal{N}_x$.

Alternatively, based on the observation that different features contribute differently to the saliency analysis on each image, Mai *et al.* [120] propose to learn the image specific rather than universal weights to fuse the saliency maps that are computed on different feature channels. To this end, the CRF aggregation model of saliency maps is trained only on the retrieved similar images to account for the dependence of aggregation on individual images[6].

Similar image retrieval works well if large-scale image collections are available. Saliency annotation is time consuming, tedious, and even intractable on such collections, however. To this end, some methods try to leverage the *unannotated* similar images. With the web-scale image collections $\mathcal{C}$, Wang *et al.* [119] propose a simple yet effective saliency estimation algorithm. The pixel-wise saliency map is computed as:

$$s(x) = \sum_{k=1}^{K} ||I(x) - \tilde{I}_k(x)||_1, \quad (25)$$

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

Siva *et al.* [121] propose a probabilistic formulation for saliency computation as a sampling problem. A patch p_x is considered to be salient if it has the low probability of being sampled from the images $\mathcal{C}_I \cup I$. In another word, higher saliency scores will be given to p_x if it is unique among a bag of patches extracted from the similar images.

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 input 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 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 a *single* input image.

Co-saliency detection is closely related to the concept of image co-segmentation that aims to segment similar objects from multiple images [130], [131]. As stated in [127], 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 [132], [133]. Second, some co-segmentation methods, *e.g.*, [131], need user inputs to guide the segmentation process under ambiguous situations. Third, salient object detection often serves as a pre-processing step, and thus more efficient algorithms are preferred than co-segmentation algorithms, especially over a large number of images.

Li and Ngan [125] propose a co-saliency detection method from an image pair that contain some objects in common. The co-saliency is defined as the inter-image correspondence, *i.e.*, low saliency values should be given to the dissimilar regions. Similarly in [126], Chang *et al.* propose to compute the co-saliency by exploiting the additional *repeatedness* property across multiple images. Specifically, the co-saliency score of a pixel is defined as the multiplication of its traditional saliency score [48] and its repeatedness likelihood over the input images.

Fu *et al.* [127] propose a cluster-based co-saliency detection algorithm by exploiting the well-established global contrast and spatial distribution concepts on a single image. Additionally, the corresponding cues over multiple images are introduced to account for the saliency co-occurrence. In specific, K_C clusters $\{C^k\}_{k=1}^{K_C}$ are first obtained from the entire input images $\{I^i\}_{i=1}^{M}$. Pixels in the image I^j are denoted as $\{x_i^j\}_{i=1}^{N_j}$. To obtain the saliency with the corresponding cue, a M-bin histogram $\mathbf{q}^k = \{q_j^k\}_{j=1}^{M}$ is adopted to describe the distribution of the cluster C^k in M images:

$$q_j^k = \frac{1}{n^k} \sum_{i=1}^{N_j} \delta(x_i^j \in C^k), j = 1, \cdots, M, \quad (26)$$

where n^k is the pixel number of the cluster C^k. The corresponding cue is then defined as:

$$w_c(C^k) = \frac{1}{1 + \text{var}(\mathbf{q}^k)}, \quad (27)$$

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

$$s_{co}(C^k) = w_c(C^k) w_g(C^k) w_s(C^k), \quad (28)$$

where $w_g(C^k)$ and $w_s(C^k)$ are the saliency derived from the global contrast and spatial distribution cues, respectively, in a similar way to the single image.

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 [122] first estimate 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:

$$s_t(\mathbf{p}_i) = \sum_{j=1}^{n} D_m(\mathbf{p}_i, \mathbf{p}_j), \quad (29)$$

where n is the number of correspondences. $D_m(\mathbf{p}_i, \mathbf{p}_j)$ captures the motion contrast between $\mathbf{p}_i$ and $\mathbf{p}_j$. For simplicity, the motion model is assumed to be homography. Specifically, M_H homographies $\{\mathbf{H}_m\}_{m=1}^{M_H}$ are estimated by RANSAC algorithms. For each homography $\mathbf{H}_m$, a set of points $\mathcal{L}_m = \{\mathbf{p}_1^m, \cdots, \mathbf{p}_{n_m}^m\}$ are considered as its inliers, with n_m being the number of inliers. The motion contrast

6. We will discuss more technical details about [120] in Sect. 2.1.4.

can be defined as:

$$D_m(\mathbf{p}_i, \mathbf{p}_j) = \epsilon(\mathbf{p}_i, \mathbf{H}_m), \tag{30}$$

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:

$$s_t(\mathbf{p}) = \sum\nolimits_{j=1}^{M} \alpha_j \epsilon(\mathbf{p}, \mathbf{H}_j). \tag{31}$$

$\alpha_j \in [0, 1]$ 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 is then assigned to its corresponding spanning region to get the final saliency map.

Liu *et al.* [123] extend their original spatial saliency features [7] 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.* [112] integrate the spatial saliency into the energy minimization framework by considering the temporal coherence constraint.

Li *et al.* [124] extend the regional contrast-based saliency to the spatio-temporal 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 are estimated based on their color, texture, and motion features in a interactive manner on an undirected un-weighted matching graph. The regional saliency is 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.* [134] through experimental analysis of the importance of depth cues for eye fixation prediction. Recently, researchers have started to study how to exploit the depth cues for salient object detection [128], [129], 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 [128], [129]. Furthermore, Niu *et al.* [128] demonstrate 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 is defined as:

$$s_{d,1}(r_i) = \begin{cases} \frac{d_{max} - \bar{d}_i}{d_{max}} & \text{if } \bar{d}_i \geq 0 \\ \frac{d_{min} - \bar{d}_i}{d_{min}} & \text{if } \bar{d}_i < 0 \end{cases} \tag{32}$$

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 are perceived popping out from the scene. The second type of regional stereo saliency is then defined as:

$$s_{d,2}(r_i) = \frac{d_{max} - \bar{d}_i}{d_{max} - d_{min}}. \tag{33}$$

The stereo saliency is linearly computed with an adaptive weight.

Salient object detection on light field. Recently the light field for salient object detection is proposed in [111]. A light field, which is 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 offers two benefits for salient object detection: 1) it allows synthesizing a stack of images focusing at different depths, 2) it provides an approximation to scene depth and occlusions.

With this additional information, Li *et al.* [111] first utilize the focusness and objectness priors to robustly choose the background and select the foreground candidates. Specifically, the layer with the estimated background likelihood score is used to estimate the background regions. Regions, coming from Mean-shift algorithm, with the high foreground likelihood score are chosen as salient object candidates. Finally, the estimated background and foreground are utilized to compute the contrast-based saliency map on the all-focus image.

#	Model	Pub	Year	Type	Code
1	**COMP** [135]	ICCV	2011	Localization	NA
2	**GSAL** [136]	CVPR	2012	Localization	NA
3	**CTXT** [137]	ICCV	2011	Segmentation	NA
4	**LCSP** [138]	IJCV	2014	Segmentation	NA
5	**BENCH** [139]	ECCV	2012	Aggregation	M
6	**SIO** [140]	SPL	2013	Optimization	NA
7	**ACT** [51]	PAMI	2012	Active	C
8	**SCRT** [52]	CVPR	2014	Active	NA
9	**WISO** [53]	TIP	2014	Active	NA

Fig. 6. Salient object detection models that are not covered in Fig. 4 and Fig. 5.

2.1.4 Other Models

In previous sections, we review models that compute a saliency map for the input image, which might be useful for some applications like image re-targeting. 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. For localization purpose, the final output of [7] is rectangles around the salient objects by converting the binary segmentations to bounding boxes.

Feng *et al.* [135] define saliency for each sliding window as its composition cost using the remaining parts of the image. Based on an over-segmentation of the input image, the local maxima, which can efficiently be found among all the sliding windows in a brute-force manner, are believed to be corresponding to salient objects.

The basic assumption of previous approaches that at least one salient object exists in the input image may not always hold as some *background images* contain no salient objects at all. In [136], Wang *et al.* investigate the problem of detecting the *existence* and the location of salient objects on thumbnail images. Specifically, 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 is learned via the Random Forest regressor 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. Yu *et al.* [95] utilize the complementary characteristics of imperfect saliency maps generated by different contrast-based saliency models. Specifically, two complementary saliency maps are 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. It is extended in [138] by learning the complementary saliency maps for salient object segmentation.

Lu *et al.* [96] exploit the convexity (concavity) prior for salient object segmentation, which assumes that the region on the convex side of a curved boundary tends to belong to the foreground. Based on this assumption, concave arcs are first found on the contours of superpixels. For a concave arc, its convexity context is defined as the windows which are tightly around the arc. An undirected weight graph is then built over the superpixels with concave arcs, where the weights between vertices are determined by the summation of concavity context on different scales in the hierarchical segmentation of the image. Finally the Normalized Cut algorithm [141] is performed to separate the salient object from the background.

In order to more effectively leverage the contextual cues, Wang *et al.* [137] propose to integrate an auto-context classifier [142] 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 and optimization 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 map to facilitate the detection of salient objects.

Denote $S_i(x)$ as the saliency value at pixel x of the i-th saliency map. In [139], Borji *et al.* propose a standard saliency aggregation method as follows:

$$S(x) = P(s_x = 1|\mathbf{f}_x) \propto \frac{1}{Z}\sum\nolimits_{i=1}^{M} \zeta(S_i(x)) \tag{34}$$

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

$$\zeta_1(z) = z;\ \zeta_2(z) = \exp(z);\ \zeta_3(z) = -\frac{1}{\log(z)}. \tag{35}$$

Inspired by the aggregation model in [139], Mai *et al.* [120] propose two aggregation solutions. The first solution also adopts the pixel-wise aggregation:

$$P(s_x = 1|\mathbf{f}_x; \lambda) = \sigma\left(\sum_{i=1}^{M} \lambda_i S_i(x) + \lambda_{M+1}\right) \tag{36}$$

where $\lambda = \{\lambda_i | i = 1, \ldots, M+1\}$ is the set of model parameters and $\sigma(z) = 1/(1+\exp(-z))$. However, it is noted that one potential problem of such direct aggregation is its ignorance of the interaction between neighboring pixels. Inspired by [8], they propose 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 are optimized on the training data and the saliency aggregation result for each pixel is the posterior probability of being labeled as salient with the trained CRF. To account for the different contributions of different kinds of saliency maps on each individual image, the training data is collected only from the most similar images to the input one.

Alternatively, Yan *et al.* [87] integrate the saliency maps computed on the hierarchical segmentations of the image into a tree-structure graphical model, where each node corresponds to a region in every hierarchy. Due to the tree structure, the saliency inference can efficiently be conducted using the belief propagation. In fact, solving the three layer hierarchical model is equivalent to applying a weighted average to all single-layer maps. Different from naive multi-layer fusion, this hierarchical inference algorithm can select optimal weights for each region instead of global weighting.

Li *et al.* [140] propose 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. Based on the correlations (similarity scores) between region pairs, the saliency value of each superpixel is optimized by quadratic programming when considering the influences of all the other superpixels. Based on the correlations (similarity scores) w_{ij} of two regions r_i and r_j, the saliency values $\{s_i\}_{i=1}^N$ (denote $s(r_i)$ as s_i for short) can be optimized by solving:

$$\begin{aligned} \min_{\{s_i\}_{i=1}^N} & \sum_{i=1}^{N} s_i \sum_{j\neq i}^{N} w_{ij} + \lambda_c \sum_{i=1}^{N} s_i e^{d_i/d_D} \\ & + \lambda_r \sum_{i=1}^{N}\sum_{j\neq i}^{N} (s_i - s_j)^2 w_{ij} e^{-d_{ij}/d_D} \\ s.t.\ \ & 0 \le s_i \le 1, \forall i, \ \text{ and } \sum_{i=1}^{N} s_i = 1. \end{aligned} \tag{37}$$

where d_D is half the image diagonal length. d_{ij} and d_i are the spatial distances from the r_i to r_j and the image center, respectively. In the optimization, the saliency value of each superpixel is optimized by quadratic programming when considering the influences of all the other superpixels. Similarly, Zhu *et al.* [109] also adopt such optimization-based framework to integrate multiple foreground/background cues as well as the smoothness terms to automatically infer the optimal saliency values.

The Bayesian framework is adopted to more effectively integrate the complementary dense and sparse reconstruction errors [102]. A fully-connected Gaussian Markov Random Field between each pair of regions is constructed to enforce the consistency between salient regions [105],

which leads to an efficient computation of the final regional saliency scores.

Active models: Inspired by the interactive segmentation models (seen Sec. 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 Sec. 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.* [51] combine 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 is segmented in polar space. Li *et al.* [52] propose 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 [53] proposes 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 adopt annotations to determine the upper-bound of the segmentation performance, propose datasets with multiple objects in scenes, and provide new insight to the inherent connections of fixation prediction and salient object segmentation.

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 [46], [143]–[145] 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.* [1]). 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 [146].

Fixation prediction models have been used to predict where people look during free-viewing of static natural scenes (*e.g.*, [17], [19], [20], [147] or dynamic scenes/vidoes (*e.g.*, [148]–[150]) by employing motion, flicker, optical flow (*e.g.*, [151]), or spatiotemporal interest points learned from image regions at fixated locations (*e.g.*, [152] 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 [5], [153]. 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 has been used to scale up current models and reach the human performance.

#	Model	Ref.	Pub	Year	Code
1	**IT**	Itti *et al.* [1]	PAMI	1998	M
2	**AIM**	Bruce & Tsotsos [6]	JOV	2006	M
3	**GB**	Harel *et al.* [164]	NIPS	2007	M + C
4	**SR**	Hou & Zhang [18]	CVPR	2007	M
5	**SUN**	Zhang *et al.* [165]	JOV	2008	M
6	**SeR**	Seo & Milanfar [61]	JOV	2009	M
7	**SIM**	Murray *et al.* [166]	CVPR	2011	M
8	**SS**	Hou *et al.* [60]	PAMI	2012	M
9	**COV**	Erdem & Erdem [62]	JOV	2013	M
10	**BMS**	Zhang *et al.* [110]	ICCV	2013	M + C

Fig. 7. Fixation prediction models. JOV = Journal of Vision.

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.* [154] propose that objects are better predictors of fixations than bottom-up saliency. Although we have shown that this study has some shortcomings [155], there is evidence that object might be important in guiding attention and fixations [146], [156]–[158]. Cerf *et al.* [159] show that faces and text attract human gaze. Li *et al.* [160] propose a saliency model (denoted as SP) that can effectively predict human fixation by incorporating the prior knowledge learned from millions of unlabeled images. Subramanian *et al.* [161], by recording eye fixations over a large affective image dataset, observe that fixations are directed toward emotional and action stimuli and duration of fixations are longer on such stimuli. Similarly, Judd *et al.* [17], by plotting image regions at the top salient locations of the human saliency map (made of fixations), observe that humans, faces, cars, text, and animals attract human gaze probably because they convey more information in a scene. Borji et al. [157] have shown that gaze direction guides eye movements in free-viewing of natural scenes. Alongside, some personal characteristics such as experience, age, gender, and culture change the way humans look at images (*e.g.*, [162], [163]). Some models can handle both fixation prediction and salient object detection (*e.g.*, BMS [110]). Some authors have thresholded their saliency maps (*e.g.*, top 20% of activation) to detect proto-objects (*e.g.*, COV [62] models).

Fig. 7 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 [144], [145].

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 [167], Normalized Cut [141], and Ratio Cut [168]. A representative graph-based image segmentation algorithm is [91], 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 [55] 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 have been generated based on the gPb edge detector [13] by converting the ultrametric contour map (UCM) into a hierarchical region tree using oriented watershed transformation (OWT).

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 have been proposed recently. Quick-Shift [169], 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 [56] generates superpixels efficiently by clustering the pixels in terms of color and spatial distance using the k-means algorithm. Turbo pixels [92], 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 [170] and object detection [171].

Interactive segmentation algorithms have also been developed in recent years, where the goal is to separate the foreground object from the background with the help of users. Some scratches [172] or a bounding box [173] 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 [174] 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 [175].

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 contains an object rather than other 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.

#	Model	Pub	Year	Output	Code
1	**OBJ** [15]	CVPR	2010	BB	M+C
2	**CPMC** [176]	CVPR	2010	SG	M
3	**CIOP** [16]	ECCV	2010	SG	M+C
4	**LOC** [177]	ICCV	2011	BB	C
5	**SELECT** [178]	ICCV	2011	BB	M
6	**SS** [179]	ECCV	2012	SG	M+C
7	**LCP** [180]	ACCV	2012	BB	NA
8	**PRIME** [181]	ICCV	2013	BB	M+C
9	**BING** [57]	CVPR	2014	BB	C
10	**GLS** [182]	CVPR	2014	SG	M+C
11	**DNN** [183]	CVPR	2014	BB	NA
12	**MCG** [184]	CVPR	2014	SG	M+C
13	**EB** [185]	ECCV	2014	BB	M+C
14	**GOP** [186]	ECCV	2014	SG	M+C
14	**DDO** [187]	PAMI	2014	SG	NA
15	**VOP** [188]	CVPRW	2012	SG	NA
16	**ODSA** [189]	ICRA	2013	SG	M+C
17	**STOP** [190]	ECCV	2014	SG	NA

Fig. 8. Object proposal generation models. For the proposal form, {BB = bounding boxes, SG = segments}.

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 is usually utilized as a pre-processing step to accelerate the object detection. On the other hand, while saliency can be used as a visual cue to determine the objectness, a generic object is not necessarily to be visually salient. Siva *et al.* [121] demonstrate the saliency map can be utilized to generate object proposals by sampling. It is worth noting that the salient object localization algorithms 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.

Following the seminal work of objectness [15], each bounding box is jointly described by a set of features. Alexe *et al.* [15] argue that some generic characteristics, such as boundary closure, high local contrast, and sometimes visual saliency, are shared by objects of any categories. Boundary edge distribution and window symmetry [177], binarized normed gradients (BING) feature resized to a fixed size (*e.g.*, 8×8) [57], and the number of contours intersecting the box's boundary [185] are investigated in later works. These features are fed into a trained classifier, *e.g.*, the Naive Bayes classifier [15] or the linear SVM [57], to determine how likely it is for a bounding box to contain an object of any class.

In addition, Ristin *et al.* [180] demonstrate that randomly sampled local patches may provide contextual information to estimate the prior distribution of object locations. Recently, the Deep Neural Network (DNN) is adopted to generate bounding box object proposals [183], where proposal generation is defined as a regression problem to the coordinates and objectness scores of output bounding boxes.

Meanwhile, some other researchers focus on ranking the segments (regions) of the input image, where a set of object hypotheses are generated as a set of figure-ground seg-

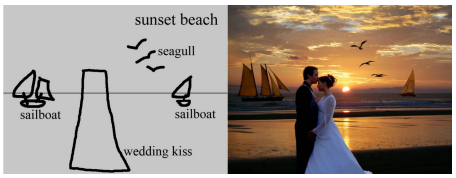

(a) Content aware resizing [59]


(b) Image collage [191]

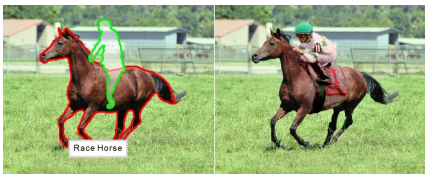
(c) View selection [192]


(d) Unsupervised learning [193]


(e) Mosaic [194]



(f) Image montage [195]



(g) Object manipulation [196]

(h) Semantic colorization [197]

Fig. 9. Sample applications of salient object detection. Images are reproduced from corresponding references.

mentations of the input image. Initially, object hypotheses are first automatically extracted by solving a sequence of constrained parametric min-cut problems (CPMC) on the pixel grids [176] or labeling the superpixels as foreground or background on the CRF framework based on seed regions [16]. Such coarse region proposals are then learned to rank using the Random Forest [176] or the structured learning framework [16]. In [186], object proposals are generated by adaptively thresholding the signed geodesic distance transform based on the constructed foreground and background masks, which are derived from learned seed superpixels.

Another possible way of generating regional object proposals is to merge the superpixels resulting from an over-segmentation of the input image. The selective search algorithm [178] gradually merges adjacent regions to form hierarchical segmentations to generate object proposals. In [181], the proposals are defined as the random partial spanning trees of an undirected graph where the vertices correspond to superpixels and the weights capture the similarity that two neighboring superpixels. In [184], 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. The local selective search strategy by grouping and the global search strategy by figure-ground segmenting the image are combined to form object proposals in a recent work [182].

Alternatively, object proposals can be generated by leveraging an auxiliary dataset. Kim and Grauman [179] introduce a global category-independent shape prior for object proposal generation. By performing a series of figure-ground segmentations using graph-cuts based on the shape priors that are derived from the associated exemplar shapes of the database, a set of object proposals is generated. Similarly, large collections of example object regions are also maintained in [187]. Objectness score of each segment of the input image can be computed based on its nearest neighbors in the database by combining multiple properties of each exemplar region.

Those models discussed above are all working on a single input image. Object proposals can also be generated to leverage the spatio-temporal cues [188], [190] and multi-modal data [189]. Similar to still images, video object proposals can be generated in a global way as figure-ground segmentations of each frame based on the seed spatio-temporal object hypotheses [188], which are obtained by linking object proposals [16] independently extracted at each frame. Alternatively, supervoxels can be randomly merged to form video object proposals [190] in a local manner. Karpathy *et al.* [189] generate object proposals from 3D meshes of indoor environments. The scene is first decomposed into a set of candidate mesh segments. Several intrinsic shape cues are then extracted to jointly compute the objectness score of each segment.

Fig. 8 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 [57]. On the other hand, region proposals are more natural for object representation. It is worth pointing out that region object proposals can be converted to bounding boxes by fitting a tight rectangle around the segment.

2.3 Applications of Salient Object Detection

The value of salient object detection models lies on their applications in many fields of computer vision, graphics, and robotics. Salient object detection models have been utilized for several applications such as object detection and recognition [198]–[204], image and video compression [205], [206], video summarization [207]–[209], photo collage/media re-targeting/cropping/thumb-nailing [191], [210], [211], image quality assessment [212]–[214], image segmentation [215]–[218], content-based image retrieval and image collection browsing [195], [219]–[221], image editing and manipulating [192], [194], [196], [197], visual tracking [222]–[228], object discovery [189], [229], and human-robot interaction [230], [231]. Fig. 9 shows some 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 aim to collect images with salient objects being annotated with bounding boxes (*e.g.*, **MSRA-A** and **MSRA-B** [7]), while later efforts annotate such salient objects with pixel-wise binary masks (*e.g.*, **ASD** [81] and **DUT-OMRON** [101]). 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 backgrounds (*e.g.*, [45], [52], [53]). 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 [53] and Li *et al.* [52] use 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. 10, which contains 20 datasets with still images and only 2 dataset for evaluating salient object detection models in video. This fact implies that more video datasets are required 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 salient object only in each scene (*e.g.*, [7]) or annotate the most salient one among several candidates (*e.g.*, [45])

3.1.1 MSRA and Its Descendants

- **MSRA-A & MSRA-B** [7][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 is resized to have a maximum side length of 400 pixels and the salient object(s) is manually annotated by rectangles (3 subjects for **MSRA-A** and 9 subjects for **MSRA-B**). Since the bounding box is inaccurate and often fails to reveal the accurate object boundaries, this dataset can be best used for salient object localization rather 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 centers. Consequently, some other datasets built upon the images from this dataset also suffer from this drawback.
- **ASD** [81][9]: This dataset (a.k.a., **MSRA1K**) is the most popular dataset in the literature. It contains 1000 images from the **MSRA-B** with a binary pixel-wise object mask 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 even using simple algorithms. Actually, recent approaches seem to reach an "upper bound" on the **ASD** dataset, which poses a pressing demand for larger datasets with more complex images (*e.g.*, multiple objects in a cluttered background).
- **MSRA5K** [115][10]: In a recent work, Jiang *et al.* fully annotate the 5,000 images from the **MSRA-B** dataset with pixel-wise binary masks.
- **MSRA10K** [86][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 [86].

3.1.2 Other Datasets

- **SOD** [234][12]: This dataset is a collection of salient object boundaries based on Berkeley Segmentation Dataset (BSD) [14]. In the annotation process, seven subjects are 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 [99], a binary salient object mask in each image is 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** [131][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 a pixel-wise annotation that may cover one or multiple salient objects. It is used for evaluating the salient object detection models in [115].
- **SED** [233][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 [139], Borji *et al.* demonstrate that saliency object detection models usually perform significantly worse on **SED2** than on the simple datasets (*e.g.*, **ASD**). In their recent work [45], Borji *et al.* further ask 70 observers to select the most salient objects among the two objects in each image of **SED2**.
- **CSSD & ECSSD** [87][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**), is proposed in [87]. It is expanded to the Extended Complex Scene Saliency Dataset (**ECSSD**) containing 1000 semantically meaningful but structurally complex images. Images in these two datasets are acquired from the **BSD** dataset [14], PASCAL VOC [242] and the Internet. The binary masks for the salient objects are produced by 5 subjects.
- **ImgSal** [223][16]: Since the problems of salient object detection and fixation prediction are tightly correlated with each other, it would be valuable to construct an

7. Some spatio-temporal salient object detection models proposed to use the surveillance video with foreground targets annotated by rectangles for quantitative evaluation
8. http://research.microsoft.com/en-us/um/people/jiansun/
9. http://ivrgwww.epfl.ch/supplementary_material/RK_CVPR09/
10. http://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/

#	Dataset	Reference	Year	Images	Objects	Annotation	Resolution	Annotators	Eye data	Image/Video
1	**MSRA-A**	[7], [232]	2007	20K	∼1	Bounding Box	400 × 300	3	-	I
2	**MSRA-B**	[7], [232]	2007	5K	∼1	Bounding Box	400 × 300	9	-	I
3	**SED1**	[139], [233]	2007	100	1	Pixel-wise	∼300 × 225	3	-	I
4	**SED2**	[139], [233]	2007	100	2	Pixel-wise	∼300 × 225	3	-	I
5	**ASD**	[7], [81]	2009	1000	∼1	Pixel-wise	400 × 300	1	-	I
6	**SOD**	[13], [234]	2010	300	∼3	Pixel-wise	481 × 321	7	-	I
7	**iCoSeg**	[131]	2010	643	∼1	Pixel-wise	∼500 × 400	1	-	I
8	**MSRA5K**	[7], [97]	2011	5K	∼1	Pixel-wise	400 × 300	1	-	I
9	**Infrared**	[235], [236]	2011	900	∼5	Pixel-wise	1024 × 768	2	15	I
10	**ImgSal**	[223]	2013	235	∼ 2	Pixel-wise	640 × 480	19	50	I
11	**CSSD**	[87]	2013	200	∼1	Pixel-wise	∼400 × 300	1	-	I
12	**ECSSD**	[87], [237]	2013	10K	∼1	Pixel-wise	∼400 × 300	1	-	I
13	**MSRA10K**	[7], [238]	2013	10K	∼1	Pixel-wise	400 × 300	1	-	I
14	**THUR15K**	[7], [238]	2013	15K	∼1	Pixel-wise	400 × 300	1	-	I
15	**DUT-OMRON**	[101]	2013	5,172	∼5	Bounding Box	400 × 400	5	5	I
16	**Bruce-A**	[6], [45]	2013	120	∼4	Pixel-wise	681 × 511	70	20	I
17	**Judd-A**	[53], [239]	2014	900	∼5	Pixel-wise	1024 × 768	2	15	I
18	**PASCAL-S**	[52]	2014	850	∼5	Pixel-wise	variable	12	8	I
19	**UCSB**	[46]	2014	700	∼5	Point-wise Clicks	405 × 405	100	8	I
20	**OSIE**	[156]	2014	700	∼5	Pixel-wise	800 × 600	1	15	I
21	**RSD**	[240]	2009	62,356	variable	Bounding Box	variable	23	-	V
22	**STC**	[241]	2011	4,870	∼1	Bounding Box	variable	1	-	V

Fig. 10. Overview of popular salient object datasets (sorted based on the year). Top: image datasets, Bottom: video datasets.

image dataset with both binary masks and human fixations. Toward this end, Li *et al.* introduce a dataset in [223] with these two kinds of information. In particular, they divide 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
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** [243][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 this dataset.
- **DUT-OMRON** [101][18]: This 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 backgrounds. In the annotation process, each image is resized to have a maximum side length of 400 pixels. Both bounding boxes and pixel-wise object masks are provided for each image. Additionally, the fixation data is 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 a feasible way to explore the latent connections between these three research fields.
- **Bruce-A & Judd-A** [45], [53][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** [6] and **Judd** also known as **MIT** [17]) 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. This means that eye movement center-bias in these datasets is due to photographer bias which is the tendency of photographers to frame salient objects at the image center [153].
- **PASCAL-S** [52][20]: In a similar effort to [45], this dataset contains annotations of the most salient objects over complex scenes taken from PASCAL VOC [242] 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.* take annotations of the PASCAL dataset and recorded eye movements of observers on them.

In addition to salient object datasets listed in Fig. 10, there exist some other datasets where objects, instead of "salient" objects, are manually annotated (*e.g.*, [43], [244]). 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 [45], by summarizing

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/

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 also be turned into salient object datasets. Intuitively, the object around the peak of the fixation density map can be selected as the most salient one (as in [53], **UCSB** [46], [156] and **OSIE** [156]). 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.

What is a representative suitable dataset for salient object evaluation? We believe that for model benchmarking, it is important to assess models over *a 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 annotation of multiple objects (as in [243]) or tracking human fixations (as in [52]) is time consuming. In particular, when processing images with multiple foreground objects and a complex background, the subjective annotations from various users may diversify from each other (as in [45]). Toward this end, existing studies often adopt several datasets with different properties for model evaluation.

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:

$$Precision = \frac{|M \cap G|}{|M|}, \quad Recall = \frac{|M \cap G|}{|G|} \tag{38}$$

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.* [81] propose the image-dependent adaptive threshold for binarizing S, which is computed as twice the mean saliency of S:

$$T_a = \frac{2}{W \times H} \sum_{x=1}^{W} \sum_{y=1}^{H} S(x, y), \tag{39}$$

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 [86]). 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 [173]. After several iterations, the binary mask can be gradually refined.

F-measure. Often, 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:

$$F_\beta = \frac{(1 + \beta^2) Precision \times Recall}{\beta^2 Precision + Recall}. \tag{40}$$

As suggested by many salient object detection works (*e.g.*, [81]), β^2 is often 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 [8]). For instance, 100% recall can be easily achieved by setting the whole region to foreground.

According to the different ways of binarizing a saliency map, 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 (Eq. 40), 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:

$$TPR = \frac{|M \cap G|}{|G|}, \quad FPR = \frac{|M \cap G|}{|M \cap G| + |\bar{M} \cap \bar{G}|} \tag{41}$$

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 distills 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 of around 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 fails to detect non-salient regions. Moreover, in some application scenarios [58], the quality of the weighted, continuous saliency maps may be of higher importance than the binary masks. For a more comprehensive comparison it is recommended to 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:

$$MAE = \frac{1}{W \times H} \sum_{x=1}^{W} \sum_{y=1}^{H} ||S(x,y) - G(x,y)||. \quad (42)$$

4 Discussions

4.1 Design Choices

In the past decades, hundreds of methods for salient object detection have been proposed and a large number of design choices have been explored. Our detailed method summarization (see Fig. 4 & 5) does suggest some clear messages about the commonly used design choices, which are valuable for the design of future algorithms:

4.1.1 Block-based vs. Region-based

From the chronologically 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 elements helps to abstract out unnecessary details and is important for high quality saliency detection [54], [89]. Third, the region itself contains some important cues which could be missing at pixel/patch level, such as shapes, aspect ratios, and perimeter [115]. As a result, the region-based method, RC [86], achieves 90% segmentation precision in the most widely used benchmark [81]. It outperforms previously best-reported results (75% segmentation precision in the pixel-based method FT [81]) 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.

4.1.2 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 3 purely intrinsic cues based methods (CA, RC, and CB) among the top 5 methods (see [139]).

There is a consistent trend of moving from local cues to global cues, possibly because the latter tends to assign similar saliency values across similar image regions rather than highlighting only the boundary regions. Furthermore, some regional priors have been widely studied in recent works, *e.g.*, the backgroundness prior, which is tightly correlated with the center prior of salient objects. (Some others might prefer to treat it as the center-bias of the dataset. See more discussions in Sec. 4.3.)

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.

4.1.3 Heuristic vs. Learning From Data

Like the early stages of other areas [245], [246], most existing saliency studies still focused on creating effective features and using heuristic models to detect salient object [54], [81], [86], [101]. To date, various features have been shown to be helpful to salient object detection (see Fig. 4 and Fig. 5), including the local contrast, global contrast, edge density, backgroundness, focuses, 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. By contrast, a classifier is capable of easily integrating multiple cues and automatically discovering the discriminative ones.

The simplicity and training-free properties of many successful salient object detection models has been an attractive 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 conflict with the easy useability? Unlike other classical computer vision problems, *e.g.*, generic object detection, classification, the data-driven approach in salient object detection seems to have surprisingly good generalization ability. Despite the huge characteristic differences among datasets evaluated [115] (see Sec. 3), the DRFI approach is only trained on a small subset of **MSRA5K**, and still consistently outperforms other methods on all other dataset. These encouraging results suggest that we might be able to further explore data-driven salient object detection without losing the simplicity and easy-to-use generality in the application side.

4.2 Salient Object Detection, Fixation Prediction, and Object Proposal Generation

The attentive visual search mechanisms have been studied in different background and problem focuses: including salient object detection [8], [86], [115], [139], fixation prediction [1], [18], [60], and object proposal generation [57], [64]–[67]. 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 humans look, *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 of these two areas have made great progress in the last few decades and enable 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). Unlike salient object detection and fixation prediction, object proposal generation models typically aim at producing 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 always a major objective.

According to the object-based attention theory [12], [247], [248], brain groups similar pixels into proto-objects and the saliency of proto-objects is estimated and incorporated together. Strictly speaking, attentional focus and gaze do not always coincide: attentional focus can be directed to new target without accompanied eye-movements [249], [250]. However, a strong correlation between fixations and salient objects exists and the definition of a salient object is highly

consistent among human subjects [52], [251]. Object proposal generation and salient object detection are highly correlated as well; saliency estimation is even explicitly used as a cue for objectness methods [15], [121].

Recent study [248] suggests that the unit of attention depends on the task, the field of view, and the observer's intention [252]. Attention might adopt a spatial-based behavior within complex 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 [253].

4.3 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/unlimitted problem/application domain, and contains a certain degree of bias. 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* [254]. 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 [52], [86]) in most frequently used salient object benchmark datasets [81]. Thus two practices in dataset construction are proffered: i) *having independent image selection and annotation process* [52], 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 a 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 [17], [144], [153].

Bias is often closely related with application task and sometimes could be deliberately utilized as a prior in a specific task to improve the performance of an algorithm. For instance, for aesthetics reasons (*e.g.*, rule of thirds), photographers tend to frame the salient object near the center of the image [153], [255]. From an application point of view, most images we are dealing with are intentionally captured by humans with the salient object away from image borders (except images from surveillance camera and driving recorder).

4.4 Promising Future Directions

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

4.4.1 Beyond Working with Single Image

Most benchmarks and saliency models discussed in this study are about a single input image. Unfortunately, salient object detection on multiple input images, *e.g.*, salient object detection on video sequences, co-salient object detection, and salient object detection with depth/light field, are ignored. The additional input data is becoming more and more cheap. For example, with the popularity of the consumer depth camera (*e.g.*, Kinect), incorporating depth cues for salient object detection will be easier. It is usually nontrivial, however, to adapt existing single-image-input models to these scenarios. Integrating additional cues such as spatio-temporal consistence and depth will be beneficial for salient object detection.

One possible reason for lacking studies on multiple input images might be the limited availability of benchmark datasets (recall the booming of image saliency models after the publication of **MSRA-B** and **ASD**). For example, as we introduced in Sec. 3, there are only two publicly available benchmark datasets for salient object detection on videos 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. As a future work, it is urgent to build up benchmark datasets for different scenarios.

Neurobiological evidence is another issue when detecting salient objects on multiple images. Take the video scenario for instance, 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 propose to use motion consistency (*e.g.*, [122], [207], [256]) as a saliency measure, while [35] argue 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.

Finally, similar to the single-input-image scenario, efficiency is usually met with high (even higher) demand when designing algorithms. For instance, 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. To sum up, these three challenges, especially the benchmark challenge, should be addressed first before the booming of video-based salient object detection models.

4.4.2 Other Directions

Traditionally, saliency models have added feature channels such as face, car, animal, text, etc 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.

Currently, we directly compare models against the annotation data, one might also consider comparing models

based on their accuracy in specific applications, for instance in image thumbnailing or object detection. The 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.*, [13]). In addition, 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 [136]. The performance of algorithms on such background images needs to be investigated further.

An emerging trend is active salient object segmentation (see Sec. 2.1.4). 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 from other fields? What is the best way to tackle the 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.

5 Summary and Conclusion

In this paper, we exhaustively review salient object detection models and closely related areas to it.

The numerous works on salient object detection call for a methodological approach for evaluating results. We review a large body of work in saliency modeling and discussed their pros and cons. We categorize the related works in three divisions including: *classic salient object/region detection and segmentation*, *fixation prediction*, and *category-independent objectness measurement or object proposals generation*.

Detecting and segmenting salient objects is 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 necessity of training or annotation.

Future works should focus on better situating salient object detection models among other related areas such as fixation prediction, object proposal generation, 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 as well. Furthermore, 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 do 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 detection algorithm 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[21]. — Jitendra Malik

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 to run robustly [86], [191], [192], [195], [197], [257] and even supports unsupervised learning [193].

References

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

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

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

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

[5] 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.

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

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

[8] 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.

[9] R. Achanta, F. Estrada, P. Wils, and S. Süsstrunk, "Salient region detection and segmentation," in *Comp. Vis. Sys.*, 2008.

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

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

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

21. http://www.cs.berkeley.edu/∼malik/student-tree-2010.pdf

[13] 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.

[14] 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.

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

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

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

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

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

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

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

[22] 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.

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

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

[25] 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.

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

[27] 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.

[28] 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.

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

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

[31] 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.

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

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

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

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

[36] 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.

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

[38] J. Vogel and B. Schiele, "A semantic typicality measure for natural scene categorization," in *Pattern Recognition*, 2004.

[39] K. A. Ehinger, J. Xiao, A. Torralba, and A. Oliva, "Estimating scene typicality from human ratings and image features," in *Annual Cognitive Science Conference*, 2011.

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

[41] N. H. Mackworth and A. J. Morandi, "The gaze selects informative details within pictures," *Percep. & Psyc.*, 1967.

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

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

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

[45] 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.

[46] K. Koehler, F. Guo, S. Zhang, and M. P. Eckstein, "What do saliency models predict?" *J. Vision*, 2014.

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

[48] S. Goferman, L. Zelnik-Manor, and A. Tal, "Context-aware saliency detection," *IEEE TPAMI*, vol. 34, no. 10, 2012.

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

[50] 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.

[51] A. K. Mishra, Y. Aloimonos, L. F. Cheong, and A. Kassim, "Active visual segmentation," *IEEE TPAMI*, vol. 34, 2012.

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

[53] A. Borji, "What is a salient object? a dataset and a baseline model for salient object detection," in *IEEE TIP*, 2014.

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

[55] D. Comaniciu and P. Meer, "Mean shift: A robust approach toward feature space analysis," *IEEE TPAMI*, vol. 24, 2002.

[56] 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, 2012.

[57] 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.

[58] S. Avidan and A. Shamir, "Seam carving for content-aware image resizing," in *ACM TOG*, vol. 26, no. 3, 2007, p. 10.

[59] 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.

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

[61] H. J. Seo and P. Milanfar, "Static and space-time visual saliency detection by self-resemblance," *J. Vision*, 2009.

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

[63] 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.

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

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

[66] J. Uijlings, K. van de Sande, T. Gevers, and A. Smeulders, "Selective search for object recognition," *IJCV*, vol. 104, 2013.

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

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

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

[70] R. Desimone and J. Duncan, "Neural mechanisms of selective visual attention," *Annual review of neuroscience*, 1995.

[71] 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.

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

[73] 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.

[74] 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.

[75] B. C. Ko and J.-Y. Nam, "Automatic object-of-interest segmentation from natural images," in *ICPR*, 2006, pp. 45–48.

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

[77] Y. Hu, D. Rajan, and L.-T. Chia, "Robust subspace analysis for detecting visual attention regions in images," in *ACM Multimedia*, 2005, pp. 716–724.

[78] R. Vidal, Y. Ma, and S. Sastry, "Generalized principal component analysis (gpca)," *IEEE TPAMI*, vol. 27, no. 12, 2005.

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

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

[81] R. Achanta, S. Hemami, F. Estrada, and S. Süsstrunk, "Frequency-tuned salient region detection," in *CVPR*, 2009.

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

[83] 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.

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

[85] Z. Yu and H.-S. Wong, "A rule based technique for extraction of visual attention regions based on real-time clustering," *IEEE TMM*, pp. 766–784, 2007.

[86] M.-M. Cheng, N. J. Mitra, X. Huang, P. H. S. Torr, and S.-M. Hu, "Global contrast based salient region detection," *IEEE TPAMI (CVPR 2011)*, 2014.

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

[88] 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.

[89] 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.

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

[91] P. F. Felzenszwalb and D. P. Huttenlocher, "Efficient graph-based image segmentation," *IJCV*, pp. 167–181, 2004.

[92] A. Levinshtein, A. Stere, K. N. Kutulakos, D. J. Fleet, S. J. Dickinson, and K. Siddiqi, "Turbopixels: Fast superpixels using geometric flows," *IEEE TPAMI*, pp. 2290–2297, 2009.

[93] A. Adams, J. Baek, and M. A. Davis, "Fast high-dimensional filtering using the permutohedral lattice," in *Computer Graphics Forum*, vol. 29, no. 2, 2010, pp. 753–762.

[94] 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.

[95] H. Yu, J. Li, Y. Tian, and T. Huang, "Automatic interesting object extraction from images using complementary saliency maps," in *ACM Multimedia*, 2010, pp. 891–894.

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

[97] H. Jiang, J. Wang, Z. Yuan, T. Liu, and N. Zheng, "Automatic salient object segmentation based on context and shape prior," in *BMVC*, 2011.

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

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

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

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

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

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

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

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

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

[107] 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 *AAAI*, 2013.

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

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

[112] E. Rahtu, J. Kannala, M. Salo, and J. Heikkilä, "Segmenting salient objects from images and videos," in *ECCV*, 2010.

[113] P. Khuwuthyakorn, A. Robles-Kelly, and J. Zhou, "Object of interest detection by saliency learning," in *ECCV*, 2010.

[114] P. Mehrani and O. Veksler, "Saliency segmentation based on learning and graph cut refinement." in *BMVC*, 2010, pp. 1–12.

[115] 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.

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

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

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

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

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

[121] 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.

[122] Y. Zhai and M. Shah, "Visual attention detection in video sequences using spatiotemporal cues," in *ACM Multimedia*, 2006, pp. 815–824.

[123] T. Liu, N. Zheng, W. Ding, and Z. Yuan, "Video attention: Learning to detect a salient object sequence," in *ICPR*, 2008.

[124] S. Bin, Y. Li, L. Ma, W. Wu, and Z. Xie, "Temporally coherent video saliency using regional dynamic contrast," *IEEE TCSVT*, vol. 23, no. 12, pp. 2067–2076, 2013.

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

[126] 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.

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

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

[129] K. Desingh, K. M. Krishna, D. Rajan, and C. Jawahar, "Depth really matters: Improving visual salient region detection with depth," in *BMVC*, 2013.

[130] 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.

[131] 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.

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

[133] 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.

[134] 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.

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

[136] 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.

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

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

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

[140] 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.

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

[142] 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.

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

[144] 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.

[145] 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.

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

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

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

[149] 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*, 2009.

[150] 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.

[151] L. Itti, "Quantitative modelling of perceptual salience at human eye position," *Visual cognition*, 2006.

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

[153] B. W. Tatler, "The central fixation bias in scene viewing: Selecting an optimal viewing position independently of motor biases and image feature distributions," *J. Vision*, 2007.

[154] W. Einhäuser, M. Spain, and P. Perona, "Objects predict fixations better than early saliency," *J. Vision*, 2008.

[155] 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," *J. Vision*, vol. 13, no. 10, p. 18, 2013.

[156] J. Xu, M. Jiang, S. Wang, M. S. Kankanhalli, and Q. Zhao, "Predicting human gaze beyond pixels," *J. Vision*, 2014.

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

[158] H.-C. Wang and M. Pomplun, "The attraction of visual attention to texts in real-world scenes," *J. Vision*, 2012.

[159] 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.

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

[161] 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," *J. Vision*, 2014.

[162] 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, 2012.

[163] H. F. Chua, J. E. Boland, and R. E. Nisbett, "Cultural variation in eye movements during scene perception," *PNAS*, 2005.

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

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

[166] 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.

[167] Z. Wu and R. M. Leahy, "An optimal graph theoretic approach to data clustering: Theory and its application to image segmentation," *IEEE TPAMI*, 1993.

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

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

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

[171] T. Malisiewicz and A. A. Efros, "Improving spatial support for objects via multiple segmentations," in *BMVC*, 2007.

[172] 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.

[173] C. Rother, V. Kolmogorov, and A. Blake, ""GrabCut": interactive foreground extraction using iterated graph cuts," *ACM TOG*, vol. 23, no. 3, pp. 309–314, 2004.

[174] J. Wang and M. F. Cohen, "Image and video matting: A survey," *Foundations and Trends in Comp. Graph. and Vis.*, 2007.

[175] M. Maire, S. X. Yu, and P. Perona, "Hierarchical scene annotation," in *BMVC*, 2013, pp. 84.1–84.11.

[176] J. Carreira and C. Sminchisescu, "Constrained parametric min-cuts for automatic object segmentation," in *CVPR*, 2010.

[177] E. Rahtu, J. Kannala, and M. Blaschko, "Learning a category independent object detection cascade," in *ICCV*, 2011.

[178] 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.

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

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

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

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

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

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

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

[186] P. Krhenbhl and V. Koltun, "Geodesic object proposals," in *ECCV*, 2014.

[187] H. Kang, A. Efros, T. Kanade, and M. Hebert, "Data-driven objectness," *IEEE TPAMI*, vol. 99, p. 1, 2014.

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

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

[190] D. Oneata, J. Revaud, J. Verbeek, and C. Schmid, "Spatio-temporal object detection proposals," in *ECCV*, 2014.

[191] 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.

[192] H. Liu, L. Zhang, and H. Huang, "Web-image driven best views of 3d shapes," *The Visual Computer*, 2012.

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

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

[195] T. Chen, M.-M. Cheng, P. Tan, A. Shamir, and S.-M. Hu, "Sketch2photo: internet image montage," *ACM TOG*, 2009.

[196] C. Goldberg, T. Chen, F.-L. Zhang, A. Shamir, and S.-M. Hu, "Data-driven object manipulation in images," *Computer Graphics Forum*, vol. 31, pp. 265–274, 2012.

[197] 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 TOG*, vol. 30, no. 6, p. 156, 2011.

[198] U. Rutishauser, D. Walther, C. Koch, and P. Perona, "Is bottom-up attention useful for object recognition?" in *CVPR*, 2004.

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

[200] F. Moosmann, D. Larlus, and F. Jurie, "Learning saliency maps for object categorization," in *ECCV Workshop*, 2006.

[201] A. Borji, M. N. Ahmadabadi, and B. N. Araabi, "Cost-sensitive learning of top-down modulation for attentional control," *Machine Vision and Applications*, 2011.

[202] A. Borji and L. Itti, "Scene classification with a sparse set of salient regions," in *IEEE ICRA*, 2011, pp. 1902–1908.

[203] 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*, 2013.

[204] Z. Ren, S. Gao, L.-T. Chia, and I. Tsang, "Region-based saliency detection and its application in object recognition," *IEEE TCSVT*, vol. PP, no. 99, pp. 1–1, 2013.

[205] C. Guo and L. Zhang, "A novel multiresolution spatiotemporal saliency detection model and its applications in image and video compression," *IEEE TIP*, 2010.

[206] L. Itti, "Automatic foveation for video compression using a neurobiological model of visual attention," *IEEE TIP*, 2004.

[207] 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 TMM*, 2005.

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

[209] 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*, 2012.

[210] S. Goferman, A. Tal, and L. Zelnik-Manor, "Puzzle-like collage," *Computer Graphics Forum*, 2010.

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

[212] 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.

[213] H. Liu and I. Heynderickx, "Studying the added value of visual attention in objective image quality metrics based on eye movement data," in *IEEE ICIP*, 2009, pp. 3097–3100.

[214] A. Li, X. She, and Q. Sun, "Color image quality assessment combining saliency and fsim," in *ICDIP*, vol. 8878, 2013.

[215] M. Donoser, M. Urschler, M. Hirzer, and H. Bischof, "Saliency driven total variation segmentation," in *ICCV*, 2009.

[216] Q. Li, Y. Zhou, and J. Yang, "Saliency based image segmentation," in *ICMT*, 2011, pp. 5068–5071.

[217] 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, 2013.

[218] M. Johnson-Roberson, J. Bohg, M. Bjorkman, and D. Kragic, "Attention-based active 3d point cloud segmentation," in *IEEE IROS*, 2010, pp. 1165–1170.

[219] 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.

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

[221] 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.

[222] S. Stalder, H. Grabner, and L. Van Gool, "Dynamic objectness for adaptive tracking," in *ACCV*, 2012.

[223] 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.

[224] 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*, 2012, pp. 357–366.

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

[226] D. A. Klein, D. Schulz, S. Frintrop, and A. B. Cremers, "Adaptive real-time video-tracking for arbitrary objects," in *IEEE IROS*, 2010, pp. 772–777.

[227] S. Frintrop and M. Kessel, "Most salient region tracking," in *IEEE ICRA*, 2009, pp. 1869–1874.

[228] G. Zhang, Z. Yuan, N. Zheng, X. Sheng, and T. Liu, "Visual saliency based object tracking," in *ACCV*, 2010.

[229] S. Frintrop, G. M. Garcıa, and A. B. Cremers, "A cognitive approach for object discovery," in *ICPR*, 2014.

[230] 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.

[231] Y. Sugano, Y. Matsushita, and Y. Sato, "Calibration-free gaze sensing using saliency maps," in *CVPR*, 2010.

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

[233] 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.

[234] V. Movahedi and J. H. Elder, "Design and perceptual validation of performance measures for salient object segmentation," in *IEEE CVPRW*, 2010.

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

[236] Q. Wang, P. Yan, Y. Yuan, and X. Li, "Multi-spectral saliency detection," *Elsevier PRL*, 2013.

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

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

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

[240] 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.

[241] Y. Wu, N. Zheng, Z. Yuan, H. Jiang, and T. Liu, "Detection of salient objects with focused attention based on spatial and temporal coherence," *Chinese Science Bulletin*, vol. 56, pp. 1055–1062, 2011.

[242] 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.

[243] 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.

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

[245] 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.

[246] K. Mikolajczyk and C. Schmid, "A performance evaluation of local descriptors," *IEEE PAMI*, 2005.

[247] 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.

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

[249] 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*, 2007.

[250] 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.

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

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

[253] 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*, 2008.

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

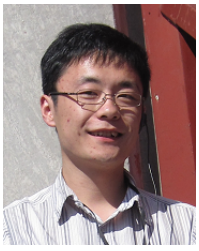
[255] R. Datta, D. Joshi, J. Li, and J. Z. Wang, "Studying aesthetics in photographic images using a computational approach," in *ECCV*, 2006, pp. 288–301.

[256] S. Li and M.-C. Lee, "Efficient spatiotemporal-attention-driven shot matching," in *ACM Multimedia*, 2007.

[257] 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.

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

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 at 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 served 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.