Noname manuscript No.
(will be inserted by the editor)

A Neurodynamic model of Saliency prediction in V1

David Berga* · **Xavier Otazu**

Received: date / Accepted: date

Abstract *Objectives:* Lateral connections in the primary visual cortex (V1) have long been hypothesized to be responsible of several visual processing mechanisms such as brightness induction, chromatic induction, visual discomfort and bottom-up visual attention (also named saliency). Many computational models have been developed to independently predict these and other visual processes, but no computational model has been able to reproduce all of them simultaneously. In this work we show that a biologically plausible computational model of lateral interactions of V1 is able to simultaneously predict saliency and all the aforementioned visual processes.
Methods: Our model's (NSWAM) architecture is based on Pennachio's neurodynamic model of lateral connections of V1. It is defined as a network of firing rate neurons, sensitive to visual features such as brightness, color, orientation and scale. We tested NSWAM saliency predictions using images from several eye tracking datasets.
Results: We show that accuracy of predictions, using shuffled metrics, obtained by our architecture is similar to other state-of-the-art computational methods, particularly with synthetic images (CAT2000-Pattern & SID4VAM) which mainly contain low level features. Moreover, we outperform other biologically-inspired saliency models that are specifically designed to exclusively reproduce saliency. *Conclusions:* Hence, we show that our biologically plausible model of lateral connections can simultaneously explain different visual proceses present in V1 (without applying any type of training or optimization and keeping the same parametrization for all the visual processes). This can be useful for the definition of a unified architecture of the primary visual cortex.

Computer Vision Center, Universitat Autonoma de Barcelona
Edifici O, Campus UAB, 08193, Bellaterra
Tel.: +34 935 81 18 28
E-mail: dberga@cvc.uab.es
Correspondence: David Berga

1 Introduction

Visual salience can be defined as "the distinct subjective perceptual quality which makes some items in the world stand out from their neighbors and immediately grab our attention" [35]. Hence, saliency could be defined as one of the properties of the visual scene that attracts our attention toward a particular set of visual features. Although not being the best option for the study of visual saliency, several studies of eye movements using different approaches have been performed. Eye movements are controlled by many different factors, e.g. low/high-level information, task, endogenous factors, etc. Hence, prediction of eye movement cannot be performed only by one of these factors. Koch and Ullman [38] propose a computational framework in which visual features are integrated to generate a saliency map. These visual features are projected to V1 and later processed distinctively on the ventral ("what") and dorsal ("where") streams. These connections are projected to the superior colliculus (SC), which would generate either top-down (relevance) or bottom-up (saliency) control of eye movements by combining neuronal activity from distinct brain areas to a unique map (priority map) [24][82].

Related Work

Given these distinct levels of processing from the human visual system (HVS), a set of computational models are proposed in order to reproduce eye movement behavior.

Itti et al. introduce a biologically-inspired model [36] in which low-level features are extracted using linear DoG filters, their conspicuity is calculated using center-surround differences (inspired by V1's simple cell computations) and integrated (pooled to the SC as a master saliency map) using winner-take-all mechanisms. Although computations of existing saliency models seem to mimic HVS mechanisms, complexity of scenes make eye-movement behavior hard to predict because of the aforementioned additional factors. Bruce & Tsotsos model [11] offered a semi-supervised mechanism to account for relevant information of the scenes in combination with the bottom-up computations of V1, predicting eye movement behavior at distinct scene contexts. Given the basis of these models, a myriad of computational models, both with artificial and biological inspiration [37][6][89][68], have implemented distinct ways to predict human eye movements obtaining better performance on its predictions [67][8][9][13]. Thus, although proposed computational eye movement prediction models could precisely resemble eye-tracking data, it is questionable to consider that these predictions accurately and specifically represent saliency [10][4][5]. We have added a table describing each model with its inspiration (Cognitive/Biological, Information-based, Probabilistic and Deep Learning, as described in [37][6][5]) and type of feature processing (with global or local features). Saliency corresponds to bottom-up attention, which derives from generating conspicuity from low-level image features [34]. The relation of eye movements in relation to saliency is mainly based on the fact that eye movements are driven by both bottom-up attention (saliency in this case) and top-down attention (also coined with the name "relevance" [23]). The problem is that "saliency models" are told to predict saliency while, in fact, they are predicting fixations (which experimentally is dissimilar, as it includes both bottom-up and top-down effects).

Table 1: Description of saliency models

Model	Authors	Year	Inspiration				Type	
			C	I	P	D	G	L
IKN	Itti et al.[36]	1998	✓				✓	✓
AIM	Bruce & Tsotsos [11]	2005	✓	✓				✓
GBVS	Harel et al.[30]	2006			✓		✓	✓
SUN	Zhang et al. [90]	2008			✓			✓
SDSR	Seo & Milanfar [73]	2009	✓		✓		✓	✓
SIM	Murray et al.[51]	2011	✓				✓	✓
AWS	Garcia-Diaz et al.[27]	2012	✓				✓	✓
OpenSALICON	Jiang et al.[76]	2015				✓		✓
ML-Net	Cornia et al.[18]	2016				✓		✓
DeepGazeII	Kümmerer et al.[41]	2016				✓		✓
SalGAN	Pan et al.[59]	2017				✓		✓
SAM	Cornia et al.[19]	2018				✓		✓

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

Motivation

Li's work [45][46][47][91] proposes that V1's computations, particularly lateral connections, are the ones responsible of the representation of the aforementioned saliency map. Following her work, the role for the early processing of the visual features relies in V1, mainly driven for this case by uniquely processing low-level visual features. These connections are later projected to the SC in order to generate bottom-up saccadic eye movements [70][71, Chapter 9][83]. Since Li's architecture only worked on lattices of oriented bars, in order to process greyscale images [62] enhanced Li's architecture by adding receptive fields of different orientations and spatial scales.

These authors show that this architecture reproduces the brightness induction visual process for both still images and dynamic visual stimulus (i.e. videos). By enhancing the architecture adding two channels (opponent red-green and blue-yellow) to the luminance channel, the same authors showed that this architecture also reproduces chromatic induction [16]. Evenmore, by studying several statistical properties of the spatial and temporal dynamics of the firing-rate activity of the architecture, they showed that it also predicts the visual phenomena of visual discomfort [64] (which is one of the the main triggers of migraines).

Considering that these works have shown that the neuronal mechanism of lateral connections is partially responsible for these effects, we aim to use the same model in order to address another process present in the primary visual cortex: visual saliency. Using this Penacchio's computational architecture, we aim to compute feature conspicuity (distinctiveness between feature maps), which will alternatively represent the function of the aforementioned saliency map.

Objectives

In this study we want to test if the computations of our model are able to reproduce eye movement behavior being consistent with eye-tracking psychophysical experimentation. Current evaluation of saliency predictions is unfair and do not consider many biases, specific to saliency. Using metrics not affected by these biases, we want to show that our architecture can obtain similar results in comparison to other state-of-the-art models (or outpeform them in some specific datasets). Concretely, we want to study whether our architecture can obtain results similar or better than other stateof-the-art models when using datasets with less top-down eye movement biases (that is, datasets which mainly reflects bottom-up saliency related eye movements). Ad-

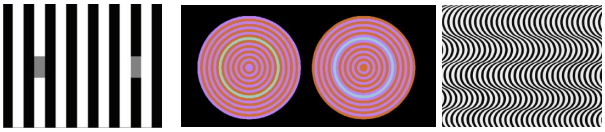
ditionally, we want to show that we can obtain these results without applying any type of training or optimization and keeping the same parametrization for all the visual processes (e.g. brightness induction, chromatic induction and visual discomfort). Hence, we want to show that a computational model of lateral connections can offer a unified architecture reproducing distinct V1 functionality, leading to a unification of several visual processes.

Unifying an architecture of several visual processes

Li's architecture is a model of a neuronal mechanism present in the primary visual cortex (and other areas). All the perceptual processes that rely on this mechanism could be computationally reproduced, at least partially, with the same architecture. As explained in previous section we showed in previous studies [62][16][64] that the proposed firing-rate neurodynamic model of V1's intra-cortical interactions extended in Penacchio's model [62] is able to simultaneously reproduce several visual processes such as brightness and color induction effects as well as visual discomfort mechanisms.

Brightness induction refers to the changes in perceived brightness of a visual target due to the luminance of its surrounding area. From this statement, the HVS can either perceive the visual target and the surrounding area with similar/equal brightness (assimilation) or to perceive brightness differences (contrast). We can observe in Fig. 1**A** how two grey patches are perceived distinctively whilst being with same brightness. Similarly, the HVS perceives the chromatic properties of a visual target distinctively depending on the chromaticities of its surrounding area. This phenomena is named chromatic induction. It appears in both "l" and "s" opponent channels ("l" for red-green and "s" for blue-yellow). This effect is observable on Fig. 1**B**, where the central ring from the reference stimulus (left) appears to be "greener" (being perceived with lower "l" chromatic properties) than the central ring from the test (right), which appears to be "bluer" instead (being perceived with higher "s" chromatic properties).

These effects were reproduced previously in a multiresolution wavelet framework with BiWaM [58] and CiWaM [57] computational models. These models' aim was to mimic V1's simple cell mechanisms by computing center-surround differences at distinct color and luminance opponencies. Being inspired by the aforementioned Li's model, Penacchio et al. [62] modeled an excitatory and inhibitory model of V1 as a more biologically plausible approach to reproduce these visual effects. Considering physiological and neurodynamic properties of V1 cells [45] at different spatial frequencies and orientations, Penacchio et al. [62] show it is possible to simultaneously reproduce psychophysical experiments of brightness [62] and chromatic [16] induction effects using a unified computational architecture.



A **B** **C**

Fig. 1: **(A)** Example of Brightness induction present at the White effect [85]. The two grey squares are the same luminance, but the the left square is perceived as darker than the right square. **(B)** Chromatic induction from Monnier & Shevell's concentric ring stimuli [50]. On the left disk we perceive a greenish ring and on the right disk we perceive a bluish, but these two rings are the exactly the same color. **(C)** Discomfortable image (credit by Nicholas Wade). If we fixate our gaze at this image, after some tens of seconds it could become uncomfortable to look at. [80].

Latest experiments showed that this computational architecture is also able to predict visual discomfort [64]. Specific visual patterns (Fig. 1**C**) are shown to cause discomfort, malaise, nausea or even migraine [63][42]. Taking into account the relative contrast energy from stimulus regions (due to its orientation, luminance, chromatic and spatial frequency distributions), we can predict whether a stimulus can cause hyperexcitability in V1, a possible cause of visual discomfort for certain images.

Hypothesis

The Hypothesis of the present work is that a computational architecture implementing a biologically plausible model of lateral connections in the primary visual cortex is able to predict low-level saliency while simultaneously reproducing all the previously commented visual processes (e.g. brightness and chromatic induction, and visual discomfort).

2 Model Description

The model is extended from previous implementation by Pennacchio et al. [62] in Matlab and C++ [1]. Here

[1] Code can be downloaded from `https://github.com/dberga/NSWAM`

we describe the main steps in relation to the computations done to the images: 2.1. Feature Extraction, 2.2. Feature Conspicuity and 2.3. Feature Integration. In this section, computations in the early visual pathways will be represented in line with a stimulus example. Overall model architecture was inspired by previous work from Murray et al.'s Saliency Induction Model (SIM) [51], defining a biologically-inspired and unsupervised low-level model for saliency prediction. Although it provided a promising approach for predicting saliency maps, we want to stress the novelty of computations of firing rate dynamics proposed in our architecture are in accordance with physiological properties of V1 cells.

2.1 From images to Sensory Signals: Feature Extraction

2.1.1 Color representation

Human retinal cone photoreceptors are sensitive to distinct wavelengths of the visual spectrum, corresponding to long, medium and short wavelengths. Similarly, traditional digital cameras capture light as values in the RGB color space (corresponding to Red, Green and Blue components). Retinal ganglion cells (RGC) encode luminance and chromatic signals as an opponent representation. This opponent representation separates channels of "Red vs Green" and "Blue vs Yellow" from cone cell responses, and luminance ("Bright vs Dark") from both cones and rod responses. Activity from these channels (R-G, B-Y and L) is then projected respectively to the lateral geniculate nucleus (LGN) and through parvo-cellular (P-), konio-cellular (K-) and magno-cellular (M-) pathways towards V1.

In order to represent this opponent colour information, we use the widely used opponent colour representation:

$$L = R + G + B, \tag{1}$$

$$rg = \frac{R - G}{L}, \tag{2}$$

$$by = \frac{R + G - 2B}{L}, \tag{3}$$

We can interpret L, rg and by components defined in Eqs. 1,2,3 as means of luminance opponency and chrominance oponencies R-G and B-Y, respectively. In Fig. 2 we illustrate an example of an image and its conversion to this representation, with higher activation on the "Red vs Green" opponent cells than the case of "Blue vs Yellow" and "Bright vs Dark" opponencies. It has been shown that this representation is related to some perceptual properties of colour perception [60].

All RGB pixel values of processed images are previously corrected with $\gamma = 1/2.2$.

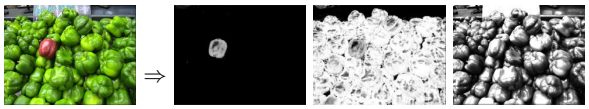


Fig. 2: Example of RGB image (left image) and its corresponding opponent color representation: **(A)** "red vs green" (rg) , **(B)** "blue vs yellow" (by) and **(C)** "luminance" (L) channels.

2.1.2 Multiscale and orientation representation

V1 cell sensitivities to distinct orientations [34] and spatial frequencies [49] are usually modeled as Gabor filters. Since Gabor transforms cannot be inverted to obtain the original image, we used the *à trous* algorithm, which is an undecimated discrete wavelet transform (DWT) [28][74, Chapter 6]. This decomposition allows to perform an inverse, where the basis functions remain similar to Gabor filters. We propose biologically plausible computations for extracting multiple orientations and multiscale feature representations of from V1's receptive field (RF) hypercolumnar organization (Fig. 3). The wavelet approximation planes $c_{s,\theta}$ (s for scale and θ for orientation) are computed by convolving the image with the filter h_s.

$$\begin{aligned} c_{s,h} &= c_{s-1} \otimes h_s, \\ c_{s,v} &= c_{s-1} \otimes h'_s. \end{aligned} \tag{4}$$

The filter h_s is obtained from h_{s-1} by doubling its size, i.e. $h_s = \uparrow h_{s-1}$, where $\uparrow$ means upsampling by introducing zeros between the coefficients. The filter (h_s) for the first scale is

$$h_1 = \frac{1}{16}\left[1\ 4\ 6\ 4\ 1\right]$$

This filter can be also transposed (h'_s) to obtain distinct approximation orientation planes $c_{s,h}$ and $c_{s,v}$. From these approximation planes, we can obtain the wavelet coefficients $\omega_{s,\theta}$ at distinct scales and orientations:

$$\begin{aligned} \omega_{s,h} &= c_{s-1} - c_{s,h}, \\ \omega_{s,v} &= c_{s-1} - c_{s,v}, \\ \omega_{s,d} &= c_{s-1} - (c_{s,h} \otimes h'_s + \omega_{s,h} + \omega_{s,v}), \\ c_s &= c_{s-1} - (\omega_{s,h} + \omega_{s,v} + \omega_{s,d}). \end{aligned} \tag{5}$$

Here, ω_h, ω_v and ω_d correspond to the coefficients with "horizontal", "vertical" and "diagonal" orientations. Initial $c_0 = I_o$ (e.g. $s = 0$) is obtained from the opponent components ($o = L, rg, by$) and c_n corresponds to the residual plane of the last wavelet component (e.g. $s = n$). The inverse transform is obtained by integrating wavelet coefficients and residual planes:

$$I'_o = \sum_{s=1,\theta=h,v,d}^{n} \omega_{s,\theta} + c_n. \tag{6}$$

Considering that for every image, $M \times N$ is the size of the feature map (resized to $N \leq 128$), the set of spatial scales is ($s = 1..S$), where $S = \lfloor log_2(N/8) \rfloor + 2$.

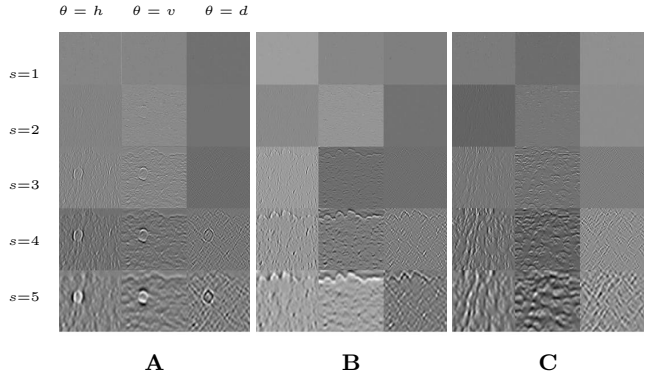


Fig. 3: Output from *à-trous DWT* of the signals shown on Fig. 2. We show values for rescaled wavelet filters, with scales $s = 1..5$ and orientations $\theta = h, v, d$ corresponding to distinct channel opponencies **(A)** $\omega_{o=rg}$, **(B)** $\omega_{o=by}$ and **(C)** $\omega_{o=L}$.

2.2 Computing V1 Dynamics: Feature Conspicuity

Feature conspicuity from previous Murray's SIM model is computed using center-surround feature computations (CS) while applying a contrast sensitivity function (eCSF). Similarly, we extract low-level feature-dependent computations corresponding to the orientation sensitivities ($\theta = 0, 90, 45/135°$) of the retinotopic positions (i) at distinct spatial frequencies (s) for ON and OFF-center cells. These ON and OFF cells activities (before the computation of lateral connections) responses are computed by taking the positive and negative values of the wavelet planes, respectively. Feature distinctiveness is computed with the Penacchio et al. network of excitatory-inhibitory firing rate neurons, simulating V1's lateral interactions (Fig. 4). Contrast enhancement or suppression emerges from lateral connections as an induction mechanism. Lateral interactions are implemented to have self-directed (J_0) and monosynaptic connections (J) between excitatory neurons. Inhibitory interactions have disynaptic connections (W) through all inhibitory interneurons, defined by:

$$J_{[is\theta,js'\theta']} = \lambda(\Delta_s)0.126e^{(-\beta/d_s)^2-2(\beta/d_s)^7-d_s^2/90}, \tag{7}$$

$$W_{[is\theta,js'\theta']} = \lambda(\Delta_s)0.14(1-e^{-0.4(\beta/d_s)^{1.5}})e^{-(\Delta_\theta/(\pi/4))^{1.5}}, \tag{8}$$

Equation 7 is applied if ($0 < d \leq 10$ and $\beta < \pi/2.69$) or [($0 < d \leq 10$ and $\beta < \pi/2.69$) and $|\theta_1| < \pi/5.9$ and $|\theta_2| < \pi/5.9$], otherwise $J_{[is\theta,js'\theta']} = 0$. We take $W_{[is\theta,js'\theta']} = 0$ if $d = 0$ or $d \geq 10$ or $\beta < \pi/1.1$ or $|\Delta\theta| \geq \pi/3$ or $|\theta_1| < \pi/1.99$, otherwise we use the expression in Equation 7. In these equations, $d = d(i, j)$ is the distance between the nodes at position i and j, and θ_1, θ_2 are the angles between the nodes and the line defined by $i - j$, with $|\theta_1| \leq |\theta_2| \leq \pi/2$. The sign of the angles is determined by the condition $|\theta_i| \leq \pi/2$. Parameter $\beta = 2\theta_1 + 2\sin|\theta_1 + \theta_2|$ and $\Delta\theta = \theta - \theta'$ (with $|\theta - \theta' \leq \pi/2$). Term $\lambda(\Delta_s)$ is related to the difference between the spatial scales ($\Delta_s = |s - s'|$) of the two connected nodes. Details of this term can be found on [62, Supporting Information S1].

In Fig.4 C and D we have shown a graphical representation of these connections. Considering these in a simulated retinotopic space (corresponding to a the visual space but at distinct RF sizes) with a radius $\Delta_s = 15 \times 2^{s-1}$ and radial distance Δ_θ (respectively accounting for the distance between RF neurons from different spatial frequencies as d_s and radial distance as β). We can see that excitatory connections J are defined between nodes with similar orientation that are relatively aligned. In contrast, inhibitory connections W are defined between nodes with similar orientation but non-aligned.

Excitatory and inhibitory membrane potentials (their derivatives) are described by

$$\begin{aligned}\dot{x}_{is\theta} = &-\alpha_x x_{is\theta} - g_y(y_{is\theta}) \\ &- \sum_{\Delta_s,\Delta_\theta \neq 0} \Psi(\Delta_s, \Delta_\theta) g_y(y_{is} + \Delta_{s\theta} + \Delta_\theta) + J_0 g(x_{is\theta}) \\ &+ \sum_{j \neq i,s',\theta'} J_{[is\theta,js'\theta']} g_x(x_{js'\theta'}) + I_{is\theta} + I_0,\end{aligned} \tag{9}$$

$$\begin{aligned}\dot{y}_{is\theta} = &-\alpha_y y_{is\theta} - g_x(x_{is\theta}) \\ &+ \sum_{j \neq i,s',\theta'} W_{[is\theta,js'\theta']} g_x(x_{js'\theta'}) + I_c \quad .\end{aligned} \tag{10}$$

Functions g_x and g_y correspond to the activation function (implemented as piece-wise linear functions)

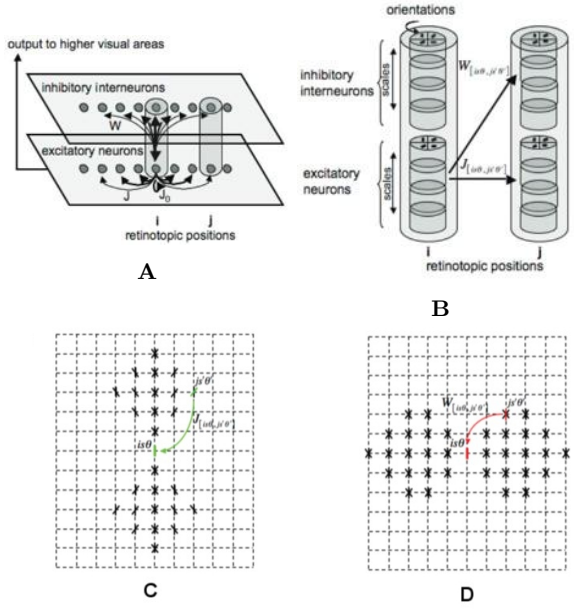


Fig. 4: Illustration of the different elements and their connections that define our computational network. **(A)** Two populations of excitatory and inhibitory nodes are defined in a 2-dimensional regular discrete lattice (here reduced to a single dimension for the sake of clarity). Nodes of these two populations are connected between them. The output of te layer of excitatory nodes is considered the output of the network. **(B)** At each retinotopic position we consider we have an hypercolumn composed by a set excitatory and inhibitory nodes tuned to different spatial orientations and scales. **(C,D)** Sketches of the weights of both the excitatory J and inhibitory W between retinotopic positions i and j. These connections are traslation invariant.

. Reprinted with permission from "A Neurodynamical Model of Brightness Induction in V1", 2013, by O. Penacchio, *PLoS ONE, 8(5):e64086*, p.5. Copyright 2013 by the Public Library of Science [62].

for transforming the membrane potentials to firing rate values. The spread of the inhibitory activity within a hypercolumn is represented as Ψ. Terms $\alpha_x = 1/\tau_x$, $\alpha_y = 1/\tau_y$ are the decay constants that define the decay of excitatory and inhibitory potentials to their resting potential values, respectively. Terms τ_x and τ_y are the mean time that excitatory and inhibitory membrane potentials, respectively, take to decay to its mean value. We have used $\alpha_x = \alpha_y = 1$ values. The variable $I_{is\theta}$ corresponds to the external input values of the image, which in our case are the wavelet coefficients that simulate the response of the classical receptive field of every node ($I_{is\theta} \equiv \omega_{is\theta}$). Inhibitory top-down activity can be introduced to the model through I_c, including a noise signal to stabilize the nonlinear equilibrium. We suggest to read further details of the model and its parameters are specified in [62, Supporting Information S1]. We compute the temporal average of ON and OFF-center cells $M(\omega_{is}^{t+})$ and $M(\omega_{is}^{t-})$ as the model output over several oscillation cycles (being the mean of g_x for a specific range of t, where t is the membrane time, which corresponds to 10 ms) from distinct color opponencies ($o = L, rg, by$). Distinctively from the induction cases described in Unifying an architecture of several visual processes, we do not combine the model output $M(\omega_{iso}^{t})$ to the coefficients ω_{iso}^{t}, instead, we consider the firing rate from the model output as our predictor of feature distinctiveness, which will define our main function for our saliency map (Eq. 11). The model output can provide detail of single neuron dynamics of firing rate, which its dynamical properties may vary across stimulus properties such as color opponency, scale and orientation.

$$\hat{S}_{iso\theta}^{t} = M(\omega_{iso\theta}^{t+}) + M(\omega_{iso\theta}^{t-}) + c_i, \tag{11}$$

2.3 Generating the saliency map: Feature Integration

After computing feature distinctiveness for the low-level feature maps, we need to integrate these conspicuity or distinctiveness maps in order to pool the neuronal activity to the projections of the SC as means of acquiring a unique map, which will represent our saliency map. First, we have computed the inverse transform from the DWT (IDWT) Eq. 6 for integrating the sensitivities for orientation (θ) and spatial frequencies (s). In this case, instead of the $\omega_{s,\theta}$, we use $\hat{S}$ as the sum of ON and OFF cells after processing the dynamical model (Eq. 11) summated for each channel:

$$\hat{S}_{io}(inverse/sum) = \sum_{s=1..S,\theta=h,v,d}^{n} \hat{S}_{iso\theta} + c_n. \tag{12}$$

Second, we have computed the euclidean norm ($\hat{S}$) for integrating the firing rate of the distinct color opponencies (Eq. 13).

$$\hat{S}_i = \sqrt{\hat{S}_{i;rg} + \hat{S}_{i;by} + \hat{S}_{i;L}}, \tag{13}$$

Third, we have normalized the resulting map ($z(\hat{S})$) by the variance of the firing rate (Eq. 14), as stated by Li [91, Chapter 5]. Finally, we convolved the saliency map with a Gaussian filter in order to simulate a smoothing caused by the deviations of $\sigma = 1$ deg given from eye

tracking experimentation, recommended by LeMeur & Baccino [44].

$$z_i(\hat{S}) = \frac{\hat{S}_i - \mu_{\hat{S}}}{\sigma_{\hat{S}}}, \tag{14}$$

where $\mu_{\hat{S}}$ and $\sigma_{\hat{S}}$ are the mean value and the standard deviation of $\hat{S}_i$ over all i pixels, respectively.

3 Experiments

In order to test the validity of our hypothesis, we tested the accuracy of NSWAM for prediction of visual saliency using fixations from eye-tracking experiments. Eye movement data (i.e. ground truth or GT) is combined across all fixations from participants' data, being represented as binary maps (called fixation maps), according to the fixation localizations in the visual space for each corresponding image, or as density distributions (alternatively named density maps) from these fixations considering eye-movement localization probabilities (Fig. 5). Fixation density maps are computed accordingly from fixation maps with a Gaussian filter [44].

Saliency Metrics

Prediction scores are calculated using spatially dependent metrics [14][15] which compare either fixation maps or fixation density maps to saliency map predictions from the models. For the case of AUC, it computes the Area Under ROC considering true positive (TP) values for the saliency predictions inside the locations from the fixation maps and false positive (FP) values for saliency outside the maps. The Normalized Scanpath Saliency (NSS) is calculated by standarizing the saliency map of the TP. Other metrics such as Correlation Coefficient (CC) or Similarity (SIM), compare correlations of pixels between fixation density maps and predicted saliency maps. Also using the fixation density maps as GT, the Kullback-Leibler divergence (KL) measures the statistical difference between the two maps (the density map of GT and the saliency map), therefore the lower score is the better.

Other metrics compare saliency maps with a baseline set of other image fixation maps in order to prevent behavioral tendencies such as center biases (see [4][31]), which are not representative data for saliency prediction. For instance, the shuffled AUC (sAUC) is calculated as the proportion between TP of the current GT and penalizes for TP of GT from other images. For the case of Information Gain (InfoGain) a Gaussian baseline of all GT (adding up fixations for all dataset to one unique map) is substracted from the prediction for penalizing for center biases.

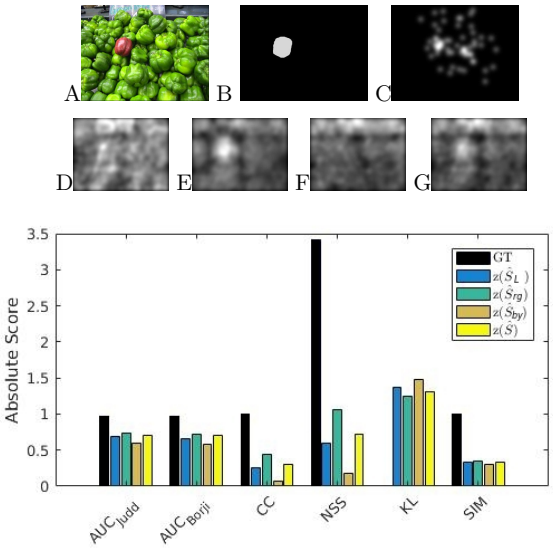


Fig. 5: **(A)** Example Image. **(B)** Mask of the salient region (manually defined). **(C)** Fixation density map (GT, i.e. ground truth) obtained by psychophysical experimentation with observational subjects. **(D,E,F,G)** Predicted saliency maps of the different color opponent channels $z(\hat{S}_L)$, $z(\hat{S}_{rg})$, $z(\hat{S}_{by})$ and the final saliency map $z(\hat{S})$ respectively. **(E)** Comparing the mask and the fixations (both by GT and the computationally predicted) we calculate different metrics from these saliency maps. Results for $z(\hat{S})$ corresponds to our model's saliency prediction (NSWAM). We can see that NSWAM obtain results very similar to other methods.

3.1 Predicting human eye movements in natural images

We have computed the saliency maps[2] for images from distinct eye-tracking datasets, corresponding to 120 real scenes (Toronto) [11], 40 nature scenes (KTH) [39], 100 synthetic patterns ($CAT2000_{Pattern}$)[7] and 230 synthetic images with specific feature contrast (SID4VAM) [4][5]. We have computed these image datasets with deep supervised artificial saliency models that specifically compute high-level features (OpenSalicon [33][76], DeepGazeII [41], SAM [19], SalGan [59]), and models that extract low-level features, corresponding to the cases with artificial (SUN [90], GBVS [**?**]) and biological inspiration (IKN [36], AIM [12], SSR [73], AWS [27]

[2] Code for model evaluations can be downloaded in `https://github.com/dberga/saliency`

and SIM [51]). The Saliency WAvelet Model (SWAM) and Neurodynamic SWAM (NSWAM) corresponds to our model excluding or including lateral interactions explained in Section 2.2.

Our results show that our model performance is similar to other saliency models, outperforming previous Murray's SIM model for the cases of SID4VAM, CAT2000 and KTH (Tables 2, 3 and 4), corresponding to synthetic and nature images, as well as showing stable metric scores for distinct contexts (similarly as AWS and GBVS). NSWAM outperforms SWAM as well as other biologically-inspired models (IKN, AIM, SSR & SIM) specially for metrics that account for center biases. These center biases are qualitatively present even for images where the salient region is conspicuous (Fig. 6, rows 8-9).

Saliency models that compute high-level visual features are shown to perform better with real image scenes (Table 5). However, the image contexts that lack of high-level visual information should be more representative indicators of saliency, due to the absence of semantically or contextually-relevant visual information (nature images), or to be characterized to uniquely contain low-level features (synthetic images) presenting clear pop-out spots to direct participants fixations (which would cause lower inter-participant differences and therefore lower center biases).

Although AWS and GBVS perform better on predicting fixations at distinct contexts, we remark the plausibility of our unified design for modeling distinct HVS' functionality. NSWAM shows a new insight of applying a more biologically plausible computation of the aforementioned steps. First, we transform image values to color opponencies, found in RGC. Second, we model LGN projections to V1 simple cells using a multiresolution wavelet transform. Third, conspicuity is computed with the Penacchio's dynamical model of the lateral interactions between these cells. Fourth, these channels are integrated to a unique map which will represent SC activity. Using a neurodynamic model with firing-rate neurons allows a more detailed understanding of the dependency of saliency on lateral connections and a potential further study in terms of single neuron dynamics using real image scenes.

	method	↑AUC$_{Judd}$	↑AUC$_{Borji}$	↑CC	↑NSS	↓KL	↑SIM	↑sAUC	↑InfoGain
	Humans	**0.943**	**0.882**	**1.000**	**4.204**	**0.000**	**1.000**	**0.860**	**2.802**
HIGH-LEVEL	**OpenSalicon**	0.692	0.673	0.284	0.956	1.549	0.375	0.615	0.052
HIGH-LEVEL	**DeepGazeII**	0.640	0.634	0.177	0.630	1.685	0.336	0.618	-0.150
HIGH-LEVEL	**SAM$_{RESNET}$**	0.727	0.673	0.305	0.967	2.610	0.388	0.600	-1.475
HIGH-LEVEL	**SAM$_{VGG}$**	0.537	0.523	0.026	0.070	11.947	0.216	0.503	-14.954
HIGH-LEVEL	**SalGan**	0.715	0.662	0.287	0.883	2.506	0.373	0.593	-1.350
LOW-LEVEL	**SUN**	0.542	0.532	0.080	0.333	16.408	0.165	0.530	-21.024
LOW-LEVEL	**GBVS**	**0.747**	**0.718**	**0.400**	**1.464**	**1.363**	**0.413**	0.628	**0.331**
LOW-LEVEL	**SSR**	0.672	0.665	0.192	0.639	1.904	0.365	0.642	-0.467
LOW-LEVEL	**AWS**	0.679	0.667	0.255	1.088	1.592	0.373	**0.672**	0.013
LOW-LEVEL	**AIM**	0.570	0.566	0.122	0.473	14.472	0.224	0.557	-18.182
LOW-LEVEL	**IKN**	0.686	0.678	0.283	0.878	1.748	0.380	0.608	-0.233
LOW-LEVEL	**SIM**	0.650	0.641	0.189	0.694	1.702	0.357	0.619	-0.148
LOW-LEVEL	***SWAM (Ours)***	***0.639***	***0.618***	***0.177***	***0.682***	***1.799***	***0.340***	***0.601***	***-0.281***
LOW-LEVEL	***NSWAM (Ours)***	***0.614***	***0.610***	***0.136***	***0.529***	***1.686***	***0.335***	***0.622***	***-0.150***

Table 2: Results for prediction metrics (columns) with SID4VAM dataset [4] with synthetic images for different computational models (rows). An up/down arrow (↑ / ↓) besides a metric name means that the highest/lowest the value of this metric, the better the prediction of the particular method. Best results for every metric is shown in bold and underlined. We can see that the GBVS method is usually the one obtaining the best results. Our models (SWAM and NSWAM) are shown in the last rows in bold and italics.

	method	↑AUC$_{Judd}$	↑AUC$_{Borji}$	↑CC	↑NSS	↓KL	↑SIM	↑sAUC	↑InfoGain
	Humans	**0.895**	**0.826**	**0.890**	**2.335**	**0.265**	**0.736**	**0.623**	**0.777**
HIGH-LEVEL	**OpenSalicon**	0.651	0.621	0.220	0.603	1.526	0.357	0.555	-1.092
HIGH-LEVEL	**DeepGazeII**	0.611	0.561	0.157	0.467	1.932	0.325	0.547	-1.657
HIGH-LEVEL	**SAM$_{RESNET}$**	**0.766**	0.711	**0.518**	**1.356**	1.747	**0.456**	0.546	-1.444
HIGH-LEVEL	**SAM$_{VGG}$**	0.625	0.581	0.123	0.320	8.581	0.322	0.508	-11.262
HIGH-LEVEL	**SalGan**	0.751	0.714	0.417	1.080	1.720	0.430	0.553	-1.384
LOW-LEVEL	**SUN**	0.549	0.539	0.068	0.193	5.860	0.280	0.526	-7.237
LOW-LEVEL	**GBVS**	0.759	**0.717**	0.399	1.056	**1.113**	0.430	0.561	**-0.503**
LOW-LEVEL	**SSR**	0.592	0.582	0.118	0.318	1.760	0.334	0.568	-1.432
LOW-LEVEL	**AWS**	0.604	0.594	0.209	0.609	1.521	0.339	**0.595**	-1.077
LOW-LEVEL	**AIM**	0.570	0.565	0.118	0.332	5.323	0.301	0.544	-6.490
LOW-LEVEL	**IKN**	0.701	0.692	0.323	0.828	1.267	0.382	0.562	-0.724
LOW-LEVEL	**SIM**	0.586	0.578	0.120	0.336	1.614	0.328	0.566	-1.225
LOW-LEVEL	***SWAM (Ours)***	***0.617***	***0.602***	***0.180***	***0.503***	***1.484***	***0.335***	***0.571***	***-1.029***
LOW-LEVEL	***NSWAM (Ours)***	***0.588***	***0.584***	***0.139***	***0.383***	***1.471***	***0.326***	***0.571***	***-1.017***

Table 3: Results for prediction metrics with CAT2000 dataset [7] training subset (Pattern) of uniquely synthetic images. Best results for every metric is shown in bold and underlined

.

	method	↑AUC$_{Judd}$	↑AUC$_{Borji}$	↑CC	↑NSS	↓KL	↑SIM	↑sAUC	↑InfoGain
	Humans	**0.969**	**0.954**	**1.000**	**3.831**	**0.000**	**1.000**	**0.903**	**2.425**
HIGH-LEVEL	**OpenSalicon**	0.821	0.771	0.522	1.655	1.113	0.429	0.716	0.232
HIGH-LEVEL	**DeepGazeII**	0.850	0.768	0.595	1.877	0.997	0.483	**0.717**	**0.422**
HIGH-LEVEL	**SAM$_{RESNET}$**	0.850	0.725	0.612	1.955	2.420	**0.516**	0.666	-1.555
HIGH-LEVEL	**SAM$_{VGG}$**	0.569	0.543	0.055	0.158	11.972	0.214	0.506	-15.522
HIGH-LEVEL	**SalGan**	**0.858**	**0.816**	**0.629**	**1.898**	**0.986**	0.510	0.716	0.387
LOW-LEVEL	**SUN**	0.694	0.682	0.242	0.755	1.589	0.290	0.645	-0.499
LOW-LEVEL	**GBVS**	0.817	0.803	0.487	1.431	1.168	0.397	0.632	0.077
LOW-LEVEL	**SSR**	0.765	0.756	0.364	1.084	1.355	0.340	0.700	-0.174
LOW-LEVEL	**AWS**	0.773	0.761	0.401	1.229	1.322	0.352	0.714	-0.106
LOW-LEVEL	**AIM**	0.727	0.716	0.292	0.883	1.612	0.314	0.663	-0.580
LOW-LEVEL	**IKN**	0.794	0.782	0.421	1.246	1.248	0.366	0.650	-0.024
LOW-LEVEL	**SIM**	0.754	0.744	0.317	0.951	1.486	0.302	0.705	-0.369
LOW-LEVEL	***SWAM (Ours)***	***0.728***	***0.716***	***0.287***	***0.868***	***1.492***	***0.305***	***0.654***	***-0.378***
LOW-LEVEL	***NSWAM (Ours)***	***0.706***	***0.694***	***0.257***	***0.764***	***1.604***	***0.278***	***0.631***	***-0.552***

Table 5: Results for prediction metrics with Toronto dataset [12], corresponding to real (indoor and outdoor) images. Best results for every metric is shown in bold and underlined.

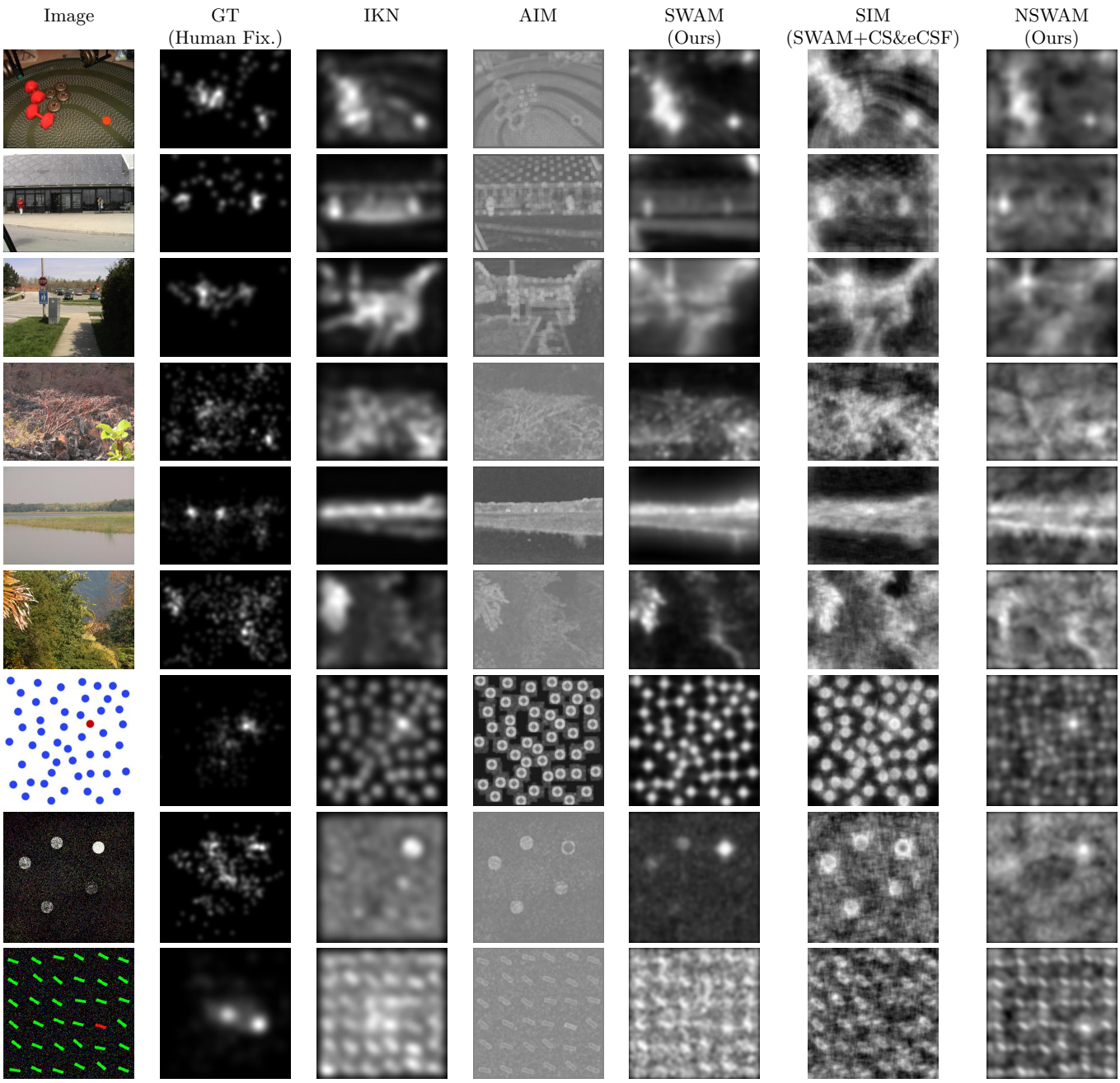


Fig. 6: For images showing distinct contexts (first column), we show the eye-tracking psychophysical experimentation (GT, column 2)
and several examples of saliency maps from Itti et al. (IKN), Bruce & Tsotsos (AIM), Saliency WAvelet Model (SWAM), Murray et al.'s model (SIM) and our Neurodynamic model (columns 5 to 7, respectively).

3.2 Psychophysical study with low-level visual features

In the previous section we studied the accuracy of the computational architecture to predict eye movements for natural images. But one of the open questions is how every low-level visual feature, e.g. contrast, size, orientation, etc, contributes to conspicuity of feature maps. Acknowledging that the HVS process visual information according to the visual context, human performance on detecting a salient object on a scene may also vary according to the visual properties of such object. With a synthetic image dataset [4][5] a specific analysis of how each individual feature influences saliency can be done. In this study we will show how fixation data is predicted when varying feature contrast, concretely on parametrizing Set Size, and Brightness, Color, Size

	method	↑AUC$_{Judd}$	↑AUC$_{Borji}$	↑CC	↑NSS	↓KL	↑SIM	↑sAUC	↑InfoGain
	Humans	**0.902**	**0.850**	**1.000**	**2.038**	**0.000**	**1.000**	**0.822**	**1.415**
HIGH-LEVEL	**OpenSalicon**	0.634	0.611	0.300	0.452	0.780	0.541	0.556	-0.278
	DeepGazeII	0.648	0.618	0.362	**0.578**	**0.678**	0.559	**0.588**	**-0.104**
	SAM$_{RESNET}$	**0.660**	0.599	0.371	0.570	3.125	0.508	0.548	-3.643
	SAM$_{VGG}$	0.525	0.525	0.058	0.074	8.800	0.354	0.501	-11.836
	SalGan	0.655	0.626	**0.391**	0.581	1.666	0.544	0.560	-1.554
LOW-LEVEL	**SUN**	0.535	0.532	0.083	0.132	0.804	0.512	0.526	-0.303
	GBVS	0.649	**0.638**	0.351	0.505	0.711	0.563	0.533	-0.177
	SSR	0.575	0.573	0.172	0.270	0.778	0.525	0.557	-0.260
	AWS	0.587	0.583	0.210	0.329	0.851	0.511	0.581	-0.362
	AIM	0.572	0.568	0.179	0.274	0.918	0.523	0.552	-0.509
	IKN	0.617	0.611	0.274	0.403	0.714	0.547	0.551	-0.173
	SIM	0.587	0.584	0.201	0.311	0.745	0.531	0.573	-0.212
	SWAM (Ours)	***0.601***	***0.596***	***0.231***	***0.346***	***0.749***	***0.529***	***0.574***	***-0.221***
	NSWAM (Ours)	***0.598***	***0.593***	***0.230***	***0.345***	***0.711***	***0.536***	***0.565***	***-0.168***

Table 4: Results for prediction metrics with KTH dataset [39] subset of uniquely nature images. Best results for every metric is shown in bold and underlined

and Orientation contrast between a target salient object Fig. 5B and the rest of distractors (feature singleton search).

In order to quantitatively estimate the accuracy of the computational model predictions, we used the shuffled AUC (sAUC) metric. It computes the area under ROC considering TP as fixations inside the saliency map, similarly to the AUC metric. In contrast to AUC, sAUC does not evaluate FP at random areas of the image but instead uses fixations inside other random images from the same dataset over several trials (10 by default). The sAUC metric gives a more accurate evaluation of predicted maps with respect human fixations but penalizing for higher model center biases (which are or can be present for distinct images in the ground truth).

3.2.1 Brightness differences

Differences in brightness are major factors for making an object to attract attention. That is, a bright object is less salient as luminance of other surround objects increase (Fig. 7). Conversely, a dark target in a bright background will be more salient as surround distractors have higher luminance [61][56]. NSWAM processes luminance signals separately from chromatic ones using the L channel (feature conspicuity from a distinctively bright object upon a dark background will be processed similarly to a dark object upon a bright background). We compare sAUC metrics for both conditions and NSWAM is shown to acquire similar performance to SIM and SWAM, with higher sAUC than IKN Fig. 8,A-B, specially for stimulus with higher contrasts ($\Delta L_{D,T} > .25$). Results on sAUC for NSWAM correlates with brightness contrast, for both cases of bright ($\rho = .941, p = 1.6 \times 10^{-3}$) and dark ($\rho = .986, p = 4.7 \times 10^{-5}$) background.

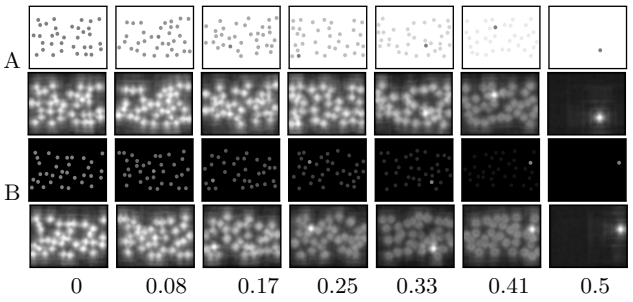


Fig. 7: Synthetic stimuli representing distinct brightness contrasts (HSL luminance differences) from target and distractors ($\Delta L_{D,T}$) with **(A)** bright background ($L_T = 0.5, L_B = 1, L_D = 0.5..1$) and **(B)** dark background ($L_T = 0.5, L_B = 0, L_D = 0..0.5$). Rows below **A,B** are NSWAM predictions.

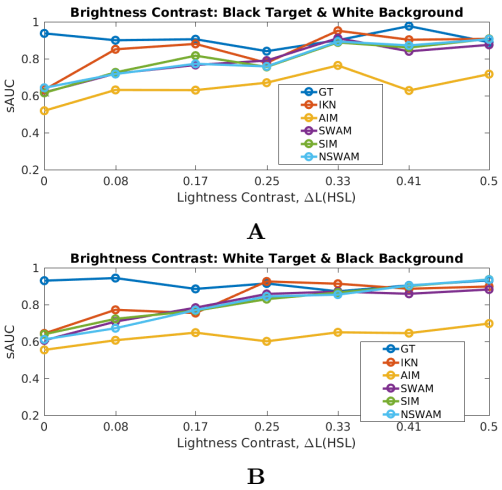


Fig. 8: Results of sAUC upon brightness contrast, $\Delta L_{D,T}$) with **(A)** bright and **(B)** dark background. We can see that our models SWAM, SIM and NSWAM are usually among the best methods.

3.2.2 Color differences

Color changes spatial and temporal behavior of eye movements, influencing conspicuity of specific objects on a scene [23][3]. Similarly to previous section, here we vary the chromaticity of the background, which can alter search efficiency [52][20]. In this section, we used stimuli similar to Rosenholtz's experimentation [69], with red and blue singletons for achromatic or saturated backgrounds Fig. 9. Here, chromatic contrast is defined as the HSL saturation differences ($\Delta S_{D,T}$) between a salient target and the rest of distractors.

Similarly to Fig. 7, NSWAM has similar sAUC to SIM for all background conditions (Fig. 10,A-D). Achromatic backgrounds contribute to salient object detection by increasing sAUC of the pop-out singleton. That

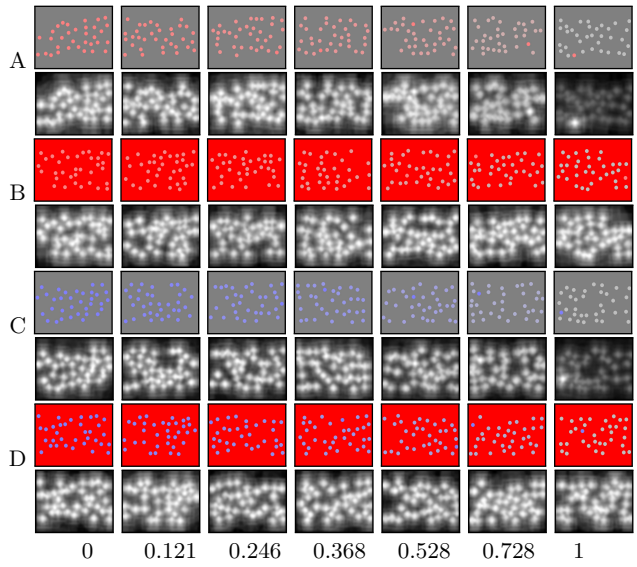


Fig. 9: Chromatic stimuli upon saturation contrast ($\Delta S_{D,T}$) between a red target ($H_T = 0°$) and a **(A)** grey background or a **(B)** saturated red background. Other cases **(C,D)** present a blue target ($H_T = 240°$) with same background properties to **(A)** and **(B)** respectively. Rows below **A-D** correspond to NSWAM's predicted saliency maps.

effect is present for visual search results and our saliency prediction. Results comparing target search fixation maps and sAUC show distinct performance upon saturation contrast depending on background conditions. Cases where stimulus background is achromatic, distinct from the feature singleton, have higher correlation than with saturated background. For the cases of grey (achromatic) background, there is a correlation between sAUC results for our model and $\Delta S_{D,T}$ with a red ($\rho = .864, p = 1.2 \times 10^{-2}$) and blue ($\rho = .944, p = 1.4 \times 10^{-3}$) target singleton. However, when background color is saturated red, while targets are either red ($\rho = .106, p = .82$) or blue ($\rho = .483, p = .27$), then saturation contrast do not correlate with sAUC.

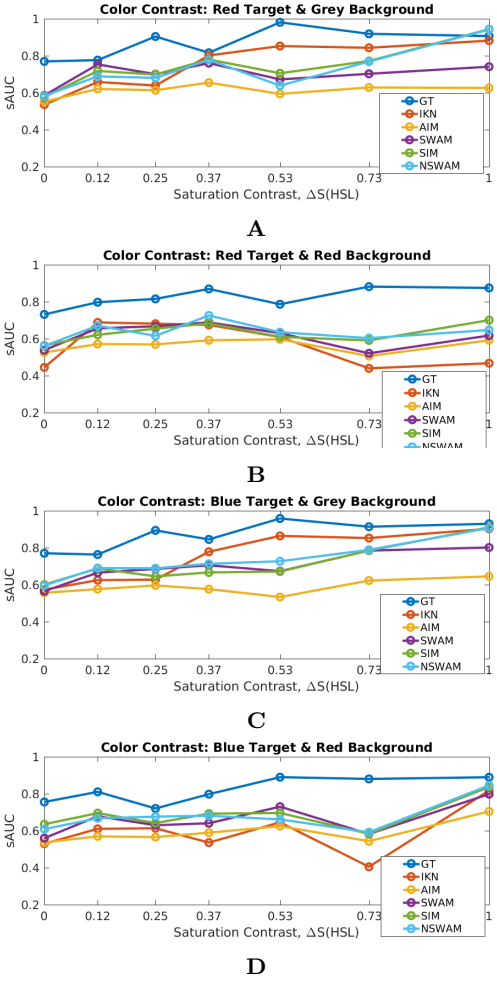


Fig. 10: Results of the sAUC metric upon saturation contrast ($\Delta S_{D,T}$) on a red singleton with **(A)** achromatic or **(B)** saturated red background, or either a blue singleton with **(C)** achromatic or **(D)** saturated red background. We can see that our models SWAM, SIM and NSWAM are usually among the best methods.

3.2.3 Size contrast

Feature distinctiveness using feature singletons have been tested by varying set size, object orientation and/or color. Here, we test how object size affects its saliency, previously tested with visual search experimentation [29][75][66]. A set of 34 symmetric objects (with a dark circle shape) are distributed randomly around the image Fig. 11, preserving equal diameter. One of the circles is defined with dissimilar size, either with higher or lower diameter with respect the rest (which are defined with a diameter of 2.5 deg). Performance for NSWAM's sAUC improves with size dissimilarity. When the diameter of the dissimilar circle is higher, sAUC is higher for that particular region. For the highest scaling fac-

tor (when the dissimilar object is bigger), NSWAM has higher sAUC compared to previous biologically-inspired models (Fig. 12). In addition, there is a significant correlation between circle diameter and our model's results of sAUC ($\rho = .955, p = 8.3 \times 10^{-4}$).

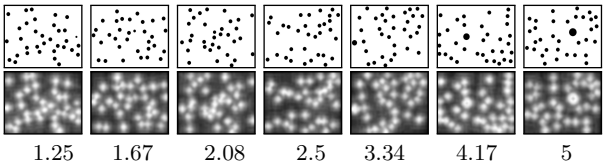


Fig. 11: Examples of circle distractors with equal diameter ($\varnothing_D = 2.5\,\text{deg}$), containing a salient one with dissimilar size ($\varnothing_T = 1.25..5\,\text{deg}$) with respect the rest. In lower row there are NSWAM's predicted saliency maps.

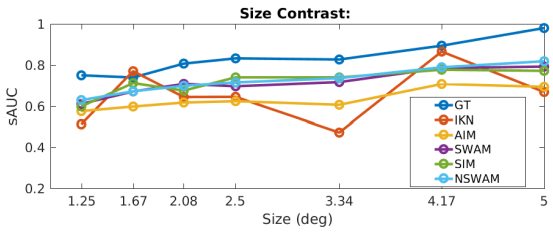


Fig. 12: Results of the sAUC metric for Size Contrast stimuli. We can see that our models SWAM, SIM and NSWAM are usually among the best methods.

3.2.4 Orientation contrast

Using visual stimuli defined by oriented bars, varying angle of objects is found to increase search efficiency when angle contrast is increased [22][55][54]. A total of 34 bars were oriented horizontally and randomly displaced around the scene (Fig. 13). The dissimilar object for this case is a bar oriented with an angle contrast with respect the rest of bars of $\Delta\Phi(1,0)$=[0, 10, 20, 30, 42, 56, 90]°. Although results of sAUC show that NSWAM overperforms SIM's saliency maps, IKN is best for capturing orientation distinctiveness (Fig. 14). In NSWAM, 3 types of orientation selective cells are modeled, corresponding to the orientation for the wavelet coefficients ($\theta = h, v, d$). A higher number of orientation selective cells would provide a higher accuracy, specially for diagonal angles (here we only provide $\theta = d$ for 45/135° combined). By modeling orientation selective cells with 2D Gabor and Log-Gabor transforms [43][25][27] it would be possible to correctly build an hypercolumnar organization with a higher number of angle sensitivities.

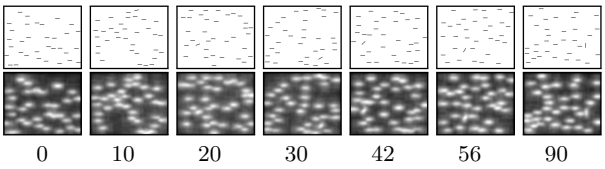


Fig. 13: An oriented bar with an orientation contrast of $\Delta\Phi = 0..90°$ with respect to a set of bars oriented at $\Phi_D = 0°$. In lower row there are NSWAM's predicted saliency maps.

We have to acknowledge that for this experimentation, distractors have been set with same horizontal configuration. Specific connectivity interactions [2] between orientation dissimilarities needs to be defined in order to reproduce orientation-dependent visual illusions and conspicuity under heterogeneous, nonlinear and categorical angle configurations (seen to be performed by V2 cells [1]), which are previously known to distinctively affect visual attention [55][54][26].

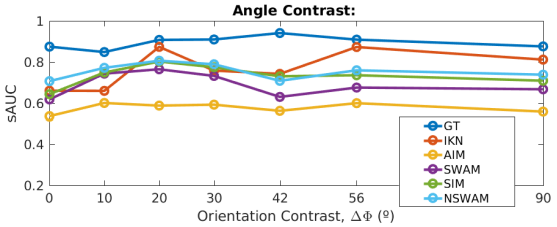


Fig. 14: Results for sAUC metric for Orientation Contrast stimuli.

3.2.5 Visual Asymmetries

Search asymmetries appear when searching target of type "a" is found efficiently among distractors of type "b", but not in the opposite case (i.e. searching for "b" among distractors of type "a") [77][86]. Previous studies pointed out this concept when searching a circle crossed by a vertical bar among plain circles and searching a plain circle among circles crossed by a vertical bar. Using these two configurations, we filled a grid of distractors according to specific scales (Fig. 15). Scale values ($s = [1.25, 1.67, 2.08, 2.5, 3.33, 4.17, 5]\,\text{deg}$) change the amount of items, with arrays of 5×7, 6×8, 8×10, 10×13, 15 × 20 and 20 × 26 objects. In Fig. 16 our model is not only more efficient than other biologically-inspired

models upon dissimilar sized objects but also on detecting conspicuous objects at distinct scales, accounting for lower or larger amount of distractors. sAUC for NSWAM showed to correlate for a conspicuous circle crossed by a vertical bar among circles ($\rho = .83$, $p = 2.1 \times 10^{-2}$) but not for a conspicuous circle among circles crossed by a vertical bar ($\rho = .15$, $p = .75$).

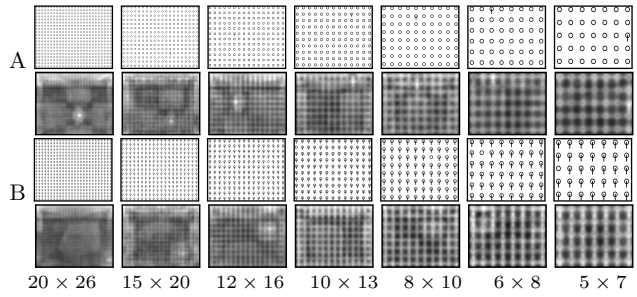


Fig. 15: Stimuli with distinct set sizes corresponding to search asymmetries present on a **(A)** salient circle crossed by a vertical bar among other circles and a **(B)** salient circle among other circles crossed by a vertical bar. Rows below **A,B** correspond to NSWAM's predicted saliency maps.

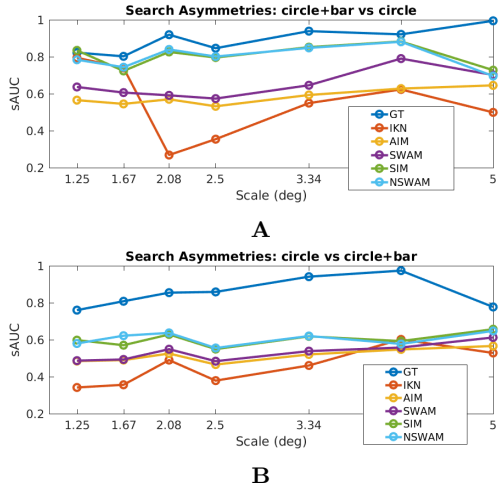


Fig. 16: Results of sAUC upon varying scale and set size of **(A)** an array of circles and a salient one crossed by a vertical bar and **(B)** an array of circles crossed by a bar and a salient circle.

3.3 Ablation Study of Feature Integration

In this section, we include some brief results testing how distinct fusion methods can efficiently integrate information to the unique saliency map (mainly what we consider as the SC function). As mentioned in the first step of Section 2.3, our original model (SWAM and NSWAM) uses the default inverse equation (Eq. 6) which can be used for obtaining the original image (if we do not sum the conspicuity maps and normalize as the other steps). For this, we tested distinct mechanisms of Feature Integration, alternative to Eq. 12:

$$\hat{S}_{io}(max) = \max_{s,\theta}(\hat{S}_{iso\theta}) + c_n. \tag{15}$$

$$\hat{S}_{io}(argmax) = arg\max_{s,\theta}(\hat{S}_{iso\theta}) + c_n \tag{16}$$

Where $\hat{S}_{io}(max)$ calculates the pointwise maximum of the retinotopic positions "i" in each scale "s" and orientation "θ", separately for each channel "o" (rg, by, L). The case of $\hat{S}_{io}(argmax)$ considers the "winner" as the whole channel map $\hat{S}_{io}$ that contains the neuron with highest activity for all multiscale dimensions (s, θ). We have computed the eye fixation prediction results for the all datasets in Figure 6, and results show that performing the inverse transform (sum) of all maps we get best scores. In addition, we added qualitative results for Fig. 17, with 3 examples of real, nature and synthetic images (being the inverse more similar overall to the GT).

Dataset	Integration	↑AUC$_{Judd}$	↑AUC$_{Borji}$	↑CC	↑NSS	↓KL	↑SIM	↑sAUC
TORONTO	inverse	**0.706**	**0.694**	**0.257**	**0.764**	**1.604**	0.278	**0.631**
	max	0.672	0.661	0.199	0.603	1.645	0.272	0.624
	argmax	0.664	0.649	0.178	0.548	1.714	**0.281**	0.618
KTH	inverse	**0.598**	**0.593**	**0.230**	**0.345**	**0.711**	**0.536**	**0.565**
	max	0.553	0.549	0.101	0.160	0.794	0.507	0.552
	argmax	0.557	0.553	0.138	0.204	0.785	0.517	0.537
CAT2000-Pattern	inverse	**0.588**	**0.584**	**0.139**	**0.383**	**1.471**	**0.326**	**0.571**
	max	0.521	0.517	0.016	0.074	1.599	0.301	0.552
	argmax	0.572	0.564	0.091	0.247	1.509	0.320	0.546
SID4VAM	inverse	0.614	**0.610**	0.136	**0.529**	**1.686**	0.335	**0.622**
	max	0.575	0.570	0.078	0.309	1.767	0.320	0.586
	argmax	**0.626**	0.609	**0.183**	0.528	1.704	**0.338**	0.575

Table 6: Results for prediction metrics for testing distinct baselines (inverse/sum, max and argmax), corresponding to mechanisms explained in Eqs. 12, 15 and 16.

4 Conclusions

In this work, we hypothesize that low-level saliency is likely to be associated by the computations of V1. Concretely, we hypothesized that a neurodynamic model of V1's lateral interactions, processing each channel separately and acquiring firing rate dynamics from real image simulations, is able to simultaneously reproduce several visual processes, including low-level visual saliency.

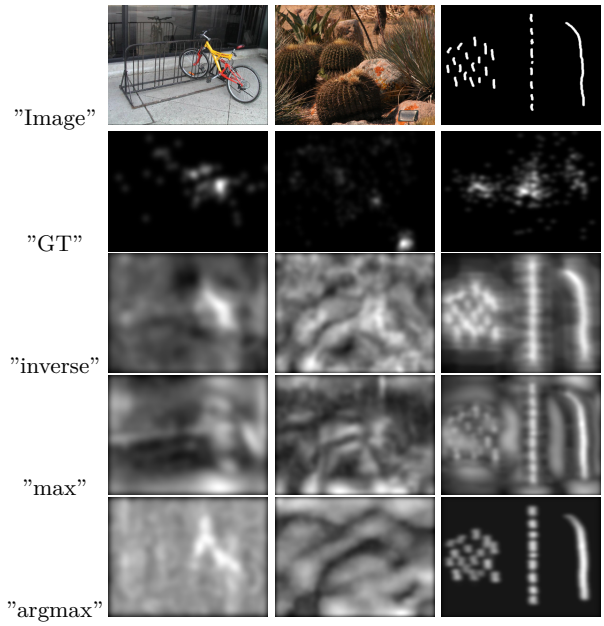


Fig. 17: Qualitative examples for distinct Feature Integration techniques (rows 3-5), with real, nature and synthetic images (columns 1-3).

Here we have to pinpoint three statements in agreement with our findings:

- First, our model of the lateral interactions in V1 show a performance similar to other state-of-the-art models on human eye fixations. In that sense, our model acquires similar results in comparison with saliency prediction baselines, specifically in metrics that penalize for center biases (sAUC and InfoGain). Additionally, our model outperforms other biologically-inspired saliency models in natural and synthetic images.
- Second, our model is consistent with human psychophysical measurements (tested for Visual Asymmetries, Brightness, Color, Size and Orientation contrast). Adding up to the stated hypothesis, our model presents highest performance at highest contrast from feature singleton stimuli (where salient objects pop-out easily).
- Three, we remark the model plausibility by mimicking HVS physiology on its processing steps and being able to reproduce other effects such as Brightness Induction [62], Color Induction [16] and Visual Discomfort [64], efficiently working without applying any type of training or optimization and keeping the same parametrization.

Other biologically plausible alternatives that predict attention using neurodynamic modeling [45][21][17] do not provide a unified model of the visual cortex able to reproduce these distinct tasks simultaneously, and specifically, using real static or dynamic images as input. We suggest that V1 computations work as a common substrate for several tasks, simultaneously.

Future work of interest would consist on predicting scan-paths for real scenes in order to provide gaze-wise temporal detail for saliency prediction and saccade programming. To do so, a foveation mechanism (such as a retinal [81] or a cortical magnification transformation towards V1 retinotopy [72]) would be needed in order to process each view of the scene distinctively. Other applications of the same model would be to generate saliency maps with dynamic scenes or videos (mainly used for visual tracking and salient object detection in several real world applications), integrating other features such as flicker or motion. In order to provide top-down computations for representing feature relevance apart from saliency, we could feed our model with a selective mechanism [78][32] for specific low-level feature maps, enabling the possibility to perform visual search tasks. As shown in Section 3.2.4, saliency computations could be more accurately represented with a higher number of 2D Gabor/Log-Gabor filters [43][25][27]. Considering the dependence of saliency to stimulus contrast, the usage of contrast-adaptive mechanisms [65] in the Feature Integration step (2.3) could dramatically improve results, specially for psychophysical pattern images. Further modeling would include intra and inter-cortical interactions between simple and complex cells in a multilayer implementation of V1. Such implementation could adequate more detailed and efficient computations of V1, projecting the excitatory recurrent dynamics from V1 (specifically from Layer 5 complex cells, also named "Meynert" cells) to SC [48][53]. Although latest hypotheses about the SC have suggested that saliency is processed in the SC and not by the visual cortex, corresponding to a distinct, feature-agnostic saliency map [79][84], we claim the importance of the mechanisms of V1 to be responsible for computing distinctiveness between the stated low-level features, which might conjunctively contribute to the generation of saliency [46][47][87]. However, modeling the computations of the pathways from the RGC to the SC would be of interest for a more integrated and complete model of eye-movement prediction, seeing the roles of the distinct projections to the SC and their computations, alternatively involved in the control of eye movements.

5 Compliance with Ethical Standards

Funding: This work was funded by the Spanish Ministry of Economy and Competitivity (DPI2017-89867-C2-1-R), Agencia de Gestió d'Ajuts Universitaris i de

Recerca (AGAUR) (2017-SGR-649), and CERCA Programme / Generalitat de Catalunya.

Conflict of Interests:The authors declare that they have no conflict of interest.

Informed Consent:Informed consent was not required as no human or animals were involved.

Human and Animal Rights:This article does not contain any studies with human or animal subjects performed by any of the authors.

References

1. A. Anzai, X. Peng, and D. C. V. Essen. Neurons in monkey visual area v2 encode combinations of orientations. *Nature Neuroscience*, 10(10):1313–1321, sep 2007.
2. M. A. Asenov. Dynamic model of interactions between orientation selective neurons in primary visual cortex. Master's thesis, University of Edinburg, Edinburgh, UK, 2016.
3. B. Bauer, P. Jolicoeur, and W. B. Cowan. Distractor heterogeneity versus linear separability in colour visual search. *Perception*, 25(11):1281–1293, nov 1996.
4. D. Berga, X. R. Fdez-Vidal, X. Otazu, V. LeborÃąn, and X. M. Pardo. Psychophysical evaluation of individual low-level feature influences on visual attention. *Vision Research*, 154:60 – 79, 2019.
5. D. Berga, X. R. Fdez-Vidal, X. Otazu, and X. M. Pardo. Sid4vam: A benchmark dataset with synthetic images for visual attention modeling. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 8789–8798, 2019.
6. A. Borji and L. Itti. State-of-the-art in visual attention modeling. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(1):185–207, jan 2013.
7. A. Borji and L. Itti. Cat2000: A large scale fixation dataset for boosting saliency research. *CVPR 2015 workshop on "Future of Datasets"*, 2015. arXiv preprint arXiv:1505.03581.
8. A. 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.
9. A. Borji, H. R. Tavakoli, D. N. Sihite, and L. Itti. Analysis of scores, datasets, and models in visual saliency prediction. In *2013 IEEE International Conference on Computer Vision*. IEEE, dec 2013.
10. N. D. Bruce, C. Wloka, N. Frosst, S. Rahman, and J. K. Tsotsos. On computational modeling of visual saliency: Examining what's right, and what's left. *Vision Research*, 116:95–112, nov 2015.
11. N. D. B. Bruce and J. K. Tsotsos. Saliency based on information maximization. In *Proceedings of the 18th International Conference on Neural Information Processing Systems*, NIPS'05, pages 155–162, Cambridge, MA, USA, 2005. MIT Press.
12. N. D. B. Bruce and J. K. Tsotsos. Saliency, attention, and visual search: An information theoretic approach. *Journal of Vision*, 9(3):5–5, mar 2009.
13. Z. Bylinskii, E. DeGennaro, R. Rajalingham, H. Ruda, J. Zhang, and J. Tsotsos. Towards the quantitative evaluation of visual attention models. *Vision Research*, 116:258–268, nov 2015.
14. Z. Bylinskii, T. Judd, F. Durand, A. Oliva, and A. Torralba. Mit saliency benchmark. http://saliency.mit.edu/.
15. Z. Bylinskii, T. Judd, A. Oliva, A. Torralba, and F. Durand. What do different evaluation metrics tell us about saliency models?, 2016.
16. X. Cerda and X. Otazu. A Multi-Task Neurodynamical Model of Lateral Interactions in V1: Chromatic Induction. *39th European Conference of Visual Perception, PERCEPTION*, 45(2):51, 2016.
17. S. Chevallier, N. Cuperlier, and P. Gaussier. Efficient neural models for visual attention. In *Computer Vision and Graphics*, pages 257–264. Springer Berlin Heidelberg, 2010.
18. M. Cornia, L. Baraldi, G. Serra, and R. Cucchiara. A Deep Multi-Level Network for Saliency Prediction. In *International Conference on Pattern Recognition (ICPR)*, 2016.
19. M. Cornia, L. Baraldi, G. Serra, and R. Cucchiara. Predicting human eye fixations via an lstm-based saliency attentive model, 2016.
20. M. V. Danilova and J. D. Mollon. Symmetries and asymmetries in chromatic discrimination. *Journal of the Optical Society of America A*, 31(4):A247, feb 2014.
21. G. Deco and E. T. Rolls. A neurodynamical cortical model of visual attention and invariant object recognition. *Vision Research*, 44(6):621–642, mar 2004.
22. J. Duncan and G. W. Humphreys. Visual search and stimulus similarity. *Psychological Review*, 96(3):433–458, 1989.
23. M. D'Zmura. Color in visual search. *Vision Research*, 31(6):951–966, jan 1991.
24. H. E. Egeth and S. Yantis. VISUAL ATTENTION: Control, representation, and time course. *Annual Review of Psychology*, 48(1):269–297, feb 1997.
25. S. Fischer, F. Šroubek, L. Perrinet, R. Redondo, and G. Cristóbal. Self-invertible 2d log-gabor wavelets. *International Journal of Computer Vision*, 75(2):231–246, jan 2007.
26. D. Gao. *A discriminant hypothesis for visual saliency: computational principles, biological plausibility and applications in computer vision.* PhD thesis, UC San Diego, 2008.
27. A. Garcia-Diaz, X. R. Fdez-Vidal, X. M. Pardo, and R. Dosil. Saliency from hierarchical adaptation through decorrelation and variance normalization. *Image and Vision Computing*, 30(1):51–64, jan 2012.
28. M. González-Audícana, X. Otazu, O. Fors, and A. Seco. Comparison between mallat's and the 'à trous' discrete wavelet transform based algorithms for the fusion of multispectral and panchromatic images. *International Journal of Remote Sensing*, 26(3):595–614, feb 2005.
29. P. Goolkasian. Size scaling and spatial factors in visual attention. *The American Journal of Psychology*, 110(3):397, 1997.
30. J. Harel, C. Koch, and P. Perona. Graph-based visual saliency. *Proc. Advances in Neural Information Processing Systems (NIPS 2007)*, 19:545–552, 2007.
31. T. R. Hayes and J. M. Henderson. Center bias outperforms image salience but not semantics in accounting for attention during scene viewing. *Attention, Perception, & Psychophysics*, Aug. 2019.
32. L. Huang and H. Pashler. A boolean map theory of visual attention. *Psychological Review*, 114(3):599–631, 2007.
33. X. Huang, C. Shen, X. Boix, and Q. Zhao. SALICON: Reducing the semantic gap in saliency prediction by adapting deep neural networks. In *2015 IEEE International Conference on Computer Vision (ICCV)*. IEEE, dec 2015.

34. D. H. Hubel and T. N. Wiesel. Receptive fields and functional architecture of monkey striate cortex. *The Journal of Physiology*, 195(1):215–243, mar 1968.
35. L. Itti. Visual salience. *Scholarpedia*, 2(9):3327, 2007.
36. L. Itti, C. Koch, and E. 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.
37. T. Judd, F. Durant, and A. Torralba. A benchmark of computational models of saliency to predict human fixations. *CSAIL Technical Reports*, jan 2012.
38. C. Koch and S. Ullman. Shifts in selective visual attention: Towards the underlying neural circuitry. In *Matters of Intelligence*, pages 115–141. Springer Netherlands, 1987.
39. G. Kootstra, B. de Boer, and L. R. B. Schomaker. Predicting eye fixations on complex visual stimuli using local symmetry. *Cognitive Computation*, 3(1):223–240, jan 2011.
40. N. Kumar, H. K. Sardana, S. N. Shome, and N. Mittal. Saliency subtraction inspired automated event detection in underwater environments. *Cognitive Computation*, 12(1):115–127, Aug. 2019.
41. M. KÃijmmerer, T. S. A. Wallis, and M. Bethge. Deepgaze ii: Reading fixations from deep features trained on object recognition, 2016.
42. A. T. Le, J. Payne, C. Clarke, M. A. Kelly, F. Prudenziati, E. Armsby, O. Penacchio, and A. J. Wilkins. Discomfort from urban scenes: Metabolic consequences. *Landscape and Urban Planning*, 160:61–68, apr 2017.
43. T. S. Lee. Image representation using 2d gabor wavelets. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 18(10):959–971, 1996.
44. O. LeMeur and T. Baccino. Methods for comparing scanpaths and saliency maps: strengths and weaknesses. *Behavior Research Methods*, 45(1):251–266, jul 2012.
45. Z. Li. A neural model of contour integration in the primary visual cortex. *Neural Computation*, 10(4):903–940, may 1998.
46. Z. Li. Contextual influences in v1 as a basis for pop out and asymmetry in visual search. *Proceedings of the National Academy of Sciences*, 96(18):10530–10535, aug 1999.
47. Z. Li. A saliency map in primary visual cortex. *Trends in Cognitive Sciences*, 6(1):9–16, jan 2002.
48. J. S. Lund and R. G. Boothe. Interlaminar connections and pyramidal neuron organisation in the visual cortex, area 17, of the macaque monkey. *The Journal of Comparative Neurology*, 159(3):305–334, feb 1975.
49. L. Maffei and A. Fiorentini. The visual cortex as a spatial frequency analyser. *Vision Research*, 13(7):1255–1267, jul 1973.
50. P. Monnier and S. K. Shevell. Chromatic induction from s-cone patterns. *Vision Research*, 44(9):849–856, apr 2004.
51. N. Murray, M. Vanrell, X. Otazu, and C. A. Parraga. Saliency estimation using a non-parametric low-level vision model. In *CVPR 2011*. IEEE, jun 2011.
52. A. L. Nagy. Interactions between achromatic and chromatic mechanisms in visual search. *Vision Research*, 39(19):3253–3266, oct 1999.
53. H. L. Nhan and E. M. Callaway. Morphology of superior colliculus- and middle temporal area-projecting neurons in primate primary visual cortex. *The Journal of Comparative Neurology*, 520(1):52âĂŞ80, Nov 2011.
54. H.-C. Nothdurft. The conspicuousness of orientation and motion contrast. *Spatial Vision*, 7(4):341–363, jan 1993.
55. H.-C. Nothdurft. The role of features in preattentive vision: Comparison of orientation, motion and color cues. *Vision Research*, 33(14):1937–1958, sep 1993.
56. H.-C. Nothdurft. Salience from feature contrast: additivity across dimensions. *Vision Research*, 40(10-12):1183–1201, jun 2000.
57. X. Otazu, C. A. Parraga, and M. Vanrell. Toward a unified chromatic induction model. *Journal of Vision*, 10(12):5–5, oct 2010.
58. X. Otazu, M. Vanrell, and C. A. Párraga. Multiresolution wavelet framework models brightness induction effects. *Vision Research*, 48(5):733–751, feb 2008.
59. J. Pan, C. Canton, K. McGuinness, N. E. O'Connor, J. Torres, E. Sayrol, and X. a. Giro-i Nieto. Salgan: Visual saliency prediction with generative adversarial networks. In *arXiv*, January 2017.
60. C. A. Parraga, G. Brelstaff, T. Troscianko, and I. R. Moorhead. Color and luminance information in natural scenes. *Journal of the Optical Society of America A*, 15(3):563–569, 1998.
61. H. Pashler, K. Dobkins, and L. Huang. Is contrast just another feature for visual selective attention? *Vision Research*, 44(12):1403–1410, jun 2004.
62. O. Penacchio, X. Otazu, and L. Dempere-Marco. A neurodynamical model of brightness induction in v1. *PLoS ONE*, 8(5):e64086, may 2013.
63. O. Penacchio and A. J. Wilkins. Visual discomfort and the spatial distribution of fourier energy. *Vision Research*, 108:1–7, mar 2015.
64. O. Penacchio, A. J. Wilkins, X. Otazu, and J. M. Harris. Inhibitory function and its contribution to cortical hyperexcitability and visual discomfort as assessed by a computation model of cortical function. *39th European Conference of Visual Perception, PERCEPTION*, 45(2):51, 2016.
65. F. J. Poirier and H. R. Wilson. A biologically plausible model of human radial frequency perception. *Vision Research*, 46(15):2443–2455, July 2006.
66. M. J. Proulx. Size matters: Large objects capture attention in visual search. *PLoS ONE*, 5(12):e15293, dec 2010.
67. N. Riche, M. Duvinage, M. Mancas, B. Gosselin, and T. Dutoit. Saliency and human fixations: State-of-the-art and study of comparison metrics. In *2013 IEEE International Conference on Computer Vision*. IEEE, dec 2013.
68. N. Riche and M. 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.
69. R. Rosenholtz, A. L. Nagy, and N. R. Bell. The effect of background color on asymmetries in color search. *Journal of Vision*, 4(3):9, mar 2004.
70. P. H. Schiller, M. Stryker, M. Cynader, and N. Berman. Response characteristics of single cells in the monkey superior colliculus following ablation or cooling of visual cortex. *Journal of Neurophysiology*, 37(1):181–194, jan 1974.
71. P. H. Schiller and E. J. Tehovnik. *Vision : from neurons to cognition*. Elsevier Science, Amsterdam New York, 2001.
72. E. L. Schwartz. Computational anatomy and functional architecture of striate cortex: A spatial mapping approach to perceptual coding. *Vision Research*, 20(8):645–669, jan 1980.
73. H. J. Seo and P. Milanfar. Static and space-time visual saliency detection by self-resemblance. *Journal of Vision*, 9(12):15–15, nov 2009.
74. T. Stathaki. *Image fusion : algorithms and applications*. Academic Press/Elsevier, Amsterdam Boston, 2008.
75. A. Tavassoli, I. van der Linde, A. Bovik, and L. Cormack. Eye movements selective for spatial frequency and orientation during active visual search. *Vision Research*, 49(2):173–181, jan 2009.

76. C. L. Thomas. Opensalicon: An open source implementation of the salicon saliency model. Technical Report TR-2016-02, University of Pittsburgh, 2016.
77. A. Treisman and J. Souther. Search asymmetry: A diagnostic for preattentive processing of separable features. *Journal of Experimental Psychology: General*, 114(3):285–310, 1985.
78. J. K. Tsotsos, S. M. Culhane, W. Y. K. Wai, Y. Lai, N. Davis, and F. Nuflo. Modeling visual attention via selective tuning. *Artificial Intelligence*, 78(1-2):507–545, oct 1995.
79. R. Veale, Z. M. Hafed, and M. 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.
80. N. Wade. *The art and science of visual illusions*. Routledge & Kegan Paul, London Boston, 1982.
81. A. B. Watson. A formula for human retinal ganglion cell receptive field density as a function of visual field location. *Journal of Vision*, 14(7):15–15, jun 2014.
82. B. White and D. P. Munoz. *The Oxford Handbook of Eye Movements*. Oxford University Press, aug 2011.
83. B. J. White, D. J. Berg, J. Y. Kan, R. A. Marino, L. Itti, and D. P. Munoz. Superior colliculus neurons encode a visual saliency map during free viewing of natural dynamic video. *Nature Communications*, 8:14263, jan 2017.
84. B. J. White, J. Y. Kan, R. Levy, L. Itti, and D. P. Munoz. Superior colliculus encodes visual saliency before the primary visual cortex. *Proceedings of the National Academy of Sciences*, 114(35):9451–9456, aug 2017.
85. M. White. A new effect of pattern on perceived lightness. *Perception*, 8(4):413–416, aug 1979.
86. J. M. Wolfe. Asymmetries in visual search: An introduction. *Perception & Psychophysics*, 63(3):381–389, apr 2001.
87. Y. Yan, L. Zhaoping, and W. Li. Bottom-up saliency and top-down learning in the primary visual cortex of monkeys. *Proceedings of the National Academy of Sciences*, page 201803854, sep 2018.
88. J. Zhan, H. Zhao, P. Zheng, H. Wu, and L. Wang. Salient superpixel visual tracking with graph model and iterative segmentation. *Cognitive Computation*, June 2019.
89. L. Zhang and W. Lin. *Selective Visual Attention*. John Wiley & Sons (Asia) Pte Ltd, mar 2013.
90. L. Zhang, M. H. Tong, T. K. Marks, H. Shan, and G. W. Cottrell. SUN: A bayesian framework for saliency using natural statistics. *Journal of Vision*, 8(7):32, dec 2008.
91. L. Zhaoping. *Understanding vision : theory, models, and data*. Oxford University Press, Oxford, United Kingdom, 2014.