Salient Object Detection: A Benchmark

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

***Abstract*— We extensively compare, qualitatively and quantitatively, 41 state-of-the-art models (29 salient object detection, 10 fixation prediction, 1 objectness, and 1 baseline) over seven challenging data sets for the purpose of benchmarking salient object detection and segmentation methods. From the results obtained so far, our evaluation shows a consistent rapid progress over the last few years in terms of both accuracy and running time. The top contenders in this benchmark significantly outperform the models identified as the best in the previous benchmark conducted three years ago. We find that the models designed specifically for salient object detection generally work better than models in closely related areas, which in turn provides a precise definition and suggests an appropriate treatment of this problem that distinguishes it from other problems. In particular, we analyze the influences of center bias and scene complexity in model performance, which, along with the hard cases for the state-of-the-art models, provide useful hints toward constructing more challenging large-scale data sets and better saliency models. Finally, we propose probable solutions for tackling several open problems, such as evaluation scores and data set bias, which also suggest future research directions in the rapidly growing field of salient object detection.**

***Index Terms*— Salient object detection, saliency, explicit saliency, visual attention, regions of interest, objectness, segmentation, interestingness, importance, eye movements.**

I. Introduction

VISUAL *attention*, the astonishing capability of human visual system to selectively process only the *salient* visual stimuli in details, has been investigated by multiple disciplines such as cognitive psychology, neuroscience, and computer vision [2]–[5]. Following cognitive theories (*e.g.*, *feature integration theory (FIT)* [6], *guided search model* [7], [8]) and early attention models (*e.g.*, Koch and Ullman [9] and Itti *et al.* [10]), hundreds of computational saliency models have been proposed to detect salient visual subsets from images and videos.

Despite the psychological and neurobiological definitions, the concept of visual saliency is becoming vague in the field of computer vision. Some visual saliency models (*e.g.*, [3], [10]–[16]) aimed to *predict human fixations* as a way to test their accuracy in saliency detection, while other models [17]–[19], which were often driven by computer vision applications such as content-aware image resizing and photo visualization [20], attempted to *identify salient regions/objects* and used explicit saliency judgments for evaluation [21]. Although both types of saliency models are expected to be applicable interchangeably, their generated saliency maps actually demonstrate remarkably different characteristics due to the distinct purposes in saliency detection. For example, fixation prediction models usually pop-out sparse blob-like salient regions, while salient object detection models often generate smooth connected areas. On the one hand, detecting large salient areas often causes severe false positives for fixation prediction. On the other hand, popping-out only sparse salient regions causes massive misses in detecting salient regions and objects.

To separate these two types of saliency models, in this study we provide a precise definition and suggest an appropriate treatment of salient object detection. Generally, a salient object detection model should, *first* detect the salient attention-grabbing objects in a scene, and *second*, segment the entire objects. Usually, the output of the model is a saliency map where the intensity of each pixel represents its probability of belonging to salient objects. From this definition, we can see that this problem in its essence is a figure/ground segmentation problem, and the goal is to only segment the salient foreground object from the background. Note that it slightly differs from the traditional image segmentation problem that aims to partition an image into perceptually coherent regions.

The value of salient object detection models lies in their applications in many areas such as computer vision, graphics, and robotics. For instance, these models have been successfully applied in many applications such as object detection and recognition [22]–[30], image and video compression [31], [32], video summarization [33]–[35], photo collage/media re-targeting/cropping/thumb-nailing [20], [36], [37], image quality assessment [38]–[41], image segmentation [42]–[45], content-based image retrieval and image collection browsing [46]–[49], image editing and manipulating [50]–[53], visual tracking [54]–[60], object discovery [61], [62], and human-robot interaction [63]–[65].

The field of salient object detection develops very fast. Many new models and benchmark datasets have been proposed

Manuscript received January 5, 2015; revised July 13, 2015 and September 19, 2015; accepted October 4, 2015. Date of publication October 7, 2015; date of current version October 23, 2015. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Christine Guillemot. *(Ali Borji and Ming-Ming Cheng equally contributed to this work.)*

A. Borji is with the Computer Science Department, University of Wisconsin, Milwaukee, WI 53211 USA (e-mail: borji@uwm.edu).

M.-M. Cheng (corresponding author) is with the Department of Engineering Science, University of Oxford, Oxford OX1 3PJ, U.K. (e-mail: cmm.thu@gmail.com).

H. Jiang is with the College of Information and Computer Sciences, University of Massachusetts–Amherst, Amherst, MA 01003 USA (e-mail: hzjiang@mail.xjtu.edu.cn).

J. Li is with the State Key Laboratory of Virtual Reality Technology and Systems, School of Computer Science and Engineering, Beihang University, Beijing 100871, China, and also with the International Research Institute for Multidisciplinary Science, Beihang University, Beijing 100871, China (e-mail: jiali@buaa.edu.cn).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TIP.2015.2487833

since our earlier benchmark conducted three years ago [1]. Yet, it is unclear how the new algorithms fare against previous models and new datasets. Are there any *real improvements* in this field or we are just fitting models to datasets? It is also interesting to test the performance of old high-performing models on the new benchmark datasets. A recent exhaustive review of salient object detection models can be found in [28].

In this study, we compare and analyze models from three categories: 1) salient object detection, 2) fixation prediction, and 3) object proposal generation.[1] The reason to include the latter two types of models is to conduct across-category comparison and to study whether models specifically designed for salient object detection show actual advantage over models for fixation prediction and object proposal generation. This is particularly important since these models have different objectives and generate visually distinctive maps. We also include a baseline model to study the effect of center bias in model comparison. In summary, we hope that such a benchmark not only allows researchers to compare their models with other algorithms but also helps identify the chief factors affecting the performance of salient object detection models.

II. Salient Object Detection Benchmark

In this benchmarking, we focus on evaluating models whose input is a single image. This is due to the fact that salient object detection on a single input image is the main research direction, while the comprehensive evaluation of models working on multiple input images (*e.g.*, co-salient object detection and spatio-temporal saliency) lacks public benchmarks.

A. Compared Models

In this study, we run 41 models in total (29 salient object detection models, 10 fixation prediction models, 1 objectness proposal model, and 1 baseline) whose codes or executables were accessible (see Fig. 1 for a complete list). The baseline model, denoted as "Average Annotation Map (AAM)," is simply the average of ground-truth annotations of all images on each dataset. Note that AAM often has a larger activation at the image center (see Fig. 2), and we can thus study the effect of center bias in model comparison.

B. Datasets

Since there exist many datasets that differ in number of images, number of objects per image, image resolution and annotation form (bounding box or accurate region mask), it is likely that models may rank differently across datasets. Hence, to come up with a fair comparison, it is necessary to run models over multiple datasets so as to draw objective conclusions. A good model should perform well over almost all datasets. Toward this end, seven datasets[2] were chosen for model comparison, including: 1) **MSRA10K** [102], 2) **THUR15K** [102], 3) **ECSSD** [77], 4) **JuddDB** [103], 5) **DUT-OMRON** [78] and 6) **SED2** [1], [104], and 7) **PASCAL-S** [105]. These datasets were selected based on the following four criteria: 1) being widely-used, 2) containing a large number of images, 3) having different biases (*e.g.*, number of salient objects, image clutter, center-bias), and 4) potential to be used as benchmarks in the future research.

#	Model	Pub	Year	Code	Time(s)	Cat.
1	**AC** [66]	ICVS	2008	C	.129	Salient Object Detection
2	**FT** [18]	CVPR	2009	C	.072	
3	**CA** [67]	CVPR	2010	M + C	40.9	
4	**MSS** [68]	ICIP	2010	C	.076	
5	**SEG** [69]	ECCV	2010	M + C	10.9	
6	**RC** [70]	CVPR	2011	C	.136	
7	**HC** [70]	CVPR	2011	C	.017	
8	**SWD** [71]	CVPR	2011	M + C	.190	
9	**SVO** [72]	ICCV	2011	M + C	56.5	
10	**CB** [73]	BMVC	2011	M + C	2.24	
11	**FES** [74]	Img.Anal.	2011	M + C	.096	
12	**SF** [75]	CVPR	2012	C	.202	
13	**LMLC** [76]	TIP	2013	M + C	140.	
14	**HS** [77]	CVPR	2013	EXE	.528	
15	**GMR** [78]	CVPR	2013	M	.149	
16	**DRFI** [79]	CVPR	2013	C	.697	
17	**PCA** [80]	CVPR	2013	M + C	4.34	
18	**LBI** [81]	CVPR	2013	M + C	251.	
19	**GC** [82]	ICCV	2013	C	.037	
20	**CHM** [83]	ICCV	2013	M + C	15.4	
21	**DSR** [84]	ICCV	2013	M + C	10.2	
22	**MC** [85]	ICCV	2013	M + C	.195	
23	**UFO** [86]	ICCV	2013	M + C	20.3	
24	**MNP** [52]	Vis.Comp.	2013	M + C	21.0	
25	**GR** [87]	SPL	2013	M + C	1.35	
26	**RBD** [88]	CVPR	2014	M	.269	
27	**HDCT** [89]	CVPR	2014	M	4.12	
28	**ST** [90]	TIP	2014	M+C	79.1	
29	**QCUT** [91]	ICPR	2014	M + C	1.82	
1	**IT** [10]	PAMI	1998	M	.302	Fixation Prediction
2	**AIM** [92]	JOV	2006	M	8.66	
3	**GB** [93]	NIPS	2007	M + C	.735	
4	**SR** [94]	CVPR	2007	M	.040	
5	**SUN** [95]	JOV	2008	M	3.56	
6	**SeR** [96]	JOV	2009	M	1.31	
7	**SIM** [97]	CVPR	2011	M	1.11	
8	**SS** [98]	PAMI	2012	M	.053	
9	**COV** [99]	JOV	2013	M	25.4	
10	**BMS** [100]	ICCV	2013	M + C	.575	
1	**OBJ** [101]	CVPR	2010	M+C	3.01	-
1	**AAM**	-	-	-	-	-

Fig. 1. Compared salient object detection, fixation prediction, object proposal generation, and baseline models sorted by their publication year {M = Matlab, C = C/C++, EXE = executable}. The average running time is tested on MSRA10K dataset (typical image resolution 400 × 300) using a desktop machine with Xeon E5645 2.4 GHz CPU and 8GB RAM. We evaluate those models whose codes or executables are available.

MSRA10K is a descendant of the MSRA dataset [17]. It contains 10,000 annotated images that covers all the 1,000 images in the popular ASD dataset [18]. **THUR15K** and **DUT-OMRON** are used to compare models on a large scale. **ECSSD** contains a large number of semantically meaningful but structurally complex natural images. The reason to include **JuddDB** and **PASCAL-S** datasets was to assess performance of models over scenes with multiple objects with high background clutter. Finally, we also evaluate models over **SED2** to check whether salient object detection algorithms can perform well on images containing more than one salient object (*i.e.*, two in **SED2**). Fig. 2 shows the AAM model output of six benchmark datasets to illustrate their different center

[1]Object proposal generation is a recently emerging trend which attempts to detect image regions that may contain objects from any object category (a.k.a, category independent object proposals).

[2]To save space, we show some plots over the **ECSSD** dataset on our online benchmark website.

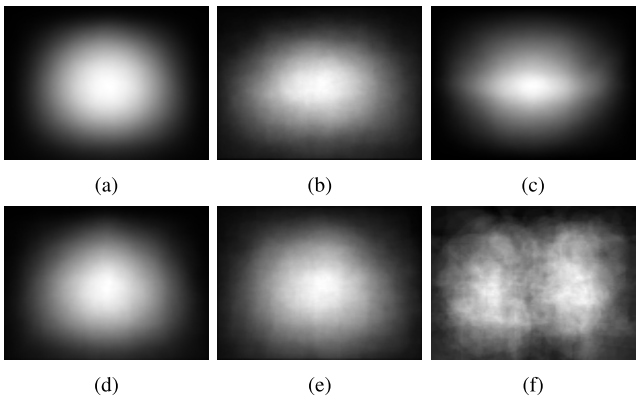


Fig. 2. Average annotation maps of six datasets used in benchmarking. (a) **MSRA10K**. (b) **PASCAL-S**. (c) **THUR15K**. (d) **DUT-OMRON**. (e) **JuddDB**. (f) **SED2**.

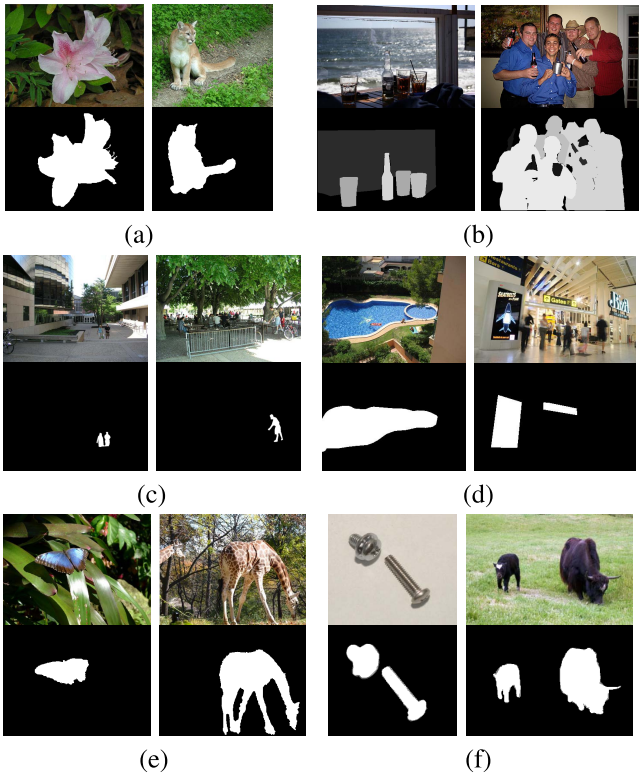


Fig. 3. Images and pixel-level annotations from six salient object datasets. (a) **MSRA10K**. (b) **PASCAL-S**. (c) **JuddDB**. (d) **DUT-OMRON**. (e) **THUR15K**. (f) **SED2**.

biases. See Fig. 3 for representative images and annotations from each dataset.

We illustrate in Fig. 4 the statistics of the seven chosen datasets. In Fig. 4(a), we show the normalized distances from the centroid of salient objects to the corresponding image centers. We can see that salient objects in **ECCSD** have the shortest distance to image centers, while salient objects in **SED2** have the longest distances. This is reasonable since images in **SED2** usually have two objects aligned around opposite image borders. Moreover, we can see that the spatial distribution of salient objects in **JuddDB** has a larger variety than other datasets, indicating that this dataset has smaller positional bias (*i.e.*, center-bias of salient objects and border-bias of background regions).

In Fig. 4(b), we aim to show the complexity of images in seven benchmark datasets. Toward this end, we apply the segmentation algorithm by Felzenszwalb and Huttenlocher [106] to see how many super-pixels (*i.e.*, homogeneous regions) can be obtained on average from salient objects and background regions of each image, respectively. In this manner, we can use this measure to reflect how challenging a benchmark dataset is since massive super-pixels often indicate complex foreground objects and cluttered background. From Fig. 4(b), we can see that **JuddDB** (followed by **PASCAL-S**) is the most challenging benchmark since it has an average number of 493 super-pixels from the background of each image. On the contrary, **SED2** contains fewer number of super-pixels in foreground and background regions, indicating that images in this benchmark often contain uniform regions and are relatively easier to process.

In Fig. 4(c), we demonstrate the average object sizes of these benchmarks, while the size of each object is normalized by the size of the corresponding image. We can see that **MSRA10K** and **ECCSD** datasets have larger objects while **SED2** has smaller ones. In particular, we can see that some benchmarks contain a limited number of image regions with large foreground objects. By jointly considering the center-bias property, it becomes very easy to achieve a high precision on these images.

C. Evaluation Measures

There are several ways to measure the agreement between model predictions and human annotations [21]. Some metrics evaluate the overlap between a tagged region and and model predictions while others try to assess the accuracy of drawn shapes with object boundary. In addition, some metrics have tried to consider both boundary and shape [107].

Here, we use four universally-agreed, standard, and easy-to-understand measures for evaluating a 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. The third measure directly computes 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 to represent 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. Moreover, we also use the fourth measure proposed by Margolin *et al.* [108] which remedies some problems with the classic F-measure for evaluating foreground-background maps obtained using segmentation algorithms.

1) Precision-Recall (PR): For a saliency map S, we can convert it to a binary mask M and compute $Precision$ and

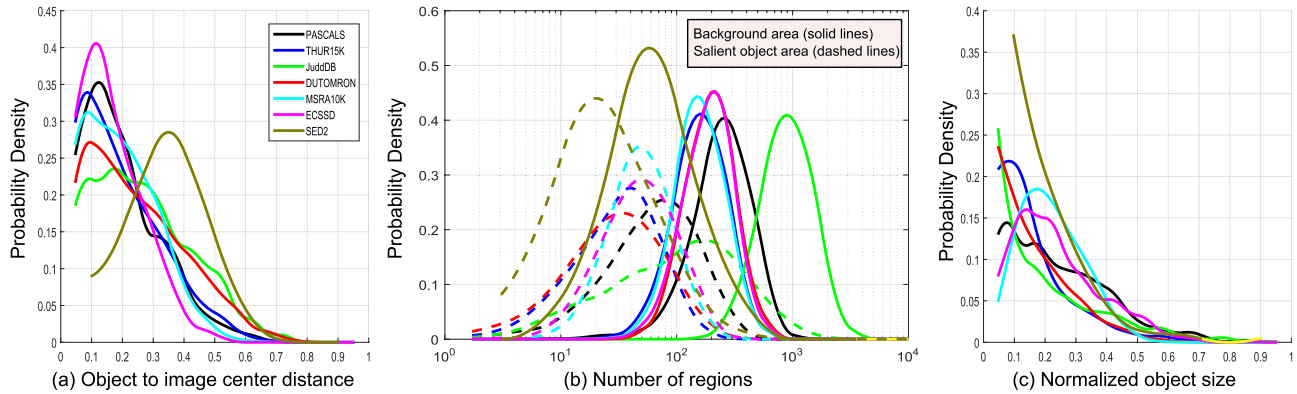


Fig. 4. Statistics of the benchmark datasets. a) distribution of normalized object distance from image center, b) distribution of number of super-pixels on salient objects and image background, and c) distribution of normalized object size. See text for precise definitions.

Recall by comparing M with ground-truth G:

$$Precision = \frac{|M \cap G|}{|M|}, \quad Recall = \frac{|M \cap G|}{|G|} \tag{1}$$

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

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

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

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

The third way of binarization is to use the SaliencyCut algorithm [70]. In this solution, a loose threshold, which typically results in good recall but relatively poor precision, is used to generate the initial binary mask. Then the method iteratively uses the GrabCut segmentation method [109] to gradually refine the binary mask. The final binary mask is used to re-compute the precision-recall value.

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

$$F_\beta = \frac{(1 + \beta^2) Precision \times Recall}{\beta^2 Precision + Recall} \tag{3}$$

As suggested by many salient object detection works (*e.g.*, [18], [70], [75]), β^2 is set to 0.3 to increase the importance of the *Precision* value. The reason for weighting precision more than recall is that recall rate is not as important as precision (see also [110]). For instance, 100% recall can be easily achieved by setting the whole region to foreground.

According to the different ways for saliency map binarization, there exist two ways to compute F-Measure. When the adaptive threshold or GrabCut algorithm is used for the binarization, we can generate a single F_β for each image and the final F-Measure is computed as the average F_β. When using fixed thresholding, the resulted PR curve can be scored by its maximal F_β, which is a good summary of the detection performance (as suggested in [111]). As defined in (3), F-Measure is the weighted harmonic mean of precision and recall, thus share the same value bounds as precision and recall values, *i.e.* [0, 1].

3) 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:

$$TPR = \frac{|M \cap G|}{|G|}, \quad FPR = \frac{|M \cap \bar{G}|}{|\bar{G}|} \tag{4}$$

where $\bar{M}$ and $\bar{G}$ denote the complement of the binary mask M and ground-truth, respectively. The ROC curve is the plot of TPR versus FPR by varying the threshold T_f.

4) Area Under ROC Curve (AUC) Score: 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 around 0.5.

5) Mean Absolute Error (MAE) Score: 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 saliency to salient pixels but fail to detect non-salient regions over methods that successfully detect non-salient pixels but make mistakes in determining the salient ones [75], [82]. Moreover, in some application scenarios [112] the quality of the weighted, continuous saliency maps may be of higher importance than the binary masks. For a more comprehensive comparison, we therefore also evaluate the mean absolute error (MAE) between the continuous saliency map $\bar{S}$ and the binary ground truth $\bar{G}$, both normalized in the range [0, 1]. The MAE score is defined as:

$$MAE = \frac{1}{W \times H} \sum_{x=1}^{W} \sum_{y=1}^{H} |\bar{S}(x, y) - \bar{G}(x, y)| \tag{5}$$

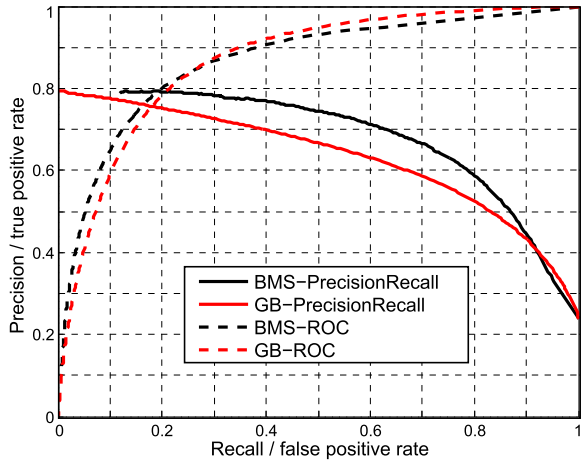


Fig. 5. PR and ROC curves for BMS [100] and GB [93] over **ECSSD**.

6) $\mathbf{F}_\beta^\mathbf{w}$*-Measure:* Here, we adopt the technique proposed by Margolin *et al.* [108] for quantitative evaluation of models. As an intuitive generalization of the F_β-measure, the new evaluation metric (F_β^w-measure) provides reliable evaluation by i) extending the basic quantities (true-positive, true-negative, false-positive, and false negative) to non-binary values, and ii) weighting errors according to their location and their neighborhood. F_β^w-measure offers unified solution for evaluation of binary and non-binary maps.

Note that these scores sometimes do not agree with each other. For example, Fig. 5 shows a comparison of two models over the **ECSSD** dataset using PR and ROC metrics. While there is not a big difference in ROC curves (thus about the same AUC), one model clearly scores better using the PR curve (thus having higher F_β). Such disparity between the ROC and PR measures has been extensively studied in [113]. Note that the number of negative examples (non-salient pixels) is typically much bigger than the number of positive examples (salient object pixels) in evaluating salient object detection models. Therefore, PR curves are more informative than ROC curves and can present an over optimistic view of an algorithm's performance [113]. Thus we mainly base our conclusions on the PR curves scores (*i.e.*, F-Measure scores), and also report other scores for comprehensive comparisons and for facilitating specific application requirements. It is worth mentioning that active research is ongoing to figure out the better ways of evaluating salient object detection and segmentation models (*e.g.* [108]).

D. Quantitative Comparison of Models

We evaluate saliency maps produced by different models on seven datasets by using all evaluation metrics:

1) Fig. 6 and Fig. 7 show PR and ROC curves;
2) Fig. 8 and Fig. 9 demonstrate AUC and MAE scores;
3) Fig. 10 and Fig. 11 show F_β^w and F_β scores of all models, respectively.[3]

[3]Three segmentation methods are used, including adaptive threshold, fixed threshold, and SaliencyCut algorithm. The influence of segmentation methods will be discussed in Sect. III-A.

In terms of both **PR** and **ROC** curves, DRFI model surprisingly outperforms all other models on seven benchmark datasets with large margins. Besides, RBD, DSR and MC (solid lines with blue, yellow, and magenta colors, respectively) achieve close performance and perform slightly better than other models.

Using the **F-measure** (*i.e.*, F_β), the five best models are: DRFI, MC, RBD, DSR, and GMR, where DRFI model consistently wins over all the 5 datasets. MC ranks the second best over 2 datasets and the third best over 2 datasets. SR and SIM models perform the worst.

With respect to the **AUC** score, DRFI again ranks the best over all seven datasets. Following DRFI, DSR model ranks the second over 4 datasets. RBD ranks the second on 1 dataset and the third on 2 datasets. While PCA ranks the third on 1 dataset in terms of AUC score, it is not on the list of top three contenders using F_β measure. IT, SR, and SUN achieve the worst performance. It is worth being mentioned that all the models perform well above chance level (AUC $= 0.5$) on seven benchmark datasets.

Rankings of models using **MAE** are more diverse than either F_β or AUC scores. DSR, RBD and DRFI rank on the top, but none of them are among top three models over **JuddDB**. MC, which performs well in terms of F_β and AUC, is not included in the top three models on any dataset. PCA performs the best on **JuddDB** but worse on others. SIM and SVO models perform the worst.

Using the $\mathbf{F}_\beta^\mathbf{w}$-measure, RBD, DRFI, and ST rank at the top. Other top contenders here are: DSR, QCUT, RC and HS. RBD model ranks better using this score than the other ones.

On average, the compared fixation prediction and object proposal generation models perform worse than salient object detection models. As two outliers, COV and BMS outperform several salient object detection models in terms of all evaluation metrics, implying that they are suitable for detecting salient proto objects. Additionally, Fig. 12 shows the distribution of F_β, ROC and MAE scores of all salient object detection models versus all fixation prediction models over all benchmark datasets. We can see a sharp separation of models especially for the F_β score, where most of the top models are salient object detection models. This result is consistent with the conclusion in [1] that fixation prediction models perform lower than salient object detection models. Though stemming from fixation prediction, research in salient object detection shares its unique properties and has truly added to what traditional saliency models focusing on fixation prediction already offer.

In particular, most of the salient object detection models outperform the baseline AAM model. Among these 29 models, AAM only outperforms 1 model over **MSRA10K**, 8 models over **ECSSD**, 3 on **THUR15K**, 11 on **JuddDB**, 9 on **PASCAL-S** and 3 on **DUT-OMRON** in terms of F_β (Fixed). Interestingly, AAM model does not outperform any model over **SED2**, which means that indeed there is less center bias in this dataset and salient object detection models can detect off-center objects. Notice that AAM ranks lowest on **SED2** compared to other datasets. Please notice that it does not

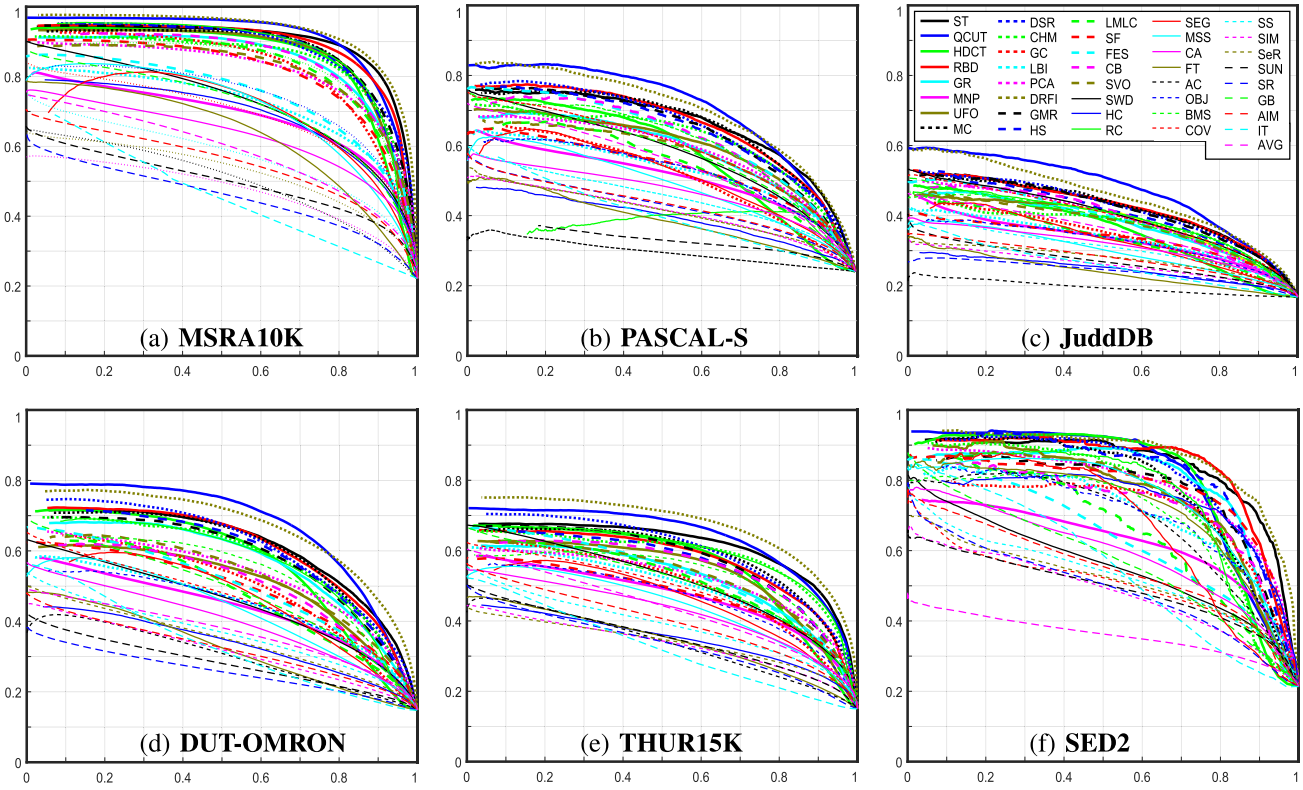


Fig. 6. Precision (vertical axis) and recall (horizontal axis) curves of saliency methods on 6 popular benchmark datasets.

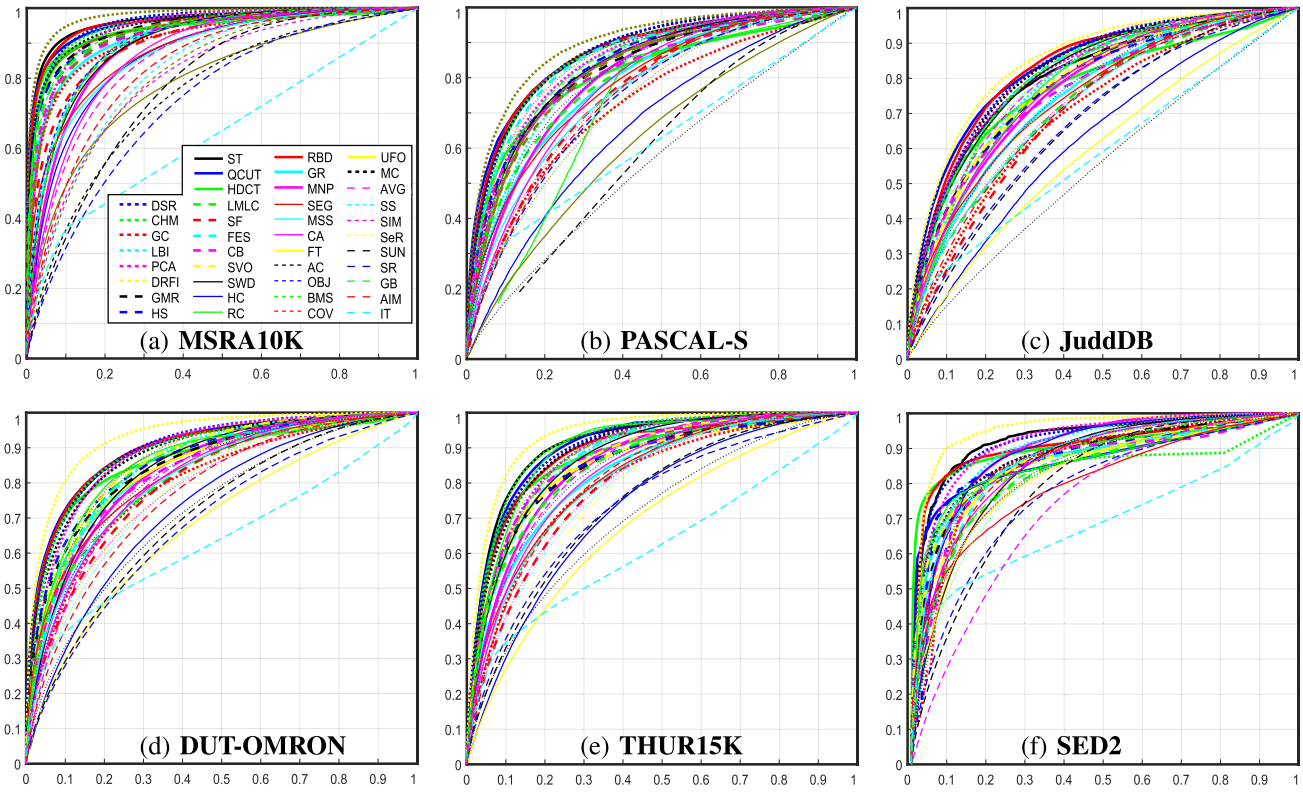


Fig. 7. ROC curves of models on 6 benchmarks. False and true positive rates are shown in x and y axes, respectively.

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

On average, over all models and scores, the performances were lower on **JuddDB**, **PASCAL-S** and **THUR15K**, implying that these datasets were more challenging. The low model

Model	PASCAL	THUR	JuddDB	DUT	MSRA	ECSSD	SED2
ST	.868	.911	.806	.895	.961	.914	.922
QCUT	.870	.907	.831	.897	.956	.909	.860
HDCT	.815	.878	.771	.869	.941	.866	.898
RBD	.867	.887	.826	.894	.955	.894	.899
GR	.794	.829	.747	.846	.925	.831	.854
MNP	.807	.854	.768	.835	.895	.820	.888
UFO	.825	.853	.775	.839	.938	.875	.845
MC	.873	.895	.823	.887	.951	.910	.877
DSR	.873	.902	.826	.899	.959	.914	.915
CHM	.864	.910	.797	.890	.952	.903	.831
GC	.728	.803	.702	.796	.912	.805	.846
LBI	.828	.876	.792	.854	.910	.842	.896
PCA	.848	.885	.804	.887	.941	.876	.911
DRFI	.897	.938	.851	.933	.978	.944	.944
GMR	.829	.856	.781	.853	.944	.889	.862
HS	.840	.853	.775	.860	.933	.883	.858
LMLC	.800	.853	.724	.817	.936	.849	.826
SF	.762	.799	.711	.803	.905	.817	.871
FES	.854	.867	.805	.848	.898	.860	.838
CB	.818	.870	.760	.831	.927	.875	.839
SVO	.826	.865	.784	.866	.930	.857	.875
SWD	.835	.873	.812	.843	.901	.857	.845
HC	.669	.735	.626	.733	.867	.704	.880
RC	.713	.896	.775	.859	.936	.892	.852
SEG	.787	.818	.747	.825	.882	.808	.796
MSS	.766	.813	.726	.817	.875	.779	.871
CA	.782	.830	.774	.815	.872	.784	.853
FT	.629	.684	.593	.682	.790	.661	.820
AC	.565	.707	.548	.721	.756	.668	.831
OBJ	.811	.839	.750	.822	.907	.818	.870
BMS	.834	.879	.788	.856	.929	.865	.852
COV	.856	.883	.826	.864	.904	.879	.833
SS	.752	.792	.754	.784	.823	.725	.826
SIM	.756	.797	.727	.783	.808	.734	.833
SeR	.736	.778	.746	.786	.813	.695	.835
SUN	.588	.746	.674	.708	.778	.623	.789
SR	.745	.741	.676	.688	.736	.633	.769
GB	.843	.882	.815	.857	.902	.865	.839
AIM	.753	.814	.719	.768	.833	.730	.846
IT	.621	.623	.586	.636	.640	.577	.682
AAM	.835	.849	.797	.814	.857	.863	.736

Fig. 8. AUC: area under ROC curve (Higher is better. The top three models are highlighted in red, green and blue).

Model	PASCAL	THUR	JuddDB	DUT	MSRA	ECSSD	SED2
ST	.224	.179	.240	.182	.122	.193	.145
QCUT	.195	.128	.178	.119	.118	.171	.148
HDCT	.229	.177	.209	.164	.143	.199	.162
RBD	.199	.150	.212	.144	.108	.173	.130
GR	.299	.256	.311	.259	.198	.285	.189
MNP	.298	.255	.286	.272	.229	.307	.215
UFO	.245	.165	.216	.173	.150	.207	.180
MC	.230	.184	.231	.186	.145	.204	.182
DSR	.204	.142	.196	.139	.121	.173	.140
CHM	.222	.153	.226	.152	.142	.195	.168
GC	.265	.192	.258	.197	.139	.214	.185
LBI	.281	.239	.273	.249	.224	.280	.207
PCA	.245	.198	.181	.206	.185	.248	.200
DRFI	.221	.150	.213	.155	.118	.166	.130
GMR	.233	.181	.243	.189	.126	.189	.163
HS	.262	.218	.282	.227	.149	.228	.157
LMLC	.284	.246	.303	.277	.163	.260	.269
SF	.253	.184	.218	.183	.175	.230	.180
FES	.218	.155	.184	.156	.185	.215	.196
CB	.272	.227	.287	.257	.178	.241	.195
SVO	.392	.382	.422	.409	.331	.404	.348
SWD	.315	.288	.292	.310	.267	.318	.296
HC	.354	.291	.348	.310	.215	.331	.193
RC	.304	.168	.270	.189	.137	.187	.148
SEG	.350	.336	.354	.337	.298	.342	.312
MSS	.251	.178	.204	.177	.203	.245	.192
CA	.299	.248	.282	.254	.237	.310	.229
FT	.307	.241	.267	.250	.235	.291	.206
AC	.286	.195	.239	.190	.227	.265	.206
OBJ	.334	.306	.359	.323	.262	.337	.269
BMS	.225	.181	.233	.175	.151	.216	.184
COV	.226	.155	.182	.156	.197	.217	.210
SS	.315	.267	.301	.277	.266	.344	.266
SIM	.414	.414	.412	.429	.388	.433	.384
SeR	.371	.345	.379	.352	.310	.404	.290
SUN	.467	.310	.319	.349	.306	.396	.307
SR	.291	.175	.200	.181	.232	.266	.220
GB	.270	.229	.261	.240	.222	.263	.242
AIM	.337	.298	.331	.322	.286	.339	.262
IT	.240	.199	.200	.198	.213	.273	.245
AAM	.316	.248	.343	.288	.260	.276	.405

Fig. 9. MAE: Mean Absolute Error (Smaller is better. The top three models are highlighted in red, green and blue).

performance of **JuddDB** can be caused by both less center bias and small objects in images. By investigating some images of these datasets for which models performed low, we found that there are several objects that can be potentially the most salient one. This makes the generation of ground-truth quite subjective and challenging, although the most salient object in **JuddDB** and **PASCAL-S** datasets has objectively been defined to be the most looked-at object measured from eye movement data.

E. Qualitative Comparison of Models

Fig. 13 shows output maps of all models for a sample image with relatively complex background. Dark blue areas are less salient while dark red indicates higher saliency values. Compared with other models, top contenders like DRFI and DSR suppress most of the background well while almost successfully detect the whole salient object. They thus generate higher precision scores and less false positive rates. Some models that include a center-bias component also result in appealing maps, *e.g.*, CB. Interestingly, region-based approaches, *e.g.*, RC, HS, DRFI, BMR, CB, and DSR always preserve the object boundary well compared with other pixel-based or patch-based models.

We can also clearly see the distinctness of different categories of models. Salient object detection models try to highlight the whole salient object and suppress the background. Fixation prediction models often produce blob-

Model	PASCAL	THUR	JuddDB	DUT	MSRA	ECSSD	SED2
ST	.476	.415	.274	.393	.660	.491	.585
QCUT	.383	.421	.293	.448	.627	.467	.533
HDCT	.384	.360	.233	.365	.582	.430	.520
RBD	.472	.421	.308	.429	.685	.490	.642
GR	.371	.276	.211	.286	.485	.355	.487
MNP	.370	.274	.199	.270	.416	.335	.417
UFO	.381	.353	.231	.334	.556	.405	.477
MC	.422	.349	.256	.355	.576	.441	.465
DSR	.439	.422	.273	.420	.656	.490	.583
CHM	.391	.378	.235	.366	.573	.424	.463
GC	.396	.367	.244	.358	.612	.437	.555
LBI	.373	.286	.210	.276	.443	.346	.435
PCA	.353	.298	.205	.303	.473	.358	.437
DRFI	.421	.444	.264	.417	.654	.516	.638
GMR	.438	.380	.274	.384	.643	.476	.584
HS	.451	.365	.251	.356	.604	.449	.572
LMLC	.439	.343	.236	.313	.594	.428	.399
SF	.280	.259	.168	.279	.440	.309	.448
FES	.306	.286	.202	.276	.388	.312	.267
CB	.391	.318	.232	.284	.528	.403	.463
SVO	.296	.229	.206	.228	.363	.329	.323
SWD	.355	.244	.198	.239	.365	.332	.317
HC	.343	.255	.186	.243	.481	.323	.507
RC	.326	.414	.254	.377	.608	.493	.567
SEG	.349	.223	.194	.234	.332	.323	.294
MSS	.241	.229	.144	.225	.345	.251	.382
CA	.333	.244	.192	.243	.379	.304	.348
FT	.241	.195	.135	.191	.334	.239	.393
AC	.167	.178	.106	.166	.172	.191	.323
OBJ	.373	.252	.207	.250	.403	.339	.383
BMS	.404	.348	.233	.358	.566	.420	.447
COV	.250	.252	.186	.254	.303	.277	.256
SS	.294	.204	.177	.210	.312	.268	.309
SIM	.338	.195	.176	.201	.293	.291	.294
SeR	.354	.225	.198	.229	.352	.295	.359
SUN	.310	.194	.156	.188	.301	.254	.308
SR	.277	.134	.091	.118	.155	.138	.184
GB	.354	.269	.210	.266	.392	.344	.335
AIM	.307	.229	.178	.216	.341	.285	.355
IT	.181	.119	.078	.122	.228	.141	.165
AAM	.374	.258	.219	.255	.381	.365	.267

Fig. 10. Evaluation results using F_β^w-measure [108].

like and sparse saliency maps corresponding to the fixation areas of humans on scenes. The objectness map is a rough indication of the salient object. The output of the latter two types of models might not be suitable for segmenting the whole salient object well.

III. Performance Analysis

Based on the performances reported above, we also conduct several experiments to provide a detailed analysis of all the benchmarking models and datasets.

A. Analysis of Segmentation Methods

In many computer vision and graphics applications, segmenting regions of interest is of great practical importance [37], [46], [49]–[51], [114], [115]. The simplest way of segmenting a salient object is to binarize the saliency map using a fixed threshold, which might be hard to choose. In this section, we extensively evaluate two additional most commonly used salient object segmentation methods, including adaptive threshold [18] and SaliencyCut [70]. Average F_β scores for salient object segmentation results on seven benchmark datasets are shown in Fig. 11. Each segmentation algorithm was fed with saliency maps produced by all 41 compared models.

Except **JuddDB**, **PASCAL-S** and **SED2** datasets, best segmentation results are all achieved via SaliencyCut method combined with a sophisticated salient object detection model (*e.g.*, DRFI, RBD, MNP). This suggests that enforcing label consistency in terms of using graph-based segmentation and global appearance statistics benefits salient object segmentations. The default SaliencyCut [70] program only outputs the most dominate salient object, This causes results for **SED2**, **PASCAL-S** and **JuddDB** benchmarks to be less optimal, as images in these two datasets (see Fig. 3) do not follow the "single none ambiguous salient object assumption" made in [70].

As also observed by most works in image segmentation literature, nearby pixels with similar appearance tend to have similar object labels. To validate this, we demonstrated in Fig. 14(a) some better segmentation results by further enforcing label consistency among nearby and similar pixels. Enforcing such label consistency often helps improve labeling pixels specially when the majority of the salient object pixels have been highlighted in the detection phase. Challenging examples might still exist, however, such as complex object topology, spindle components, and similar appearance with respect to image background. More results of using the best combination, DRFI saliency maps and SaliencyCut segmentation, are demonstrated for images with various complexities, as shown in Fig. 14(b).

A failure case of SaliencyCut segmentation along with intermediate results is also shown in the last row of Fig. 14(a). Due to the complex topology of the salient object, label consistency in a local range considered in the SaliencyCut algorithm may not work well. Additionally, the appearance of the object looks very distinct due to the existence of shading and reflection, which makes the segmentation of the whole object very challenging. Therefore, only a part of the object is finally segmented.

B. Analysis of Center Bias

In this section, we study the center-bias challenge since it has caused a major problem in evaluating fixation prediction and salient object detection models. Some studies usually add a Gaussian center prior to models when comparing them. This might not be fair as several salient object detection models already contain center-bias at different levels. Alternatively, we randomly choose 1000 images with no/less center bias from the **MSRA10K** dataset. First, the distance of salient object centroid to the image center is computed for each image. Those images for which such distance is bigger than a threshold are then chosen. Some sample images with no/less

Model	PASCAL-S			THUR15K			JuddDB			DUT-OMRON			MSRA10K			ECSSD			SED2		
	Fixed	AdpT	SCut	Fixed	AdpT	SCut	Fixed	AdpT	SCut	Fixed	AdpT	SCut	Fixed	AdpT	SCut	Fixed	AdpT	SCut	Fixed	AdpT	SCut
ST	.660	.601	.671	.631	.580	.648	.455	.394	.459	.631	.577	.635	.868	.825	.896	.752	.690	.777	.818	.805	.768
QCUT	.695	.654	.613	.651	.625	.620	.509	.454	.480	.683	.647	.647	.874	.843	.843	.779	.738	.747	.810	.801	.672
HDCT	.604	.572	.611	.602	.571	.636	.412	.378	.422	.609	.572	.643	.837	.807	.877	.705	.669	.740	.822	.802	.758
RBD	.652	.607	.667	.596	.566	.618	.457	.403	.461	.630	.580	.647	.856	.821	.884	.718	.680	.757	.837	.825	.750
GR	.596	.508	.604	.551	.509	.546	.418	.338	.378	.599	.540	.580	.816	.770	.830	.664	.583	.677	.798	.753	.639
MNP	.522	.510	.630	.495	.523	.603	.367	.337	.405	.467	.486	.576	.668	.724	.822	.568	.555	.709	.621	.778	.765
UFO	.606	.554	.622	.579	.557	.610	.432	.385	.433	.545	.541	.593	.842	.806	.862	.701	.654	.739	.742	.781	.729
MC	.661	.622	.670	.610	.603	.600	.460	.420	.434	.627	.603	.615	.847	.824	.855	.742	.704	.745	.779	.803	.630
DSR	.646	.619	.650	.611	.604	.597	.454	.421	.410	.626	.614	.593	.835	.824	.833	.737	.717	.703	.794	.821	.632
CHM	.631	.586	.634	.612	.591	.643	.417	.368	.424	.604	.586	.637	.825	.804	.857	.722	.684	.735	.750	.750	.658
GC	.535	.472	.553	.533	.517	.497	.384	.321	.342	.535	.528	.506	.794	.777	.780	.641	.612	.593	.729	.730	.616
LBI	.538	.525	.629	.519	.534	.618	.371	.353	.416	.482	.504	.609	.696	.714	.857	.586	.563	.738	.692	.776	.764
PCA	.593	.567	.634	.544	.558	.601	.432	.404	.368	.554	.554	.624	.782	.782	.845	.646	.627	.720	.754	.796	.701
DRFI	.679	.615	.690	.670	.607	.674	.475	.419	.447	.665	.605	.669	.881	.838	.905	.787	.733	.801	.831	.839	.702
GMR	.643	.607	.654	.597	.594	.579	.454	.409	.432	.610	.591	.591	.847	.825	.839	.740	.712	.736	.773	.789	.643
HS	.637	.559	.647	.585	.549	.602	.442	.358	.428	.616	.565	.616	.845	.800	.870	.731	.659	.769	.811	.776	.713
LMLC	.555	.505	.614	.540	.519	.588	.375	.302	.397	.521	.493	.551	.801	.772	.860	.659	.600	.735	.653	.712	.674
SF	.544	.488	.461	.500	.495	.342	.373	.319	.219	.519	.512	.377	.779	.759	.573	.619	.576	.378	.764	.794	.509
FES	.619	.605	.534	.547	.575	.426	.424	.411	.333	.520	.555	.380	.717	.753	.534	.645	.655	.467	.617	.785	.174
CB	.623	.561	.636	.581	.556	.615	.444	.375	.435	.542	.534	.593	.815	.775	.857	.717	.656	.761	.730	.704	.657
SVO	.586	.361	.621	.554	.441	.609	.414	.279	.419	.557	.407	.609	.789	.585	.863	.639	.357	.737	.744	.667	.746
SWD	.577	.523	.642	.528	.560	.649	.434	.386	.454	.478	.506	.613	.689	.705	.871	.624	.549	.781	.548	.714	.737
HC	.423	.383	.464	.386	.401	.436	.286	.257	.280	.382	.380	.435	.677	.663	.740	.460	.441	.499	.736	.759	.646
RC	.466	.351	.470	.610	.586	.639	.431	.370	.425	.599	.578	.621	.844	.820	.875	.741	.701	.776	.774	.807	.649
SEG	.534	.344	.627	.500	.425	.580	.376	.268	.393	.516	.450	.562	.697	.585	.812	.568	.408	.715	.704	.640	.669
MSS	.503	.485	.399	.478	.490	.200	.341	.324	.089	.476	.490	.193	.696	.711	.362	.530	.536	.203	.743	.783	.298
CA	.489	.472	.586	.458	.494	.557	.353	.330	.394	.435	.458	.532	.621	.679	.748	.515	.494	.625	.591	.737	.565
FT	.408	.367	.357	.386	.400	.238	.278	.250	.132	.381	.388	.259	.635	.628	.472	.434	.431	.257	.715	.734	.436
AC	.326	.279	.265	.382	.431	.068	.227	.199	.049	.354	.383	.040	.520	.566	.014	.411	.410	.038	.684	.729	.140
OBJ	.544	.444	.596	.498	.482	.593	.368	.282	.413	.481	.445	.578	.718	.681	.840	.574	.456	.698	.685	.723	.731
BMS	.617	.596	.624	.568	.578	.594	.434	.404	.416	.573	.576	.580	.805	.798	.822	.683	.659	.690	.713	.760	.627
COV	.589	.604	.535	.510	.587	.398	.429	.427	.315	.486	.579	.373	.667	.755	.394	.641	.677	.413	.518	.724	.212
SS	.469	.451	.552	.415	.482	.523	.344	.321	.397	.396	.443	.502	.572	.642	.675	.467	.441	.574	.533	.696	.641
SIM	.434	.407	.599	.372	.429	.568	.295	.292	.384	.358	.402	.539	.498	.585	.794	.433	.391	.672	.498	.685	.725
SeR	.433	.406	.566	.374	.419	.536	.316	.285	.388	.385	.411	.532	.542	.607	.755	.419	.391	.596	.521	.714	.702
SUN	.359	.294	.467	.387	.432	.486	.303	.291	.285	.321	.360	.445	.505	.596	.670	.388	.376	.478	.504	.661	.613
SR	.447	.442	.497	.374	.457	.002	.279	.270	.001	.298	.363	.000	.473	.569	.001	.381	.385	.001	.504	.700	.002
GB	.581	.567	.651	.526	.571	.650	.419	.396	.455	.507	.548	.638	.688	.737	.837	.624	.613	.765	.571	.746	.695
AIM	.450	.375	.593	.427	.461	.559	.317	.260	.360	.361	.377	.495	.555	.575	.750	.449	.357	.571	.541	.718	.693
IT	.414	.453	.255	.373	.437	.005	.297	.283	.000	.378	.449	.005	.471	.586	.158	.407	.414	.003	.579	.697	.008
AAM	.549	.536	.578	.458	.569	.620	.392	.367	.411	.406	.514	.534	.580	.692	.779	.597	.627	.756	.388	.524	.640

Fig. 11. F_β statistics on each dataset, using varying fixed thresholds, adaptive threshold, and SaliencyCut (Higher is better. The top three models are highlighted in red, green and blue).

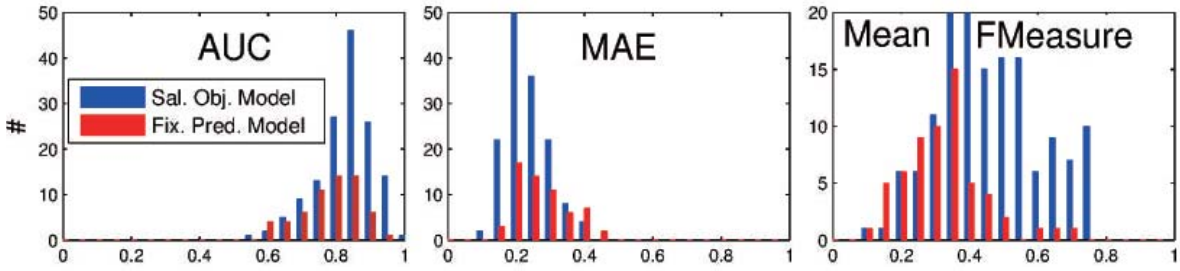


Fig. 12. Histogram of AUC, MAE, and Mean F_β scores for salient object detection models (blue) versus fixation prediction models (red) collapsed over all datasets.

center-bias, as well as an illustration of the threshold of choosing images, are shown in Fig. 15. The average annotation of less center-biased images shows two peaks on the left and on the right of the image, which is suitable for testing the performance of salient object detection models on off-center images.

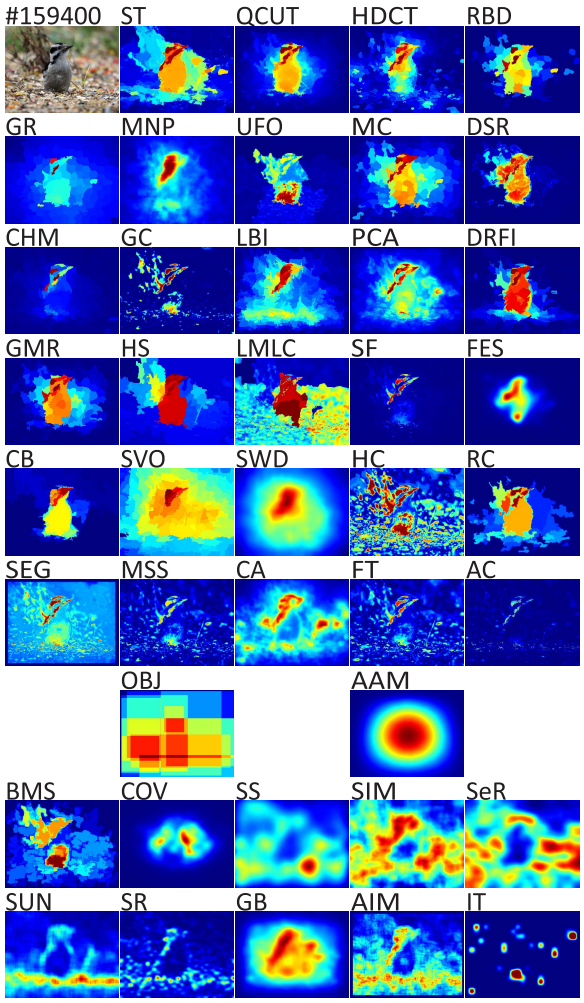


Fig. 13. Estimated saliency maps from various salient object detection models, object proposal generation model, average annotation map, and fixation prediction models.

We evaluate all the compared 41 models on these 1000 images. PR and ROC curves, F_β, AUC, and MAE scores are all shown in Fig. 16. DRFI and DSR again perform the best. Overall, most models' performance decrease when testing on no/less center biased images (*e.g.*, the AUC score of MC declines from 0.951 to 0.888), while a few others show increase. For example, the AUC score of SVO raises from 0.930 to 0.942 and it gets the second ranking. Some models, *e.g.*, HS (with the second ranking in terms of F_β score), performs better according to their rank changes w.r.t the whole **MSRA10K** dataset. DRFI still wins over other models here with a large margin. The difference in F_β, AUC, and MAE scores are not very large for this model over all data and 1000 less center-biased images (difference are 0.05, 0.05, and 0.009, respectively). This means that this model is not taking advantage of center-bias much. In the contrast, CB model uses a great deal of location prior and that is why it's performance drops heavily when applied to the off-center images (difference are 0.122, 0.122, and 0.029, respectively).

Additionally, it can be observed from Fig. 2(f), there is less center bias over the **SED2** dataset where there is less activation in the center of its average annotation map. We can therefore study the center bias on it. Similarly, DRFI and

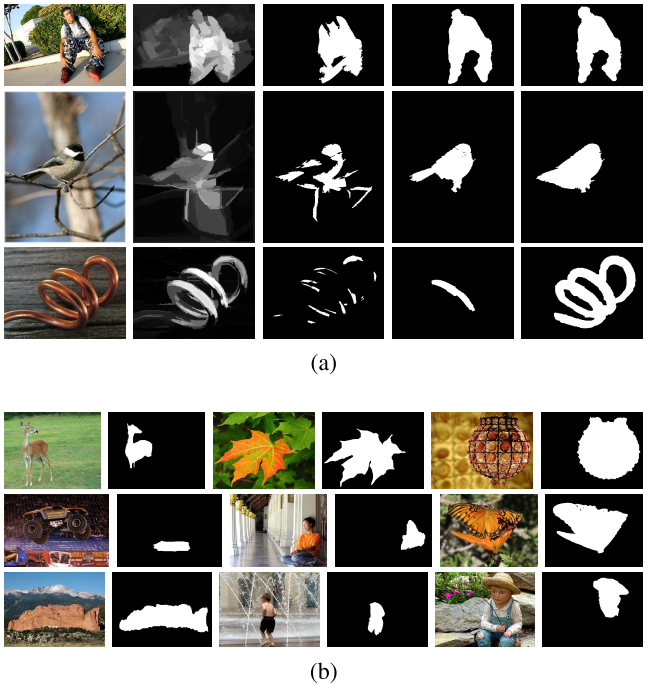


Fig. 14. Samples of salient object segmentation results. (a) Left to right: image, saliency map, AdpT: Adaptive Threshold, SCut: SaliencyCut and gTruth: Ground Truth. (b) DRFI model output fed to the SaliencyCut algorithm.

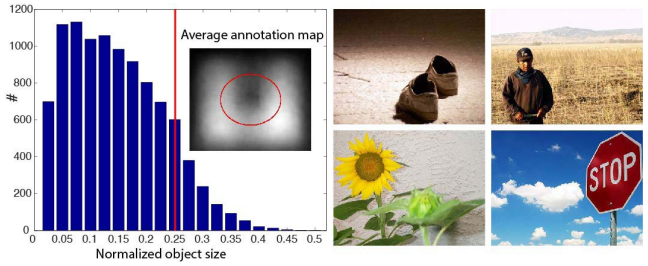


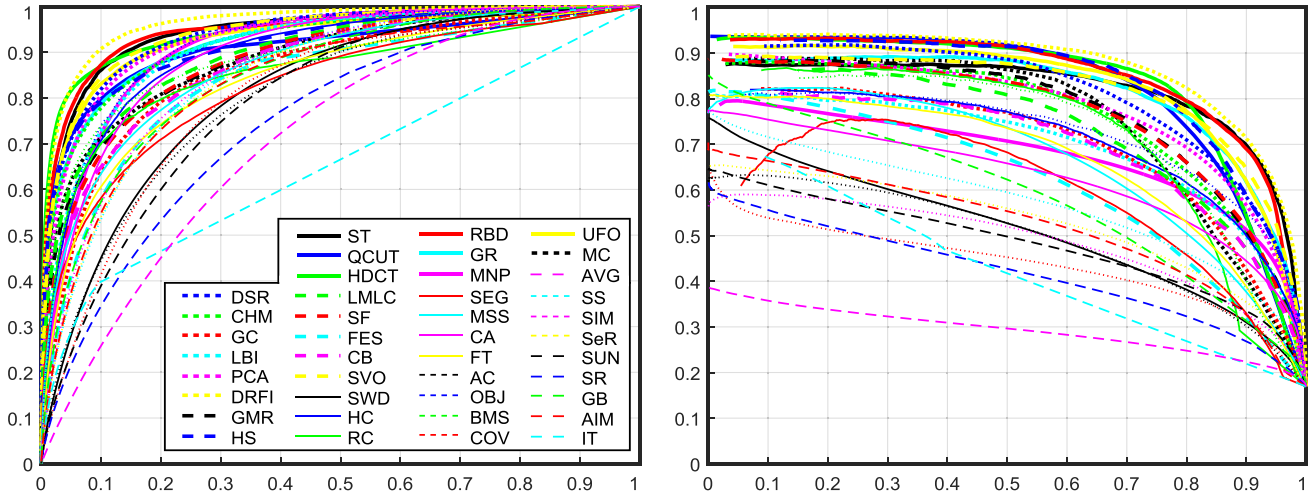
Fig. 15. Left: Histogram of object center over all images, threshold (red line = 0.247), and annotation map over 1000 less center-biased images from **MSRA10K** dataset. Right: Four less center-biased images. The overlaid circle illustrates the center-bias threshold.

DSR outperform other models in terms of F_β, AUC, and MAE scores, indicating they are more robust to the location variations of salient objects. HS again ranks second according to the F_β score. Fig. 17 shows best and worst off-centered stimuli for DRFI and DSR models.

Overall, all the models perform well above the chance level over either the less center-biased subset of **MSRA10K** or **SED2**. It is also worth noticing that the AAM model performs significantly worse on these two datasets, as well as **JuddDB**, validating our motivation of studying center bias on them.

C. Analysis of Salient Object Existence

The existence of a salient object in the image is somewhat neglected by the community. Almost all of existing salient object detection models assume that there is at least one salient object in the input image. This impractical assumption might lead to less optimal performance on "background images", which do not contain any dominant salient objects, as studied



Method	ST	QCUT	HDCT	RBD	GR	MNP	UFO	MC	DSR	CHM	GC	LBI	PCA	DRFI	GMR	HS	LMLC	SF	FES	CB
Max	.798	.808	.822	.811	.791	.661	.805	.764	.776	.746	.697	.685	.750	.831	.754	.815	.720	.747	.621	.693
AUC	.941	.911	.941	.943	.925	.912	.929	.888	.938	.920	.860	.910	.928	.964	.886	.918	.896	.885	.839	.872
MAE	.135	.115	.122	.106	.183	.188	.128	.171	.117	.138	.164	.197	.162	.127	.148	.150	.201	.150	.160	.207

Method	SVO	SWD	HC	RC	SEG	MSS	CA	FT	AC	OBJ	BMS	COV	SS	SIM	SeR	SUN	SR	GB	AIM	IT	AAM
Max	.792	.521	.700	.744	.629	.666	.620	.671	.521	.708	.739	.463	.571	.515	.546	.498	.444	.590	.540	.460	.328
AUC	.942	.813	.898	.855	.828	.868	.896	.843	.800	.915	.879	.805	.852	.858	.849	.795	.750	.850	.836	.655	.716
MAE	.325	.291	.176	.177	.300	.167	.199	.183	.177	.243	.146	.176	.225	.363	.273	.276	.184	.208	.265	.165	.406

Fig. 16. Results of center-bias analysis over 1000 less center-biased images chosen from the **MSRA10K** dataset. Top: ROC and PR curves, Bottom: Max F_β, AUC, and MAE scores for all models.

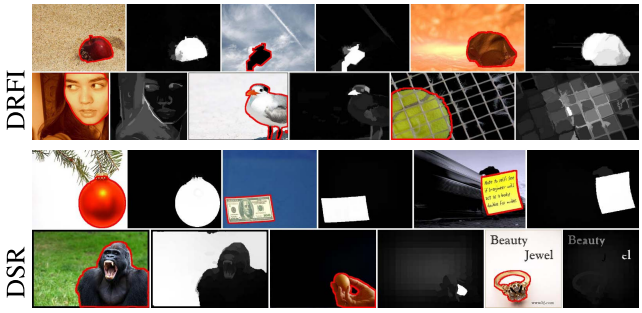


Fig. 17. Top and Bottom rows for each model illustrate best and worst cases in off-centered images.

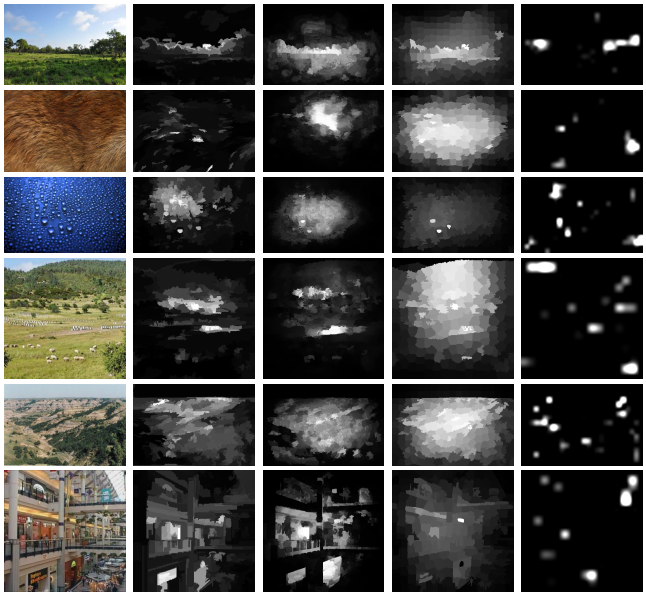


Fig. 18. Sample background-only images and prediction maps of DRFI, DSR, MC, and IT models.

in [116]. Just recently, Zhang *et al.* [117] introduced a fast method for a more challenging task of counting (subitizing) salient objects in a scene.

We can see from Fig. 18 that no dominated salient object exists in background images consisting of only textures or cluttered backgrounds. A good model should generate a dark (blank) saliency map on a background image, *i.e.*, without any activation as there are no salient objects. Fig. 18 shows saliency maps using three top salient object detection models and a classical fixation prediction model on background images. Top salient object detection models like DRFI, DSR, and MC do not perform well and often generate activations on the background images even though only regular textures exist (the second and third rows of Fig. 18). This is reasonable as they always assume there exist salient objects in the input image and will try their best to find one. These models can be distracted by the clutter in the background since high contrast always exist on the cluttered region. Most of existing salient object detection models compute saliency based on contrast values. These cluttered regions are thus more likely considered as salient. It is worth pointing out that ground truth of *eye fixations* do exist on such background images.

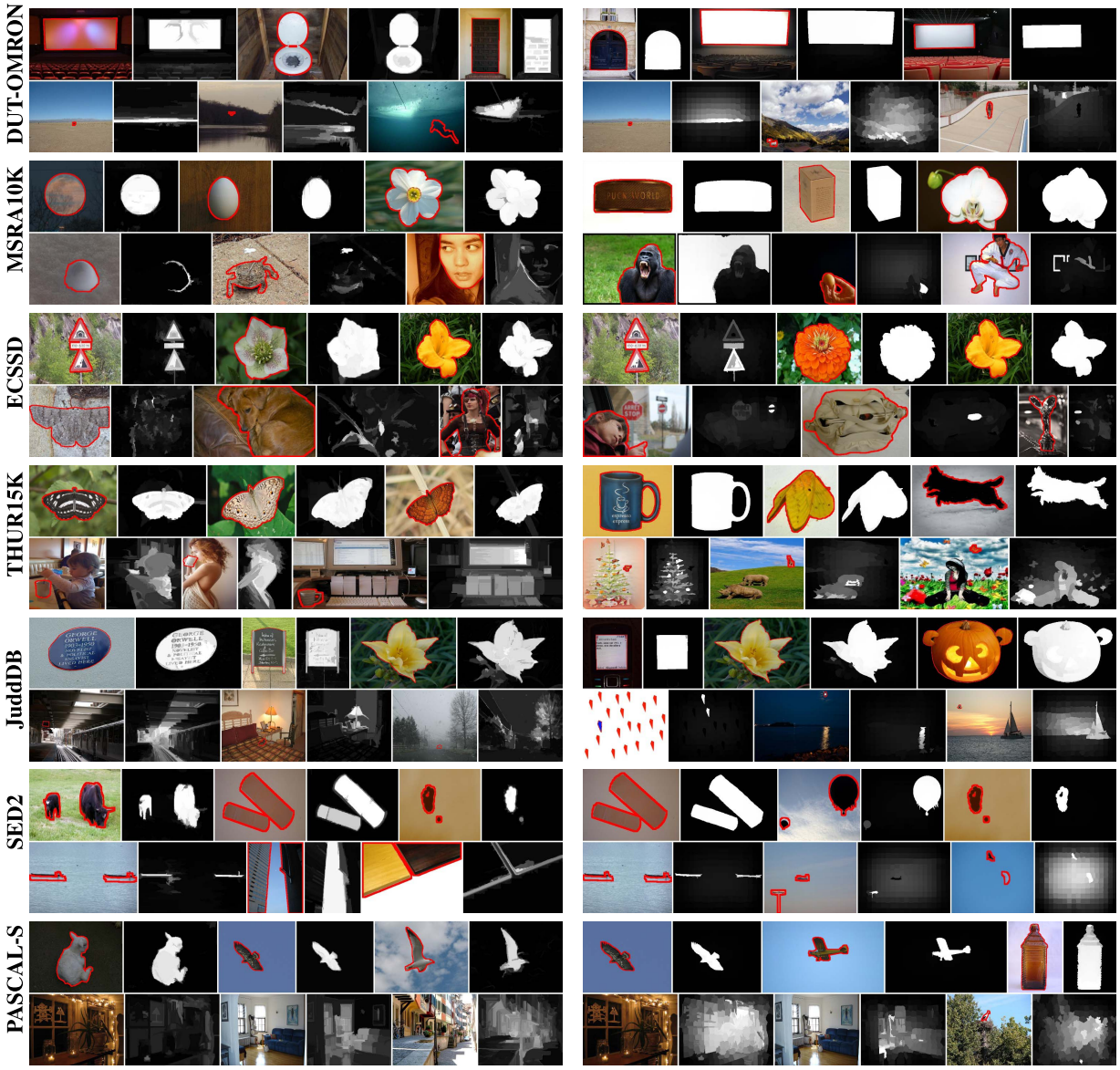


Fig. 19. Best (1st rows for each model on a dataset) and worst (2nd rows) cases of DRFI and MC. Ground-truth object(s) is denoted by a red contour. (a) DRFI (b) MC.

In addition to salient object existence, quantitative evaluations of models on background images is an open problem as well. Note that it is not feasible to calculate PR and ROC curves (and thus F_β and AUC scores) on background images since the ground truth positive labeling is empty. MAE score is not informative either as most salient object detection methods explicitly normalize the saliency maps in the range of [0,255] as a post-processing step. By demonstrating qualitative results of salient object detection models on some background images, we aim to motivate future works focusing on salient object detection on background images.

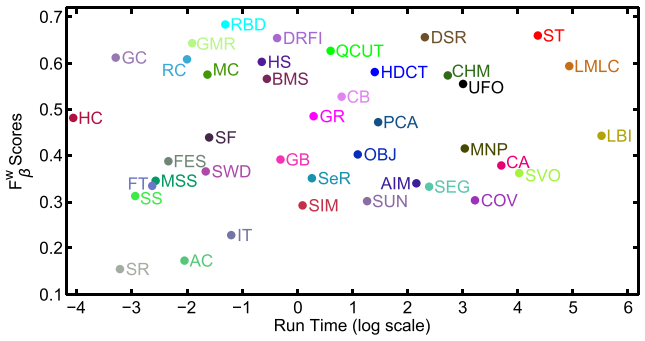


Fig. 20. F_β^w scores versus (log scale of) runtime of different methods based on the quantitative results of MSRA10K dataset.

D. Analysis of Worst and Best Cases for Top Models

To understand what are the challenges for existing salient object detection models, we illustrate the three best and the three worst cases for top models over all seven benchmark datasets. The stimuli for 11 top models were sorted according to the F_β scores. We only give a demonstration of DRFI and MC models in Fig. 19 due to limited space. See our online

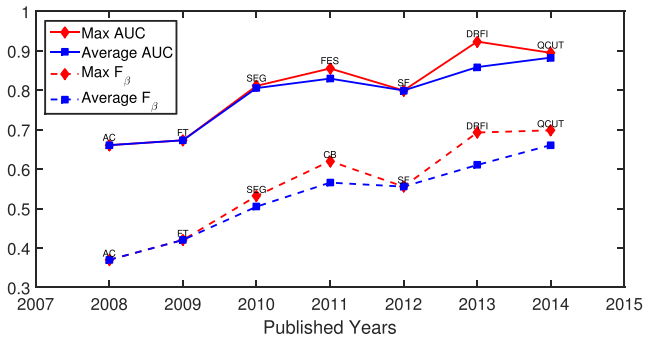


Fig. 22. Maximum and average AUC and F_β scores of different salient object methods versus their publication years. Model accuracy shows an increasing trend.

challenge website for additional illustrations.

It can be noticed from Fig. 19 that models share the same easy and difficult stimuli. Both DRFI and MC perform substantially well on the cases where a dominated salient object exists in a relatively clean background. Since most existing salient object detection models do not utilize any high-level prior knowledge, they may fail when a complex scene has a cluttered background or when the salient object is *semantically* salient (*e.g.*, DRFI fails on images with faces in **MSRA10K**). Another reason causing poor saliency detection is object size. Both DRFI and MC models have difficulty in detecting small objects (See hard cases on **DUT-OMRON** and **JuddDB**).

Particularly, since saliency cues adopted by DRFI are mainly based on contrast, this model fails on scenes where salient objects share close appearance with the background (*e.g.*, the hard cases of **MSRA10K** and **ECSSD**). Another possible reason is related to the failure in segmenting the image. MC relies on the pseudo-background prior that the image border areas are background. That is why it fails on scenes where the salient object touches the image border, *e.g.*, the gorilla image in **MSRA10K** dataset (4th row of the right column of Fig. 19).

E. Runtime Analysis

Runtime of compared models are shown in Fig. 1 over all 10K images of **MSRA10K** (typical image resolution of 400×300) using an Intel Xeon E5645 2.40GHz CPU with 8 GB RAM. A 2D scatter plot of F_β^w scores versus running time of different methods based on the quantitative results of MSRA10K dataset is shown in Fig. 20, which is helpful to demonstrate the trade-off between efficacy and efficiency of compared models.

Of all compared methods, the HC model is the fastest (about 0.017 seconds per image) followed by GC and SR models. The best model in our benchmark (DRFI) needs about 0.697 seconds to process one image. We can also observe that RC, GMR, MC, and RBD share similar trade-offs between F_β^w scores and runtime.

IV. Discussions and Conclusions

From the results obtained so far, we summarize in Fig. 21 the rankings of models based on average performance over datasets in terms of segmentation methods, center bias, salient object existence, and run time.[4] Based on the rankings, we conclude that:

"*DRFI, QCUT, RBD, ST, DSR, and MC are the top 6 models for salient object detection.*"

To gauge the progress in this field, we show in Fig. 22, the maximum and average AUC and F_β scores of different salient object detection methods versus their publication years. We find a continuous ascending success rate over the last couple of years which raises the hope that even better salient object detection models are possible in the future.

By investigating the performances and the design choices of all compared models, our extensive evaluations do suggest some clear messages about commonly used design choices, which could be valuable for developing future algorithms. We refer readers to our recent survey [28] for a comprehensive review of different design choices adopted for salient object detection.

- From the elements perspective, top five models (except QCUT) are built upon superpixels (regions). On the one hand, compared with pixels, more effective features (*e.g.*, color histogram) can be extracted from regions. On the other hand, compared with patches, the boundary of the salient object is better preserved for region-based approaches, leading to more accurate detection performance. Moreover, since the number of superpixels is far less than the number of pixels or patches, region-based methods has the potential to run faster.
- All the top six models explicitly consider the background prior, which assumes that the area in the narrow border of the image belongs to the background. Compared with the location prior of a salient object, such a background prior performs more robust.
- The leading method in our benchmark (*i.e.*, DRFI), discriminatively trains a regression model to predict region saliency according to a 93-dimensional feature vector. Instead of purely relying on the cues extracted only from the input image, DRFI resorts to human annotations to automatically discover feature integration rules. The high performance of this simple learning-based method encourages pursuing data-driven approaches for salient object detection.

However, even considering top performing models, salient object detection still seems far from being solved. To achieve more appealing results, three challenges should be addressed. First, in our large-scale benchmark (see Sec. II), all top performing algorithms use the location prior cues, limiting their adaptation to general cases. Second, although the ranking of top scoring models are quite consistent across datasets, performance scores (F_β and AUC) drop significantly from easier datasets to more difficult ones. The third challenge regards the run time of models. Some models need around one minute to process a 400×300 image (*e.g.*, CA: 40.9s,

[4]We have created a unified repository for sharing code and data where researchers can run models with a single click or can add new models for benchmarking purposes. All codes, data, and results are available in our online benchmark website: http://mmcheng.net/salobjbenchmark/

Method	ST	QCUT	HDCT	RBD	GR	MNP	UFO	MC	DSR	CHM	GC	LBI	PCA	DRFI	GMR	HS	LMLC	SF	FES	CB
F_β	3	2	9	5	14	29	13	4	6	10	20	26	16	1	7	8	19	21	18	12
F_β^w	2	6	11	1	17	20	15	10	4	12	8	18	19	3	5	7	13	25	26	16
AUC	3	4	9	5	23	22	20	6	2	8	31	15	7	1	11	12	26	28	19	21
MAE	7	1	6	3	27	29	9	12	2	5	17	25	18	4	10	19	26	15	8	23
AdpT	7	1	8	6	19	23	15	4	3	9	24	20	14	2	5	16	27	21	12	18
SCut	1	8	6	3	26	18	12	15	21	10	31	9	16	2	17	7	19	36	34	11
CB	7	5	2	4	9	25	6	11	10	15	20	22	13	1	12	3	18	14	27	21
Time	38	23	27	14	22	34	33	12	30	32	2	40	28	18	10	16	39	13	7	24
Overall Rank	4	2	9	3	18	23	12	6	5	8	19	20	15	1	7	10	17	29	22	14

Method	SVO	SWD	HC	RC	SEG	MSS	CA	FT	AC	OBJ	BMS	COV	SS	SIM	SeR	SUN	SR	GB	AIM	IT	AAM
F_β	17	22	32	11	25	28	31	34	37	27	15	24	33	39	36	40	41	23	35	38	30
F_β^w	30	27	24	9	31	34	28	37	39	21	14	33	36	35	29	38	40	22	32	41	23
AUC	13	18	33	17	30	29	27	40	38	24	16	10	35	34	36	39	37	14	32	41	25
MAE	40	34	31	14	37	16	30	28	22	36	13	11	32	41	39	38	20	24	35	21	33
AdpT	38	22	32	13	39	26	28	34	40	29	10	11	30	35	33	41	36	17	37	31	25
SCut	13	4	32	14	24	38	29	37	39	20	23	35	30	25	27	33	40	5	28	41	22
CB	8	35	19	16	26	24	28	23	34	31	17	38	30	36	32	37	40	29	33	39	41
Time	37	11	1	9	31	6	36	5	8	25	17	35	4	20	21	26	3	19	29	15	-
Overall Rank	27	21	35	13	30	33	28	38	41	26	11	25	31	36	34	37	39	16	32	40	24

Fig. 21. Summary rankings of models under different evaluation metrics over all datasets (excluding SED2). The overall rankings of different methods are computed based on the average (the higher the better) of AUC, (1-MAE), Max F_β, AdpT, ScutT, and F_β^w scores. The top three models under each evaluation metric are highlighted in red, green and blue.

SVO: 56.5s, and LMLC 140s).

One area for future research would be designing scores for tackling dataset biases and evaluation of saliency segmentation maps with respect to ground-truth annotations similar to [108]. In this benchmark, we only focused on single-input scenarios. Although some RGBD datasets exist [118], benchmark datasets for multiple input images (*e.g.*, salient object detection on videos, co-salient object detection [28]) are still lacking. Another future direction will be following active segmentation algorithms (*e.g.*, [103], [105], [119]) by segmenting a salient object from a seed point. For example, a simple model proposed by Borji [103] which segments the most salient object (at the peak of a map generated by a fixation prediction model as the seed point) using superpixels outperforms several salient object detection models on scenes with multiple salient objects (JuddDB). This indicates that several models are affected by a bias imposed by some former datasets (i.e., ASD) which is the existence of only one object in the image. Aggregation of saliency models for building a strong prediction model (similar to [1], [120], [121], and behavioral investigation of saliency judgments by humans (*e.g.*, [21], [122]) are two other interesting directions. The relationship (similarity and difference) between salient object detection and related fields such as object detection, object proposals, general segmentation, and fixation prediction[5] and the ways these areas can benefit from each other still remains to be explored further.

Inspired by the overwhelming performance of deep learning methods in other vision tasks like image classification [123], [124] and object detection [125], deep convolutional neural networks (CNNs) are also studied in recent works [126]–[129]. The leading performance of DRFI demonstrates the effectiveness of data-driven feature integration. Through deep architectures, more powerful representations can be learned than hand-crafted features for salient object detection tasks even if CNNs are trained for image classification. It indicates the promising direction of investigating deep learning methods for salient object detection in the future.

Saliency models (whether predicting where humans look in a scene or which objects they choose as salient [21], [130]) play an important role in the way we represent and understand scenes at the high level. Saliency models continue to be useful in a variety of domains encompassing human-robot interaction, image processing, and computer vision. So far modeling effort has been focused on improving the performance of existing datasets. State of the art models do very well even on large scale salient object datasets. We believe that it is now the time to consider how saliency detection can help other challenging tasks in computer vision for problems such as describing a scene (e.g., language and vision [131]–[136]), scene understanding (e.g., [134], [137]–[139]), and even object and scene classification (e.g., [123], [138], [140]).

Salient object detection is a very active research area in computer vision with several papers emerging each year in major conferences and journals. In fact, several models have been introduced since the initial submission of this work. Some, we have included in our benchmark[6] during the review process and some newer ones (such as [126]–[129] mainly based on the deep CNNs) will be considered in our online saliency detection benchmark. We will extensively review and discuss these models in our ongoing work [28].

ACKNOWLEDGMENT

Authors would like to thank anonymous reviewers for their helpful comments on the paper. Ali Borji was supported by Defense Advanced Research Projects Agency (NO. HR0011-10-C-0034), the National Science Foundation (CRCNS grant number BCS-0827764), the General Motors Corporation, and the Army Research Office (NO. W911NF-08-1-0360). Ming-Ming Cheng is supported by the grants from NSFC (NO.

[5]Please see our fixation prediction benchmark at http://saliency.mit.edu

[6]We encourage researchers to actively engage in this benchmark and help us gauge the future progress in this field and address potential challenges.

61572264). Jia Li is supported by the grants from NSFC (NO. 61370113), and Fundamental Research Funds for the Central Universities.

REFERENCES

[1] A. Borji, D. N. Sihite, and L. Itti, "Salient object detection: A benchmark," in *Proc. 12th ECCV*, 2012, pp. 414–429.

[2] A. Borji and L. Itti, "State-of-the-art in visual attention modeling," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 1, pp. 185–207, Jan. 2013.

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

[4] M. Hayhoe and D. Ballard, "Eye movements in natural behavior," *Trends Cognit. Sci.*, vol. 9, no. 4, pp. 188–194, 2005.

[5] L. Itti and C. Koch, "Computational modelling of visual attention," *Nature Rev. Neurosci.*, vol. 2, no. 3, pp. 194–203, 2001.

[6] A. M. Treisman and G. Gelade, "A feature-integration theory of attention," *Cognit. Psychol.*, vol. 12, no. 1, pp. 97–136, 1980.

[7] J. M. Wolfe, K. R. Cave, and S. L. Franzel, "Guided search: An alternative to the feature integration model for visual search," *J. Experim. Psychol., Human Perception Perform.*, vol. 15, no. 3, pp. 419–433, 1989.

[8] J. M. Wolfe, "Guidance of visual search by preattentive information," in *Neurobiology of Attention*. Amsterdam, The Netherlands: Elsevier, 2005, pp. 101–104.

[9] C. Koch and S. Ullman, "Shifts in selective visual attention: Towards the underlying neural circuitry," in *Matters of Intelligence*. New York, NY, USA: Springer-Verlag, 1987, pp. 115–141.

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

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

[12] J. Li, Y. Tian, T. Huang, and W. Gao, "Probabilistic multi-task learning for visual saliency estimation in video," *Int. J. Comput. Vis.*, vol. 90, no. 2, pp. 150–165, Nov. 2010.

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

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

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

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

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

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

[19] Y. Tian, J. Li, S. Yu, and T. Huang, "Learning complementary saliency priors for foreground object segmentation in complex scenes," *Int. J. Comput. Vis.*, vol. 111, no. 2, pp. 153–170, 2014.

[20] J. Wang, L. Quan, J. Sun, X. Tang, and H.-Y. Shum, "Picture collage," in *Proc. IEEE Comput. Soc. Conf. CVPR*, vol. 1. Jun. 2006, pp. 347–354.

[21] A. Borji, D. N. Sihite, and L. Itti, "What stands out in a scene? A study of human explicit saliency judgment," *Vis. Res.*, vol. 91, pp. 62–77, Oct. 2013.

[22] U. Rutishauser, D. Walther, C. Koch, and P. Perona, "Is bottom-up attention useful for object recognition?" in *Proc. IEEE Comput. Soc. Conf. CVPR*, Jun./Jul. 2004, pp. II-37–II-44.

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

[24] F. Moosmann, D. Larlus, and F. Jurie, "Learning saliency maps for object categorization," in *Proc. ECCV Workshop*, 2006, pp. 1–15.

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

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

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

[28] A. Borji, M.-M. Cheng, H. Jiang, and J. Li. (Nov. 2014). "Salient object detection: A survey." [Online]. Available: http://arxiv.org/abs/1411.5878

[29] Z. Ren, S. Gao, L.-T. Chia, and I. W.-H. Tsang, "Region-based saliency detection and its application in object recognition," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 24, no. 5, pp. 769–779, May 2013.

[30] M. Guo, Y. Zhao, C. Zhang, and Z. Chen, "Fast object detection based on selective visual attention," *Neurocomputing*, vol. 144, pp. 184–197, Nov. 2014.

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

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

[33] 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 Trans. Multimedia*, vol. 7, no. 5, pp. 907–919, Oct. 2005.

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

[35] 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 Process., Image Commun.*, vol. 28, no. 3, pp. 241–253, 2012.

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

[37] H. Huang, L. Zhang, and H.-C. Zhang, "Arcimboldo-like collage using Internet images," *ACM Trans. Graph.*, vol. 30, no. 6, 2011, Art. ID 155.

[38] 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 *Proc. IEEE ICIP*, Sep./Oct. 2007, pp. II-169–II-172.

[39] H. Liu and I. Heynderickx, "Studying the added value of visual attention in objective image quality metrics based on eye movement data," in *Proc. 16th IEEE ICIP*, Nov. 2009, pp. 3097–3100.

[40] A. Li, X. She, and Q. Sun, "Color image quality assessment combining saliency and FSIM," in *Proc. SPIE, 5th ICDIP*, vol. 8878. 2013, pp. 88780I-1–88780I-5.

[41] W. Zhang, A. Borji, Z. Wang, P. Le Callet, and H. Liu, "The application of visual saliency models in objective image quality assessment: A statistical evaluation," *IEEE Trans. Neural Netw. Learn. Syst.*, to be published. DOI: 10.1109/TNNLS.2015.2461603

[42] M. Donoser, M. Urschler, M. Hirzer, and H. Bischof, "Saliency driven total variation segmentation," in *Proc. 12th IEEE ICCV*, Sep./Oct. 2009, pp. 817–824.

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

[44] 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, pp. 378–391, Apr. 2013.

[45] M. Johnson-Roberson, J. Bohg, M. Björkman, and D. Kragic, "Attention-based active 3D point cloud segmentation," in *Proc. IEEE/RSJ IROS*, Oct. 2010, pp. 1165–1170.

[46] T. Chen, M.-M. Cheng, P. Tan, A. Shamir, and S.-M. Hu, "Sketch2Photo: Internet image montage," *ACM Trans. Graph.*, vol. 28, no. 5, 2009, Art. ID 124.

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

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

[49] L. Li, S. Jiang, Z.-J. Zha, Z. Wu, and Q. Huang, "Partial-duplicate image retrieval via saliency-guided visual matching," *IEEE Multimedia*, vol. 20, no. 3, pp. 13–23, Jul./Sep. 2013.

[50] A. Y.-S. Chia *et al.*, "Semantic colorization with Internet images," *ACM Trans. Graph.*, vol. 30, no. 6, 2011, Art. ID 156.

[51] H. Liu, L. Zhang, and H. Huang, "Web-image driven best views of 3D shapes," *Vis. Comput.*, vol. 28, no. 3, pp. 279–287, 2012.

[52] R. Margolin, L. Zelnik-Manor, and A. Tal, "Saliency for image manipulation," *Vis. Comput.*, vol. 29, no. 5, pp. 381–392, 2013.

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

[54] S. Stalder, H. Grabner, and L. J. Van Gool, "Dynamic objectness for adaptive tracking," in *Proc. ACCV*, 2012, pp. 43–56.

[55] J. Li, M. D. Levine, X. An, X. Xu, and H. He, "Visual saliency based on scale-space analysis in the frequency domain," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 4, pp. 996–1010, Apr. 2013.

[56] 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*. Berlin, Germany: Springer-Verlag, 2012, pp. 357–366.

[57] A. Borji, S. Frintrop, D. N. Sihite, and L. Itti, "Adaptive object tracking by learning background context," in *Proc. IEEE Comput. Soc. Conf. CVPR*, Jun. 2012, pp. 23–30.

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

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

[60] G. Zhang, Z. Yuan, N. Zheng, X. Sheng, and T. Liu, "Visual saliency based object tracking," in *Proc. 9th ACCV*, 2010, pp. 193–203.

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

[62] S. Frintrop, G. M. García, and A. B. Cremers, "A cognitive approach for object discovery," in *Proc. IEEE ICPR*, Aug. 2014, pp. 2329–2334.

[63] D. Meger *et al.*, "Curious george: An attentive semantic robot," *Robot. Auto. Syst.*, vol. 56, no. 6, pp. 503–511, 2008.

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

[65] A. Borji and L. Itti, "Defending Yarbus: Eye movements reveal observers' task," *J. Vis.*, vol. 14, no. 3, p. 29, 2014.

[66] R. Achanta, F. Estrada, P. Wils, and S. Süsstrunk, "Salient region detection and segmentation," in *Computer Vision Systems*. Berlin, Germany: Springer-Verlag, 2008.

[67] S. Goferman, L. Zelnik-Manor, and A. Tal, "Context-aware saliency detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 10, pp. 1915–1926, Oct. 2012.

[68] R. Achanta and S. Süsstrunk, "Saliency detection using maximum symmetric surround," in *Proc. 17th IEEE ICIP*, Sep. 2010, pp. 2653–2656.

[69] E. Rahtu, J. Kannala, M. Salo, and J. Heikkilä, "Segmenting salient objects from images and videos," in *Proc. 11th ECCV*, 2010, pp. 366–379.

[70] M.-M. Cheng, N. J. Mitra, X. Huang, P. H. S. Torr, and S.-M. Hu, "Global contrast based salient region detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 3, pp. 569–582, Mar. 2015.

[71] L. Duan, C. Wu, J. Miao, L. Qing, and Y. Fu, "Visual saliency detection by spatially weighted dissimilarity," in *Proc. IEEE Conf. CVPR*, Jun. 2011, pp. 473–480.

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

[73] H. Jiang, J. Wang, Z. Yuan, T. Liu, and N. Zheng, "Automatic salient object segmentation based on context and shape prior," in *Proc. BMVC*, 2011, pp. 110.1–110.12.

[74] H. R. Tavakoli, E. Rahtu, and J. Heikkilä, "Fast and efficient saliency detection using sparse sampling and kernel density estimation," in *Proc. 17th Scandin. Conf. Image Anal.*, 2011, pp. 666–675.

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

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

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

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

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

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

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

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

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

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

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

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

[87] C. Yang, L. Zhang, and H. Lu, "Graph-regularized saliency detection with convex-hull-based center prior," *IEEE Signal Process. Lett.*, vol. 20, no. 7, pp. 637–640, Jul. 2013.

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

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

[90] Z. Liu, W. Zou, and O. Le Meur, "Saliency tree: A novel saliency detection framework," *IEEE Trans. Image Process.*, vol. 23, no. 5, pp. 1937–1952, May 2013.

[91] C. Aytekin, S. Kiranyaz, and M. Gabbouj, "Automatic object segmentation by quantum cuts," in *Proc. IEEE 22nd ICPR*, Aug. 2014, pp. 112–117.

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

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

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

[95] L. Zhang, M. H. Tong, T. K. Marks, H. Shan, and G. W. Cottrell, "Sun: A Bayesian framework for saliency using natural statistics," *J. Vis.*, vol. 8, no. 7, p. 32, Dec. 2008.

[96] H. J. Seo and P. Milanfar, "Static and space-time visual saliency detection by self-resemblance," *J. Vis.*, vol. 9, no. 12, p. 15, 2009.

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

[98] X. Hou, J. Harel, and C. Koch, "Image signature: Highlighting sparse salient regions," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 1, pp. 194–201, Jan. 2012.

[99] E. Erdem and A. Erdem, "Visual saliency estimation by nonlinearly integrating features using region covariances," *J. Vis.*, vol. 13, no. 4, p. 11, Mar. 2013.

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

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

[102] 2015. *THUR15000 dataset*. [Online]. Available: http://mmcheng.net/gsal/

[103] A. Borji, "What is a salient object? A dataset and a baseline model for salient object detection," *IEEE Trans. Image Process.*, vol. 24, no. 2, pp. 742–756, Feb. 2015.

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

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

[106] P. F. Felzenszwalb and D. P. Huttenlocher, "Efficient graph-based image segmentation," *Int. J. Comput. Vis.*, vol. 59, no. 2, pp. 167–181, 2004.

[107] V. Movahedi and J. H. Elder, "Design and perceptual validation of performance measures for salient object segmentation," in *Proc. IEEE Comput. Soc. Conf. CVPRW*, Jun. 2010, pp. 49–56.

[108] R. Margolin, L. Zelnik-Manor, and A. Tal, "How to evaluate foreground maps?" in *Proc. IEEE Conf. CVPR*, Jun. 2014, pp. 248–255.

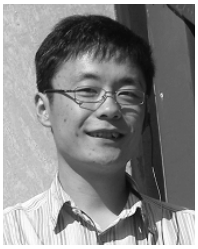
[109] C. Rother, V. Kolmogorov, and A. Blake, "'GrabCut': Interactive foreground extraction using iterated graph cuts," *ACM Trans. Graph.*, vol. 23, no. 3, pp. 309–314, 2004.

[110] T. Liu *et al.*, "Learning to detect a salient object," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 2, pp. 353–367, Feb. 2011.

[111] D. R. Martin, C. C. Fowlkes, and J. Malik, "Learning to detect natural image boundaries using local brightness, color, and texture cues," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 26, no. 5, pp. 530–549, May 2004.
[112] S. Avidan and A. Shamir, "Seam carving for content-aware image resizing," *ACM Trans. Graph.*, vol. 26, no. 3, 2007, Art. ID 10.
[113] J. Davis and M. Goadrich, "The relationship between precision-recall and ROC curves," in *Proc. 23rd ICML*, 2006, pp. 233–240.
[114] J.-Y. Zhu, J. Wu, Y. Wei, E. Chang, and Z. Tu, "Unsupervised object class discovery via saliency-guided multiple class learning," in *Proc. IEEE Conf. CVPR*, Jun. 2012, pp. 3218–3225.
[115] J. He *et al.*, "Mobile product search with bag of hash bits and boundary reranking," in *Proc. IEEE Conf. CVPR*, Jun. 2012, pp. 3005–3012.
[116] P. Wang, J. Wang, G. Zeng, J. Feng, H. Zha, and S. Li, "Salient object detection for searched Web images via global saliency," in *Proc. IEEE Conf. CVPR*, Jun. 2012, pp. 3194–3201.
[117] J. Zhang *et al.*, "Salient object subitizing," in *Proc. IEEE Conf. CVPR*, May 2015, pp. 4045–4054.
[118] H. Peng, B. Li, W. Xiong, W. Hu, and R. Ji, "RGBD salient object detection: A benchmark and algorithms," in *Proc. 13th ECCV*, 2014, pp. 92–109.
[119] A. K. Mishra, Y. Aloimonos, L.-F. Cheong, and A. A. Kassim, "Active visual segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 4, pp. 639–653, Apr. 2012.
[120] L. Mai, Y. Niu, and F. Liu, "Saliency aggregation: A data-driven approach," in *Proc. IEEE Conf. CVPR*, Jun. 2013, pp. 1131–1138.
[121] O. Le Meur and Z. Liu, "Saliency aggregation: Does unity make strength?" in *Proc. 12th ACCV*, 2014, pp. 18–32.
[122] 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. Vis.*, vol. 13, no. 10, p. 18, 2013.
[123] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2012, pp. 1097–1105.
[124] C. Szegedy *et al.*, (2014). "Going deeper with convolutions," [Online]. Available: http://arxiv.org/abs/1409.4842
[125] R. B. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Columbus, OH, USA, Jun. 2014, pp. 580–587.
[126] R. Zhao, W. Ouyang, H. Li, and X. Wang, "Saliency detection by multi-context deep learning," in *Proc. IEEE Conf. CVPR*, May 2015, pp. 1265–1274.
[127] S. He, R. W. H. Lau, W. Liu, Z. Huang, and Q. Yang, "SuperCNN: A superpixelwise convolutional neural network for salient object detection," *Int. J. Comput. Vis.*, pp. 1–15, Apr. 2015. DOI 10.1007/s11263-015-0822-0
[128] Y. Lin, S. Kong, D. Wang, and Y. Zhuang, "Saliency detection within a deep convolutional architecture," in *Proc. Workshops 28th AAAI Conf. Artif. Intell.*, 2014, pp. 31–37.
[129] G. Li and Y. Yu, "Visual saliency based on multiscale deep features," *CVRR*, pp. Jun. 2015, pp. 5455–5463.
[130] A. Borji and J. Tanner. (Mar. 2015). "Reconciling saliency and object center-bias hypotheses in explaining free-viewing fixations." [Online]. Available: http://arxiv.org/abs/1503.08853
[131] G. Kulkarni *et al.*, "Baby talk: Understanding and generating simple image descriptions," in *Proc. IEEE Conf. CVPR*, Jun. 2011, pp. 1601–1608.
[132] A. Farhadi, I. Endres, D. Hoiem, and D. Forsyth, "Describing objects by their attributes," in *Proc. IEEE Conf. CVPR*, Jun. 2009, pp. 1778–1785.
[133] L. Itti and M. A. Arbib, "Attention and the minimal subscene," in *Action to Language via the Mirror Neuron System*. Cambridge, U.K.: Cambridge Univ. Press, 2006.
[134] S. Antol *et al.* (May 2015). "VQA: Visual question answering." [Online]. Available: http://arxiv.org/abs/1505.00468
[135] H. Fang *et al.* (Nov. 2014). "From captions to visual concepts and back." [Online]. Available: http://arxiv.org/abs/1411.4952
[136] C. L. Zitnick, D. Parikh, and L. Vanderwende, "Learning the visual interpretation of sentences," in *Proc. IEEE ICCV*, Dec. 2013, pp. 1681–1688.
[137] K. Yun, Y. Peng, D. Samaras, G. J. Zelinsky, and T. L. Berg, "Studying relationships between human gaze, description, and computer vision," in *Proc. IEEE Conf. CVPR*, Jun. 2013, pp. 739–746.
[138] X. Chen *et al.* (Apr. 2015). "Microsoft COCO captions: Data collection and evaluation server." [Online]. Available: http://arxiv.org/abs/1504.00325
[139] D. Geman, S. Geman, N. Hallonquist, and L. Younes, "Visual Turing test for computer vision systems," *Proc. Nat. Acad. Sci. USA*, vol. 112, no. 12, pp. 3618–3623, 2015.
[140] O. Russakovsky *et al.*, "ImageNet large scale visual recognition challenge," *Int. J. Comput. Vis.*, pp. 1–42, Apr. 2014. DOI: 10.1007/s11263-015-0816-y

Ali Borji received the B.S. degree in computer engineering from the Petroleum University of Technology, Tehran, Iran, in 2001, the M.S. degree in computer engineering from Shiraz University, Shiraz, Iran, in 2004, and the Ph.D. degree in cognitive neurosciences from the Institute for Studies in Fundamental Sciences, Tehran, Iran, in 2009. He spent four years as a Post-Doctoral Scholar with iLab, University of Southern California, from 2010 to 2014. He is currently an Assistant Professor with the 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 the Ph.D. degree from Tsinghua University, in 2012. Then he did two years research fellow, with Prof. P. Torr in Oxford. He is currently an Associate Professor with Nankai University. His research interests includes computer graphics, computer vision, and image processing. He has received the Google Ph.D. Fellowship Award, the IBM Ph.D. Fellowship Award, and the new Ph.D. Researcher Award from the Chinese Ministry of Education.



Huaizu Jiang received the B.S. and M.S. degrees from Xi'an Jiaotong University, China, in 2005 and 2009, respectively. He is a Ph.D. student at the University of Massachusetts, Amherst. 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 the B.E. degree from Tsinghua University, in 2005, and the Ph.D. degree from the Chinese Academy of Sciences, in 2011. In 2011 and 2013, he served as a Research Fellow and Visiting Assistant Professor with Nanyang Technological University, Singapore. He is currently an Associate Professor with Beihang University, Beijing, China. His research interests include visual attention/saliency modeling, multimedia analysis, and vision from big data.