SID4VAM: A Benchmark Dataset with Synthetic Images for Visual Attention Modeling

David Berga[1] Xosé R. Fdez-Vidal[2] Xavier Otazu[1] Xosé M. Pardo[2]

[1]Computer Vision Center, Universitat Autònoma de Barcelona, Spain
[2]CiTIUS, Universidade de Santiago de Compostela, Spain

{dberga,xotazu}@cvc.uab.es {xose.vidal,xose.pardo}@usc.es

Abstract

A benchmark of saliency models performance with a synthetic image dataset is provided. Model performance is evaluated through saliency metrics as well as the influence of model inspiration and consistency with human psychophysics. SID4VAM is composed of 230 synthetic images, with known salient regions. Images were generated with 15 distinct types of low-level features (e.g. orientation, brightness, color, size...) with a target-distractor pop-out type of synthetic patterns. We have used Free-Viewing and Visual Search task instructions and 7 feature contrasts for each feature category. Our study reveals that state-of-the-art Deep Learning saliency models do not perform well with synthetic pattern images, instead, models with Spectral/Fourier inspiration outperform others in saliency metrics and are more consistent with human psychophysical experimentation. This study proposes a new way to evaluate saliency models in the forthcoming literature, accounting for synthetic images with uniquely low-level feature contexts, distinct from previous eye tracking image datasets.

1. Introduction

Although eye movements are indicators of *"where people look at"*, a more complex question arises as a consequence for understanding bottom-up visual attention: *Are all eye movements equally valuable for determining saliency?* According to the initial hypotheses in visual attention [53, 58], we could define visual saliency as the perceptual quality that makes our human visual system (HVS) to gaze towards certain areas that pop-out on a scene due to their distinctive visual characteristics. Therefore, this capacity (saliency) cannot be influenced by top-down factors, which seemingly guide eye movements regardless of stimulus characteristics [60]. Accounting for prior knowledge of whether a stimulus area is salient or not, when it becomes salient, and why, are issues that need to be accounted for saliency evaluation [7, 2].

Common frameworks for predicting saliency have been created since Koch & Ullman's seminal work [31]. This framework defined a theoretical basis for modeling the early visual stages of the HVS in order to obtain a representation of the saliency map. By extracting sensory signals as feature maps, processing the conspicuous objects and selecting the maximally-active locations through winner-take-all (WTA) mechanisms, it is possible to obtain a unique/master saliency map. However, it was hypothesized that visual attention combines both bottom-up (saliency) and top-down (relevance) mechanisms in a central representation (priority) [14, 17]. These top-down specificities (e.g. world, object, task, etc.) were later accounted in the selective tuning model as a hierarchy of WTA-like processes [54]. Despite the neural correlates simultaneously involved in saliency have been investigated [55], the direct relation between saliency and eye movements defined in a unique computational framework requires further study. Itti et al. initially introduced a computational biologically-inspired model [27] composed of 3 main steps: First, feature maps are extracted using oriented linear DoG filters for each chromatic channel. Second, feature conspicuity is computed using center-surround differences. Third, conspicuity maps are integrated with linear WTA mechanisms. This architecture has been the main inspiration for current saliency models [62, 43], that alternatively use distinct mechanisms (accounting for different levels of processing, context or tuning depending on the scene) but preserving same or similar structure for these steps. Although current state-of-the-art models precisely resemble eye-tracking fixation data [6, 9], we question if these models represent saliency. We will test this hypothesis with a novel synthetic image dataset.

1.1. Related Work

In order to determine whether an object or a feature attracts attention, initial experimentation was assessing feature discriminability upon display characteristics (e.g. dis-

play size, feature contrast...) during visual search tasks [53, 58]. Parallel search occurs when features are processed preattentively, therefore search targets are found efficiently regardless of distractor properties. Instead, serial search happens when attention is directed to one item at a time, requiring a "binding" process to allow each object to be discriminated. For this case, search time decrease with higher target-distractor contrast and/or lower set size (following the Weber Law [16]). More recent studies replicated these experiments by providing real images with parametrization of feature contrast and/or set size (iLab USC, UCL, VAL Hardvard, ADA KCL), combining visual search or visual segmentation tasks, however not providing eye tracking data (Table 1B). Rather, current eye movement datasets provide fixations and scanpaths from real scenes during free-viewing tasks. These image datasets are usually composed of real image scenes (Table 1A), either from indoor / outdoor scenes (Toronto, MIT1003, MIT300), nature scenes (KTH) or semantically-specific categories such as faces (NUSEF) and several others (CAT2000). A complete list of eye tracking datasets is in Winkler & Subramanian's overview [57]. CAT2000 training subset of "Pattern" images (CAT2000$_p$) provides eye movement data with psychophysical / synthetic image patterns during 5 sec of free-viewing. However, no parametrization of feature contrast nor stimulus properties is given. A synthetic image dataset could provide information of how attention is dependent on feature contrast and other stimulus properties with distinct tasks. We describe in Section 2 how we do so with our novel SID4VAM's dataset.

Table 1: Characteristics of eye tracking datasets

A: Real Images

Dataset	Task	# TS	# PP	PM	DO
Toronto [8]	FV	120	20		✓
MIT1003 [30]	FV	1003	15		✓
NUSEF [41]	FV	758	25		✓
KTH [32]	FV	99	31		✓
MIT300 [29]	FV	300	39		✓
CAT2000 [5]	FV	4000	24		✓

B: Psychophysical Pattern / Synthetic Images

Dataset	Task	# TS	# PP	PM	DO
iLab USC [26]	-	~540	-	✓	
UCL [64]	VS & SG	2784	5	✓	
VAL Harvard [59]	VS	4000	30	✓	
ADA KCL [50]	-	~430	-	✓	
CAT2000$_p$ [5]	FV	100	18		✓
SID4VAM (Ours)	FV & VS	230	34	✓	✓

TS: total number of stimuli, PP: participants, PM: Parametrization, DO: Fixation data is available online, FV: Free-Viewing, VS: Visual Search, SG: visual segmentation

Being inspired by Itti et al's architecture, a myiriad of computational models has been proposed with distinct computational approaches, from biological, mathematical and physical inspiration [62, 43]. By processing global and/or local image fatures for calculating feature conspicuity, these models are able to generate a master saliency map to predict human fixations (Table 2). Taking up Judd et al. [29] and Borji & Borji's [4] reviews, we have grouped saliency model inspiration in five general categories according to its saliency computation endeavour:

- Cognitive/Biological (C): Saliency is usually generated by mimicking HVS neuronal mechanisms or either specific patterns found in human eye movement behavior. Feature extraction is generally based on Gabor-like filters and its integration with WTA-like mechanisms.
- Information-Theoretic (I): These models compute saliency by selecting the regions that maximize visual information of scenes.
- Probabilistic (P): Probabilistic models generate saliency by optimizing the probability of performing certain tasks and/or finding certain patterns. These models use graphs, bayesian, decision-theoretic and other approaches for their computations.
- Spectral/Fourier-based (F): Spectral Analysis or Fourier-based models derive saliency by extracting or manipulating features in the frequency domain (e.g. spectral frequency or phase).
- Machine/Deep Learning (D): These techniques are based on training existing machine/deep learning architectures (e.g. CNN, RNN, GAN...) by minimizing the error of predicting fixations of images from existing eye tracking data or labeled salient regions.

1.2. Problem formulation

Visual saliency is a term coined on a perceptual basis. According to this principle, a correct modelization of saliency should consider specific experimental conditions upon a visual attention task. The output of such a model can vary for stimulus or task, but must arise as a common behavioral phenomena in order to validate the general hypothesis definition from Treisman, Wolfe, Itti and colleagues [53, 58, 26]. Eye movements have been considered the main behavioral markers of visual attention. But understanding saliency means not only to prove how visual fixations can be predicted, but to simulate which patterns of eye movements are gathered from vision and its sensory signals (here avoiding any top-down influences). This challenge offers eye tracking researchers to consider several experimental issues (with respect contextual, contrast, temporal, oculomotor and task-related biases) when capturing bottom-up attention, largely explained by Borji et al. [4], Bruce et al. [7] and lately by Berga et al. [2]. Computational models advance several ways to predict, to some extent, human visual

Table 2: Description of saliency models

Model	Authors	Year	Inspiration					Type	
			C	I	P	F	D	G	L
IKN	Itti et al.[27, 26]	1998	✓					✓	✓
AIM	Bruce & Tsotsos [8]	2005	✓	✓					✓
GBVS	Harel et al.[21]	2006			✓			✓	✓
SDLF	Torralba et al. [52]	2006			✓			✓	✓
SR & PFT	Hou & Zhang[23]	2007				✓		✓	
PQFT	Guo & Zhang[20]	2008				✓		✓	
ICL	Hou & Zhang [24]	2008		✓	✓			✓	✓
SUN	Zhang et al. [63]	2008			✓				✓
SDSR	Seo & Milanfar [48]	2009	✓		✓			✓	✓
FT	Achanta et al.[1]	2009				✓		✓	
DCTS/SIGS	Hou et al.[22]	2011				✓		✓	
SIM	Murray et al.[39]	2011	✓					✓	✓
WMAP	Lopez-Garcia et al.[38]	2011	✓			✓		✓	✓
AWS	Garcia-Diaz et al.[18]	2012	✓					✓	✓
CASD	Goferman et al.[19]	2012	✓	✓	✓		✓	✓	✓
RARE	Riche et al.[45]	2012		✓				✓	✓
QDCT	Schauerte et al.[47]	2012				✓		✓	
HFT	Li et al.[37]	2013				✓		✓	
BMS	Zhang & Sclaroff [61]	2013			✓			✓	
SALICON	Jiang et al.[28, 51]	2015					✓		✓
ML-Net	Cornia et al.[12]	2016					✓		✓
DeepGazeII	Kümmerer et al.[33]	2016					✓		✓
SalGAN	Pan et al.[40]	2017					✓		✓
ICF	Kümmerer et al.[33]	2017			✓			✓	✓
SAM	Cornia et al.[13]	2018					✓		✓
NSWAM	Berga & Otazu [3]	2018	✓					✓	✓
Sal-DCNN	Jiang et al. [34]	2019				✓	✓	✓	✓

Inspiration: {C: Cognitive/Biological, I: Information-Theoretic, P: Probabilistic, F: Fourier/Spectral, D: Machine/Deep Learning} Type: {G: Global, L: Local}

fixations. However, the limits of the prediction capability of these saliency models arise as a consequence of the validity of the evaluation from eye tracking experimentation. We aim to to provide a new dataset with uniquely synthetic images and a benchmark, studying for each saliency model:

1. How model inspiration and feature processing influences model predictions?

2. How does temporality of fixations affect model predictions?

3. How low-level feature type and contrast influences model's psychophysical measurements?

2. SID4VAM: Synthetic Image Dataset for Visual Attention Modeling

Fixations were collected from 34 participants in a dataset[1] of 230 images[2]. Images were displayed in a resolution of 1280 × 1024 px and fixations were captured at about 40 pixels per degree of visual angle using SMI RED binocular eye tracker. The dataset had been splitted in two tasks: Free-Viewing (FV) and Visual Search (VS). For the FV task, participants had to freely look at the image during 5 seconds. On each stimuli there was a salient area of interest (AOI). For the VS task, participants had the instruction to visually locate the AOI, setting the salient region as the different object. For this task, the trigger for prompting the transition to next image was by gazing inside the AOI or pressing a key (for reporting absence of target). We can observe the stimuli generated for both tasks on Figs. 1-2.

The dataset was divided in 15 stimulus types, 5 corresponding to FV and 10 to VS. Some of these blocks had distinct subsets of images (due to the alteration of either target or distractor shape, color, configuration and background properties), abling a total of 33 subtypes. Each of these blocks was individually generated as a low-level feature category, which had its own type of feature contrast between the salient region and the rest of distractors / background. FV categories were mainly based for analyzing preattentive effects (Fig. 1): 1) Corner Salience, 2) Visual Segmentation by Bar Angle, 3) Visual Segmentation by Bar Length, 4) Contour Integration by Bar Continuity and 5) Perceptual Grouping by Distance. VS categories were based on a feature-singleton search stimuli, where there was a unique salient target and a set of distractors and/or altered background (Fig. 2). These categories were: 6) Feature and Conjunctive Search, 7) Search Asymmetries, 8) Search in a Rough Surface, 9) Color Search, 10) Brightness Search, 11) Orientation Search, 12) Dissimilar Size Search, 13) Orientation Search with Heterogeneous distractors, 14) Orientation Search with Non-linear patterns, 15) Orientation search with distinct Categorization. Stimuli for SID4VAM's dataset was inspired by previous psychophysical experimentation [64, 58, 50].

Dataset stimuli were generated with 7 specific instances of feature contrast (Ψ), corresponding to hard ($\Psi_h = \{1..4\}$) and easy ($\Psi_e = \{5..7\}$) difficulties of finding the salient regions. These feature contrasts had their own parametrization (following Berga et al's psychophysical formulation [2, Section 2.4]) corresponding to the feature differences between the salient target and the rest of distractors (e.g. differences of target orientation, size, saturation, brightness...) or global effects (e.g. overall distractor scale, shape, background color, background brightness).

3. Experiments

Fixation maps from eye tracking data are generated by distributing each fixation location to a binary map. Fixation density maps are created by convolving a gaussian filter to the fixation maps, this simulates a smoothing caused by the deviations of σ=1 deg given from eye tracking experimentation, recommended by LeMeur & Baccino [36].

Typically, location-based saliency metrics (AUC_{Judd}, AUC_{Borji}, NSS) increase their score fixation locations fall inside (TP) the predicted saliency maps. Conversely, scores decrease fixation locations are not captured by saliency maps (FN) or when saliency maps exist in locations with no present fixations (FP). In distribution-based metrics (CC,

[1]Download the dataset in http://www.cvc.uab.es/neurobit/?page_id=53

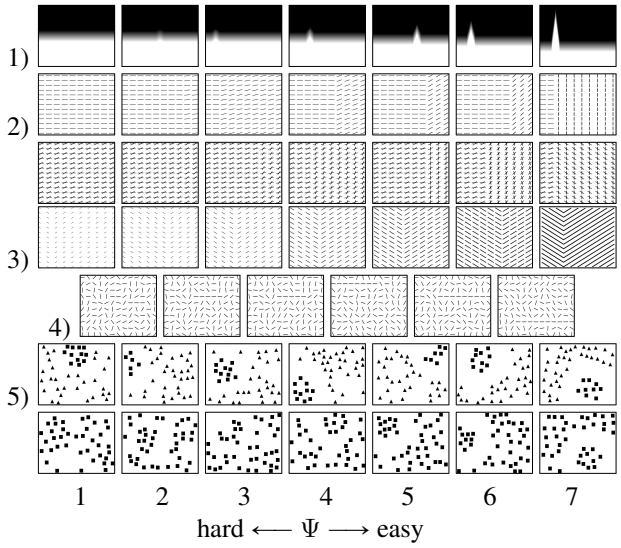


Figure 1: Free-Viewing stimuli

SIM, KL), saliency maps score higher when they have higher correlations with respect to fixation density map distributions. We have to point out that shuffled metrics (sAUC, InfoGain) consider FP values when saliency maps coincide with other fixation map locations or a baseline (here, corresponding to the center bias), which are not representative data for saliency prediction. Prediction metrics and its calculations are largely explained by Bylinskii et al. [11]. Our saliency metric scores and pre-processing used for this experimentation[2] have been replicated from the official saliency benchmarking procedure [10]. Psychometric evaluation of saliency predictions has been done with the Saliency Index (SI) [49, 50]. This metric evaluates the energy of a saliency map inside (S_t) a salient region (which would enclose a salient object) compared to the energy outside (S_b) the salient region. This metric allows evaluation of a saliency map when the salient region is known, considering in absolute terms the distribution of saliency of a particular AOI / mask. Here we show the formula of the SI

$$SI(S_t, S_b) = \frac{S_t - S_b}{S_b}.$$

Saliency maps have been computed from models shown on Table 2. Model evaluations have been divided according to its inspiration and prediction scores have been evaluated with saliency metrics and in psychophysical terms.

3.1. Model results on predicting fixations

Previous saliency benchmarks [6, 42, 9, 7, 10] reveal that Deep Learning models such as SALICON, ML-Net SAM-ResNet, SAM-VGG, DeepGazeII or SalGan score high-

[2]Code for metrics: https://github.com/dberga/saliency

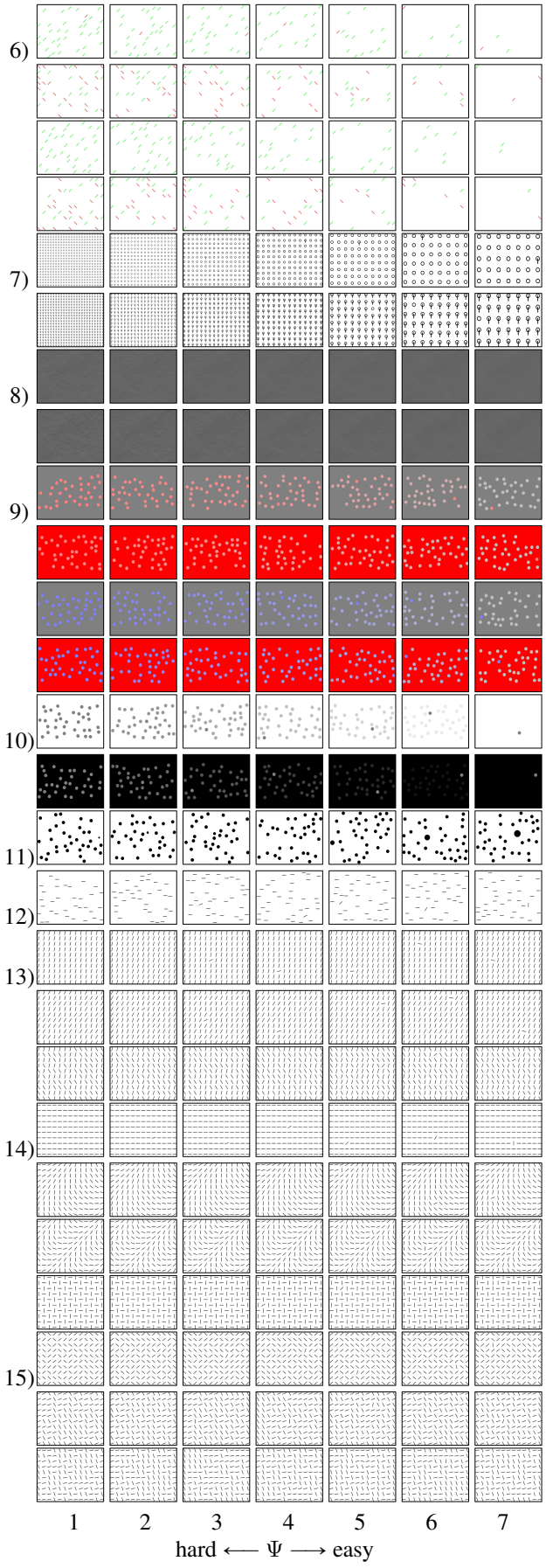


Figure 2: Visual Search stimuli

est on both shuffled and unshuffled metrics. In this section we aim to evaluate whether saliency maps that scored highly on fixation prediction do so with a synthetic image dataset and if their inspiration influences on their performance. We present metric scores of saliency map predictions of the whole dataset in Table 3 and plots in Fig. 3. Saliency metric scores reveal that overall Spectral/Fourier-based saliency models predict better fixations on a synthetic image dataset.

Table 3: Saliency metric scores for SID4VAM

Model	AUCj	AUCb	CC	NSS	KL	SIM	sAUC	InfoGain
GT	***0.943***	***0.882***	***1.000***	***4.204***	***0.000***	***1.000***	***0.860***	***2.802***
Baseline-CG	***0.703***	***0.697***	***0.281***	***0.722***	***1.577***	***0.372***	***0.525***	***-0.189***
IKN	0.686	0.678	0.283	0.878	1.748	0.380	0.608	-0.233
SIM	0.650	0.641	0.189	0.694	1.702	0.357	0.619	-0.148
AWS	0.679	0.667	0.255	1.088	1.592	0.373	0.672	0.013
NSWAM	0.614	0.610	0.136	0.529	1.686	0.335	0.622	-0.150
AIM	0.570	0.566	0.122	0.473	14.472	0.224	0.557	-18.182
ICL	0.737	0.717	0.343	1.100	1.788	0.405	0.624	-0.313
RARE	0.707	0.622	0.204	1.046	1.736	0.444	0.633	-0.158
CASD	0.733	0.669	0.408	1.904	2.395	0.403	0.652	-1.046
GBVS	0.747	0.718	0.400	1.464	1.363	0.413	0.628	0.331
SDLF	0.620	0.607	0.156	0.585	3.954	0.322	0.596	-3.244
SUN	0.542	0.532	0.080	0.333	16.408	0.165	0.530	-21.024
SDSR	0.672	0.665	0.192	0.639	1.904	0.365	0.642	-0.467
BMS	0.677	0.643	0.274	1.143	2.306	0.397	0.627	-0.958
ICF	0.618	0.566	0.141	0.700	3.274	0.306	0.564	-2.300
SR	0.748	0.694	0.420	1.916	1.432	0.431	0.685	0.348
PFT	0.705	0.692	0.398	1.885	2.227	0.377	0.684	-0.893
PQFT	0.701	0.693	0.387	1.774	2.197	0.373	0.684	-0.856
FT	0.521	0.518	0.072	0.331	7.552	0.129	0.517	-8.498
DCTS	0.729	0.724	0.439	2.004	1.363	0.396	0.708	0.337
WMAP	0.729	0.709	0.468	2.136	2.283	0.397	**0.709**	-0.981
QDCT	0.717	0.706	0.425	1.986	1.677	0.391	0.695	-0.105
HFT	**0.771**	**0.746**	**0.538**	**2.161**	**1.295**	**0.467**	0.682	**0.448**
SalGAN	0.715	0.662	0.287	0.883	2.506	0.373	0.593	-1.350
OpenSALICON	0.692	0.673	0.284	0.956	1.549	0.375	0.615	0.052
DeepGazeII	0.639	0.606	0.176	0.714	2.023	0.346	0.597	-0.587
SAM-VGG	0.537	0.523	0.026	0.070	11.947	0.216	0.503	-14.954
SAM-ResNet	0.727	0.673	0.305	0.967	2.610	0.388	0.600	-1.475
ML-Net	0.700	0.676	0.283	0.883	2.169	0.373	0.595	-0.837
Sal-DCNN	0.726	0.650	0.288	0.961	3.676	0.359	0.580	-3.05

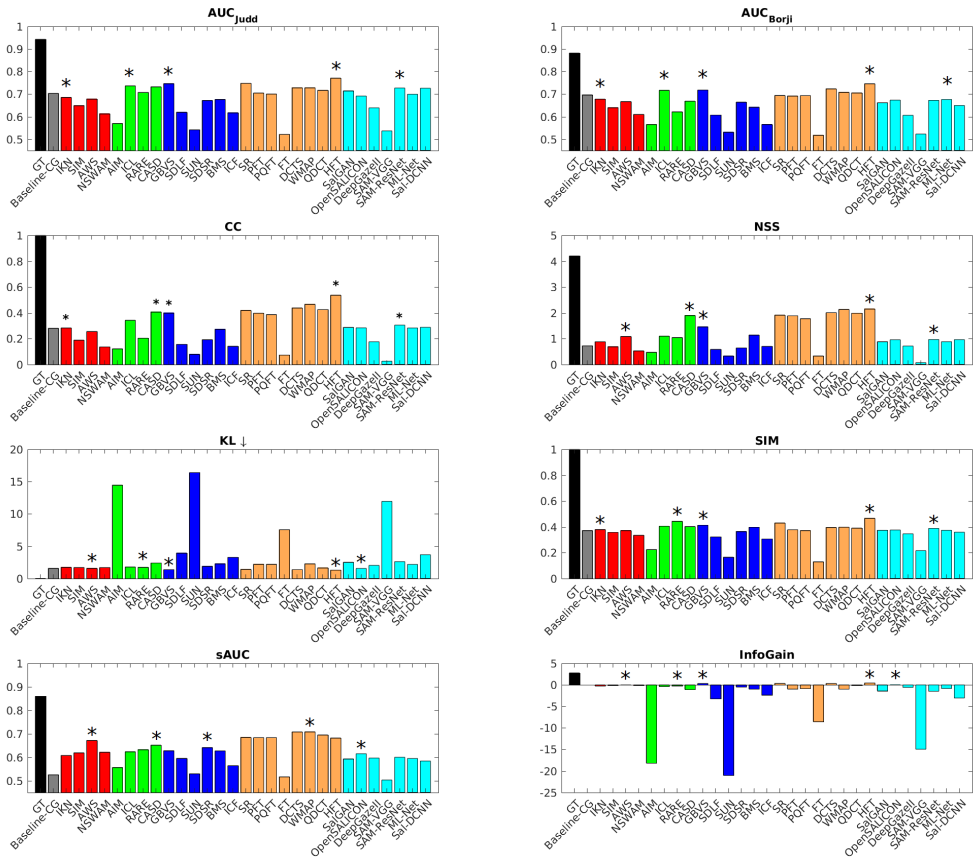
Cognitive/Biological , Information-Theoretic , Probabilistic, Fourier/Spectral, Machine/Deep Learning

Models such as HFT and WMAP remarkably outperform other saliency models. From other model inspirations, AWS score higher than other Cognitive/Biologically-inspired models, GBVS and CASD outperform other Probabilistic/Bayesian and Information-theoretic saliency models respectively. For Deep Learning models, SAM$_{ResNet}$ and OpenSALICON are the ones with highest scores. Although there are present differences in terms of model performances and model inspiration, similarities in model mechanisms can reveal phenomena of increasing and decreasing prediction statistics. This phenomena is present for Spectral/Fourier-based and Cognitive/Biologically-inspired models, withwhom all present similar performance and balanced scores throughout the distinct metric scores. It is to consider that sAUC and InfoGain metrics are more reliable compared to other metrics (which the baseline center gaussian sometimes acquires higher performance than most saliency models). In these terms, models shown on Fig. 4 are efficient saliency predictors for this dataset. We can also point out that models which process uniquely local feature conspicuity scored lower on SID4VAM fixation predictions, whereas the ones that processed global conspicuity scored higher. This phenomena might be related with the distinction of foveal (near the fovea) and ambient (away from the fovea) fixations, relative to the fixation order and the spatial locations of fixations [15]. The evaluation of gaze-wise model predictions has been done by grouping fixations of every instance separately. We have plotted results of the $sAUC$ saliency metric for each model (Fig. 5) and it is observable that model performance decrease upon fixation number, meaning that saliency is more likely to be predicted during first fixations. For evaluating the temporal relationship between human and model performance ($sAUC$), we have performed Spearman's (ρ) correlation tests for each fixation and it can be observed that IKN, ICL, GBVS, QDCT and ML-Net follow a similar slope as the GT, contrary to the case of the baseline center gaussian.

3.2. Model results on psychophysical consistency

Previous studies [4, 7, 2] found that several factors such as feature type, feature contrast, task, temporality of fixations and the center bias alternatively contribute to eye movement guidance. The HVS has specific contrast sensitivity to each stimulus feature, so that saliency models should adapt in the same way in order to be plausible in psychometric parameters. Here we will show how saliency prediction varies significantly upon feature contrast and the type of low-level features found in images. In Fig. 6a is found that saliency models increase SI with feature contrast "Ψ" following the distribution of human fixations. Most prediction SI scores show a higher slope with easy targets (salient objects with higher contrast with respect the rest, when $\Psi > 4$), being CASD and HFT the models that have higher SI at higher contrasts.

Contextual influences (here represented as distinct low-level features that appear in the image) contribute distinctively on saliency induced from objects that appear on the scene [25]. We suggest that not only the semantic content that appears on the scene affects saliency but the feature characteristics do significantly impact how salient objects are. This phenomena is observable in Fig. 6b and occurs for both human fixations and model predictions, specifically with highest SI for human fixations in 1) Corner Salience, 6) Feature and Conjunctive Search, 7) Search Asymmetries, 10) Brightness Search, 12) Dissimilar Size Search and 13) Orientation Search with Heterogeneous distractors. HFT and CASD have highest SI when GT is higher (when human fixations are more probable to fall inside the AOI), even outperforming GT probabilities for the cases of 1) and 7). We show in Fig. 7a that overall Saliency Index of most saliency models is distinct when we vary the type



*=Best performance for each model inspiration

Figure 3: Plots for saliency metric scores.

of feature contrast (easy vs hard) and the performed stimulus task (free-viewing vs visual search). Spectral/Fourier based models outperform other saliency models also in SI metric. Similarly with saliency metrics shown on previous subsection, AWS, CASD, BMS, HFT and SAM-ResNet are the most efficient models for each model inspiration category respectively. It is observable in Fig. 7b that saliency models have higher performance for easy targets, with increased overall model performance differences with respect hard targets (Fig. 7c). Similarly, visual search targets show lower difficulty (higher SI) to find predicted fixations inside the AOI than the free-viewing cases (Fig. 7d-e). Also distinct SI curves upon feature contrast are reported, revealing that contrast sensitivities are distinct for each low-level feature. Spearman's correlation tests on Fig. 6b show which models correlate with human performance over feature contrast and which one do so with the baseline (designating higher center biases). These results show that models such as AWS, CASD, BMS, DCTS or DeepGazeII highly correlate with human contrast sensitivities and do not correlate with the baseline center gaussian. Matching human contrast sensitivities on low-level visual features would be an interesting point of view to make future saliency models accurately predict saliency as well as to better understand how the HVS processes visual scenes.

4. SIG4VAM: Generating synthetic image patterns for training saliency models

We have also provided a synthetic image generator (SIG4VAM)[3], able to generate similar psychophysical images with other types of patterns. A larger set of images can be created by parametrizing factors such as stimulus size, number of distractors, feature contrast, etc. For instance, if the same 15 types (33 subtypes) of stimuli are selected instead with 28 contrast (Ψ) instances, then is generated a dataset with $33 \times 28 = 924$ stimulus. Adding to that, synthetic images with high-level features can be created using SIG4VAM (Fig. 8), by changing background properties, setting specific object instances for targets/distractors, as well as their low-level properties (orientation, brightness, color, etc.). SID4VAM has been proposed as a possible ini-

[3]Code for generating synthetic stimuli: https://github.com/dberga/sig4vam

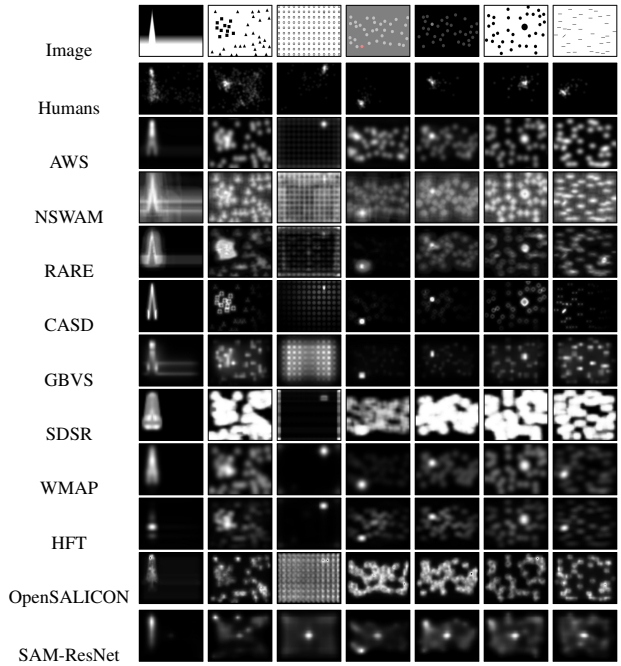


Figure 4: Examples of dataset stimuli and saliency map predictions. Only two models for each inspiration category that presented highest performance with shuffled saliency metric scores (sAUC and InfoGain) are shown.

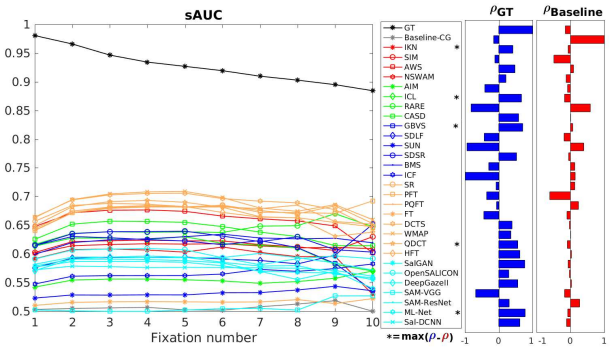


Figure 5: sAUC gaze-wise prediction scores.

tial test set for saliency prediction, where data of fixations and binary masks are available for benchmarking. Training sets can be obtained with SIG4VAM (GT of binary masks of pop-out/salient regions are automatically-generated), abling to fit contrast sensitivies and obtaining loss functions upon scores of fixation probability distribution [11] and salient region detection metrics [56] (e.g. SI, PR, MAE, S-/F-measures, etc.). Latest strategies [46] that synthetically modify real scenes have shown dramatic changes in scores of object detection tasks, using "object transplanting" (superposing an object on distinct locations on the scene). In these terms, SIG4VAM could be extended for evaluating

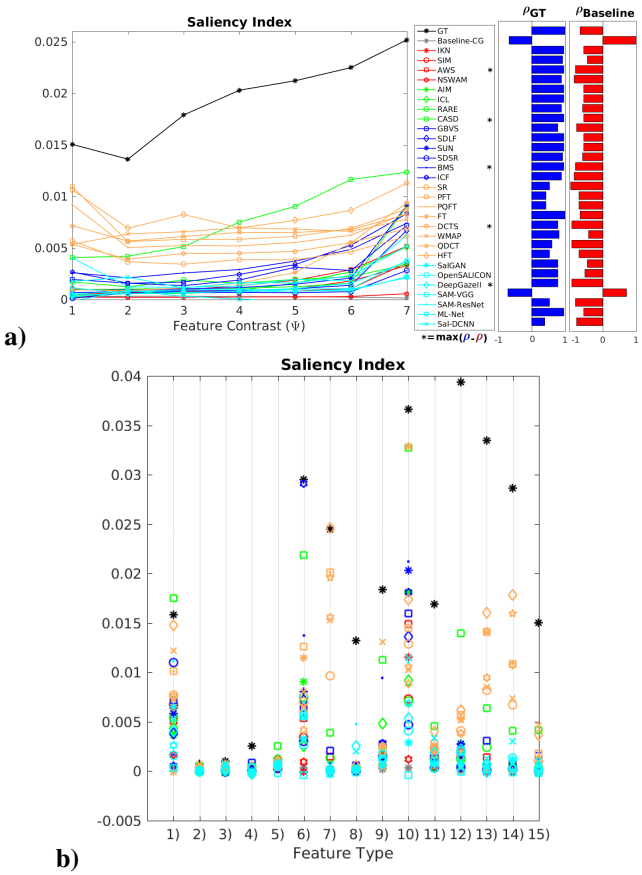


Figure 6: Results of Saliency Index of model predictions upon Feature Contrast **(a)** and Feature Type **(b)**.

predictions of models over distinct contexts and tasks.

5. Discussion

Previous saliency benchmarks show that eye movements are efficiently predicted with latest Deep Learning saliency models. This is not the case with synthetic images, also for models pre-trained with sets of psychophysical patterns (e.g. SAM with CAT2000). This suggest that their computations of saliency do not arise as a general mechanism. These methods have been trained with eye tracking data (real images containing high-level features) and although several factors guide eye movements have been shown [58] that low-level saliency (i.e. pop-out effects) is one of the most influential for determining bottom-up attention. Another possibility is that we randomly parametrized salient object location, lowering the center bias effect. With this benchmark we can evaluate how salient is a particular object by parametrizing its low-level feature contrast with respect to the rest of distractors and/or background. Therefore, the evaluation of saliency can be done accounting for feature contrast, analyzing the importance to the objects that are

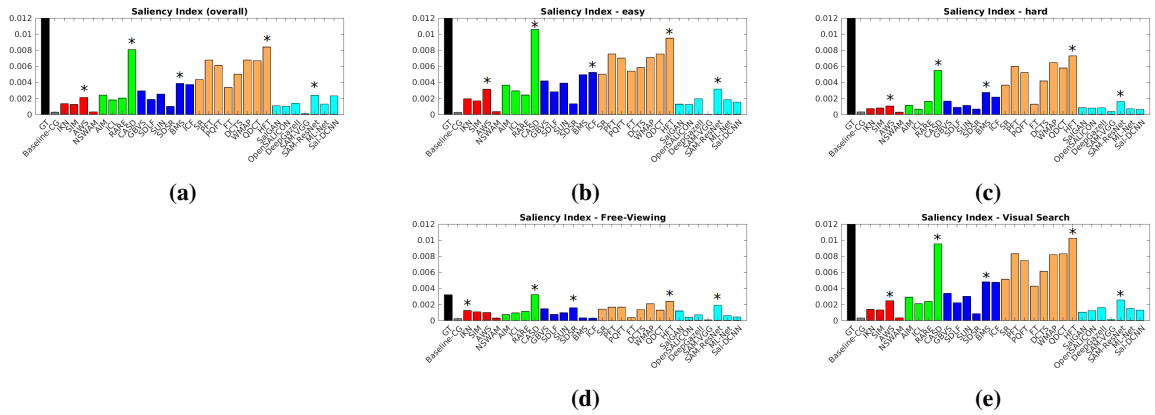


Figure 7: Results of Saliency Index metric scores from dataset model predictions **(a)**, for easy/hard difficulties **(b-c)** and Free-Viewing/Visual Search tasks **(d-e)**.

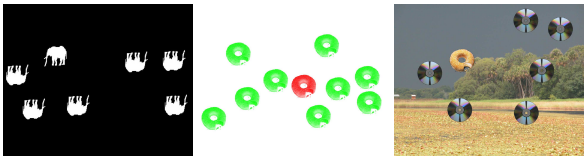


Figure 8: Examples of generating synthetic images with high-level features (i.e. objects as target/distractors), changing low-level feature properties **(a-b)** or background **(c)**.

easier to detect or preattetively. Previous saliency benchmarks usually evaluate eye tracking data spatially across all fixations, we also propose the evaluation of saliency across fixations, which is an issue of further study. Future steps for this study would include the evaluation of saliency in dynamic scenes [44, 35] using synthetic videos with both static or dynamic camera. This would allow us to investigate the impact of temporally-variant features (e.g. flicker and motion) over saliency predictions. Another analysis to consider is the impact of the spatial location of salient features (in eccentricity terms towards the image center), which might affect each model distinctively. Each of the steps in saliency modelization (i.e. feature extraction, conspicuity computation and feature fusion) might have a distinct influence over eye movement predictions. Acknowledging that conspicuity computations are the key factor for computing saliency, a future evaluation of how each mechanism contributes to model performance might be of interest.

6. Conclusion

Contrary to the current state-of-the-art, we reveal that saliency models are far away from acquiring HVS performance in terms of predicting bottom-up attention. We prove this with a novel dataset SID4VAM, which contains uniquely synthetic images, generated with specific low-level feature contrasts. In this study, we show that overall Spectral/Fourier-based saliency models (i.e. HFT and WMAP) clearly outperform other saliency models when detecting a salient region with a particular conspicuous object. Other models such as AWS, CASD, GBVS and SAM-ResNet are the best predictor candidates for each saliency model inspiration categories respectively (Cognitive/Biological, Information-Theoretic, Probabilistic and Deep Learning). In particular, visual features learned with deep learning models might not be suitable for efficiently predicting saliency using psychophysical images. Here we pose that saliency detection might not be directly related to object detection, therefore training upon high-level object features might not be significatively favorable for predicting saliency in these terms. Future saliency modelization and evaluation should account for low-level feature distinctiveness in order to accurately model bottom-up attention. Here we remark the need for analyzing other factors such as the order of fixations, the influences of the task and the psychometric parameters of the salient regions.

7. Acknowledgements

This work was funded by the MINECO (DPI2017-89867-C2-1-R, TIN2015-71130-REDT), AGAUR (2017-SGR-649), CERCA Programme / Generalitat de Catalunya, in part by Xunta de Galicia under Project ED431C2017/69, in part by the Consellería de Cultura, Educación e Ordenación Universitaria (accreditation 20162019, ED431G/08) and the European Regional Development Fund, and in part by Xunta de Galicia and the European Union (European Social Fund). We also acknowledge the generous GPU support from NVIDIA.

References

[1] Radhakrishna Achanta, Sheila Hemami, Francisco Estrada, and Sabine Susstrunk. Frequency-tuned salient region detection. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, jun 2009. 3

[2] David Berga, Xos R. Fdez-Vidal, Xavier Otazu, Vctor Leboran, and Xose M. Pardo. Psychophysical evaluation of individual low-level feature influences on visual attention. *Vision Research*, 154:60 – 79, 2019. 1, 2, 3, 5

[3] David Berga and Xavier Otazu. A neurodynamical model of saliency prediction in v1. *In Review*, 2018. arXiv:1811.06308. 3

[4] Ali Borji and Laurent Itti. State-of-the-art in visual attention modeling. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(1):185–207, jan 2013. 2, 5

[5] Ali Borji and Laurent Itti. Cat2000: A large scale fixation dataset for boosting saliency research. *CVPR 2015 workshop on "Future of Datasets"*, 2015. 2

[6] Ali Borji, D. N. Sihite, and L. Itti. Quantitative analysis of human-model agreement in visual saliency modeling: A comparative study. *IEEE Transactions on Image Processing*, 22(1):55–69, jan 2013. 1, 4

[7] Neil D.B. Bruce, Calden Wloka, Nick Frosst, Shafin Rahman, and John K. Tsotsos. On computational modeling of visual saliency: Examining what's right, and what's left. *Vision Research*, 116:95–112, nov 2015. 1, 2, 4, 5

[8] Neil D. B. Bruce and John K. Tsotsos. Saliency based on information maximization. In *18th International Conference on Neural Information Processing Systems (NIPS'05)*, pages 155–162. MIT Press, 2005. 2, 3

[9] Z. Bylinskii, E.M. DeGennaro, R. Rajalingham, H. Ruda, J. Zhang, and J.K. Tsotsos. Towards the quantitative evaluation of visual attention models. *Vision Research*, 116:258–268, nov 2015. 1, 4

[10] Zoya Bylinskii, Tilke Judd, Ali Borji, Laurent Itti, Frédo Durand, Aude Oliva, and Antonio Torralba. Mit saliency benchmark. http://saliency.mit.edu/. 4

[11] Zoya Bylinskii, Tilke Judd, Aude Oliva, Antonio Torralba, and Fredo Durand. What do different evaluation metrics tell us about saliency models? *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1–1, 2018. 4, 7

[12] Marcella Cornia, Lorenzo Baraldi, Giuseppe Serra, and Rita Cucchiara. A Deep Multi-Level Network for Saliency Prediction. In *International Conference on Pattern Recognition (ICPR)*, 2016. 3

[13] Marcella Cornia, Lorenzo Baraldi, Giuseppe Serra, and Rita Cucchiara. Predicting Human Eye Fixations via an LSTM-based Saliency Attentive Model. *IEEE Transactions on Image Processing*, 27(10):5142–5154, 2018. 3

[14] Howard E. Egeth and Steven Yantis. VISUAL ATTENTION: Control, representation, and time course. *Annual Review of Psychology*, 48(1):269–297, feb 1997. 1

[15] Michelle L. Eisenberg and Jeffrey M. Zacks. Ambient and focal visual processing of naturalistic activity. *Journal of Vision*, 16(2):5, mar 2016. 5

[16] G. T. Fechner. *Elements of Psychophysics, Volume 1*. Holt, Rinehart and Winston, the University of Michigan, 1966. 2

[17] JH Fecteau and DP Munoz. Salience, relevance, and firing: a priority map for target selection. *Trends in Cognitive Sciences*, 10(8):382–390, aug 2006. 1

[18] Anton Garcia-Diaz, Xose R. Fdez-Vidal, Xose M. Pardo, and Raquel Dosil. Saliency from hierarchical adaptation through decorrelation and variance normalization. *Image and Vision Computing*, 30(1):51–64, jan 2012. 3

[19] Stas Goferman, Lihi Zelnik-Manor, and Ayellet Tal. Context-aware saliency detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(10):1915–1926, oct 2012. 3

[20] Chenlei Guo, Qi Ma, and Liming Zhang. Spatio-temporal saliency detection using phase spectrum of quaternion fourier transform. In *2008 IEEE Conference on Computer Vision and Pattern Recognition*, jun 2008. 3

[21] Jonathan Harel, Christof Koch, and Pietro Perona. Graph-based visual saliency. In B. Schölkopf, J. C. Platt, and T. Hoffman, editors, *Advances in Neural Information Processing Systems 19*, pages 545–552. MIT Press, 2007. 3

[22] Xiaodi Hou, J. Harel, and C. Koch. Image signature: Highlighting sparse salient regions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(1):194–201, jan 2012. 3

[23] Xiaodi Hou and Liqing Zhang. Saliency detection: A spectral residual approach. In *2007 IEEE Conference on Computer Vision and Pattern Recognition*, jun 2007. 3

[24] Xiaodi Hou and Liqing Zhang. Dynamic visual attention: searching for coding length increments. In D. Koller, D. Schuurmans, Y. Bengio, and L. Bottou, editors, *Advances in Neural Information Processing Systems 21*, pages 681–688. Curran Associates, Inc., 2009. 3

[25] Alex D. Hwang, Hsueh-Cheng Wang, and Marc Pomplun. Semantic guidance of eye movements in real-world scenes. *Vision Research*, 51(10):1192–1205, may 2011. 5

[26] Laurent Itti and Christof Koch. A saliency-based search mechanism for overt and covert shifts of visual attention. *Vision Research*, 40(10-12):1489–1506, jun 2000. 2, 3

[27] Laurent Itti, Christof Koch, and Ernst Niebur. A model of saliency-based visual attention for rapid scene analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(11):1254–1259, 1998. 1, 3

[28] Ming Jiang, Shengsheng Huang, Juanyong Duan, and Qi Zhao. SALICON: Saliency in context. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, jun 2015. 3

[29] Tilke Judd, Fredo Durant, and Antonio Torralba. A benchmark of computational models of saliency to predict human fixations. *CSAIL Technical Reports*, jan 2012. 2

[30] Tilke Judd, Krista Ehinger, Fredo Durand, and Antonio Torralba. Learning to predict where humans look. In *2009 IEEE International Conference on Computer Vision*, sep 2009. 2

[31] Christof Koch and Shimon Ullman. Shifts in selective visual attention: Towards the underlying neural circuitry. In *Matters of Intelligence*, pages 115–141. Springer, 1987. 1

[32] Gert Kootstra, Bart de Boer, and Lambert R. B. Schomaker. Predicting eye fixations on complex visual stimuli using local symmetry. *Cognitive Computation*, 3(1):223–240, jan 2011. 2

[33] Matthias Kummerer, Thomas S.A. Wallis, Leon A. Gatys, and Matthias Bethge. Understanding low- and high-level contributions to fixation prediction. In *2017 IEEE International Conference on Computer Vision (ICCV)*, oct 2017. 3

[34] Mai Xu Zulin Wang Lai Jiang, Zhe Wang. Image saliency prediction in transformed domain: A deep complex neural network method. February 2019. 3

[35] Victor Leboran, Anton Garcia-Diaz, Xose R. Fdez-Vidal, and Xose M. Pardo. Dynamic whitening saliency. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(5):893–907, may 2017. 8

[36] Olivier LeMeur and Thierry Baccino. Methods for comparing scanpaths and saliency maps: strengths and weaknesses. *Behavior Research Methods*, 45(1):251–266, jul 2012. 3

[37] Jian Li, Martin D. Levine, Xiangjing An, Xin Xu, and Hangen He. Visual saliency based on scale-space analysis in the frequency domain. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(4):996–1010, apr 2013. 3

[38] Fernando Lopez-Garcia, Xose Ramon, Xose Manuel, and Raquel Dosil. Scene recognition through visual attention and image features: A comparison between SIFT and SURF approaches. In *Object Recognition*. InTech, apr 2011. 3

[39] Naila Murray, Maria Vanrell, Xavier Otazu, and C. Alejandro Parraga. Saliency estimation using a non-parametric low-level vision model. In *CVPR 2011*, jun 2011. 3

[40] Junting Pan, Cristian Canton, Kevin McGuinness, Noel E. O'Connor, Jordi Torres, Elisa Sayrol, and Xavier and Giro-i Nieto. Salgan: Visual saliency prediction with generative adversarial networks. In *CVPR 2017 Scene Understanding Workshop (SUNw)*, January 2017. 3

[41] Subramanian Ramanathan, Harish Katti, Nicu Sebe, Mohan Kankanhalli, and Tat-Seng Chua. An eye fixation database for saliency detection in images. In *Computer Vision – ECCV 2010*, pages 30–43. Springer, 2010. 2

[42] Nicolas Riche, Matthieu Duvinage, Matei Mancas, Bernard Gosselin, and Thierry Dutoit. Saliency and human fixations: State-of-the-art and study of comparison metrics. In *2013 IEEE International Conference on Computer Vision (ICCV)*, dec 2013. 4

[43] Nicolas Riche and Matei Mancas. Bottom-up saliency models for still images: A practical review. In *From Human Attention to Computational Attention*, pages 141–175. Springer New York, 2016. 1, 2

[44] Nicolas Riche and Matei Mancas. Bottom-up saliency models for videos: A practical review. In *From Human Attention to Computational Attention*, pages 177–190. Springer New York, 2016. 8

[45] Nicolas Riche, Matei Mancas, Bernard Gosselin, and Thierry Dutoit. Rare: A new bottom-up saliency model. In *2012 19th IEEE International Conference on Image Processing*, sep 2012. 3

[46] Amir Rosenfeld, Richard Zemel, and John K. Tsotsos. The elephant in the room, 2018. arXiv:1808.03305. 7

[47] Boris Schauerte and Rainer Stiefelhagen. Quaternion-based spectral saliency detection for eye fixation prediction. In *Computer Vision – ECCV 2012*, pages 116–129. Springer, 2012. 3

[48] Hae Jong Seo and Peyman Milanfar. Static and space-time visual saliency detection by self-resemblance. *Journal of Vision*, 9(12):15–15, nov 2009. 3

[49] Alireza Soltani and C. Koch. Visual saliency computations: Mechanisms, constraints, and the effect of feedback. *Journal of Neuroscience*, 30(38):12831–12843, sep 2010. 4

[50] Michael W. Spratling. Predictive coding as a model of the v1 saliency map hypothesis. *Neural Networks*, 26:7–28, feb 2012. 2, 3, 4

[51] Christopher Lee Thomas. Opensalicon: An open source implementation of the salicon saliency model. Technical Report TR-2016-02, University of Pittsburgh, 2016. 3

[52] Antonio Torralba, Aude Oliva, Monica S. Castelhano, and John M. Henderson. Contextual guidance of eye movements and attention in real-world scenes: The role of global features in object search. *Psychological Review*, 113(4):766–786, 2006. 3

[53] Anne M. Treisman and Garry Gelade. A feature-integration theory of attention. *Cognitive Psychology*, 12(1):97–136, jan 1980. 1, 2

[54] John K. Tsotsos, Scan M. Culhane, Winky Yan Kei Wai, Yuzhong Lai, Neal Davis, and Fernando Nuflo. Modeling visual attention via selective tuning. *Artificial Intelligence*, 78(1-2):507–545, oct 1995. 1

[55] Richard Veale, Ziad M. Hafed, and Masatoshi Yoshida. How is visual salience computed in the brain? insights from behaviour, neurobiology and modelling. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 372(1714):20160113, jan 2017. 1

[56] Wenguan Wang, Qiuxia Lai, Huazhu Fu, Jianbing Shen, and Haibin Ling. Salient object detection in the deep learning era: An in-depth survey, 2019. arXiv:1904.09146. 7

[57] Stefan Winkler and Ramanathan Subramanian. Overview of eye tracking datasets. In *2013 Fifth International Workshop on Quality of Multimedia Experience (QoMEX)*, jul 2013. 2

[58] J. M. Wolfe. Guided search 4.0: A guided search model that does not require memory for rejected distractors. *Journal of Vision*, 1(3):349–349, mar 2010. 1, 2, 3, 7

[59] Jeremy M. Wolfe, Evan M. Palmer, and Todd S. Horowitz. Reaction time distributions constrain models of visual search. *Vision Research*, 50(14):1304–1311, jun 2010. 2

[60] Steven Yantis and Howard E. Egeth. On the distinction between visual salience and stimulus-driven attentional capture. *Journal of Experimental Psychology: Human Perception and Performance*, 25(3):661–676, 1999. 1

[61] Jianming Zhang and Stan Sclaroff. Saliency detection: A boolean map approach. In *2013 IEEE International Conference on Computer Vision*, dec 2013. 3

[62] Liming Zhang and Weisi Lin. *Selective Visual Attention*. John Wiley & Sons (Asia) Pte Ltd, mar 2013. 1, 2

[63] Lingyun Zhang, Matthew H. Tong, Tim K. Marks, Honghao Shan, and Garrison W. Cottrell. SUN: A bayesian framework for saliency using natural statistics. *Journal of Vision*, 8(7):32, dec 2008. 3

[64] Li Zhaoping and Keith A. May. Psychophysical tests of the hypothesis of a bottom-up saliency map in primary visual cortex. *PLoS Computational Biology*, 3(4):e62, 2007. 2, 3