Quantitative Analysis of Human-Model Agreement in Visual Saliency Modeling: A Comparative Study

Ali Borji, *Member, IEEE,* Dicky N. Sihite, and Laurent Itti, *Member, IEEE*

***Abstract*— Visual attention is a process that enables biological and machine vision systems to select the most relevant regions from a scene. Relevance is determined by two components: 1) top-down factors driven by task and 2) bottom-up factors that highlight image regions that are different from their surroundings. The latter are often referred to as "visual saliency." Modeling bottom-up visual saliency has been the subject of numerous research efforts during the past 20 years, with many successful applications in computer vision and robotics. Available models have been tested with different datasets (e.g., synthetic psychological search arrays, natural images or videos) using different evaluation scores (e.g., search slopes, comparison to human eye tracking) and parameter settings. This has made direct comparison of models difficult. Here, we perform an exhaustive comparison of 35 state-of-the-art saliency models over 54 challenging synthetic patterns, three natural image datasets, and two video datasets, using three evaluation scores. We find that although model rankings vary, some models consistently perform better. Analysis of datasets reveals that existing datasets are highly center-biased, which influences some of the evaluation scores. Computational complexity analysis shows that some models are very fast, yet yield competitive eye movement prediction accuracy. Different models often have common easy/difficult stimuli. Furthermore, several concerns in visual saliency modeling, eye movement datasets, and evaluation scores are discussed and insights for future work are provided. Our study allows one to assess the state-of-the-art, helps to organizing this rapidly growing field, and sets a unified comparison framework for gauging future efforts, similar to the PASCAL VOC challenge in the object recognition and detection domains.**

***Index Terms*— Bottom-up attention, eye movement prediction, model comparison, visual attention, visual saliency.**

I. Introduction

VISUAL attention is a low-cost preprocessing step by which artificial and biological visual systems select the most relevant information from a scene, and relay it to higher-level cognitive areas that perform complex processes such as scene understanding, action selection, and decision making. In addition to being an interesting scientific challenge, modeling visual attention has many engineering applications, including in: **computer vision** (e.g., object recognition [1]–[3], object detection [4], [5], target tracking [6], image compression [7], and video summarization [8]); **computer graphics** (e.g., image rendering [9], image thumb-nailing [10], automatic collage creation [11], and dynamic lighting [12]); **robotics** (e.g., active gaze control [13], [14], robot localization and navigation [15], and human-robot interaction [16], [18]); and **others** (e.g., advertising [17] and retinal prostheses [19]).

Modeling visual saliency has attracted much interest recently and there are now several frameworks and computational approaches available. Some are inspired by cognitive findings, some are purely computational, and others are in between. However, since models have used different evaluation scores and datasets while applying various parameters, model evaluation against the state-of-the-art is becoming an increasingly complex challenge. In this paper, inspired by the PASCAL VOC object detection/recognition challenge [20], we aim to compare visual attention models in a unified framework over several scoring methods and datasets. Such a comparison helps better understand modeling parameters and provides insights towards further developing more effective models. It also helps better focus and calibrate the research effort by avoiding repetitive work and discarding less promising directions. It will also benefit experimentalists to choose the right tool/model for their applications. Since our main purpose is to compare models, rather than discuss attention concepts and models in detail, we refer the interested reader to general reviews for more information (e.g., Itti and Koch [21], Heinke and Humphreys [22], Frintrop *et al.* [23], and Borji and Itti [24]).

There is often a confusion between saliency and attention. Visual attention is a broad concept covering many topics (e.g., bottom-up/top-down, overt/covert, spatial/spatio-temporal, and space-based/object-based attention). Visual saliency, on the other hand, has been mainly referring to bottom-up processes that render certain image regions more conspicuous: For instance, image regions with different features from their surroundings (e.g., a single red dot among several blue dots). Bottom-up saliency has been studied in search tasks such as finding an odd item among distractors in pop-out and conjunction search arrays, as well as in eye movement prediction on free-viewing of images or videos. In contrast to bottom-up,

Manuscript received August 19, 2011; revised January 31, 2012; accepted May 3, 2012. Date of publication July 30, 2012; date of current version December 20, 2012. This work was supported in part by the Defense Advanced Research Projects Agency under Contract HR0011-10-C-0034, the National Science Foundation (CRCNS) under Grant BCS-0827764, the General Motors Corporation, and the Army Research Office under Grant W911NF-08-1-0360 and Grant W911NF-11-1-0046. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Alex ChiChung Kot.

The authors are with the Department of Computer Science, University of Southern California, Los Angeles, CA 90089 USA (e-mail: aliborji@gmail.com; sihite@usc.edu; itti@usc.edu).

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.2012.2210727

top-down attention deals with high-level cognitive factors that make image regions relevant, such as task demands, emotions, and expectations. It has been studied in natural behaviors such as sandwich making [25], driving [26], and interactive game playing [27]. In the real-world, bottom-up and top-down mechanisms are combined to direct visual attention. Correspondingly, models of visual attention often focus either on bottom-up (known as saliency models) or on top-down factors of visual attention. Due to the relative simplicity of bottom-up processing (compared to top-down), the majority of existing models has focused on bottom-up attention. For a review on attention in natural behavior, please refer to [28].

In addition to the dissociation between bottom-up and top-down, visual attention studies (and likewise models) can be categorized based on several other factors. Some studies have addressed explaining fixations/saccades in free viewing of static images while others have approached dynamic stimuli, such as observing movies or playing video games [29], [28]. This distinction has divided models into spatial (still images) or spatio-temporal models (over video stimuli). The majority of spatio-temporal models are also applicable to saliency estimation over static images. Although static models are also applicable to videos by processing each single frame, they have not been fundamentally built to account for such stimuli.

Models can be categorized as being space-based or object-based. Object-based models try to segment or detect objects to predict salient regions. This is supported by the finding that objects predict fixations better than early saliency [30]. In contrast, in space-based models, all operations happen at the image level (pixels or image patches), or in the image spectral phase domain. For these space-based models, the goal is to create saliency maps that may predict which locations have higher probability of attracting human attention (as measured, e.g., by subjective rankings of interesting and salient locations, reaction times in visual search, or eye movements). Salient region detection in object-based models adds a segmentation problem where the goal is to not only locate but also segment the most salient objects within a scene from the background. Perhaps because object segmentation remains a difficult machine vision problem, there are not as many object-based models as space-based models.

Another distinction is between overt and covert attention. Overt attention is the process of directing the the eyes towards a stimulus, while covert attention is that of mentally focusing onto one of several possible sensory stimuli (without necessarily moving the eyes). Many bottom-up saliency models have blurred the distinction between overt and covert attention and have focused onto detecting salient image regions, which in turn could attract one or both types of attention. Indeed, as detailed below, few models offer explicit mechanisms for the control of head/body/gaze movements.

Considering the above definitions, here we compare those visual saliency models that belong to the majority class of models, namely, those models that are: 1) bottom-up; 2) spatial or spatio-temporal; 3) space-based; 4) able to generate a topographic saliency map for an arbitrary digital image or a video; and 5) addressing free-viewing of images or videos (not solely visual search or salient object segmentation).

TABLE I
COMPARED VISUAL SALIENCY MODELS.

No.	Acronym: Model	Year	S	P	Resolution
1	**Gauss**: Gaussian-blob	-	I	M	51×51
2	**IO**: Human inter-observer	-	I	M	$W \times H$
3	**Variance**: [31]	-	I	C	$\frac{1}{16}W \times \frac{1}{16}H$
4	**Entropy**: [74]	-	I	C	$\frac{1}{16}W \times \frac{1}{16}H$
5	**Itti-CIO2**: Itti *et al.* [31], [32]	1998	I	C	$\frac{1}{16}W \times \frac{1}{16}H$
6	**Itti-Int**: Itti *et al.* [31], [32]	1998	I	C	$\frac{1}{16}W \times \frac{1}{16}H$
7	**Itti-CIO**: Itti *et al.* [33], [32]	2000	I	C	$\frac{1}{16}W \times \frac{1}{16}H$
8	**Itti-M**: Itti *et al.* [34]	2003	V	C	$\frac{1}{16}W \times \frac{1}{16}H$
9	**Itti-CIOFM**: Itti *et al.* [34]	2003	B	C	$\frac{1}{16}W \times \frac{1}{16}H$
10	**Torralba**: [35]	2003	I	M	$W \times H$
11	**VOCUS**: Frintrop *et al.* [4]	2005	B	C	$\frac{1}{4}W \times \frac{1}{4}H$
12	**Surprise-CIO**: [36]	2005	I	C	$\frac{1}{16}W \times \frac{1}{16}H$
13	**Surprise-CIOFM**: [36]	2005	B	C	$\frac{1}{16}W \times \frac{1}{16}H$
14	**AIM**: Bruce and Tsotsos [37]	2005	I	M	$\frac{1}{2}W \times \frac{1}{2}H$
15	**STB**: saliency toolbox [1]	2006	I	M	$\frac{1}{16}W \times \frac{1}{16}H$
16	**Le Meur**: Le Meur *et al.* [38], [39]	2006	B	X	$W \times H$
17	**GBVS**: Harel *et al.* [40]	2006	I	M	$W \times H$
18	**HouCVPR**: Hou *et al.* [41]	2007	I	M	64×64
19	**Rarity-L**: local rarity [42]	2007	I	M	$W \times H$
20	**Rarity-G**: global rarity [42]	2007	I	M	$W \times H$
21	**HouNIPS**: Hou *et al.* [43]	2008	I	M	$W \times H$
22	**Kootstra**: Kootstra and Shomacker [44]	2008	I	E	$W \times H$
23	**SUN**: Zhang *et al.* [45]	2008	I	M	246×331
24	**Marat**: Marat *et al.* [46]	2009	B	X	$W \times H$
25	**PQFT**: Guo *et al.* [47]	2009	I	M	400×400
26	**Yin Li**: Yin Li *et al.* [48]	2009	I	M	$W \times H$
27	**SDSR**: Seo and Milanfar [49]	2009	B	M	$W \times H$
28	**Judd**: Judd *et al.* [50]	2009	I	M	$W \times H$
29	**Bian**: Bian *et al.* [51]	2009	I	M	$\frac{1}{16}W \times \frac{1}{16}H$
30	**E-Saliency**: Avraham *et al.* [52]	2010	I	X	$W \times H$
31	**Yan**: Yan *et al.* [53]	2010	I	M	$W \times H$
32	**AWS**: Diaz *et al.* [54]	2010	I	E	$\frac{1}{2}W \times \frac{1}{2}H$
33	**Jia Li**: Jia Li *et al.* [55]	2010	I	E	$\frac{1}{16}W \times \frac{1}{16}H$
34	**Tavakoli**: Tavakoli *et al.* [56]	2011	I	M	$W/16 \times H/16$
35	**Murray**: Murray *et al.* [57]	2011	I	M	$W \times H$

S: Stimuli I: Image, V: Video, B: Both Image and Video. P: Programming Language M: MATLAB, C: C/C++, E: Executables, X: Sent Saliency Maps. W: Image Width and H: Image Height.

II. COMPARISON PLAN

First, we briefly explain experimental settings in Sec. II-A. Then, datasets including widely-used synthetic patterns and eye movement datasets over static scenes (natural, abstract, and cartoon images) and videos are described in Sec. II-B. Next, in Sec. II-C, three popular evaluation scores are explained. We then discuss some challenges in model comparison and our way to tackle them (Sec. II-D). Finally, experimental results of thorough model evaluation are shown in Sec. III followed by learned lessons in Sec. IV.

A. Settings

The first step in this study was to collect saliency models. Some models were already shared online. For others, we contacted their creators for software; the authors then either sent source code for us to compile or sent executables. Some authors, however, preferred to run their models on our stimuli and to send back saliency maps. In the end, we were able to

evaluate the 35 models listed in Table I, sorted by publication year. This table also shows stimulus types that models are applicable to and their implementation language. In addition to developed models by the authors, we also implemented two other simple yet powerful models, to serve as baseline: The Gaussian Blob (Gauss) and Human Inter-Observer (IO) models. The Gaussian blob model is simply a 2D Gaussian shape drawn at the center of the image; it is expected to predict human gaze well if such gaze is strongly clustered around the image center. The human inter-observer model outputs, for a given stimulus, a map built by integrating eye fixations from other subjects than the one under test, while they watched that stimulus. The map is then smoothed by convolving with a Gaussian filter. This inter-observer "model" is expected to provide an upper bound on prediction accuracy of computational models, to the extent that different humans may be the best predictors of each other. Since maps made by models have different resolutions, we resized them (using nearest neighbor interpolation) to the size of the original images onto which eye movements have been recorded. Map resolutions as well as model acronyms used in the rest of the paper are listed in Table I. Please note that, besides models compared here, some other models may exist that might perform well, but are not publicly available or easily accessible (e.g., [58]). We leave such models for future investigations.

B. Stimuli

Attention models have first been validated by predicting accuracy and reaction times of human subjects in target detection in visual search arrays. In addition, many models have commonly been validated against eye movement data.

1) Synthetic Stimuli: Early attention studies and models used simple synthetic patterns such as searching for a target or detecting an odd item among distractor items to find out important feature channels in directing attention and how they are combined [59]. For instance, it has been shown that reaction time for a simple pop-out search task remains constant as a function of set size (number of all items on the screen), while in conjunction search tasks (searching for a target that is different in two features) reaction time increases linearly with set size [59]. In [21], [60], authors enumerate and discuss features that influence attention. For a computational perspective on implementation of these features in saliency models, please refer to [6], [21], [24].

Fig. 1, shows a collection of 54 diverse synthetic patterns where one item (a target) differs from all other (distractor) items (pop-out, search asymmetry, texture, semantics, size, grouping, curvature, etc.). Such stimuli have been widely used for qualitative evaluation of saliency and attention models in the past. Patterns are sorted from easy to hard for models (Fig. 5) from left to right and top to bottom. They can be categorized into: orientation pop-out (3, 9, 21, 25, 38, 43, 51, 54), texture pop-out (6, 12, 14, 24, 36, 39, 47), curvature pop-out (35, 48), size pop-out (8, 10, 17, 30, 52), grouping (2, 13, 26, 28, 34), color pop-out (1, 4, 16, 19, 20, 27, 29, 31, 32, 33, 41, 44, 50, 53), intensity pop-out (11, 18, 37, 42), search asymmetry (5;15, 22;46, 40;49), and other complex search arrays (7, 23). In some patterns, targets are embedded in noise (e.g., speckle noise: 11, 20, 31 and orientation noise: 19, 41). We aimed to assess the pure target detection performance of models. This is why we included harder displays, even though humans may perform poorly on them (hence a great model of human attention should also perform poorly, but some models might transcend human abilities with such images).

2) Natural Scenes: Space-based models have often been tested for eye fixation prediction over still image datasets and spatio-temporal models have been evaluated against video data.

a) Image datasets: Since statistics of different datasets vary, we employed three popular image datasets often used for saliency evaluation: 1) Bruce and Tsotsos [37] (one of the earliest and most widely used datasets). It contains 120 images mainly indoor and in-city scenes. Due to the small size of this dataset and the small number of subjects, its sole usage is less encouraged; 2) Kootstra and Shomacker [44] (which contains a wide variety of images); and 3) Judd *et al.* [50] (which is the largest dataset available to date containing 1003 images). It contains many images with human faces and has a high degree of photographer bias and a smaller number of subjects. Le Meur [38] dataset has only 27 images with the highest number of eye-tracking subjects (40). We avoided to use this dataset as its images are highly center-biased (See Sec. II-D).

Because of the specialty of datasets (different optimal weights for features over different datasets [61]), a fair evaluation is to compare models over several datasets (Sec. III).

b) Video datasets: Unfortunately, there are not many publicly available video datasets with associated eye-tracking data. This calls for collecting more eye movement data over videos. Here, we run models over two datasets: **1)** A large popular benchmark dataset for comparison of spatio-temporal saliency, called CRCNS-ORIG [62], which is freely accessible. Fig. 2 shows a sample frame from each video of CRCNS-ORIG dataset embedded with eye fixations. **2)** A recent project called DIEM (Dynamic Images and Eye Movements) has investigated where people look during dynamic scene viewing such as during film trailers, music videos, or advertisements [63][1]. Fig. 3 shows sample frames of DIEM with fixations overlaid.

Please refer to [24] for more details on available datasets. Our choice of datasets emphasizes popularity, thoroughness, and variety in the stimuli.

We applied spatial and spatio-temporal models over static (still images) and dynamic (video) stimuli to compare accuracy of both types of models over both types of stimuli. This way we can analyze the usefulness of temporal information by comparing accuracy of models built from simple features plus the motion channel (e.g., the Itti-CIOFM model) with other high-performing models without temporal information.

[1]DIEM has so far collected data from over 250 participants watching 85 different videos. All of this data is freely available. We selected 20 videos and about 1,000 frames from each to make a benchmark for model comparison. Selected videos cover different concepts/topics. We only used right-eye positions of subjects to make model evaluation tractable. Frames of this dataset were scaled down to 640×480 while maintaining aspect ratio.

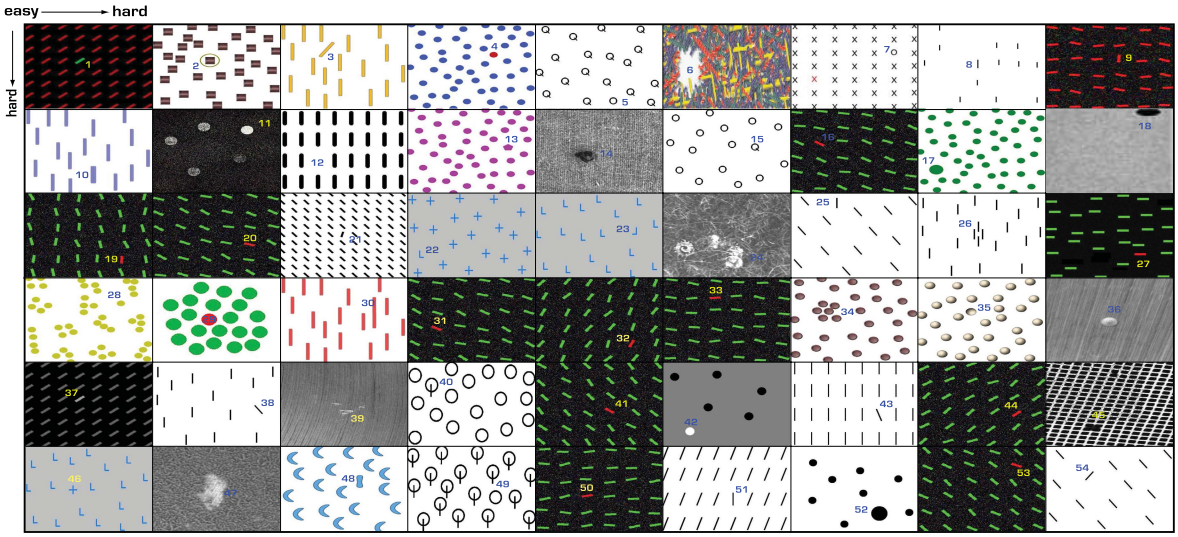

Fig. 1. Synthetic patterns. Stimuli are numbered in blue/yellow from 1 to 54 in row-first order. Numbers are positioned close to the target locations and are for illustration purposes only. Stimuli are sorted according to their average easiness of oddity detection for saliency models (see Fig. 5).

Fig. 2. One sample frame (frame no. 100) from 50 videos of CRCNS-ORIG eye movement dataset. Eye movements are embedded on images in yellow. For some videos, eye fixations are shown in blue for better illustration. Video names in order (from left to right, top to bottom) are: 1) *Beverly*01; 2) *Beverly*03; 3) *Beverly*05; 4) *Beverly*06; 5) *Beverly*07; 6) *Beverly*08; 7) *Gamecube*02; 8) *Gamecube*04; 9) *Gamecube*05; 10) *Gamecube*06; 11) *Gamecube*13; 12) *Gamecube*6; 13) *Gamecube*7; 14) *Gamecube*18; 15) *Gamecube*23; 16) *Monica*03; 17) *Monica*04; 18) *Monica*05; 19) *Monica*06; 20) *Saccadetest*; 21) *Standard*01; 22) *Standard*02; 23) *Standard*03; 24) *Standard*04; 25) *Standard*05; 26) *Standard*06; 27) *Standard*07; 28) $Tv-action01$; 29) $Tv-ads01$; 30) $Tv-ads02$; 31) $Tv-ads03$; 32) $Tv-ads04$; 33) $Tv-announce01$; 34) $Tv-music01$; 35) $Tv-news01$; 36) $Tv-news02$; 37) $Tv-news03$; 38) $Tv-news04$; 39) $Tv-news05$; 40) $Tv-news06$; 41) $Tv-news09$; 42) $Tv-sports01$; 43) $Tv-sports02$; 44) $Tv-sports03$; 45) $Tv-sports04$; 46) $Tv-sports05$; 47) $Tv-talk01$; 48) $Tv-talk03$; 49) $Tv-talk04$; and 50) $Tv-talk05$. Note that a different number of subjects observed videos. For results of model comparisons on these videos, please see Fig. 8.

Fig. 3. Sample frames from 20 videos of DIEM [63] dataset. Yellow dots show right eye positions of all human subjects. Please see Fig. 8 for results.

Another approach will be extending all spatial models to the temporal domain before comparison. This, however goes beyond our scope in this paper and should be addressed by the model creators.

C. Evaluation Scores

Here, three evaluation scores for comparison of models are explained. The motivation for analyzing models with more than one metric is to ensure that the main qualitative

conclusions are independent of the choice of metric. In the following, G denotes a ground-truth saliency map which is a map built by inserting 1's at fixation locations and convolving the result with a Gaussian for smoothing. An estimated saliency map which is computed by a saliency model is denoted by S.

1) Linear Correlation Coefficient (CC): The linear correlation coefficient measures the strength of a linear relationship between two variables: $CC(G,S) = \frac{cov(G,S)}{\sigma_G \sigma_S}$ where σ_G and σ_S are the standard deviations of the G and S maps, respectively [64], [65]. When CC is close to $+1/-1$ there is almost a perfectly linear relationship between the two variables.

2) Normalized Scanpath Saliency (NSS): NSS [66], [67] is the average of the response values at human eye positions (x_h^i, y_h^i) in a model's saliency map (S) that has been normalized to have zero mean and unit standard deviation. $NSS = 1$ indicates that the subjects' eye positions fall in a region whose predicted saliency is one standard deviation above average. Thus, when $NSS \geq 1$ the saliency map exhibits significantly higher saliency values at human fixated locations compared to other locations. Meanwhile $NSS \leq 0$ indicates that the model performs no better than picking a random position, and hence is at chance in predicting human gaze.

3) Area Under Curve (AUC): AUC is the area under the Receiver Operating Characteristics (ROC) curve [68]. Using this score, human fixations are considered as the positive set and some points from the image are sampled, either uniformly or non-uniformly [45] (for discounting center-bias), to form the negative set. The saliency map S is then treated as a binary classifier to separate the positive samples from the negatives. By thresholding over the saliency map and plotting *true positive rate* vs. *false positive rate* an ROC curve is achieved for each image. Then ROC curves are averaged over all images and the area underneath the final ROC curve is calculated [37], [69]. Perfect prediction corresponds to a score of 1 while a score of 0.5 indicates chance level.

For more details on evaluation scores please refer to [24][2].

D. Challenges and Open Problems

Here we discuss challenges that have emerged as more saliency models have been proposed. These are open issues that must be considered, not only for research but also for performing a fair comparison of all models.

1) Center Bias (CB): Perhaps the biggest challenge in model comparison is the issue of center-bias. Center-bias means that a majority of fixations happen to be near the image center. Several reasons for this have previously been proposed. For instance, it could be due to a tendency of photographers to put interesting (and hence salient [70]) objects at the image center; or it could be because of a viewing strategy by which subjects first inspect the image center, maybe to rapidly gather a global view of the scene [71], [72]. Some models have implicitly (e.g., GBVS [40]) or explicitly (e.g., Judd [50]) used center-preference (location prior) to better account for eye movements. This, however, makes fair comparison challenging. Three remedies are possible: 1) Every model adds a Gaussian of a certain size to its output. This approach has the drawback that it is hard to impose to the large community of researchers. 2) Collecting a dataset with no center-bias. This is difficult because even if we have an approach to uniformly distribute image content, viewing strategy still exists. 3) Designing suitable evaluation metrics, which is what we consider as the most reasonable approach, and which we use here.

To eliminate center-bias effects, Zhang *et al.* [45] used the shuffled AUC metric instead of the uniform AUC metric. They defined shuffled AUC as: For an image and a human subject, the positive sample set is composed of the fixations of that subject on that image, while the negative set, instead of uniformly random points, is composed of the union of all fixations of all subjects across all other images, except for the positive set. This score allows for a stronger assessment of the non-trivial off-center fixations, which are the ones that are more challenging and more interesting to predict. Alternatively, Qi and Koch [61], defined an unbiased AUC score as the ratio of normal AUC to the AUC score of the inter-observer model.

Here, along with using the shuffled AUC score, we apply models to images with low center-bias. This second-order study provides another way of differentiating models behavior over (difficult) fixations which deviate from center. Please note that this does not necessarily mean that center-bias is not a fact of human attention behavior. To this end, we propose a new measure called **Center-Bias Ratio (CBR)** to quantify the amount of center-bias in an image or a set of images. First, for an image, a heat map is generated by pooling fixations from all subjects without Gaussian smoothing. Then, the ratio of fixations inside each central circle to the overall number of fixations in the image is calculated. By varying the radius, a vector of ratios is derived. If there are more fixations at the center, the first values of this vector should be very high. By applying a fixed threshold, one can make a decision whether an image is center-biased or not.

Fig. 4 shows distribution of fixations for three datasets and their center-bias ratio. The five most and five least center-biased images from datasets are also shown. Judd *et al.*, and Bruce and Tsotsos datasets are highly center-biased (at 40% circle, from center to image corner, they explain more than 80% of fixations) and Kootstra and Shomacker has the least center-bias amongst three. This might be because this dataset has many symmetric objects (e.g., flowers) off the center.

To test how many images pass a CB criterion, at the radius level of 40%, we selected an image from a dataset if its CBR was less than 0.7. This way, 10, 58, and 120 images from Bruce and Tsotsos, Kootstra and Schomaker, and Judd *et al.* datasets passed the selection criteria, respectively (Overall 15% of 1250 images)[3].

[2]In addition to above scores, Kullback-Leibler (KL) (the divergence between the saliency distributions at human fixations and at randomly shuffled fixations; used in [36], [45]), [61], and the string-edit distance (difference between the sequence of fixations generated by a saliency model versus human fixations) [73], [74]) have also been used for model evaluation. Note that all of these scores (except the Shuffled AUC) are influenced by the center-bias. We draw conclusions based on the average model behavior on these scores.

[3]We also used another dataset from Le Meur *et al.* [38] but none of the images passed the threshold. Link: http://www.irisa.fr/temics/staff/lemeur/.

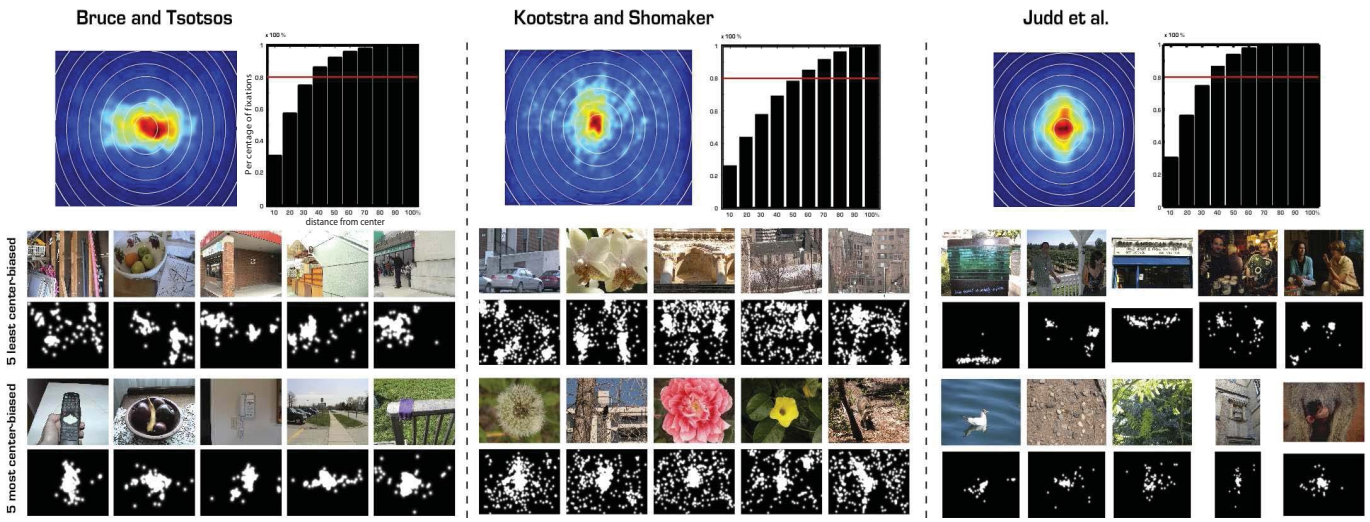

Fig. 4. Results of center-bias analysis over three datasets. The first row shows the heatmap of all fixations over all images for each dataset. White rings show 10% increase in radius from the image center, and the bar chart at the right of the heatmap shows the percentage of fixations that happen in each ring. The red horizontal bar shows the 80% density level. Five most and least center-biased images from each dataset along with eye fixations are shown at the bottom.

2) Border Effect: Another challenge is the treatment of image borders. Zhang *et al.* [45] showed that KL and ROC scores are corrupted by edge effects. When an image filter lies partially off the edge of an image, the filter response is not well defined. They varied the size of a black border added around a dummy white saliency map (of size 120×160 pixels) and showed that as the border size increases, ROC and KL scores increase as well. Since human eye fixations are rarely near the edges of test images, edge effects primarily change the distribution of saliency of the random samples. For the dummy saliency map, a baseline map (uniform white) gives a ROC value of 0.5, adding a four-pixel black border yields 0.62, and an eight-pixel black border yields 0.73. The same 3 border sizes would yield KL scores of 0, 0.12, and 0.25. Note that a black border effect due to variations in handling invalid filter responses at the image borders is similar to the center-bias issue and could be handled the same way. But the first is a problem with datasets while the second one regards a problem in modeling.

3) Scores: Some issues concern scores. For instance, as a limitation of ROC, Qi and Koch [61], compared two saliency maps with different degrees of false alarm rates. Interestingly, while one map had a clear dense activation at fixations (with almost no background activation), its standard AUC (=0.975) was not dramatically better compared to the other map (with activations at both fixations and background) with much higher false alarm rate (AUC = 0.973). Because of the normalization to the entire map, this problem did not affect NSS score.

4) Model Parameters: Another problem regarding fair model comparison is adjusting parameters in models. For instance, it has been shown that smoothing the final saliency map of a model affects the scores [75]. In models described in Table I, some authors mentioned the best set of parameters, and some manually tunned their model on our stimuli and sent back the saliency maps.

To tackle center-bias, border effects, and scoring issues, instead of only using one score, we decided to use three, with an emphasis on analysis of results using the shuffled AUC score which is more robust to center-bias and borders. A model that works well should score high (if not the best) at almost any score. Regarding model parameters, over some cross-validation data, we tried to tune models for best performance by qualitatively checking saliency maps or quantitatively by calculating scores. However, as further discussed in section IV, ultimately the model parameter issue will be best handled through an online challenge where participants can tune their own models before submitting results.

III. Experimental Results

Having laid out the evaluation framework, we are ready to compare saliency models in this section.

A. Results Over Synthetic Images

Fig. 5 shows ranking of models over synthetic patterns. The location of each target in each stimulus was tagged manually (bottom-right panel). Then the accuracy of a saliency model to capture the target was calculated using the NSS score. This score is more suitable here because there is only one target position in each image and if a model could accurately predict that location it would get a high score. The higher NSS thus means better target detection. The top-left panel in this figure shows performance of all models over all stimuli sorted both ways. The bottom-left panel ranks stimuli in terms of simplicity of target detection averaged over models (see also Fig. 1). The top-right panel shows sorted NSS scores (averaged over stimuli) of models.

On average (over stimuli), all models performed significantly above chance. Overall, models based on FIT theory performed higher on synthetic patterns (e.g., compared to statistical and information-theoretic models). STB, AWS, GBVS, VOCUS, Bian, and Itti-CIO models achieved the best NSS scores. From these models, Itti-CIO and hence its descendants STB and VOCUS are directly based on the FIT framework. Similar to these, AWS and GBVS models have used multi-scale color, intensity and orientation channels. One high-performing model which is not based on FIT is Bian's model,

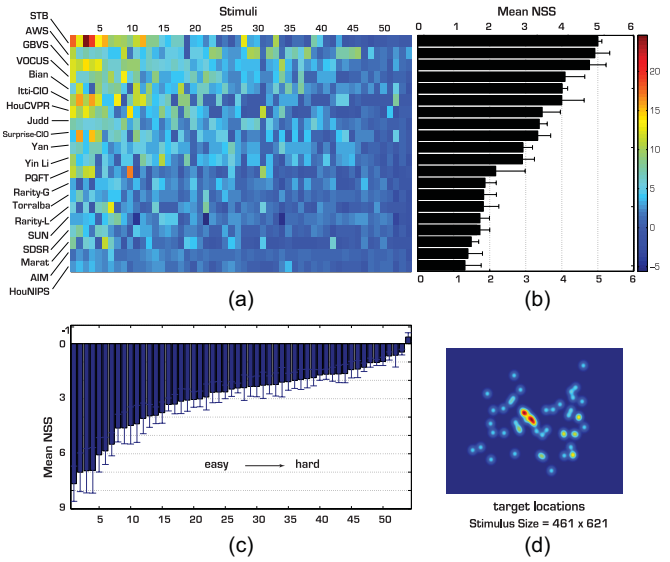

Fig. 5. Ranking models over synthetic patterns shown in Fig. 1. (a) Individual NSS scores of models for stimuli. (b) Sorts models averaged over stimuli and (c) sorts stimuli averaged over models. Error bars indicate standard error of the mean (SEM). (d) Spatial distribution of target locations.

which works in the frequency domain. Inspecting the high-performing models, we noticed that they all generate maps with a high peak at the target location and less activation elsewhere, which results in high NSS values. Models including AIM and HouNIPS seem impaired by the border effect, which affects their normalization; indeed, these models perform poorly on all our search-array stimuli. We expected that some models might actually surpass human vision in some of these images, i.e., they might mark as salient some targets which are hard to be immediately seen by humans. For example, AWS is doing quite well on hard image 23. Although some stimuli were easier for many models, no single stimulus was easy for all models. For example, stimulus 1, a simple red/green color pop-out was easy for models which include a separate color channel but remained challenging for several statistical models which are based on natural scene statistics (AIM, HouNIPS, Rarity-G). One important conclusion of our study therefore is that to date no model performs perfectly over all synthetic stimuli tested here. Fig. 6 illustrates saliency maps of models over the best and worst synthetic stimuli (averaged over all models) as well as some other sample synthetic stimuli.

Although in this section we focused on evaluating the consistency of saliency models with a number of classic psychophysical results related to bottom-up attention, there are several other tests that a model could be verified against, including: nonlinearity against orientation contrast, efficient (parallel) and inefficient (serial) search, orientation asymmetry, presence-absence asymmetry and Weber's law, and influence of background on color asymmetries (see [76], [77]). Some models have been partially tested against such stimuli [37], [54], [58], [77].

B. Results Over Natural Scenes

Fig. 7(a) shows ranking of models for fixation prediction over still images. For statistical significance testing of mean scores between two models, we used the t-test at the significance level of $p \leq 0.05$. Although the ranking order is not exactly the same over all three datasets, some general patterns can be observed. Using the CC score, over all three datasets, GBVS works the best. The Yan, Kootstra, and Gauss models are among the best six. High CC scores for the Gauss model indicate that there is high density of fixations at the image center over all three datasets. Higher CC for Gauss over the Judd *et al.* dataset (no significant difference between Gauss and GBVS; $p = 0.1$) means higher central eye concentration over this dataset. Similarly, using NSS, GBVS did the best and the Yan, Judd, AWS, and Kootstra models were among the six best. High performance for Gauss with NSS again indicates a high center-preference over datasets (Gauss ranked third over the Judd *et al.* dataset). Scores of models over the Kootstra and Shomacker dataset are smaller than over other datasets. This might be partially due to difficulty of stimuli in this dataset. For instance, many of them are outdoor natural scenes as opposed to close-up shots of objects or animals. Consistent with previous research, an important point here is that CC and NSS scores are sensitive to center-preference (high scores for Gauss model), therefore their usage is not encouraged for future work. Using shuffled AUC, the Gauss model is the worst (not significantly higher than chance) over all three datasets as we expected. Indeed, the shuffled AUC measure explicitly discounts center bias by sampling random points from human fixations. With shuffled AUC, the AWS model is significantly better than all other models over the three datasets, followed by HouNIPS model. The AIM and Judd models were the other two models that did well. One interesting observation is that AWS is able to predict human fixations over the Kootstra and Shomacker dataset at the level of human inter-observer similarity (no significant difference between model's score and Human inter-observer score). Rarity-L, Entropy, and STB are three models that did worst over CC and NSS scores. In terms of AUC scores, Gauss, STB, and Marat are at the bottom.

Except for the aforementioned case of AWS over the Kootstra and Shomacker dataset, the main conclusion of this study is that a significant gap still exists between the best models and human inter-observer agreement. The spread of models scores is also quite narrow, and for NSS over the Kootstra and Shomacker and Judd *et al.* datasets the gap between IO and the best model is greater than that between the best and worst models. This indicates that even though much progress has been made in modeling saliency over the past 13 years, dramatic and qualitatively better new models still remain to be discovered that will better approach human eye fixations. To the disappointment of the authors, many recent models overall perform worse that the Itti-CIO2 model published in 1998 [31], indicating the importance of using a comprehensive comparison framework for measuring progress. We further examine these issues in the Discussion section (Sec. IV).

An important note from our comparisons is that most models that did well overall, performed reasonably well over every combination of dataset and score. An exception is GBVS which performed the best over three datasets using the CC and NSS scores but not as well (though still quite well) with AUC. The performance drop of the GBVS model could be

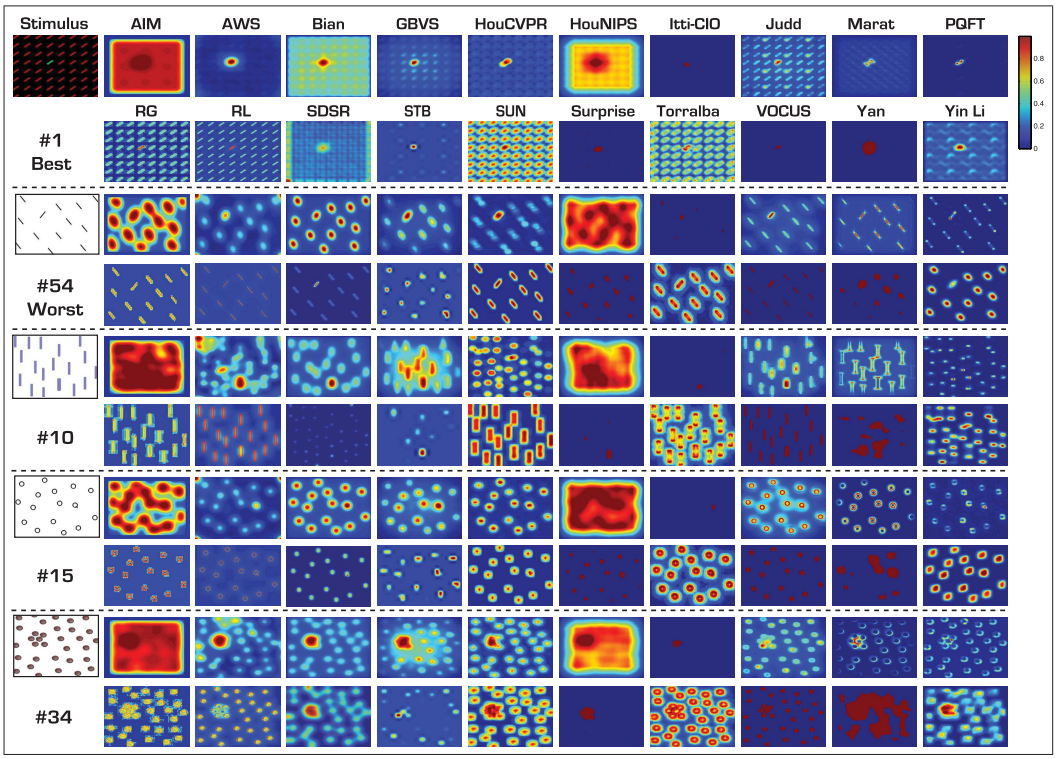

Fig. 6. Prediction maps of saliency models for the best, worst, and other sample synthetic stimuli. Best and worst stimuli are determined based on difficulty of models (on average) to detect the odd item among distractors in a search array.

because it takes advantage of center-bias. Some of the models which scored well on the synthetic patterns (Fig. 5) scored poorly on natural image datasets (e.g., STB and Itti-CIO). To some extent, we find that this may be due to the fact that these models are developed based on FIT framework which has been originally proposed to explain synthetic patterns. The Itti-CIO model also generates very sparse maps which do not reflect well the substantial inter-observer variations present in the human eye movement data.

In addition to using the shuffled AUC score, we conducted another experiment to compare models over stimuli with less center-bias. We selected 100 images from the Judd *et al.* dataset with least center-bias ratio (using 40% circle) and calculated scores for those images. Results are shown in Fig. 7(b). The rationale for focusing on images that yield many off-center fixations is that such fixations may convey more information about the processes by which attention is drawn to salient peripheral stimuli (as opposed to central fixations, which may be stimulus-driven or part of a viewing strategy; see section II-D). Indeed, we verified that the Gauss model performed poorly on this dataset. Consistent with CC and NSS scores over three datasets, here GBVS again scored the best, and the ranking of models is almost the same as when using all images across these two scores. With shuffled AUC, the ranking is almost the same as with the original datasets, with AWS, HouNIPS, and AIM at the top. Similar to the original datasets, the AUC performance of GBVS is not among the best. Note how, with shuffled AUC (which is emerging as the most reliable score), all models are closer to the IO performance in the least center-biased dataset. This new approach to dataset design helps us mitigate the above remark about the need for a qualitative jump in eye movement prediction: The off-center fixations, which arguably are the most important and difficult to predict, are captured quite well by many models.

Our next analysis is ranking models over different classes of stimuli from the Kootstra and Shomacker dataset. The intuition behind this experiment is that since different models use different features, and different classes of images may exhibit different feature distributions, it is likely that models may selectively perform higher over different types of images. Fig. 4, middle column, shows sample images from the Kootstra and Shomacker dataset. Images of this dataset fall into 5 categories: 1) Animals, 2) Automan (cars and humans), 3) Buildings, 4) Flowers, and 5) Nature. The shuffled AUC scores of all models are shown in Table II for each category. This table also shows scores of the inter-observer (IO) model as well as average scores of models (using three scores) across 5 categories. Interestingly, again the AWS model did the best over all categories (it was only significantly better than other models in the Flowers category). HouNIPS, Judd, SDSR, Yan, and AIM were also at the top. Gauss, STB, Marat, and Entropy ranked at the bottom. The least performance among categories belongs to Nature stimuli (using all 3 scores), probably because stimuli in the Nature category are more noisy and there are less solid objects or dense salient regions. All models scored below AUC $= 0.6$ in that category, and humans are also less consistent over nature stimuli (smaller AUC score for IO model). The best performance of models is over the Automan category, which consists of in-city scenes containing cars and humans, and IO also scored highest in this category. Model performance differences over categories suggests that

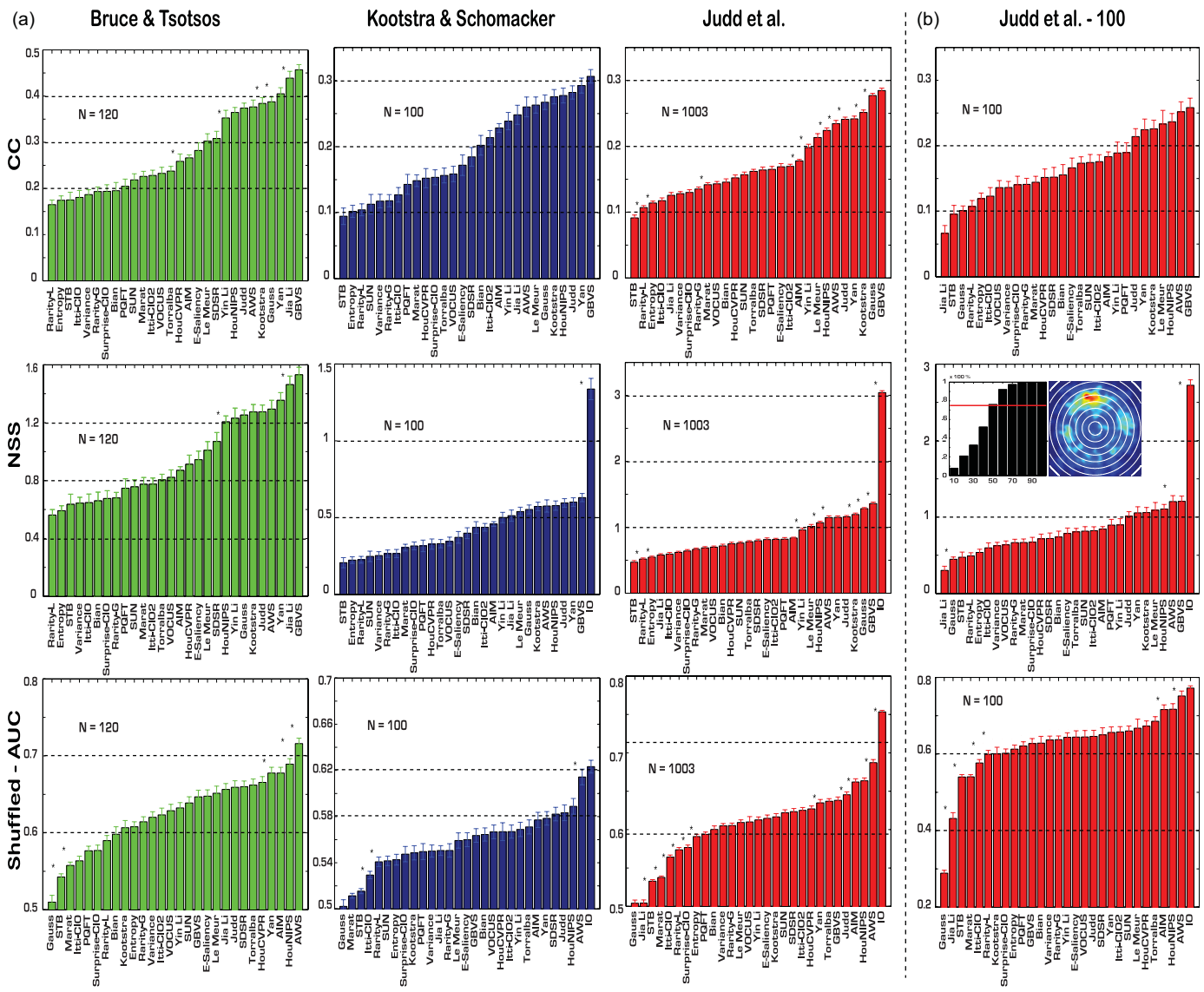

Fig. 7. (a) Ranking visual saliency models over three image datasets using three evaluation scores: Correlation coefficient (CC), normalized scanpath saliency (NSS), and shuffled AUC. Left column: Bruce and Tsotsos [37]. Middle column: Kootstra and Shomacker [44]. Right column: Judd *et al.* [50]. Stars indicate statistical significance using t-test (95%, $p \leq 0.05$) between consecutive models. Note that no star between two models that are not immediately close to each other does not necessarily mean that they are not significantly different. In fact, it is highly probable that a model that is significantly better than the one in its left, also scores significantly better than all other models on its left. Error bars indicate standard error of the mean (SEM): $(\sigma/\sqrt{N})$, where σ is the standard deviation and N is the number of images. We do not show CC results for IO model because comparing the map built from fixations of one subject with the map built from fixations of all other subjects using CC, does not generate a high value (both maps are convolved with a Gaussian). This is because few fixations of only one subject do not generate a diffused map, which is favored by CC score. We also could not calculate IO score over Bruce and Tsotsos dataset since fixations are not separated for each subject. (b) This column sorts models over 100 least center-biased images from the Judd *et al.* dataset (see Section II-D). The heatmap at the top-most panel shows distribution of fixations over selected images. Judd model uses center feature, gist, and horizon line, and object detectors for cars, faces, and human body. Itti-CIO2 is the approach proposed by Itti *et al.* [31] that uses the following normalization scheme: For each feature map, find the global max M and find the average m of all other local maxima. Then just weight the map by $(M-m)^2$. In the Itti-CIO method [33], normalization is: Convolve each map by a difference of Gaussian(DoG) filter, cut off negative values, and iterate this process for a few times. This normalization operation results in sparse saliency maps. In the literature, majority of models have been compared against Itti-CIO.

customizing models based on image category might further improve fixation prediction accuracy. Some models indeed rely on detecting the "gist" of a scene (e.g., whether it is indoors or outdoors) to establish a spatial prior on saliency [35]; these could be further combined with learning techniques (e.g., [61]) to modulate features contributing to saliency based on a scene's gist or category. Such research might benefit from deeper psychological studies of eye movement patterns over different categories of scenes.

Fig. 8(a) sorts models over the CRCNS-ORIG dataset using three scores. Rankings are almost the same over CC and NSS scores with GBVS, Gauss, Marat, HouNIPS, Judd, and Bian models at the top. Using the AUC score, AWS, HouNIPS, Bian and Human inter-observer are the best. The reason why, when using shuffled AUC, the inter-observer model is slightly lower than the three mentioned models is likely because the number of subjects is small and hence a map from other subjects may not be a good predictor of the remaining test subject.

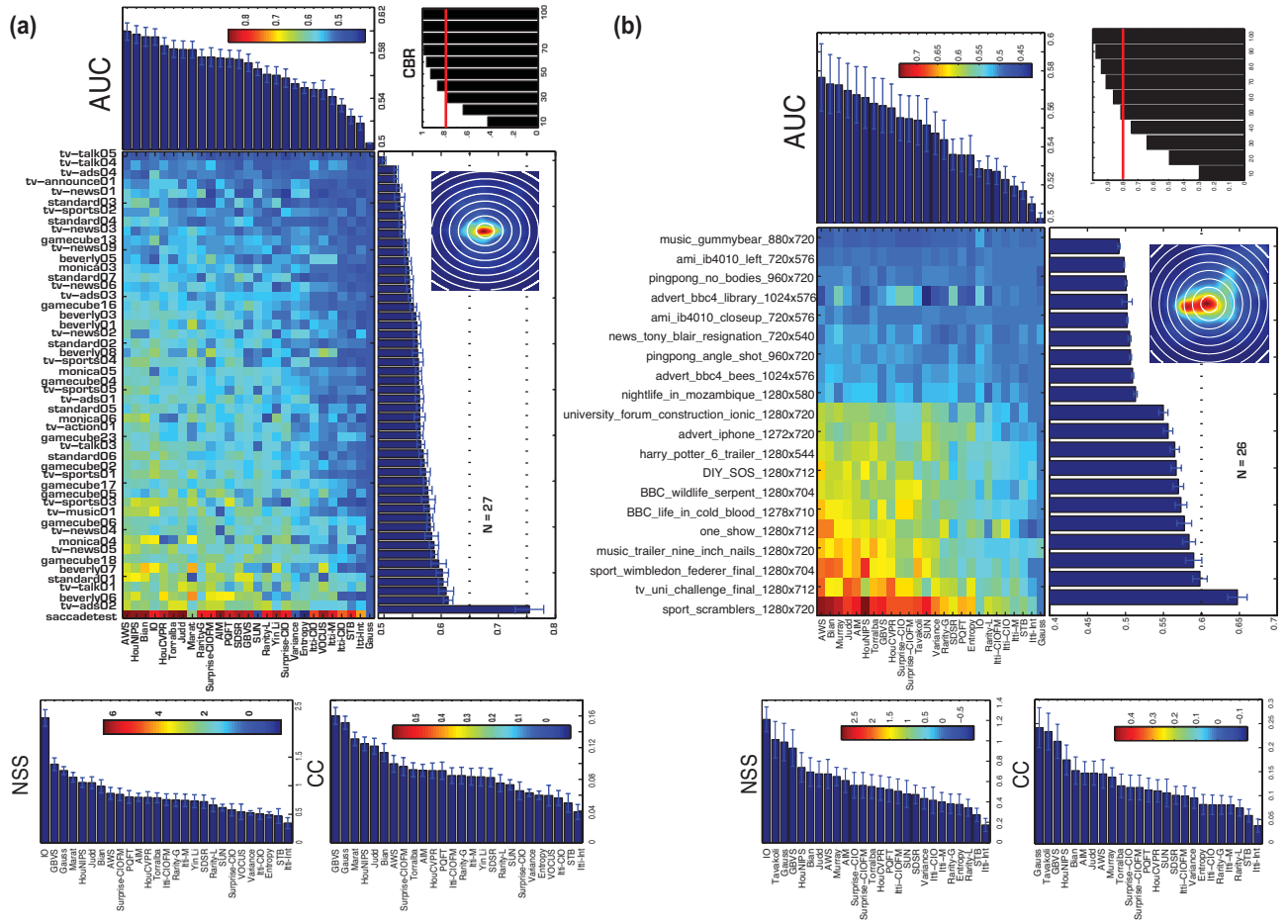

Fig. 8. (a) Ranking visual saliency models over CRCNS-ORIG dataset [62]. (b) Ranking models over DIEM dataset [63]. Only these models had motion channel: Itti-M, Itti-CIOFM, Surprise-CIOFM, Marat, and PQFT.

Why then is the human inter-observer significantly better than other models when using NSS? This is likely because even if in few occasions humans look at the same location, this generates a very large NSS value. The human inter-observer map in this dataset is a very sparse map and a hit results in a very large NSS score. Also, note that the inter-observer model is not significantly better than the three best computational models using AUC. Interestingly, only the motion channel of the Itti model (Itti-M) worked better than many models over video stimuli (specially using CC and NSS scores). Itti-Int was the worst among all models with STB, Entropy, Itti-CIO, Variance, VOCUS, and Surprise-CIO: all these models indeed only use static features. CC values are smaller here compared with still images because there are fewer fixations (due to smaller numbers of subjects).

All models achieved higher scores (all three) over the *saccadetest* video clip, which is a circular moving blob on a static blue background (see Fig. 2). Other stimuli on which models did well include *gamecube05*, *gamecube17*, *tv-news04*, *gamecube06*, and *gamecube23*, which tend to depict only one central moving actor of interest. Lowest scores belong to *standard04*, *tv-announce01*, *tv-talk05*, and *standard03*, which are very cluttered scenes with many actors and moving objects. Inspecting the difficult video clips suggests that eye fixations in these clips are often driven by complex cognitive processes; for instance, in *tv-talk-05*, fixations switch from one speaker to the other following their subtle lip movements, while the overall saliency of both their faces remains high throughout the clip. Much more thus needs to be studied in modeling such cognitive influences on saliency, as small dynamic changes pixel-wise (like moving lips) can yield dramatic differences in human gaze allocation (see, e.g., [28], [29]). Eye fixation distributions of CRCNS-ORIG dataset shows higher density at the center compared to still image datasets (about 42% at the inner-most circle (10% radius) and about 83% at 40% radius). This could also be verified by the high scores of the Gauss model over CC and NSS scores. Over this dataset, similar to image datasets, NSS and AUC scores of many models are much smaller than human inter-observer scores. Generally, models that performed well over static images also achieved higher accuracies over the CRCNS-ORIG dataset. Interestingly, overall, models with a motion channel rank towards the middle, i.e., they do not seem to work better than the best models which only use static features, though they still work better than the lowest-performing static models.

Ranking of models over DIEM video dataset is shown in Fig. 8(b). The IO, Tavakoli, Gauss, GBVS, HouNIPS, Bian, and Judd models ranked on top using CC and NSS scores. Using shuffled AUC, however, AWS, Bian,

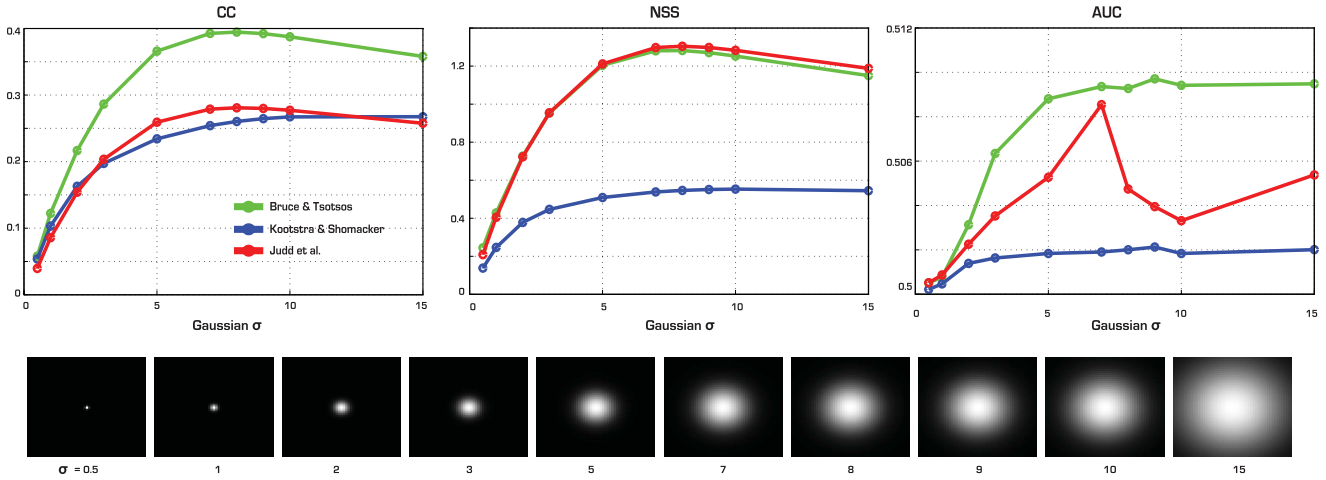

Fig. 9. Analysis of Gaussian blob size parameter. CC, NSS, and shuffled AUC scores over Gaussian blobs at the image center with increasing size from small to large (bottom-row). Size of each blob is 50×50 pixels.

TABLE II

MODEL COMPARISON OVER CATEGORIES OF KOOTSTRA AND SHOMACKER DATASET USING SHUFFLED AUC SCORE. SECOND NUMBER IN EACH PAIR OF VALUES IS SEM. THE THREE BEST MODELS FOR EACH CATEGORY ARE SHOWN IN BOLD. LAST THREE ROWS SHOW THE AVERAGE PERFORMANCE OF ALL MODELS USING THREE SCORES

	Buildings	Nature	Animals	Flowers	Automan
Size	16	40	12	20	12
IO	0.62 ± 0.03	0.58 ± 0.04	0.65 ± 0.04	0.62 ± 0.04	0.70 ± 0.03
Gauss	0.50 ± 0.04	0.50 ± 0.04	0.50 ± 0.07	0.50 ± 0.07	0.50 ± 0.07
AIM	0.58 ± 0.02	0.55 ± 0.05	0.58 ± 0.05	0.58 ± 0.06	0.63 ± 0.05
AWS	**0.60 ± 0.04**	**0.58 ± 0.06**	**0.63 ± 0.07**	**0.62 ± 0.06**	**0.68 ± 0.05**
E-Saliency	0.56 ± 0.04	0.53 ± 0.05	0.57 ± 0.06	0.54 ± 0.07	0.63 ± 0.06
Bian	0.52 ± 0.07	0.55 ± 0.05	**0.60 ± 0.08**	0.56 ± 0.08	0.61 ± 0.09
Entropy	0.54 ± 0.04	0.52 ± 0.03	0.51 ± 0.05	0.56 ± 0.04	0.57 ± 0.04
GBVS	0.56 ± 0.03	0.55 ± 0.05	0.57 ± 0.04	0.55 ± 0.06	0.60 ± 0.07
Kootstra	0.56 ± 0.03	0.53 ± 0.04	0.54 ± 0.06	0.54 ± 0.07	0.58 ± 0.05
HouCVPR	0.58 ± 0.03	0.54 ± 0.05	0.59 ± 0.05	0.55 ± 0.06	0.62 ± 0.05
HouNIPS	**0.58 ± 0.03**	**0.56 ± 0.05**	0.59 ± 0.07	**0.59 ± 0.06**	**0.66 ± 0.07**
Itti-CIO	0.52 ± 0.02	0.52 ± 0.03	0.54 ± 0.03	0.51 ± 0.03	0.54 ± 0.02
Itti-CIO2	0.55 ± 0.04	0.55 ± 0.03	0.58 ± 0.05	0.54 ± 0.04	0.64 ± 0.03
Jia Li	0.56 ± 0.04	0.53 ± 0.04	0.57 ± 0.06	0.52 ± 0.08	0.60 ± 0.05
Judd	0.57 ± 0.04	**0.56 ± 0.05**	0.58 ± 0.06	0.58 ± 0.06	0.63 ± 0.05
Le Meur	0.55 ± 0.05	0.55 ± 0.05	0.55 ± 0.05	0.55 ± 0.05	0.62 ± 0.07
Marat	0.51 ± 0.02	0.50 ± 0.02	0.51 ± 0.02	0.51 ± 0.02	0.51 ± 0.01
PQFT	0.53 ± 0.06	0.53 ± 0.05	0.52 ± 0.06	0.58 ± 0.05	0.58 ± 0.05
Rarity-G	0.53 ± 0.03	0.53 ± 0.03	0.55 ± 0.02	0.56 ± 0.04	0.57 ± 0.04
Rarity-L	0.54 ± 0.02	0.53 ± 0.03	0.54 ± 0.04	0.53 ± 0.04	0.57 ± 0.05
SDSR	**0.58 ± 0.04**	0.56 ± 0.05	**0.62 ± 0.06**	0.55 ± 0.06	**0.65 ± 0.07**
SUN	0.53 ± 0.06	0.53 ± 0.05	0.50 ± 0.06	**0.58 ± 0.05**	0.59 ± 0.07
Surprise-CIO	0.53 ± 0.03	0.54 ± 0.04	0.55 ± 0.03	0.53 ± 0.05	0.55 ± 0.02
Torralba	0.56 ± 0.03	0.54 ± 0.04	0.55 ± 0.05	0.58 ± 0.06	0.62 ± 0.05
Variance	0.54 ± 0.03	0.53 ± 0.04	0.52 ± 0.05	0.57 ± 0.06	0.59 ± 0.04
VOCUS	0.56 ± 0.03	0.54 ± 0.04	0.58 ± 0.05	0.56 ± 0.06	0.63 ± 0.06
STB	0.51 ± 0.01	0.51 ± 0.01	0.53 ± 0.04	0.51 ± 0.02	0.51 ± 0.01
Yan	0.57 ± 0.03	0.55 ± 0.06	0.60 ± 0.05	0.56 ± 0.06	0.65 ± 0.06
Yin Li	0.55 ± 0.03	0.55 ± 0.05	0.59 ± 0.06	0.57 ± 0.06	0.60 ± 0.07
Average-AUC	0.55 ± 0.02	0.54 ± 0.01	0.56 ± 0.3	0.55 ± 0.02	0.60 ± 0.04
Average-CC	0.17 ± 0.06	0.17 ± 0.07	0.22 ± 0.84	0.19± 0.82	0.24 ± 0.07
Average-NSS	0.33 ± 0.12	0.30 ± 0.13	0.57 ± 0.2	0.46 ± 0.20	0.54 ± 0.18

Murray, Judd, AIM, and HouNIPS scored best. The *sport_scramblers_1280x720* video was the easiest on average for models over three scores because it has mainly one moving object. Models that performed poorly over the CRCNS-ORIG dataset are also ranked at the bottom on DIEM dataset. Several videos clips in this dataset yield very poor model scores for all models. Here again, those clips include significant cognitive factors; for example, in the ping-pong videos, a reactive saliency model often trails behind human fixations which tend to be more predictive [78]. Adding stronger predictive abilities to models is a very hard problem as the predictions occur in the 3D world, thus requiring extensive machine vision to recover 3D structure from videos.

C. Analysis of Gaussian Blob Size

Another important factor in model comparison is the size of the Gaussian blob. We changed the sigma (σ) parameter of the Gauss model and evaluated the scores over three datasets shown in Fig. 9. Two points should be noticed here: 1) Using all three scores, maximum performance happens for the Gaussian σ equal to 6, 7, or 8. In our experiments, Gaussian $\sigma = 10$ was used for model comparison and 2) Over shuffled AUC, as it was expected, values do not change for different Gaussians over three datasets (between 0.5 and 0.512). This again shows that shuffled AUC is invariant to center-bias.

D. Time Complexity Analysis of Models

In addition to correctly predicting atypical image locations attracting human attention, a saliency model should be also very fast. For some species, attention is tightly linked to their survival (e.g., quick detection and response to a predator). Some complex processes such as cluttered scene understanding will not be feasible or will be very slow without employing an effective attentional strategy. Thus, it is important that attention should kicky orient other complex processes to important dimensions of stimuli. The average time required to compute saliency map of an image for models is shown in Table III. Average time was calculated over 100 images with resolution 511×681 from Bruce and Tsotsos dataset. All models were executed on a computer running Linux Mandriva with 4GB RAM and Quad core 2.8 GHz Intel CPU. The Itti-CIO model is the fastest ($\sim$17 ms/image) followed by VOCUS and HouCVPR models (less than 300 ms/image). Note that in this table, what matters is ranking, while absolute durations may be reduced with more powerful machines. The Judd *et al.*, model has high saliency prediction accuracy but is very slow (about 100 sec/image) since it needs to calculate several fairly complex channels (person, face, car, gist, horizontal line, etc). Most of the models need less than 16 sec to calculate saliency.

E. Illustrative Figures

The three best and three worst stimuli (measured by shuffled AUC score) for each model are shown in Fig. 10. Many models share their three best and three worst images. For the Gauss model stimuli that have fixations at the center happen to be the best and those that have fixations off the center are the worst. Since no model uses face detection (except for Judd *et al.*) and text detection channels, most models have difficulty predicting fixations over stimuli with these types of features. This means that an important point in building more successful models is to look for cognitive factors that drive visual attention (e.g., gaze direction of human characters in images, meaning of text messages [69], etc.).

IV. Discussion and Conclusion

In this paper, we briefly reviewed several state-of-the-art visual saliency models and quantitatively compared them over 54 synthetic patterns, three radically different still image datasets, and two benchmark video datasets. We also analyzed datasets in terms of center-bias and models in terms of time complexity. Here, we list the main conclusions of our comparison study:

1) All existing datasets are highly center-biased. Developing less center-biased datasets in the future can help fair model comparison[4].
2) The majority of existing eye movements datasets are small with small numbers of subjects. Further attempts are necessary to collect larger datasets with more observers (to obtain a better notion of average human performance) with higher stimulus variability. This need is more pressing for collecting fixations over videos.
3) The CC and NSS scores suffer from the center-bias issue and their use in future model comparisons is not encouraged. On the other hand, the shuffled AUC score tackles center bias and border effects, and is the best option for model comparison.
4) There is still a gap between current models and human performance. This gap is smaller for off-center fixations and for some datasets, but overall exists. As discussed above, discovering and adding more top-down features to models will hopefully boost their performance.
5) Saliency models based on FIT theory work better in locating a target over synthetic patterns.
6) Models that did well over static natural scenes in general also did well over the video datasets. The majority of these models are based on statistical techniques.
7) The top performing model in our experiments with static and dynamic natural scenes is AWS (focusing on the shuffled AUC score); it also performed second best with synthetic images.
8) Consistent with [79], we also noticed that models that generate blurrier maps achieve higher scores (e.g., GBVS, AIM, and Itti-CIO2). This should be considered by authors and future comparisons.
9) Models incorporating motion did not perform better than the best static models over video datasets. Extension of the best existing static models to the spatio-temporal domain may further scale up those models.
10) Some categories are harder for models (e.g., Nature stimuli) while some others containing less cluttered scenes and scenes with fewer objects are easier (e.g., scenes containing humans and cars).
11) Best and worst stimuli are the same for many models, which means that models have common difficulty in prediction of saliency over specific classes of stimuli (Fig. 10). This suggests some hints for future research.
12) Some models are fast and effective (e.g., HouNIPS, Bian, HouCVPR, Torralba, and Itti-CIO2) providing a tradeoff between accuracy and speed necessary for many applications.

[4]We share a dataset at: https://sites.google.com/site/saliencyevaluation/.

TABLE III

Average Saliency Computation Time (Sorted) for Models in Seconds for a 511 × 681 Image. The Two Fastest Models Are Written in C++ Code. HouCVPR Is in Matlab

Model	Judd	Yin Li	SUN	AIM	Yan	AWS	GBVS	Rarity-L	Rarity-G	SDSR	STB	Torralba	PQFT	HouNIPS	Bian	HouCVPR	VOCUS	Itti-CIO
Time	98.58	55.07	51.92	15.6	13.05	12.08	10.14	4.14	3.6	2.35	2.29	2.24	1.7	1.14	1.1	0.30	0.025	**0.017**

One remaining problem in fair model evaluation is the effect of internal model parameters, such as number of filters, type of filters, Gabor or DoG filter parameters, choice of the nonlinearities, blurring, and normalization schemes. Proper tuning of these parameters is important, and doing so may dramatically affect the performance of a system. Here, we tried our best to produce highly predictive maps for models.

Despite significant success of the models evaluated here, there is still significant room to further improve attention accuracy due to a remaining large gap between models and human observer agreement, as has also been shown using smaller datasets and less systematic comparisons in previous studies (e.g., [41], [50], [45]). Here, we suggest several directions that could help bridge this gap.

One direction to extend current models is adding top-down factors. Context [3], [35], [45], gain modulation of features for target detection [4], [5], [80], and use of target detectors tuned to specific objects [50], [69] has been used for modeling top-down attentional effects. Here, we comment more on these factors. While almost all bottom-up models have employed simple feature channels believed to be computed by early visual areas, they do not rule out the existence of top-down influences in free-viewing tasks where these models have been applied to. For instance, in free viewing of spatio-temporal stimuli, such as videos, semantic processing of scenes, and extraction of high-level knowledge plays a significant role in guiding attention and eye movements. Some semantic cues involve social interactions in images, living beings, faces (and eyes, nose, and mouth within faces), text, etc. Also it appears that attentional-bias is independent of illumination, orientation as well as scale of the salient object/concept [81]. A large dataset containing many example images with such factors (758 images viewed by on average 25.3 viewers) has recently

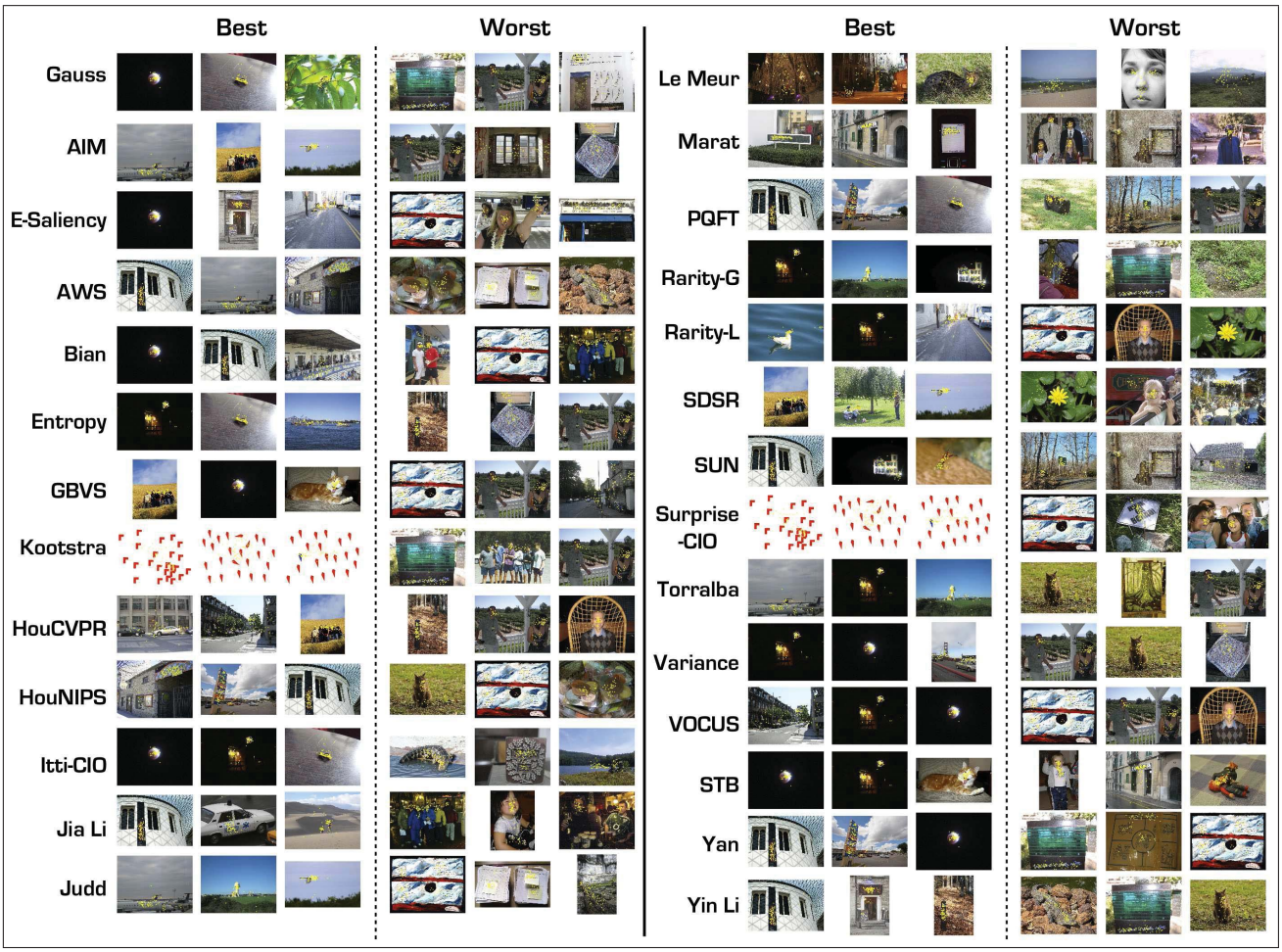

Fig. 10. Three best and three worst stimuli using shuffled AUC score for all models over Judd *et al.* dataset. Note that some images are best for many models and at the same time some worst cases repeat across many models. Yellow dots represent human fixations.

been collected by Ramanathan *et al.* [82]. They also observed that unpleasant concepts, such as reptiles or blood and injury, considerably influence visual attention whenever present. The fact that recognized concepts drive visual attention adds support to the theory that visual attention and object recognition are concurrent processes, and this is an interesting topic of research in the cognitive science community. Therefore, adding top-down factors to bottom-up models can be an important topic for future research in saliency modeling. Indeed the list of top-down factors is not limited to the above factors and several others, including task demands (e.g., real-world tasks), memories, experiences, expectations, and internal states play an important roles in directing overt attention and gaze.

We suggest taking inspiration from early visual cortex for developing more biologically inspired models of attention. For instance, the AWS model takes advantage of a basic idea, decorrelation of neural responses in representing a visual stimuli [83], [84]. In this regard, shown by our results, having many features (similar to Judd [50]) might not be as efficient as discovering the basic principles/features of attention guidance (employed by models, such as Itti-CIO, AWS, HouNIPS, and GBVS). An idea in this direction is validating models of saliency against eye movements of humans over distorted images (e.g., rotated, mirrored, or inversed images) or by considering detailed low-level neural findings revealed by neurophysiological studies (e.g., [85]).

Another future direction will be combining several different saliency models to achieve higher performances. Since each of these models is based on different mechanisms, it is likely that combining them may result in higher fixation prediction. This trend has been followed previously in biometrics (e.g., face identifications) as well as character recognition [86]. Such direction may not extend our understanding of visual attention, but if successful it may have several practical applications.

There are several other open questions for future investigations. As already mentioned, text is an important feature that is proven to attract attention [69]. But since text detection in natural scenes is an open problem and few approaches exist for that, it has not been added to current models. Basically using more features leads to better fixation prediction performance with the cost of lowering speed. One solution is parallel implementation of models (e.g., feature extraction on GPU (e.g., [87], [88]). Most models have focused on predicting locations that human observers look at, while few (e.g., [34]) have investigated other aspects of eye fixations, such as saccade dynamics, sequencing (Wang *et al.* [89]), retinal sampling, inhibition of return, the role of context, etc [90]. More work needs to be done in this direction.

A less explored application of saliency modeling is using it for understanding cluttered scenes by sequentially processing important regions or objects. Another promising direction is in developing models that can predict locations that humans find interesting, for instance by clicking and see how such models differ from traditional saliency models for fixation prediction [91]. Also, more attempts are still needed to determine important features attracting eye fixations. Extending models to include some understanding of 3D scene structure is a challenging yet pressing problem, as solving it may allow the creation of new models with significantly better predictive abilities (e.g., the expected landing point of a ball might be more salient than the ball itself). It would be also interesting to customize a saliency model for each person. For instance, by learning habits, preferences, etc. of each human subject. This way, it is theoretically possible to surpass the human inter-observer model. Interaction between attention and object recognition and their mutual benefit has been overstated, but still there are not many works in this area.

REFERENCES

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

[2] A. Salah, E. Alpaydin, and L. Akrun, "A selective attention-based method for visual pattern recognition with application to handwritten digit recognition and face recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 3, pp. 420–425, Mar. 2002.

[3] C. Kanan and G. Cottrell, "Robust classification of objects, faces, and flowers using national image," in *Proc. Comput. Vis. Pattern Recognit.*, 2010, pp. 1–8.

[4] S. Frintrop, "VOCUS: A visual attention system for object detection and goal-directed search," Ph.D. thesis, Dept. Comput. Sci., Univ. Bonn, Bonn, Germany, 2006.

[5] V. Navalpakkam and L. Itti, "An integrated model of top-down and bottom-up attention for optimizing detection speed," in *Proc. Conf. Comput. Vis. Pattern Recognit.*, 2006, pp. 2049–2056.

[6] S. Frintrop, "General object tracking with a component-based target descriptor," in *Proc. Int. Conf. Robot. Autom.*, 2010, pp. 4531–4536.

[7] 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.

[8] Y. Ma, X. Hua, L. Lu, and H. 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.

[9] D. DeCarlo and A. Santella, "Stylization and abstraction of photographs," *ACM Trans. Graph.*, vol. 21, no. 3, pp. 769–776, 2002.

[10] L. Marchesotti, C. Cifarelli, and G. Csurka, "A framework for visual saliency detection with applications to image thumbnailing," in *Proc. Int. Conf. Comput. Vis.*, 2009, pp. 2232–2239.

[11] J. Wang, J. Sun, L. Quan, X. Tang, and H. Y. Shum, "Picture collage," in *Proc. Comput. Vis. Pattern Recognit.*, vol. 1. 2006, pp. 347–354.

[12] M. S. El-Nasr, T. Vasilakos, C. Rao, and J. Zupko, "Dynamic intelligent lighting for directing visual attention in interactive 3-D scenes," *IEEE Trans. Comput. Intell. AI Games*, vol. 1, no. 2, pp. 145–153, Jun. 2009.

[13] B. Mertsching, M. Bollmann, R. Hoischen, and S. Schmalz, "The neural active vision system," in *Handbook of Computer Vision and Applications*. New York: Academic, 1999, pp. 543–568.

[14] A. Borji, M. N. Ahmadabadi, B. N. Araabi, and M. Hamidi, "Online learning of task-driven object-based visual attention control," *Image Vis. Comput.*, vol. 28, no. 7, pp. 1130–1145, 2010.

[15] C. Siagian and L. Itti, "Biologically inspired mobile robot vision localization," *IEEE Trans. Robot.*, vol. 25, no. 4, pp. 861–873, Aug. 2009.

[16] C. Breazeal, "A context-dependent attention system for a social robot," in *Proc. Int. Joint Conf. Artif. Intell.*, 1999, pp. 1146–1151.

[17] R. Rosenholtz, A. Dorai, and R. Freeman, "Do predictions of visual perception aid design?" *ACM Trans. Appl. Percept.*, vol. 8, no. 2, pp. 1–27, 2011.

[18] M. Ajallooeian, A. Borji, B. N. Araabi, M. N. Ahmadabadi, and H. Moradi, "Fast hand gesture recognition based on saliency maps: An application to interactive robotic marionette playing," in *Proc. IEEE Conf. Robot Human Interact. Commun.*, Oct. 2009, pp. 841–847.

[19] N. Parikh, L. Itti, and J. Weiland, "Saliency-based image processing for retinal prostheses," *J. Neural Eng.*, vol. 7, no. 1, pp. 016006-1–016006-10, 2010.

[20] *The PASCAL Visual Object Classes* (2005). [Online]. Available: http://pascallin.ecs.soton.ac.uk/challenges/VOC/

[21] L. Itti and C. Koch, "Computational modelling of visual attention," *Nature Rev. Neurosci.*, vol. 2, pp. 194–203, Mar. 2001.

[22] D. Heinke and G. W. Humphreys, "Computational models of visual selective attention: A review," in *Connectionist Models in Psychology*, G. Houghton, Ed. Florence, KY: Psychology Press, 2004, pp. 273–312.

[23] S. Frintrop, E. Rome, and H. I. Christensen, "Computational visual attention systems and their cognitive foundations: A survey," *ACM Trans. Appl. Percept.*, vol. 7, no. 1, pp. 1–46, 2010.

[24] A. Borji and L. Itti, "State-of-the-art in visual attention modeling," *IEEE Trans. Pattern Anal. Mach. Intell.*, to be published.

[25] D. Ballard, M. Hayhoe, and J. Pelz, "Memory representations in natural tasks," *J. Cognit. Neurosci.*, vol. 7, no. 1, pp. 66–80, 1995.

[26] M. F. Land and D. N. Lee, "Where we look when we steer," *Nature*, vol. 369, pp. 742–744, Jun. 1994.

[27] R. J. Peters and L. Itti, "Beyond bottom-up: Incorporating task-dependent influences into a computational model of spatial attention," in *Proc. Comput. Vis. Pattern Recognit.*, 2007, pp. 1–8.

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

[29] J. M. Henderson and A. Hollingworth, "High-level scene perception," *Annu. Rev. Psychol.*, vol. 50, pp. 243–271, Feb. 1999.

[30] W. Einhauser, M. Spain, and P. Perona, "Objects predict fixations better than early saliency," *J. Vis.*, vol. 8, no. 14, pp. 1–26, 2008.

[31] 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.

[32] *iLab Neuromorphic Vision C++ Toolkit (iNVT)*. (2010, Nov. 17) [Online]. Available: http://ilab.usc.edu/toolkit/

[33] L. Itti and C. Koch, "A saliency-based search mechanism for overt and covert shifts of visual attention," *Vis. Res.*, vol. 40, nos. 10–12, pp. 1489–1506, 2000.

[34] L. Itti, N. Dhavale, and F. Pighin, "Realistic avatar eye and head animation using a neurobiological model of visual attention," *Proc. SPIE*, vol. 5200, pp. 64–78, Dec. 2003.

[35] A. Torralba, "Modeling global scene factors in attention," *J. Opt. Soc. Amer.*, vol. 20, no. 7, pp. 1407–1418, 2003.

[36] L. Itti and P. Baldi, "Bayesian surprise attracts human attention," in *Advances in Neural Information Processing Systems*. Cambridge, MA: MIT Press, 2006.

[37] N. D. B. Bruce and J. K. Tsotsos, "Saliency based on information maximization," in *Advances in Neural Information Processing Systems*. New York: Academic, 2005.

[38] O. Le Meur, P. Le Callet, D. Barba, and D. Thoreau, "A coherent computational approach to model bottom-up visual attention," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 28, no. 5, pp. 802–817, May 2006.

[39] O. Le Meur, P. Le Callet, and D. Barba, "Predicting visual fixations on video based on low-level visual features," *Vis. Res.*, vol. 47, no. 19, pp. 2483–2493, 2007.

[40] J. Harel, C. Koch, and P. Perona, "Graph-based visual saliency," in *Advances in Neural Information Processing Systems*, vol. 19. Cambridge, MA: MIT Press, 2006, pp. 545–552.

[41] X. Hou and L. Zhang, "Saliency detection: A spectral residual approach," in *Proc. Comput. Vis. Pattern Recognit.*, 2007, pp. 1–8.

[42] M. Mancas, "Computational attention: Modelisation and application to audio and image processing," Ph.D. thesis, Faculté Polytechnique de Mons, Arrondissement of Mons, Belgium, 2007.

[43] X. Hou and L. Zhang, "Dynamic visual attention: Searching for coding length increments," in *Advances in Neural Information Processing Systems*. Cambridge, MA: MIT Press, 2008.

[44] G. Kootstra and L. R. B. Schomaker, "Prediction of human eye fixations using symmetry," in *Proc. 31st Annu. Conf. Cognit. Sci. Soc.*, 2009.

[45] 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, pp. 1–20, 2008.

[46] 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," *Int. J. Comput. Vis.*, vol. 82, no. 3, pp. 231–243, 2009.

[47] 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.

[48] Y. Li, Y. Zhou, J. Yan, and J. Yang, "Visual saliency based on conditional entropy," in *Proc. Asian Conf. Comput. Vis.*, 2009, pp. 246–257.

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

[50] T. Judd, K. Ehinger, F. Durand, and A. Torralba, "Learning to predict where humans look," in *Proc. Int. Conf. Comput. Vis.*, 2009, pp. 1–7.

[51] P. Bian and L. Zhang, "Biological plausibility of spectral domain approach for spatiotemporal visual saliency," in *Advances in Neuro-Information Processing* (Lecture Notes in Computer Science). New York: Springer-Verlag, 2009.

[52] T. Avraham and M. Lindenbaum, "Esaliency (extended saliency): Meaningful attention using stochastic image modeling," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 4, pp. 693–708, Apr. 2010.

[53] J. Yan, J. Liu, Y. Li, and Y. Liu, "Visual saliency detection via rank-sparsity decomposition," in *Proc. Int. Conf. Image Process.*, 2010, pp. 1089–1092.

[54] A. Garcia-Diaz, X. R. Fdez-Vidal, X. M. Pardo, and R. Dosil, "Saliency from hierarchical adaptation through decorrelation and variance normalization," *Image Vis. Comput.*, vol. 30, no. 1, pp. 51–64, 2012.

[55] 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, 2010.

[56] 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.

[57] N. Murray, M. Vanrell, X. Otazu, and C. A. Parraga, "Saliency estimation using a non-parametric low-level vision model," in *Proc. Comput. Vis. Pattern Recognit.*, Jun. 2011, pp. 433–440.

[58] D. Gao, V. Mahadevan, and N. Vasconcelos, "The discriminant center-surround hypothesis for bottom-up saliency," in *Proc. Neural Inf. Process. Syst.*, 2007, pp. 1–8.

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

[60] J. M. Wolfe and T. S. Horowitz, "What attributes guide the deployment of visual attention and how do they do it?" *Nature Rev. Neurosci.*, vol. 5, pp. 1–7, Jun. 2004.

[61] Q. Zhao and C. Koch, "Learning a saliency map using fixated locations in natural scenes," *J. Vis.*, vol. 11, no. 3, pp. 1–15, 2011.

[62] *CRCNS - Collaborative Research in Computational Neuroscience - Data Sharing* (2005). [Online]. Available: http://crcns.org/data-sets/eye/eye-1

[63] *DIEM Dataset*. [Online]. Available: http://thediemproject.wordpress.com/

[64] T. Jost, N. Ouerhani, R. von Wartburg, R. Mäuri, and H. Häugli, "Assessing the contribution of color in visual attention," *Comput. Vis. Image Understand.*, vol. 100, nos. 1–2, pp. 107–123, 2005.

[65] U. Rajashekar, A. C. Bovik, and L. K. Cormack, "Visual search in noise: Revealing the influence of structural cues by gaze-contingent classification image analysis," *J. Vis.*, vol. 6, no. 4, pp. 379–386, 2006.

[66] R. Peters, A. Iyer, L. Itti, and C. Koch, "Components of bottom-up gaze allocation in natural images," *Vis. Res.*, vol. 45, no. 18, pp. 2397–2416, 2005.

[67] 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.

[68] D. Green and J. Swets, *Signal Detection Theory and Psychophysics*. New York: Wiley, 1966.

[69] M. Cerf, P. Frady, and C. Koch, "Faces and text attract gaze independent of the task: Experimental data and computer model," *J. Vis.*, vol. 9, no. 12, pp. 1–15, 2009.

[70] L. Elazary and L. Itti, "A Bayesian model for efficient visual search and recognition," *Vis. Res.*, vol. 50, no. 14, pp. 1338–1352, 2010.

[71] B. W. Tatler, "The central fixation bias in scene viewing: Selecting an optimal viewing position independently of motor biases and image feature distributions," *J. Vis.*, vol. 14, no. 7, pp. 1–17, 2007.

[72] P. Tseng, R. Carmi, I. G. M. Cameron, D. P. Munoz, and L. Itti, "Quantifying center bias of observers in free viewing of dynamic natural scenes," *J. Vis.*, vol. 9, no. 7, pp. 1–16, 2009.

[73] S. A. Brandt and L. W. Stark, "Spontaneous eye movements during visual imagery reflect the content of the visual scene," *J. Cognit. Neurosci.*, vol. 9, no. 1, pp. 27–38, 1997.

[74] C. M. Privitera and L. W. Stark, "Algorithms for defining visual regions-of-interest: Comparison with eye fixations," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 22, no. 9, pp. 970–982, Sep. 2000.

[75] 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.

[76] H. C. Nothdurft, "Salience of feature contrast," in *Neurobiology of Attention*, L. Itti, G. Rees, and J. K. Tsotsos, Eds. San Diego, CA: Elsevier, 2005.

[77] S. Han and N. Vasconcelos, "Biologically plausible saliency mechanisms improve feedforward object recognition," *Vis. Res.*, vol. 50, no. 22, pp. 2295–2307, 2010.

[78] N. Mennie, M. Hayhoe, and B. Sullivan, "Look-ahead fixations: Anticipatory eye movements in natural tasks," *Experim. Brain Res.*, vol. 179, no. 3, pp. 427–442, 2007.

[79] T. Judd, "Understanding and predicting where people look," Ph.D. thesis, Dept. Electr. Eng. Comput. Sci., MIT, Cambrige, 2011.

[80] 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.

[81] T. Judd, F. Durand, and A. Torralba, "Fixations on low-resolution images," *J. Vis.*, vol. 11, no. 4, pp. 1–14, 2011.

[82] S. Ramanathan, H. Katti, N. Sebe, M. Kankanhalli, and T. S. Chua, "An eye fixation database for saliency detection in images," in *Proc. Eur. Conf. Comput. Vis.*, 2010, pp. 30–43.

[83] B. A. Olshausen and D. J. Field, "How close are we to understanding V1?" *Neural Comput.*, vol. 17, no. 8, pp. 1665–1699, 2005.

[84] E. Simoncelli and B. Olshausen, "Natural image statistics and neural representation," *Annu. Rev. Neurosci.*, vol. 24, pp. 1193–1216, Mar. 2001.

[85] Z. Li, "A saliency map in primary visual cortex," *Trends Cognit. Sci.*, vol. 6, no. 1, pp. 9–16, 2002.

[86] J. Kittler, M. Hatef, R. P. W. Duin, and J. Matas, "On combining classifiers," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 20, no. 3, pp. 226–239, Mar. 1998.

[87] B. Han and B. Zhou, "High speed visual saliency computation on GPU," in *Proc. Int. Conf. Image Process.*, 2007, pp. 361–364.

[88] T. Xu, T. Pototschnig, K. Kühnlenz, and M. Buss, "A high-speed multi-GPU implementation of bottom-up attention using CUDA," in *Proc. Int. Conf. Robot. Auotm.*, 2009, pp. 41–47.

[89] W. Wang, C. Chen, Y. Wang, T. Jiang, F. Fang, and Y. Yao, "Simulating human saccadic scanpaths on natural images," in *Proc. Comput. Vis. Pattern Recognit.*, Jun. 2011, pp. 441–448.

[90] B. W. Tatler, M. M. Hayhoe, M. F. Land, and D. H. Ballard, "Eye guidance in natural vision: Reinterpreting salience," *J. Vis.*, vol. 11, no. 5, pp. 1–23, May 2011.

[91] C. M. Masciocchi, S. Mihalas, D. Parkhurst, and E. Niebur, "Everyone knows what is interesting: Salient locations which should be fixated," *J. Vis.*, vol. 9, no. 11, pp. 1–22, 2009.

[92] C. Koch and S. Ullman, "Shifts in selective visual attention: Toward the underlying neural circuitry," *Human Neurobiol.*, vol. 4, no. 4, pp. 219–227, 1985.

[93] D. Reisfeld, H. Wolfson, and Y. Yeshurun, "Context-free attentional operators: The generalized symmetry transform," in *Proc. Int. Joint Conf. Vis.*, vol. 14. 1995, pp. 119–130.

[94] A. Oliva, A. Torralba, M. S. Castelhano, and J. M. Henderson, "Top-down control of visual attention in object detection," in *Proc. Int. Conf. Image Proces.*, 2003, pp. 1–4.

[95] L. Zhang, M. H. Tong, and G.W. Cottrell, "SUNDAy: Saliency using natural statistics for dynamic analysis of scenes," in *Proc. 31st Annu. Cognit. Sci. Soc.*, 2009, pp. 2944–2949.

[96] W. Kienzle, M. O. Franz, B. Schölkopf, and F. A. Wichmann, "Center-surround patterns emerge as optimal predictors for human saccade targets," *J. Vis.*, vol. 9, no. 5, pp. 1–15, 2009.

[97] G. Heidemann, "Focus-of-attention from local color symmetries," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 26, no. 7, pp. 817–830, 2004.

[98] B. C. Russell, A. Torralba, K. P. Murphy, and W. T. Freeman, "LabelMe: A database and web-based tool for image annotation," *Int. J. Comput. Vis.*, vol. 77, nos. 1–3, pp. 157–173, 2008.

[99] J. Yuen, B. C. Russell, C. Liu, and A. Torralba, "LabelMe video: Building a video database with human annotations," in *Proc. Int. Conf. Comput. Vis.*, 2009, pp. 1–8.

[100] P. Reinagel and A. Zador, "Natural scene statistics at the centre of gaze," *Network*, vol. 10, no. 4, pp. 341–350, 1999.

[101] D. J. Simons and C. F. Chabris, "Gorillas in our midst: Sustained inattentional blindness for dynamic events," *Perception*, vol. 28, no. 9, pp. 1059–1074, 1999.