Modeling Bottom-Up and Top-Down Attention with a Neurodynamic Model of V1

David Berga[a,*], Xavier Otazu[a,b]

[a]*Computer Vision Center, Edifici O, Campus UAB, 08193 Bellaterra (Barcelona), Spain*
[b]*Universitat Autònoma de Barcelona, Dept. Ciències de la Computació, Edifici Q, Campus UAB, 08193 Bellaterra (Barcelona), Spain*

ARTICLE INFO

Keywords:
scanpath
saccade
attention
search
firing-rate
V1

ABSTRACT

Previous studies suggested that lateral interactions of V1 cells are responsible, among other visual effects, of bottom-up visual attention (alternatively named visual salience or saliency). Our objective is to mimic these connections with a neurodynamic network of firing-rate neurons in order to predict visual attention. Early visual subcortical processes (i.e. retinal and thalamic) are functionally simulated. An implementation of the cortical magnification function is included to define the retinotopical projections towards V1, processing neuronal activity for each distinct view during scene observation. Novel computational definitions of top-down inhibition (in terms of inhibition of return and selection mechanisms), are also proposed to predict attention in Free-Viewing and Visual Search tasks. Results show that our model outpeforms other biologically-inpired models of saliency prediction while predicting visual saccade sequences with the same model. We also show how temporal and spatial characteristics of inhibition of return can improve prediction of saccades, as well as how distinct search strategies (in terms of feature-selective or category-specific inhibition) can predict attention at distinct image contexts.

1. Introduction

The human visual system (HVS) structure has evolved in a way to efficiently discriminate redundant information [1, 2, 3]. In order to filter or select the information to be processed in higher areas of visual processing in the brain, the HVS guides eye movements towards regions that appear to be visually conspicuous or distinct in the scene. This phenomenon was observed during visual search tasks [4, 5], where detecting early visual features (such as orientation, color or size) was done in parallel (pre-attentively) or required either a serial "binding" step depending on scene context. Koch & Ullman [6] came up with the hypothesis that neuronal mechanisms involved in selective visual attention generate a unique "master" map from visual scenes, coined with the term "saliency map". From that, Itti, Koch & Niebur [7] presented a computational implementation of the aforementioned framework (IKN), inspired by the early mechanisms of the HVS. It was done by extracting properties of the image as feature maps (using a pyramid of difference-of-gaussian filters at distinct orientations, color and intensity), obtaining feature-wise conspicuity by computing center-surround differences as receptive field responses and integrating them on a unique map using winner-take-all mechanisms. Such framework served as a starting point for saliency modeling [8, 9], which derived in a myriad of computational models, that differed in their computations but conserved a similar pipeline. From a biological perspective, further hypotheses suggested that primates' visual system structure was mainly connected to the efficient coding principle. Later studies considered that maximizing information of scenes was the key factor on forming visual feature representations. To test that, Bruce & Tsotsos [10] implemented a saliency model (AIM) by extracting sparse representations of image statistics (using independent component analysis). These representations were found to be remarkably similar to cells in V1, which follow similar spatial properties to Gabor filters [11].

While the current concept of saliency maps is to predict probabilities of specific spatial locations as candidates of eye movements, it is also crucial to understand how to predict individual fixations or saccade sequences (also named "scanpaths"). Scanpath predictions were formerly done through probabilistic measures of saccade amplitude statistics. These followed a similar heavy-tailed distribution to a Cauchy-Levy (in reference to random walks or "Levy flights", minimizing global uncertainty) [12], with highest probability of fixations at a low saccade amplitude. This procedure was implemented in Boccignone & Ferraro's model [13], taking saliency from IKN. Later, LeMeur & Liu [14] proposed

*Corresponding author
dberga@cvc.uab.es (D. Berga)
ORCID(s): 0000-0001-7543-2770 (D. Berga); 0000-0002-4982-791X (X. Otazu)

a more biologically-plausible approach, accounting for oculomotor biases and inhibition of return effects. It used a graph-based saliency model (GBVS, also inspired by IKN) [15], with a higher probability to catch grouped fixations (which tend to be in stimulus center).

In order to evaluate model predictions with eye movement data, certain patterns underlying human eye movement behavior need to be accounted for a more detailed description and analysis of visual attention. These effects are found to be dependent on context, discriminability, temporality, task and memory during scene viewing and visual search [16, 17]. Attention and spatial selection, therefore, is also dependent on the neuronal activations from both bottom-up and top-down mechanisms. These processes are known to compete [18] to form a unique representation, termed priority map [19]. These hypotheses suggest that attention is separated in distinct stages (pre-attentive as bottom-up and attentive as top-down) and that contributions towards guiding eye movements are simultaneously affected by distinct mechanisms in the HVS [20]. This competition for visual priority is biased by a term called relevance (as opposed to saliency), where top-down attention is driven by task demands, working and semantic memory as well as episodic memory, emotion and motivation (3 of which seem to be unique for each individual and momentum)[21]. At that end, it is stated [22, 23] that visual selection relies on activations from higher-level layers towards lower-level receptive fields. Therefore, modelization of attention should consider as well the influences of task and many other top-down effects.

1.1. Objectives

Initial hypotheses by Li [24, 25] suggested that visual saliency is processed by the lateral interactions of V1 cells. In their work, pyramidal cells and interneurons in the primary visual cortex (V1, Brodmann Area 17 or striate cortex) and their horizontal intracortical connections are seen to modulate activity in V1. Li's neurodynamic model [26] of excitatory and inhibitory firing-rate neurons was able to determine how contextual influences of visual scenes contribute to the formation of saliency. In this model, interactions between neurons tuned to specific orientation sensitivities served as predictors of pop-out effects and search asymmetries [27]. Li's neurodynamic model was later extended by Penacchio et al. [28] proposing the aforementioned lateral interactions to also be responsible for brightness induction mechanisms. By considering neuron orientation selectivity at distinct spatial scales, this model can act as a contrast enhancement mechanism of a particular visual area depending of induced activity from surrounding regions. Latest work from Berga & Otazu [29] has shown that the same model (without changing its parametrization) is able to predict saliency using real and synthetic color images. We propose to extend the model providing saliency computations with foveation, concerning distinct viewpoints during scene observation (mapping retinal projections towards V1 retinotopy) as a main hypothesis for predicting visual scanpaths. Furthermore, we also test how the model is able to provide predictions considering recurrent feedback mechanisms of already visited regions, as well as from visual feature and exemplar search tasks with top-down inhibition mechanisms.

1.2. A unified model of V1 predicts several perceptual processes

Here we present a novel neurodynamic model of visual attention and we remark its biological plausability as being able to simultaneously reproduce other effects such as Brightness Induction [28], Chromatic Induction [30] and Visual Discomfort [31] effects in previous work. Brightness and Chromatic induction stand for the variation of perceived luminance and color of a visual target depending on its luminance and/or chromatic properties as well as for its surrounding area respectively. Thus, a visual target can be perceived as being different (contrast) or similar (assimilation) to its physical properties by varying its surrounding context. With the simulations of our model, the output of V1's neuronal activity (coded as firing-rates during several cycles of excitatory-inhibitory V1 interneuron interactions), is used as predictor of induction and saliency respectively. These responses will act as a contrast enhancement mechanism, which for the case of saliency, are integrated towards projections in the superior colliculus (SC) for eye movement control. Therewith, our model has also been able to reproduce visual discomfort, as relative contrast energy of particular region on a scene is found to produce hyperexcitability in V1 [32, 33], one of possible causes of producing certain conditions such as malaise, nausea or even migraine. Previous neurodynamic [34, 35, 36, 37, 38, 39] and saliency models [8, 9, 40] have been able to predict eye movements. However, most of these models have been built specifically for visual saliency in a free-viewing task, a characteristic that denies their biological plausibility for modeling distinct visual processing mechanisms or other visual processes simultaneously. On behalf of model biological plasusibility on V1 function and its computations, we present a unified model of lateral connections in V1, able to predict attention (both in free-viewing and visual search) from real and synthetic color images while mimicking physiological properties of the neural circuitry stated previously.

2. Model

2.1. Retinal and LGN responses

The HVS perceives the light at distinct wavelengths of the visual spectrum and separates them to distinct channels for further processing in the cortex. First, retinal photoreceptors (or RP, corresponding to rod and cone cells) are photosensitive to luminance (rhodopsin-pigmented) and color (photopsin-pigmented) [41, 42]. Mammal cone cells are photosensitive to distinct wavelengths between a range of $\sim 400 - 700nm$, corresponding to three cell types, measured to be maximally responsive to Long (L, $\lambda_{max} \simeq 560$nm), Medium (M, $\lambda_{max} \simeq 530$nm) and Short (S, $\lambda_{max} \simeq 430$nm) wavelengths respectively [43]. RP signals are received by retinal ganglion cells (or RGC) forming an opponent process [44]. This opponent process allows to model midget, bistratified and parasol cells as "Red vs Green", "Blue vs Yellow", and "Light vs Dark" channels. In order to simulate these chromatic and light intensity opponencies using digital images, we transformed the RGB color space to the CIELAB (Lab or $L^*a^*b^*$) space (including a gamma correction of γ_{RGB}=1/2.2), as exemplified in Fig. 1.

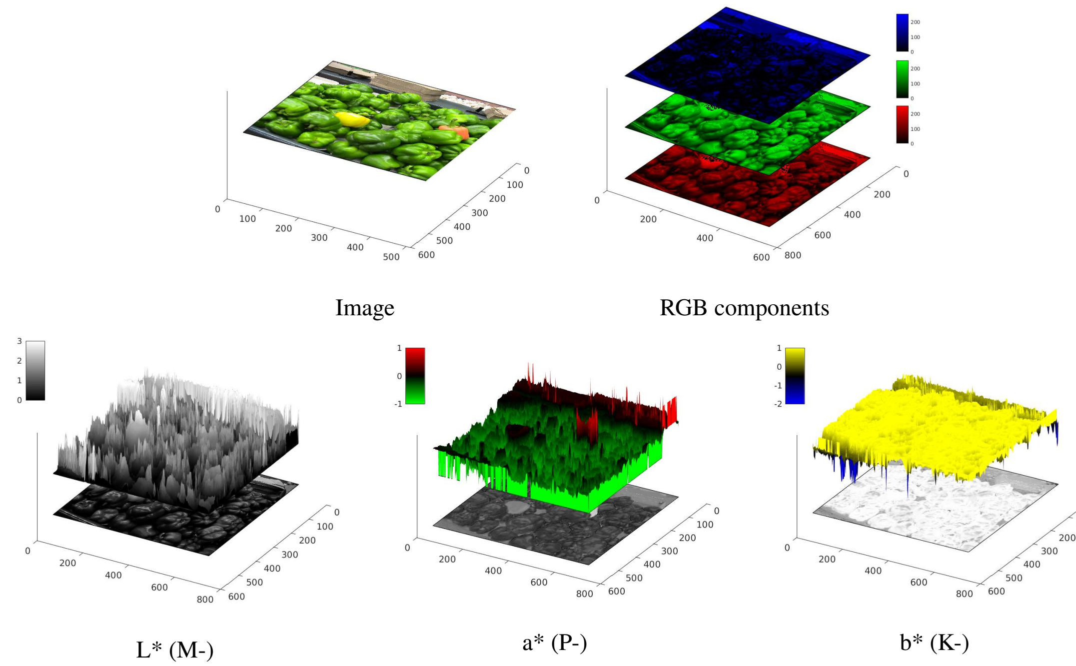


Figure 1: Example of CIELAB components of color opponencies given a sample image, corresponding to L^* (Intensity), a^* (Red-Green) and b^* (Blue-Yellow).

$$\begin{aligned} L^* &= R + G + B, \\ a^* &= \frac{R - G}{L^*}, \\ b^* &= \frac{R + G - 2B}{L^*}. \end{aligned} \tag{1}$$

The L^*, a^* and b^* channels form a cubic color space [45] with RGB opponencies ($+L$=lighter, $-L$=darker, $+a$=reddish, $-a$=greenish, $+b$=yellowish and $-b$=blueish).

Later, receptive fields in RGC [44] are activated in a center-surround fashion, receiving ON-OFF responses, being connected to horizontal (H-cell) and bipolar cell (B-cell) upstream circuitry. B-cells are hyperpolarized (OFF) or depolarized (ON) according to RP activity. In conjunction, H-cells send excitatory (center) and inhibitory feedback (surround) to RP. Midget (R-G), bistratified (B-Y) and parasol (L-D) RGC signals are sent through the optic nerve towards Parvo-, Konio- and Magno-cellular pathways in LGN respectively.

2.2. V1 Hypercolumnar organization

RGC center-surround responses are sent to LGN and projected to V1 cells. V1's cortical hypercolumns encode similar features of orientation-selective cells at different spatial frequencies. Simple cells found in V1 receptive fields (RFs) are sensitive to center-surround responses at distinct orientations, whereas complex cells overlap ON and OFF regions (and can be modeled as a combination of simple cell responses). Parvo- (P- or β), Konio- (K- or γ) and Magno-cellular (M- or α) pathways send signals separately towards distinct layers of the striate cortex (correspondingly projecting to $4C\beta$ & 6 from "P-", 2/3 & 4A from "K-" and $4C\alpha$ & 6 from "M-" cell pathways) for parallel and recurrent processing in V1.

We modeled the input to V1's simple cell responses with a 2D "a-trous" wavelet transform [46]. Discrete wavelet transforms allow to process signals by extracting information of orientation and scale-dependent features in the visual space (feature maps), which we used for filtering each of the aforementioed opponencies separately, shown in Fig. 2. Although these computations cannot be considered exact to each separate process of RGC and LGN, the transform seemingly resembles bottom-up activity projected to V1. The "a-trous" transform is undecimated and invertible, and allows to perform a transform where its basis functions remain similar to Gabor filters.

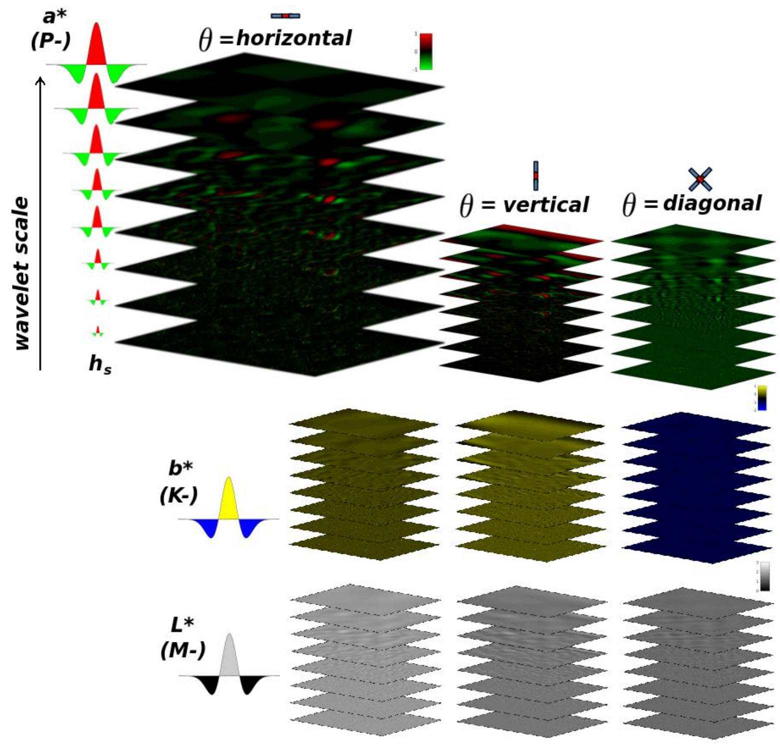


Figure 2: Representation of wavelet coefficients ($\omega_{iso\theta}$), in conjunction with the output of "a-trous" wavelet transform applied to components ($o = L^*, a^*, b^*$) shown in Fig. 1.

The "a trous" wavelet transform can be defined as:

$$
\begin{aligned}
\omega_{s,h} &= c_{s-1} - c_{s,h}, \\
\omega_{s,v} &= c_{s-1} - c_{s,v}, \\
\omega_{s,d} &= c_{s-1} - (c_{s,h} \otimes h'_s + \omega_{s,h} + \omega_{s,v}), \\
c_s &= c_{s-1} - (\omega_{s,h} + \omega_{s,v} + \omega_{s,d}).
\end{aligned}
\tag{2}
$$

where

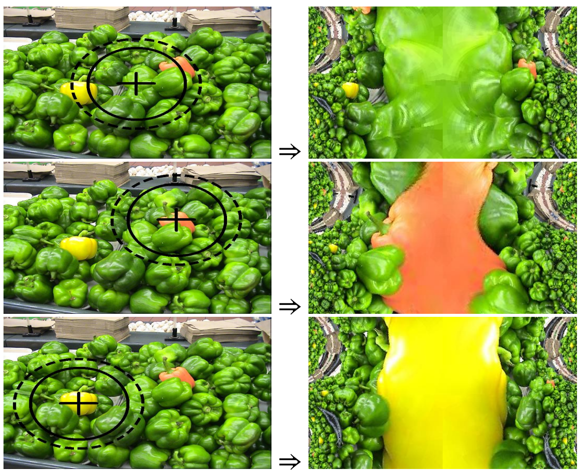
$$
\begin{aligned}
c_{s,h} &= c_{s-1} \otimes h_s, \\
c_{s,v} &= c_{s-1} \otimes h'_s.
\end{aligned}
\tag{3}
$$

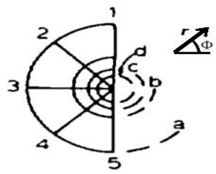
By transposing the wavelet filter (h_s, expressed in Fig. 2) and dilating it at distinct spatial scales ($s = 1...S$), we can obtain a set of wavelet approximation planes ($c_{s,\theta}$), that are combined for calculating wavelet coefficients ($\omega_{s,\theta}$) at distinct orientation selectivities ($\theta = h, v, d$). From these equations, three orientation selectivities can be extracted, corresponding to horizontal ($\theta_h \simeq \{0 \pm 30||180 \pm 30\}^o$), vertical ($\theta_v \simeq \{90 \pm 30||270 \pm 30\}^o$) and diagonal ($\theta_d \simeq \{45 \pm 15||135 \pm 15||225 \pm 15||315 \pm 15\}^o$) angles. For the case of scale features, sensititivies to size (in degree of visual angle) correspond to $2^{s_0(s-1)}/\{pxva\}$, where "$pxva$" is the number of pixels for each degree of visual angle according to experimentation, and s_0=8, is the minimum size of the wavelet filter (h_0) defining the first the scale frequency sensitivity. Initial $c_0 = I_o$ is obtained from the CIE L*a*b* components and c_n corresponds to the residual plane of the last wavelet component (e.g. $s = n$). The image inverse (I'_o) can be obtained by integrating the wavelet $\omega_{s,\theta}$ and residual planes c_n:

$$I'_o = \sum_{s=1,\theta=h,v,d}^{n} \omega_{s,\theta} + c_n. \tag{4}$$

2.3. Cortical mapping

The human eye is composed by RP but these are not homogeneously or equally distributed along the retina, contrarily to digital cameras. RP are distributed as a function of eccentricity with respect to the fovea (or central vision)[47]. Fovea's diameter is known to comprise ∼5deg of diameter in the visual field, extended by the parafovea (∼5-9deg), the perifovea (∼9-17deg) and the macula (∼17deg). Central vision is known to provide maximal resolution at ∼1deg of the fovea, whereas in periphery (∼60-180- deg) there is lower resolution for the retinotopic positions that are further away from the fovea. These effects are known to affect color, shape, grouping and motion perception of visual objects (even at few degrees of eccentricity), making performance on attentional mechanisms eccentricity-dependent [48]. Axons from the nasal retina project to the contralateral LGN, whereas the ones from the temporal retina are connected with the ipsilateral LGN. These projections [49] make the left visual field send inputs of the LGN towards the right hemifield of V1 (Fig. 3-Right), similarly for the case of the right visual field to the left hemifield of V1.





$$W(r, \Phi) = \lambda \ log(re^{i\Phi} + e_0), \tag{5}$$

$$Z(X, Yi) = e^{(W/\lambda)} - e_0. \tag{6}$$

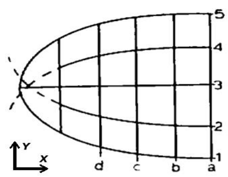


Figure 3: **Left:** Examples of applying the cortical magnification function (transforming the visual space to the cortical space) at distinct views of the image presented in Fig. 1. **Right:** Illustration of how polar coordinates (Z-plane) of azimuth $\Phi = (1, 2, 3, 4, 5)$ in the left visual field at distinct eccentricities $r = (d, c, b, a)$ are transformed to the cortical space (W-plane) in mm (X and Yi axis values). Equations 5 & 6 express the monopole direct and inverse cortical mapping transformations (parameters set as λ = 12mm and e_0 = 1deg [25, Section 2.3.1]). Illustration sketch was adapted from E.L. Schwartz [50], *Biol.Cybernetics 25*, p.184. Copyright (1977) by Springer-Verlag.

We have modeled these projections with a cortical magnification function [50][25, Section 2.3.1] using 128 mm of simulated cortical surface (see an example in Fig. 3-Left). The visual space is transformed to a cortically-magnified space (with its correspondence of millimeter for each degree of visual angle) with a logarithmic mapping function. The pixel-wise cartesian visual space is transformed to polar coordinates in terms of eccentricity and azimuth for a specific foveation instance, then transformed to coordinates in mm of cortical space. Acknowledging that the visual space for digital images is represented with either a squared or rectangular shape, we computed the continuation of cortical coordinates by symmetrically mirroring existing coordinates of the image with their correspondence of visual space outside boundaries in the cortical space. In that manner, we exclude possible effects of zero-padding over recurrent processing while preserving 2D shapes for our feature representations. For this case, these effects were minimized by the inverse and repeating the same process at specific interaction cycles. Schwartz's mapping has been applied over the wavelet coefficients represented in Fig. 2, as basis functions are convolved in the visual space, later magnified to the cortical space for representing V1 signals. These signals will serve as input to excitatory pyramidal cells, projected to their respective iso-orientation domains at distinct RF sizes.

2.4. V1 Neuronal Dynamics

Li's hypotheses suggest that V1 computations are responsible of generating a bottom-up saliency map [24, 25]. These hypotheses state that intracortical interactions between orientation-selective neurons in V1 are able to explain contextually-dependent perceptual effects present in pre-attentive vision [26, 27, 51, 52, 53, 54], relative to contour integration, visual segmentation, visual search asymmetries, figure-ground and border effects, among others. Pop-out effects that form the saliency map are believed to be the result of horizontal connections in V1, that interact with each other locally and reciprocally. These connections are formed by excitatory cells and inhibitory interneurons [55, 56], processing information from pyramidal cell signals in layers of V1. Spatial organization of these cells accounts for selectivity in their orientation columns, their RF size and axonal field localization. The aforementioned interactions between orientation-selective cells was defined by Li's model [26] of excitatory-inhibitory firing-rate neural dynamics, later extended by Penacchio et al. [28]. Here, contrast enhancement or suppression in neural responses emerge from lateral connections as an induction mechanism. Latest implementation done by Berga & Otazu [29] for saliency prediction used colour images, where chromatic (P-,K-) and luminance (M-) opponent channels were individually processed in order to compute firing-rate dynamics of each pathway separately. With cortical magnification, each gaze can significantly vary contextual information and therefore the output of the model.

Our excitatory-inhibitory model[1] is described in Table 1. Horizontal connections (lateral and reciprocal) are schematized in Fig. 4 and Table 1C, where excitatory cells have self-directed (J_0) and monosynaptic connections (J) between each other, whereas dysynaptically connected through (W) inhibitory interneurons. Axonal field projections (window) follow a concentric toroid of radius $\Delta_s = 15 \times 2^{s-1}$ and radial distance Δ_θ (accounting for RF size d_s and radial distance β). Membrane potentials of excitatory ($\dot{x}_{is\theta}$) and inhibitory ($\dot{y}_{is\theta}$) cells are obtained with partial derivative equations defined in Table 1D, composed by a chain of functions that consider firing-rates (obtained by piece-wise linear functions g_x and g_y) and membrane potentials from previous membrane cycles (modulated by α_x, α_y constants), current lateral connection potentials (J and W) and spread of inhibitory activity within hypercolumns (ψ). Background inputs (I_{noise} and I_{norm}) correspond to simulating random noise and divisive normalization signals (i.e. accounting for local nonorientation-specific cortical normalization and nonlinearities). Top-down inhibitory control mechanisms (I_c) are further explained in Table 1E and in Section 2.6. See the whole model pipeline in Fig. 6.

[1] Model implementation in MATLAB: `https://github.com/dberga/NSWAM`

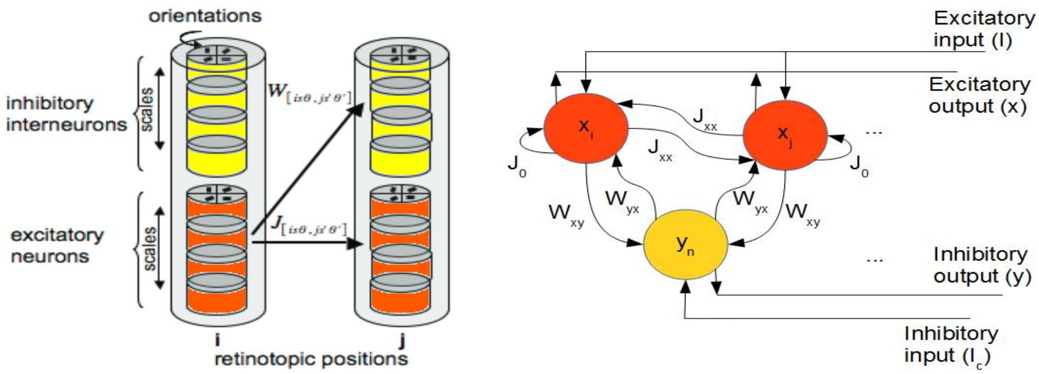


Figure 4: **Left:** Representation of cortical hypercolumns with scale and orientation selectivity interactions. **Right:** Model's intracortical excitatory-inhibitory interactions, membrane potentials (**orange** "$\dot{x}$" for excitatory and yellow "$\dot{y}$" for inhibitory) and connectivities ("$\boldsymbol{J}$" for monosynaptic excitation and "W" for dysynaptic inhibition).

Input signals ($I^t_{i;so\theta}$) have been defined as the wavelet coefficients ($\omega^t_{iso\theta}$), splitted between ON and OFF components (representing ON and OFF-center cell signals from RGC and LGN) depending on the value polarity (+ for positive and − for negative coefficient values) from the RF. These signals are processed separately during 10τ ($\tau = 1$ membrane time = $10ms$), including a rest interval (using an empty input) of 3τ to simulate intervals between each saccade shift. The model output has been computed as the firing-rate average g_x of the ON and OFF components ($M(\omega^{t+}_{iso\theta})$ and $M(\omega^{t-}_{iso\theta})$) during the whole viewing time, corresponding to a total of 10 membrane time (being the mean of g_x for a specific range of t).

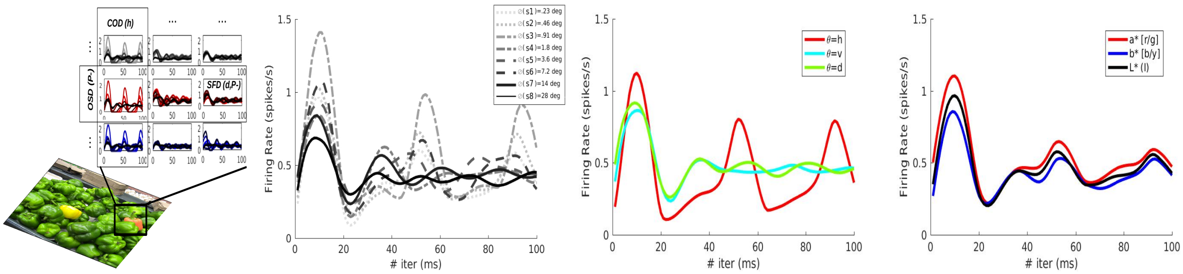


Figure 5: Firing rates plotted for 10 membrane time (100 iterations) accounting for neurons (ON+OFF values) inside a specific region (**1st col.**). Mean firing rates for all scales (Spatial Frequency Dynamics, **2nd col.**), orientations (Orientation Selectivity Dynamics, **3rd col.**), and color channels (Chromatic Opponency Dynamics, **4th col.**).

Combining the output of all components by

$$\hat{S}^t_{i;o} = \sum_{s=1..S;\theta=h,v,d}^{n_s} M(\omega^{t+}_{iso\theta}) + \sum_{s=1..S;\theta=h,v,d}^{n_s} M(\omega^{t-}_{iso\theta}) + c_i \quad , \tag{7}$$

we can describe the changes of the model (resulting from the simulated lateral interactions of V1) with respect the original wavelet coefficients $\omega^t_{iso\theta}$. Our result ($S^t_{i;o}$) will define the saliency map as an average conspicuity map or feature-wise distinctiveness (RF firing rates across scales and orientations for each pathway). These changes in firing-rate alternatively define the contrast enhancement seen on the brightness and chromatic induction cases [28, 30, 31], where the model output is combined with the wavelet coefficients $\{M(\omega^t_{iso})\omega^t_{iso}\}$ instead. The network is in total, composed of 1.18×10^6 neurons (accounting for 3 opponent channels, both ON/OFF polarities and RF sizes of $128 \times 64 \times 3{\times}8$).

Table 1: Overview of the model, following Nordlie et. al.'s format [57]. Further explanation for model variables and parameters is in [28, Supporting Information S1].

A	**Model Summary**
Populations	Excitatory (x), Inhibitory (y)
Topology	–
Connectivity	Feedforward: one-to-all, Feedback: one-to-all, Lateral: all-to-all (including self-connections)
Neuron model	Dynamic rate model
Channel models	–
Synapse model	Piece-wise linear synapse
Plasticity	–
Input	External current in lower (I) or higher (I_c) cortical areas and random noise (I_0)
Measurements	Firing-rate (g_x and g_y)

B	**Populations**	
Name	**Elements**	**Size**
x	Sigmoidal-like neuron	$K_x = M \times N \times \Theta \times S = 64 \times 128 \times 3 \times 8$
y	Sigmoidal-like neuron	$K_y = K_x$

C	**Connectivity**		
Name	**Source**	**Target**	**Pattern**
J_{xx}	x	x	Excitatory, toric, all to all, non-plastic
J_0	x	x	Excitatory, constant $J_0 = 0.8$
W_{xy}	x	y	Inhibitory, toric, all to all, non-plastic
W_{yx}	y	x	Inhibitory, toric, all to all, non-plastic

D	**Neuron and Synapse Model**
Name	V1 neuron
Type	Dynamic rate model
Synaptic dynamics	$$J_{[is\theta,js'\theta']} = \lambda(\Delta_s)0.126e^{(-\beta/d_s)^2-2(\beta/d_s)^7-d_s^2/90} \quad (8)$$ $$W_{[is\theta,js'\theta']} = \lambda(\Delta_s)0.14(1-e^{-0.4(\beta/d_s)^{1.5}})e^{-(\Delta_\theta/(\pi/4))^{1.5}} \quad (9)$$
Membrane potential	$$\dot{x}_{is\theta} = -\alpha_x x_{is\theta} - g_y(y_{is\theta}) - \sum_{\Delta_s,\Delta_\theta\neq 0} \psi(\Delta_s,\Delta_\theta) g_y(y_{is}+\Delta_{s\theta}+\Delta_\theta) + J_0 g(x_{is\theta}) + \sum_{j\neq i,s',\theta'} J_{[is\theta,js'\theta']} g_x(x_{js'\theta'}) + I_{is\theta} + I_0, \quad (10)$$ $$\dot{y}_{is\theta} = -\alpha_y y_{is\theta} - g_x(x_{is\theta}) + \sum_{j\neq i,s',\theta'} W_{[is\theta,js'\theta']} g_x(x_{js'\theta'}) + I_c \quad (11)$$

E	**Input**
Type	**Description**
Sensory (bottom-up)	Input to excitatory neurons, $I^t_{i;o} = \omega^t_{iso\theta}$
Control (top-down)	Input to inhibitory interneurons, $I_c = 1.0 + I_{noise} + I_{vs} + I_{ior}$

F	**Measurements**
Mean Firing-rate of excitatory neurons for τ=10 membrane time ($M(\omega^{p=[+,-]}_{iso\theta})$).	

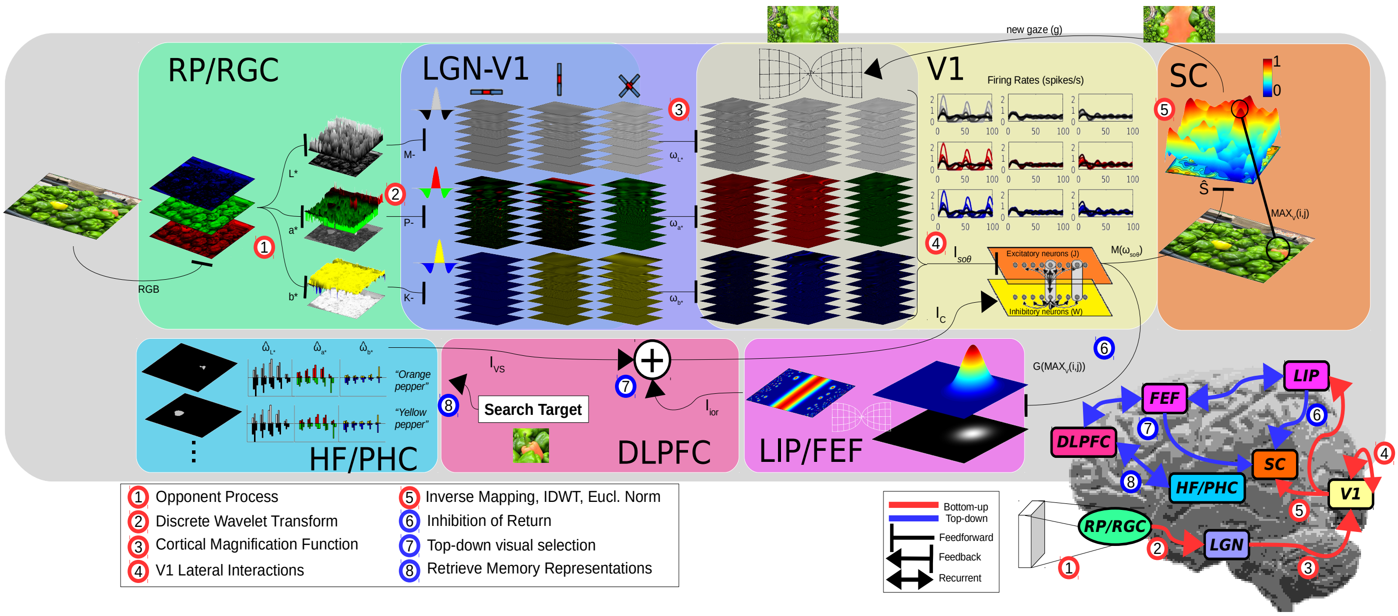


Figure 6: Diagram illustrating how visual information is processed by NSWAM-CM, including a brain drawing of each bottom-up and top-down attention mechanisms and their localization in the cortex (**Bottom-Right**).

2.5. Projections to the SC

Latest hypotheses about neural correlates of saliency [58, 59] state that the superior colliculus is responsible for encoding visual saliency and to guide eye movements [20, 60]. Acknowledging that the superficial layers of the SC (sSC) receive inputs from the early stages of visual processing (V1, retina), the SC selects these as the root of bottom-up activity to be selected in the intermediate and deep layers (iSC, dSC). In accordance to the previous stated hypotheses [24], saccadic eye movements modulated by saliency therefore are computed by V1 activity, whereas recurrent and top-down attention is suggested to be processed by neural correlates in the parieto-frontal cortex and basal ganglia. All these projections are selected as a winner-take-all mechanism in SC [24, 25, 27] to a unique map, where retinotopic positions with the highest activity will be considered as candidates to the corresponding saccade locations. These activations in the SC are transmitted to guide vertical and horizontal saccade visuomotor nerves [61]. We have defined the higher active neurons (Equation 12) as the locations for saccades in the visual space (i,j) by decoding the inverse of the cortical magnification (Equation 6) of their respective retinotopic position ("i" neuron at X,Yi).

$$MAX_W(X,Yi) = argmax(\hat{S}) \rightarrow MAX_Z(r,\Phi) \rightarrow MAX_V(i,j), \tag{12}$$

The behavioral quantity of the unique 2D saliency map has been defined by computing the inverse of the previous processes using the model output for each pathway separately. Retinotopic positions have been transformed to coordinates in the visual space using the inverse of the cortical magnification function (Equation 6). Output signals (V1 sensitivities to orientation and spatial frequencies) are integrated by computing the inverse discrete wavelet transform to obtain unique maps for each channel opponency (Equation 4). A unique representation (Equation 13) of final neuronal responses for each pathway (P-, K- and M- as a^*, b^* and L^*) is generated with the euclidean norm (adding responses of all channels as in Murray et al.[62] model). The resulting map is later normalized by the variance (Equation 14) of the firing rate [25, Chapter 5]. This map represents the final saliency map, that describes the probability distribution of fixation points in certain areas of the image. In addition to this estimation, the saliency map has been convolved with a gaussian filter simulating a smoothing caused by the deviations of $\sigma = 1$ deg given from eye tracking experimentation, recommended by LeMeur & Baccino [63].

$$\hat{S}_i = \sqrt{\hat{S}_{i;a^*} + \hat{S}_{i;b^*} + \hat{S}_{i;L^*}}, \tag{13}$$

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

2.6. Attention as top-down inhibition

An additional purpose of our work is the modeling of attentional mechanisms beyond pre-attentive visual selection. Instead of analyzing the scene serially, the visual brain uses a set of attentional biases to recognize objects, their relationships and their importance with respect to the task, all given in a set of visual representations. Similarly to the saliency map, the priority map can be interpreted as a unique 2D representation for eye movement guidance formed in the SC, here including top-down (not guided by the stimulus itself) and recurrent information as visual relevance. This phenomenon suggests that executive, long-term and short-term/working memory correlates also direct eye movement control [20, 64]. Previous hypotheses model these properties by forming the priority map through selective tuning [22, 65]. Selective tuning explains attention mechanisms as a hierarchy of winner-take-all processes. This hypothesis suggests that top-down attention can be simulated by spatially inhibiting specific layers of processing. Latest hypotheses [66] confirm that striate cortical activity gain can be modulated by SC responses, with additional modulations arising from pulvinar to extrastriate visual areas. In addition, it has also been stated [67] that V1 influences both saliency and top-down learning during visual detection tasks. By functionally simulating the aforementioned top-down mechanisms as inhibitory gates of top-down feedback control in our model [26], we are able to perform task-specific visual selection (VS) and inhibition of return (IoR) mechanisms.

Top-down selection: Goal-directed or memory-guided saccades imply executive control mechanisms that account for task requirements during stimulus perception. The dorsolateral prefrontal cortex (DLPFC) is known to be responsible for short-term spatial memory, to retrieve long-term memory signals of object representations (through projections towards the para- and hippocampal formations) as well as to perform reflective saccade inhibition, among other functions. These inhibitory signals, later projected to the frontal eye field (FEF), are able to direct gaze during search and smooth pursuit tasks [64, 68, 69] (also suggested to be crucial for planning intentional or endogenously-guided saccades), where its signals are sent to the SC. By feeding our model with inhibitory signals (I_c shown in Fig. 4 and Table 1E) we can simulate top-down feedback control mechanisms in V1 (initially proposed by Li [26, Sec. 3.7]). In this case, a new term $I_{\{vs\}}$ is added to the top-down inhibition of our V1 cortical signals that will be projected to the SC during each gaze.

$$I_{\{vs\}} = \alpha_{\{vs\}} \cdot \begin{cases} argmax_{p,s,o,\theta}(\omega) & \text{, feature-selective } (VS_M) \\ (\sum_{i=1}^{N} \omega_{pso\theta})/N & \text{, category-specific } (VS_C) \end{cases} \tag{15}$$

In this implementation, we can perform distinct search tasks such as feature search (by manually selecting the features, or selecting features with maximal responses, similarly to a boolean selection [23]), exemplar and categorical object search (by processing the mean of responses $\hat{\omega}$ from wavelet coefficients of a single or several image samples "N"). These low-level computations would serve as cortical activations to be stored as weights in our low-level memory representations, that will be used as inhibitory modulation for the task execution.

Inhibition of Return: During scene viewing, saccadic eye movements show distinct patterns of fixations [70], directed by exploratory purposes or either towards putting the attentional focus on specific objects in the scene. For the former case, the HVS needs to ignore already visited regions (triggering anti-saccades away from these memorized regions, as a consequence of inhibition) during a period of time before gazing again towards them. This phenomena is named inhibition of return [71], and similarly involves extracting sensory information and short-term memory during scene perception. As mentioned before, DLPFC is responsible of memory-guided saccades, and this function might be done in conjunction with the parietal cortex and the FEF. The parietal areas (LIP and PEF)[64, 68, 72] are known to be responsible of visuospatial integration and preparation of saccade sequences. These areas conjunctively interact with the FEF and DLPFC for planning these reflexive visually-guided saccades. Acknowledging that LIP receives inputs from FEF and DLPFC, the role of each cannot be disentangled as a unique functional correlate for the IoR. Following the above, we have modeled return mechanisms as top-down cortical inhibition feedback control accounting for previously-viewed saccade locations. Thus, we added an inhibition input $I_{\{IoR\}}$ at the start of each saccade, which will determine our IoR mechanism:

$$I^{g,t=0}_{\{IoR\}} = MAX(\hat{S}) \cdot G(MAX_V(x,y)) + I^{g-1}_{\{IoR\}},$$
$$I^{g,t>0}_{\{IoR\}} = \alpha_{\{IoR\}}(I^{t-1}_{\{IoR\}}) \prod_{i=1}^{10\tau} e^{log(\beta_{\{IoR\}})/\tau}. \tag{16}$$

This term is modulated with a constant power factor $\alpha_{\{IoR\}}$ and a decay factor $\beta_{\{IoR\}}$, which in every cycle will progressively reduce inhibition. The spatial region of the IoR has been defined as a gaussian function centered to the previous gaze (g), with a spatial standard deviation $\sigma_{\{IoR\}}$ dependent on a specific spatial scale and a peak with an amplitude of the maximal RF firing rate of our model's output ($\hat{S}$). Inhibitory activity is accumulated to the same map and can be shown how is progressively reduced during viewing time (Fig. 14). Alternatively illustrated in Itti et al.'s work [7], the IoR can be applied to static saliency models by substracting the accumulated inhibitory map to the saliency map during each gaze ($\hat{S} - I^{g}_{\{IoR\}}$).

3. Materials and Methods

3.1. Procedure

Experimental data has been extracted from eye tracking experimentation. Four datasets were analyzed, corresponding to 120 real indoor and outdoor images (Toronto [10]), 40 nature scene images (KTH [73]), 100 synthetic image patterns ($CAT2000_P$ [74]) and 230 psychophysical images (SID4VAM [17, 75]). Generically, experimentation for these type of datasets [76] capture fixations from about 5 to 55 subjects, looking at a monitor inside a luminance controlled room while being restrained with a chin rest, located at a relative distance of 30-40 pixels per degree of visual angle ($pxva$). The tasks performed mostly consist of freely looking at each image during 5000 ms, looking at the "most salient objects" or searching for specific objects of interest. We have selected these datasets to evaluate prediction performance at distinct scene contexts. Indicators of psychophysical consistency of the models has been presented, evaluating prediction performance upon fixation number and feature contrast. Visual search performance has been evaluated by computing predictions of locating specific objects of interest. For the case of stimuli from real image contexts (Fig. 18) we have used salient object segmented regions from Toronto's dataset [10], extracted from Li et al. [77]. Finally, for the case of evaluating fixations performed with synthetic image patterns, we used fixations from SID4VAM's psychophysical stimuli.

3.2. Model evaluation

Current eye tracking experimentation represent indicators of saliency as the probability of fixations on certain regions of an image[2]. Metrics used in saliency benchmarks [40] consider all fixations during viewing time with same importance, making saliency hypotheses unclear of which computational procedures perform best using real image datasets. Previous psychophysical studies [16, 17] revealed that fixations guided by bottom-up attention are influenced by the type of features that appear in the scene and their relative feature contrast. From these properties, the order of fixations and the type of task can drive specific eye movement patterns and center biases, relevant in this case.

The AUC metric (Area Under ROC/Receiver Operant Characteristic) represents a score of a curve comprised of true positive values (TP) against false positive (FP) values. The TP are set as human fixations inside a region of the saliency map, whereas FP are those predicted saliency regions that did not fall on human fixation instances. For our prediction evaluation we computed the sAUC (shuffled AUC), where FP are expressed as TP from fixations of other image instances. This metric prioritizes model consistency and penalizes for prediction biases that appear over eye movement datasets, such as oculomotor and center biases (not driven by pre-attentional factors). We also calculated the Information Gain (InfoGain) metric for model evaluation, which compares FP in the probability density distribution of human fixations with the model prediction, while substracting a baseline distribution of the center bias (all fixations grouped together in a single map). Saliency metrics, largely explained by Bylinskii et al. [78], usually compare model predictions with human fixations during the whole viewing time, regardless of fixation order. In our study is also represented the evolution of prediction scores for each gaze. For the case of scanpaths, we evaluated saccade sequences by analyzing saccade amplitude (SA) and saccade landing (SL) statistics. These are calculated using euclidean distance

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

between fixation coordinates (distance between saccade length for SA and distance between locations of saccades for SL).

Initial investigations on visual attention [4, 5] during visual search tasks formulated that reaction times of finding a target (defined in a region of interest/ROI) among a set of distractors are dependent on set size as well as target-distractor feature contrast. In order to evaluate performance on visual search, we utilised two metrics that account for the ground truth mask of specific regions for search and the saliency map (in this context, it could be considered as a "relevance" map) or predicted saccade coordinates (from locations with highest neuronal activity). The Saliency Index (SI) [17, 75, 79] calculates the amount of energy of a saliency map inside a ROI (S_t) with respect to the one outside (S_b), calculated as: $SI = (S_t - S_b)/S_b$. For the case of saccades in visual search, we considered to calculate the probability of fixations inside the ROI (PFI).

4. Results

4.1. Results on predicting Saliency

In this section, probability density maps (GT) have been generated using fixation data of all participants from Toronto, KTH, CAT2000 and SID4VAM eye tracking datasets (model scores and examples in Figs 7-10). Several saliency predictions have been computed from different biologically-inspired models. Our Neurodynamic Saliency Wavelet Model has been computed without (NSWAM) and with foveation (NSWAM-CM), as a mean of cortically-mapped saliency computations through a loop of 1, 2, 5 and 10 saccades. The loop consists on obtaining a saliency map for each view of the scene, and obtaining an unique map for each saccade instance by computing the mean of all saliency maps.

Based on the shuffled metric scores, traditional saliency models such as AIM overall score higher on real scene images (Fig. 7), scoring $sAUC_{AIM}$=.663, and $InfoGain_{IKN}$=.024. For the case of nature images (Fig. 8), our non-foveated and foveated versions of the model (NSWAM and NSWAM-CM) scored highest on both metrics ($InfoGain_{NSWAM}$=.168 and $sAUC_{NSWAM-CM10}$=.567). As mentioned before, fixation center biases are present when the task and/or stimulus do not induce regions that are enough salient to produce bottom-up saccades. In addition, in real image datasets (Toronto and KTH), not all images contain particularly salient regions. This is seemingly presented in our models' saliency maps from 1st to 10th fixations (Figs. 7-8, rows 5-8), where salient regions are presented to be less evident across fixation order.

In synthetic image patterns ($CAT2000_P$), both of our model versions outperforms other models $sAUC_{NSWAM,NSWAM-CM}$=.567. Center biases are present in such dataset (see Fig. 9, "Human Fix." heatmaps), seemingly reproduced by IKN in the illustration ($InfoGain_{IKN}$=-.724). For the case of SID4VAM dataset (Fig. 10), salient regions are labeled with specific feature type and contrast, and fixation patterns present lower center biases (due to mainly being based a singleton search type of task with a unique salient target with random location). Our model presents highest scores on both metrics ($sAUC_{NSWAM,NSWAM-CM2}$=.622 and $InfoGain_{NSWAM-CM10}$=-.131).

In Figs. 7-10 are compared the average score per gaze of human fixations and saliency model predictions. It can be observed that prediction scores for all models decrease as a function of gaze number. Scores of probability density distributions of human fixations (in comparison to fixation locations) decrease around 10% the sAUC after 10 saccades. This decrease of performance is not reproduced by any of the presented models, instead, most of them show a flat or slightly increasing slopes for the case of sAUC scores and logarithmically increasing scores for InfoGain. NSWAM and NSWAM-CM present similar results upon fixation number.

Model	sAUC	InfoGain
Human Fix.	.904	2.42
IKN [7]	.649	-.024*
AIM [10]	.663*	-.579
NSWAM	.631	-.552
NSWAM-CM1	.636	-.818
NSWAM-CM2	.644	-.738
NSWAM-CM5	.650	-.701
NSWAM-CM10	.655	-.692

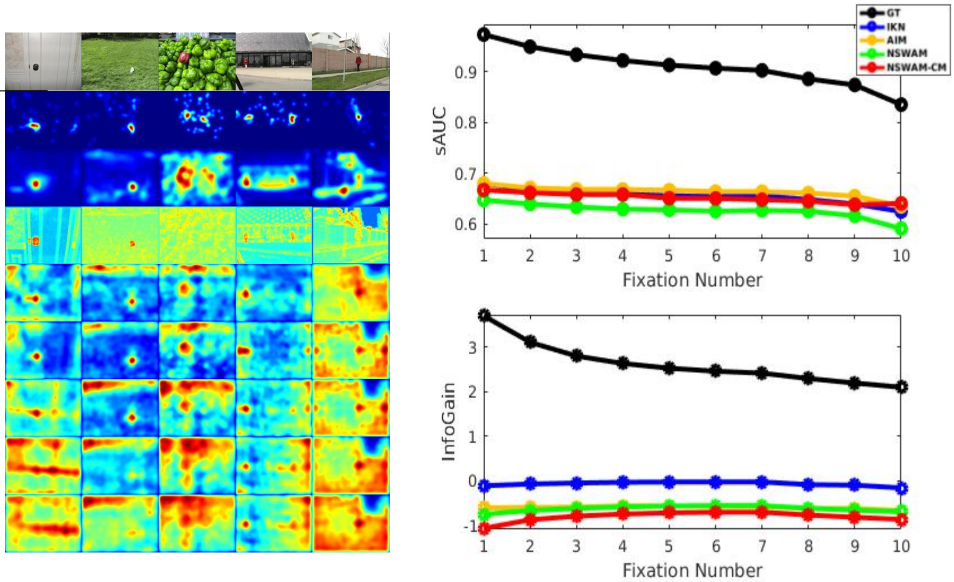


Figure 7: Results on saliency for Toronto (Bruce & Tsotsos [10]) Eye Tracking Dataset. **Left**: Saliency metric scores. **Middle**: Examples of saliency maps. **Right**: Shuffled scores per fixation number.

Model	sAUC	InfoGain
Human Fix.	.822	1.41
IKN [7]	.551	-.172
AIM [10]	.552	-.509
NSWAM	.565	-.168*
NSWAM-CM1	.564	-.227
NSWAM-CM2	.566	-.213
NSWAM-CM5	.566	-.211
NSWAM-CM10	.567*	-.209

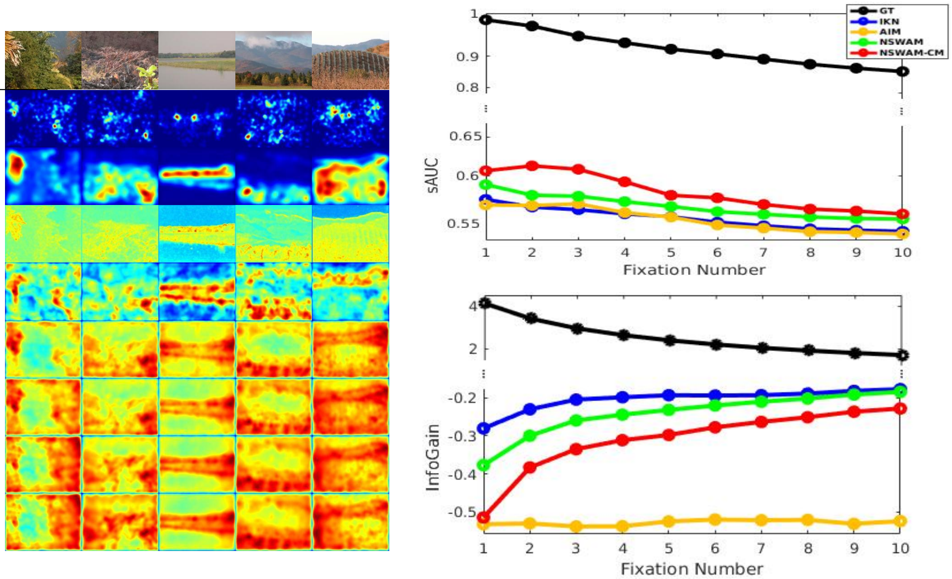


Figure 8: Results on saliency for KTH (Kootra et al'.s [73]) Eye Tracking Dataset. **Left**: Saliency metric scores. **Middle**: Examples of saliency maps. **Right**: Shuffled scores per fixation number.

Model	sAUC	InfoGain
Human Fix.	.623	.777
IKN [7]	.562	-.724*
AIM [10]	.544	-6.49
NSWAM	.567*	-1.01
NSWAM-CM1	.561	-1.24
NSWAM-CM2	.563	-1.14
NSWAM-CM5	.565	-1.09
NSWAM-CM10	.567*	-1.07

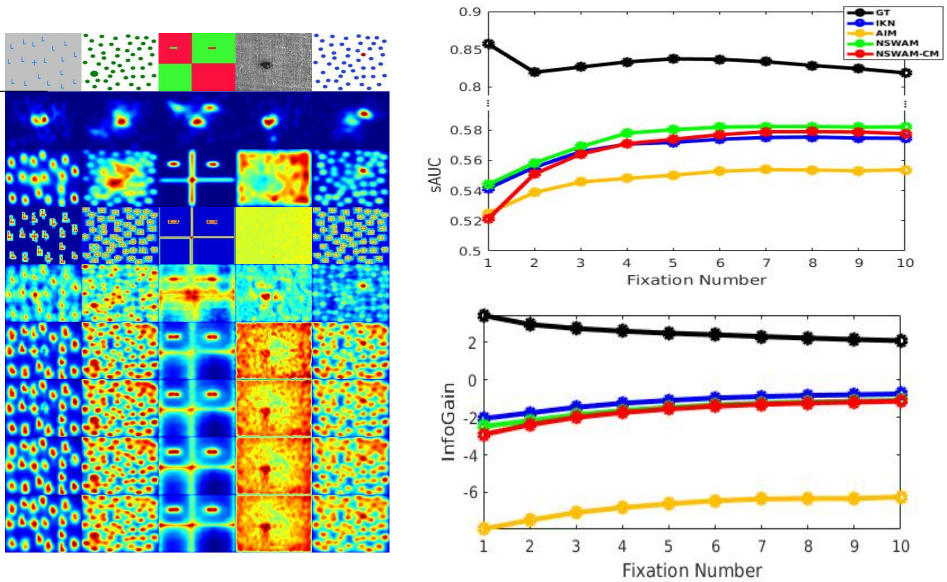


Figure 9: Results on saliency for $CAT2000_{Pattern}$ (Borji & Itti [74]) Dataset. **Left**: Saliency metric scores. **Middle**: Examples of saliency maps. **Right**: Shuffled scores per fixation number.

Model	sAUC	InfoGain
Human Fix.	.860	2.80
IKN [7]	.608	-.233
AIM [10]	.557	-18.2
NSWAM	.622*	-.149
NSWAM-CM1	.617	-.204
NSWAM-CM2	.622*	-.164
NSWAM-CM5	.620	-.139
NSWAM-CM10	.618	-.131*

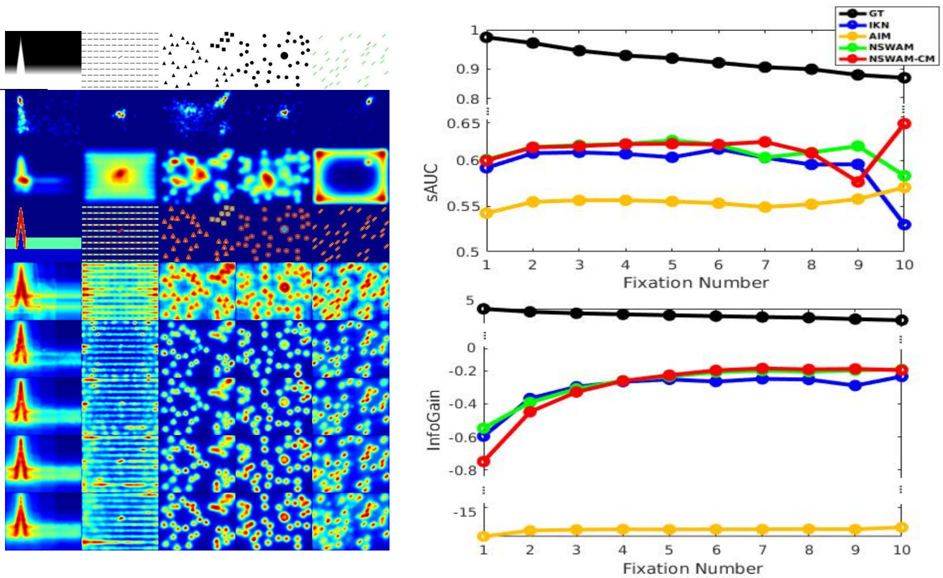


Figure 10: Results on saliency for SID4VAM (Berga et al. [17]) Eye Tracking Dataset. **Left**: Saliency metric scores. **Middle**: Examples of saliency maps. **Right**: Shuffled scores per fixation number.

In SID4VAM, stimuli are categorized with specific difficulty (according to the relative target-distractor feature contrast). With these, we computed the score for each relative contrast instance (Ψ) in Fig. 11. After computing every low-level stimulus instance with the presented models and evaluating results with the same metrics, our saliency model (NSWAM and NSWAM-CM) presents better performance than AIM and IKN and also increases score at higher feature contrasts.

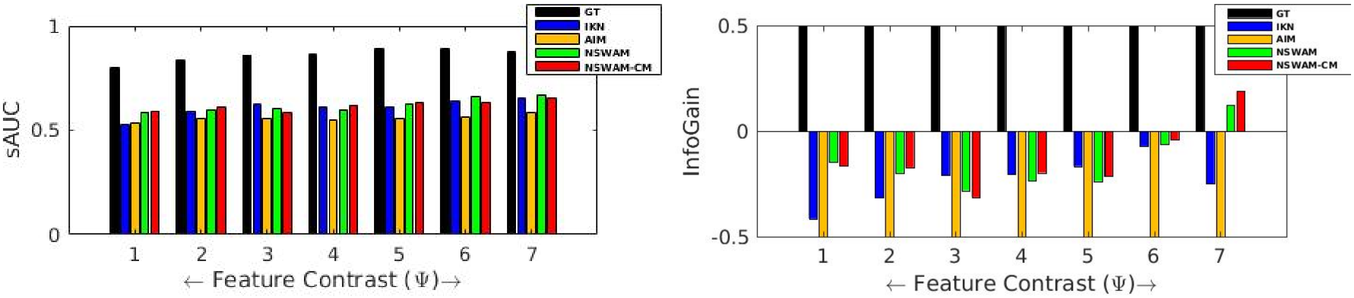


Figure 11: sAUC and InfoGain scores for each relative target-distractor feature contrast

4.1.1. Discussion

Quantitatively, systematic tendencies in free-viewing (center biases, inter-participant differences, etc.[80]) should not be likely to be considered as indicators of saliency. Although shuffled metrics try to penalize for these effects, benchmarks do not compensate for these tendencies from model evaluations (these are particular for each dataset task and stimulus properties). Acknowledging that first saccades determine bottom-up eye movement guidance [81, 82], it is a phenomenon also present in our experimental data (in terms of the decrease of performance with respect fixation region probability compared to fixation locations). In that aspect, evaluating first fixations with more importance could define new benchmarks for saliency modeling, similarly with stimuli where feature contrast in salient objects is quantified. Ideal conditions (following the Weber law) determine that if there is less difficulty for finding the salient region (higher target-distractor contrast), saliency will be focused on that region. Conversely, fixations would be distributed on the whole scene if otherwise. Our model presents better performance than other biologically-inspired ones accounting for these basis.

4.2. Results on predicting scanpaths

Illustration of scanpaths from datasets presented in Section 4.1 were computed with scanpath models in Fig. 13. Scanpaths are predicted by NSWAM-CM during the first 10 saccades, by selecting maximum activity of our model for every saccade. We have plotted our model's performance in addition to Boccignone&Ferraro's and LeMeur&Liu's predictions (Fig. 12). Saccade statistics show an initial increment of saccade amplitude, decreasing as a function of fixation number. Errors of SA and SL (ΔSA and ΔSL) are calculated as absolute differences between model predictions and human fixations. Values of ΔSL appear to be lower and similar for all models during initial fixations.

Prediction errors are shown to be sustained or increasing for CLE and NSWAM-CM (maybe due to their lack of processing higher level features, experimental center biases, etc.). Errors on ΔSA predictions are lower for LeMeur&Liu's model, retaining similar saccades (except for synthetic images of SID4VAM). Although these errors are representative in terms of saccade sequence, we also computed correlations of models' SA with GT (ρSA). In this last case, NSWAM-CM presents most higher correlation values for all datasets ($\rho SA_{Toronto}$=-.38, p=.09; ρSA_{KTH}=.012, p=.96; $\rho SA_{CAT2000_P}$=.28, p=.16; $\rho SA_{SID4VAM}$=.96, $p=1.26\times10^{-71}$) than other models. Most of them seem to accurately predict SA for SID4VAM (which contains mostly visual search psychophysical image patterns), with ρSA between .7 and .8. Our scanpath model tend to predict eye movements with large mean saccade amplitudes $\{M(SA)_{Toronto}$=7.8±3.5; $M(SA)_{KTH}$=13±6.1; $M(SA)_{CAT2000_P}$=15.7±6.7; $M(SA)_{SID4VAM}$=15.7±6.9 deg$\}$, whereas human fixations combine both short and large saccades $\{M(SA)_{Toronto}$=4.6±1; $M(SA)_{KTH}$=6.7±.5; $M(SA)_{CAT2000_P}$=5.1±.9; $M(SA)_{SID4VAM}$=5.8±1.5 deg$\}$. In that aspect, our prediction errors might arise from not correctly predicting focal fixations.

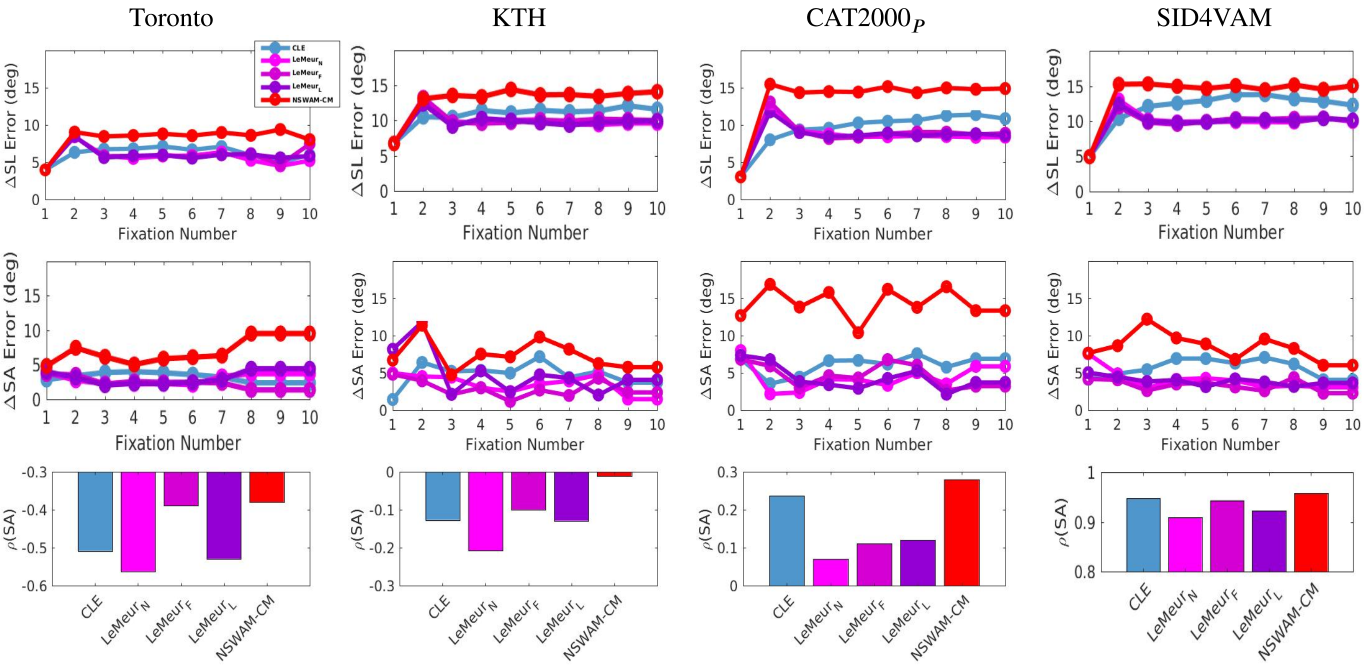


Figure 12: **1st row:** Prediction errors in Saccade Landing (ΔSL) for real indoor/outdoor (Toronto), nature (KTH) and synthetic (CAT2000$_P$ and SID4VAM) image datasets. **2nd row:** Prediction errors in Saccade Amplitude (ΔSA) on same datasets. **3rd row:** Correlations of Saccade Amplitude (ρSA) with respect human fixations.

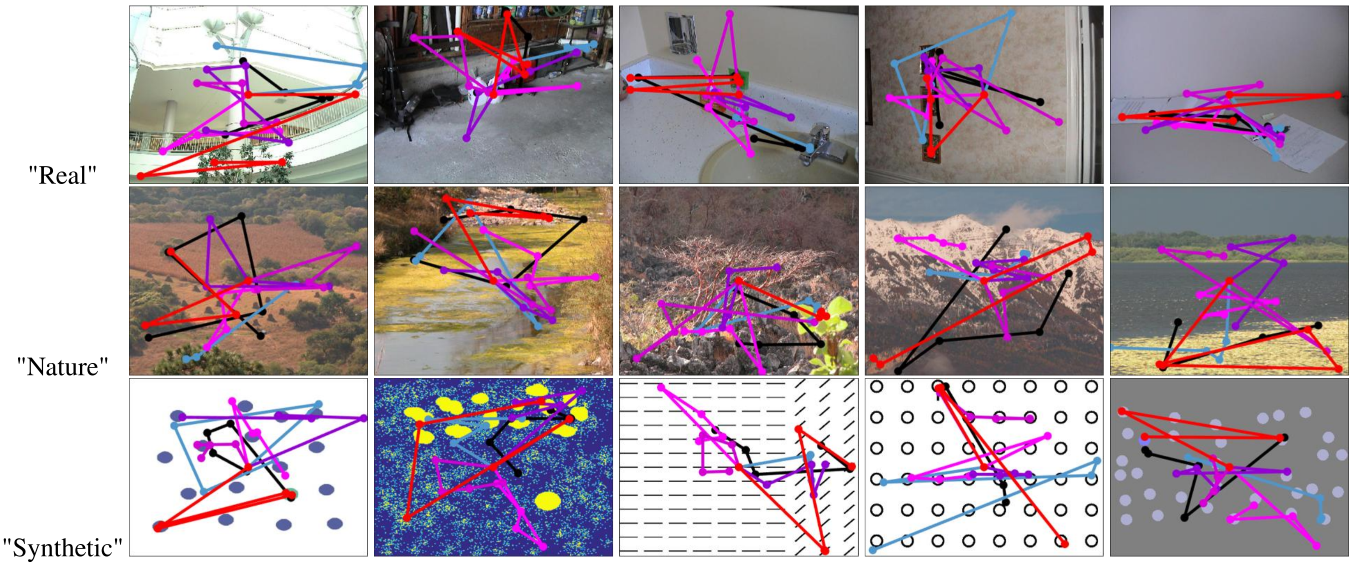


Figure 13: Examples of visual scanpaths for a set of real (**1st row**, [10]), nature (**2nd row**,[73]) and synthetic (**3rd row**,[17, 74, 75]) images. Model scanpaths correspond to **Human Fixations (single sample)**, **CLE [13]**, **LeMeur**$_{Natural}$, **LeMeur**$_{Faces}$, **LeMeur**$_{Landscapes}$ **[14]** and **NSWAM-CM (ours)**.

We simulated the inhibition factor for all datasets by substracting the inhibition factor $I_{\{IoR\}}$ to our models' saliency maps (NSWAM+IoR). After computing prediction errors in SA and SL for a single sample (Fig. 15-Top), best predictions seem to appear at decay values of $\beta_{\{IoR\}}$ between .93 and .98, which corresponds to 1 to 5 saccades (similarly explained by Samuel & Kat [83] and Berga et al. [17], where takes from 300-1600 ms for the duration of the

IoR, corresponding to 1 to 5 times the fixation duration). For the case of the $\sigma_{\{IoR\}}$, lowest prediction error (again, both in SA and SL) is found from 1 to 3 deg (in comparison, LeMeur & Liu [14] parametrized it by default as 2 deg). Results on ΔSA statistics have similar / slightly increasing performance until ($\beta_{\{IoR\}}$ <1) a single fixation time, decreasing at highest decay $\beta_{\{IoR\}}$ $\geq$5th saccade. For ΔSL values, errors in datasets such as KTH and SID4VAM are decreased at higher decay. For the latter, ΔSA errors are shown to decrease progressively at highest decay values ($\beta_{\{IoR\}}$ $\geq$.93). Lastly, when parametrizing the spatial properties of the IoR, saccade prediction performance is highest at lower size (with a near-constant error in SA and SL increasing about 1 deg for $\sigma_{\{IoR\}}$=1 to 8 deg on all datasets).

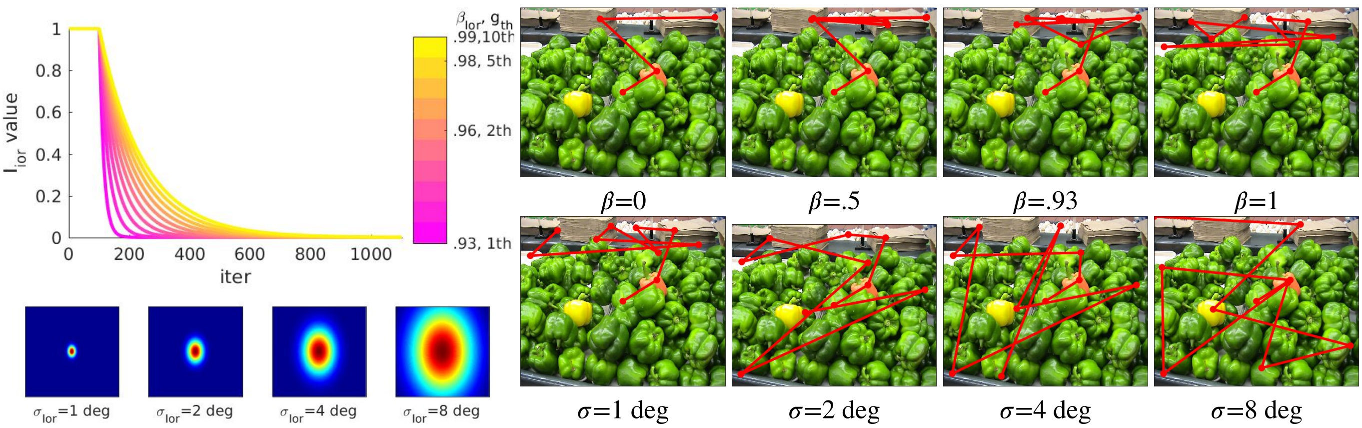
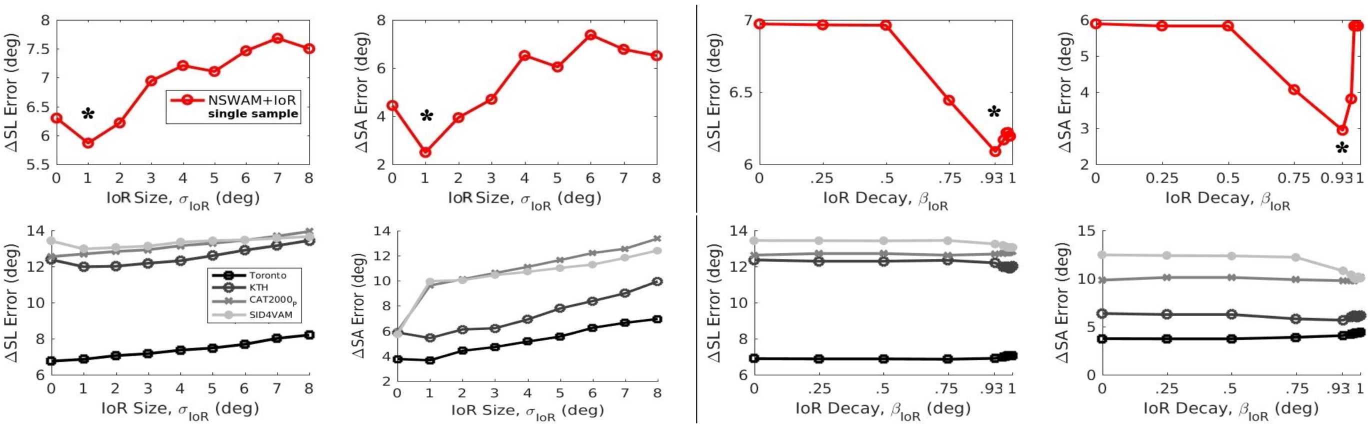


Figure 14: **Left**: Evolution of inhibition factor for 100 mem.time (about 1000 iterations), corresponding approximately to performing 10 saccades to the model (top). Spatial representation of the IoR with distinct size (bottom). **Right**: Examples of scanpaths for different IoR decay factor (top, $\sigma_{\{IoR\}}$=2 deg, $\beta_{\{IoR\}}$=$\{0, .5, .9, 1\}$) or distinct IoR size (bottom, $\sigma_{\{IoR\}}$=$\{1, 2, 4, 8\}$ deg,$\beta_{\{IoR\}}$=1).



*: Lowest error (ΔSL or ΔSA) at specific parametrization

Figure 15: Statistics of scanpath prediction (ΔSL and ΔSA) by the parametrization of IoR decay ($\beta_{\{IoR\}}$) and IoR size ($\sigma_{\{IoR\}}$) in a single sample (**Top row**, from image scanpaths in Fig. 13) and saliency datasets (**Bottom row**).

4.2.1. Discussion

Our model predictions on SA correlate better (i.e. obtain higher ρSA values) than other scanpath models (in terms of how SA evolves over fixations), however, prediction errors are higher in both SL and SA. We believe that these errors are caused by incorrectly predicting locations of fixations, but not for failing on predictions of the saccade sequence per se. These locations are mainly influenced by systematic tendencies in free-viewing (derived by center biases and/or focal fixations in a particular region of the image). Cortical magnification mechanisms might be responsible for processing higher saliency at regions outside the fovea, generating tendencies of uniquely capturing large saccades. These can be solved by processing high-level feature computations near the fovea, which would increase the probability of fixations at lower SA. Nonetheless, we have to stress that first fixations are long known for being determinants of bottom-up attention [17, 81]. Instead, higher inter-participant differences [80] and center biases [84] increase as functions of fixation number, suggested as worse candidates for predicting attention. These parameters appear to specifically affect each stimuli differently (and accounting that each stimulus may convey specific semantic importance between each contextual element), which may relate to top-down attention but not to the image characteristics per se. We also want to stress the importance of foveation in our model. This is a major procedure for determining saccade characteristics (including oculomotor tendencies) and saliency computations, as it determines current human actions during scene visualization. The decrease of spatial resolution at increasing eccentricity provides the aforementioned properties, innate in human vision and invariant to scene semantics.

Adding an IoR mechanism has been seen to affect model activity and therefore scanpath predictions. In Fig. 14-Left we show how our inhibition factor ($I_{\{Ior\}}$) decreases over simulation time in relation to the parametrized decay $\beta_{\{IoR\}}$, as well as the projected RF size with respect the gaussian parameter $\sigma_{\{IoR\}}$. These variables (decay and size) affect either location of saccades and its sequence, modulating firing rate activity to already visited locations. It is shown in Fig. 14-Right that the initial saccade is focused on the salient region and then it spreads to a specific location in the scene, not repeating with higher value of inhibition decay or field size. In the next section we show how our model can preproduce eye movements beyond free-viewing tasks by modulating of inhibitory top-down signals.

4.3. Results on feature and exemplar search

We have compared our model predictions with bottom-up only (NSWAM | NSWAM-CM) and with top-down inhibitory modulation (NSWAM+VS | NSWAM-CM+VS) for singleton search stimuli (for both real [10] and synthetic targets [75]). Top-down selection is applied to our low-level feature dimensions (scale, orientation, channel opponency and its polarity). In VS_M, inhibition is parametrized considering the feature with the highest activity inside the stimulus ROI (Equation 15-Top). Besides, inhibitory control in VS_C has been set as the mean wavelet coefficients instead (Equation 15-Bottom).

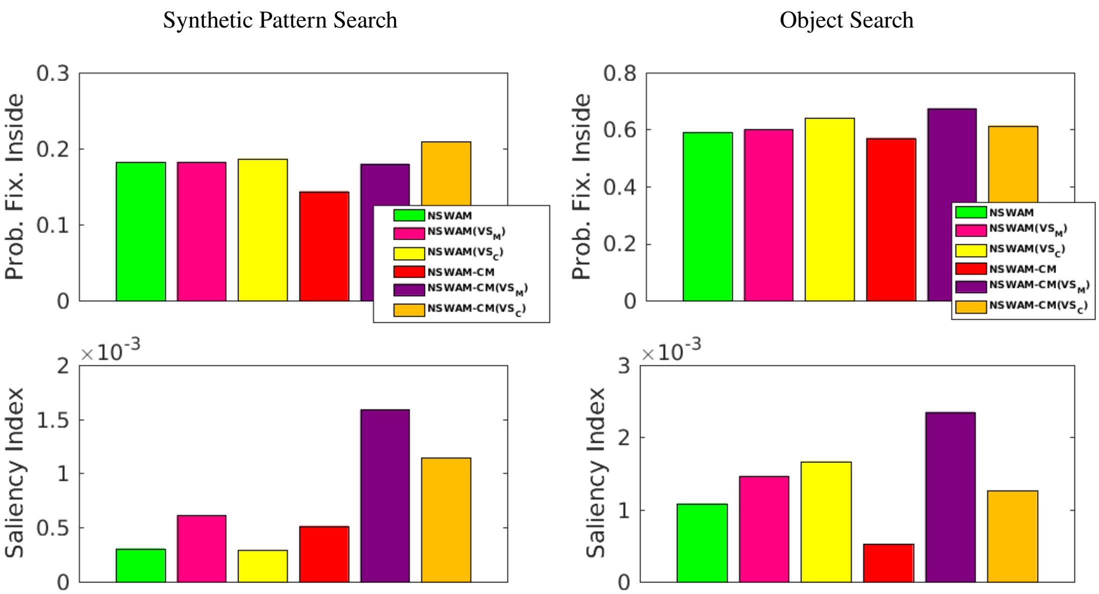


Figure 16: Probability of Fixations Inside the ROI (**Bottom row**) and statistics of Saliency Index (**Top row**) for synthetic image patterns (**Left**) and salient object detection regions from real image scenes (**Right**).

Results of our model predictions with top-down attention (NSWAM+VS | NSWAM-CM+VS) present higher scores for both SI and PFI (Fig. 16) than the case of bottom-up attention only (NWAM | NSWAM-CM), specially for the case of using cortical magnification NSWAM-CM+VS. Here, there is an increase of fixations inside the ROI: $\Delta(PFI)_{+VS_M}$ $\simeq$1%, $\Delta(PFI)_{-CM+VS_M}$ $\simeq$10% and $\Delta(PFI)_{VS_C}$ $\simeq$6%, $\Delta(PFI)_{-CM+VS_C}$ $\simeq$4% when searching real objects (Fig.16-Top/Right) and $\Delta(PFI)_{+VS_M}$ $\simeq$0%, $\Delta(PFI)_{-CM+VS_M}$ $\simeq$4% and $\Delta(PFI)_{+VS_C}$ $\simeq$1%, $\Delta(PFI)_{-CM+VS_C}$ $\simeq$7% when searching synthetic patterns (Fig.16-Top/Left). The SI is also seen to increase for both types of images, with differences of $\Delta(SI)_{+VS_M}$=3.8×10^{-4}, $\Delta(SI)_{-CM+VS_M}$=1.8×10^{-3} and $\Delta(SI)_{+VS_C}$=5.9×10^{-4}, $\Delta(SI)_{-CM+VS_C}$=7×10^{-4} for object search (Fig.16-Bottom/Right) and $\Delta(SI)_{+VS_M}$=3.1×10^{-4}, $\Delta(SI)_{-CM+VS_M}$=1.1× 10^{-3} and $\Delta(SI)_{+VS_C}$=1.3 × 10^{-5}, $\Delta(SI)_{-CM+VS_C}$=6 × 10^{-4} for psychophysical pattern search (Fig.16-Bottom/Left).

Some object localization examples are shown in Fig. 18, where the relevance maps (NSWAM+VS|NSWAM-CM+VS) seemingly capture the regions inside the ROI/mask compared to the cases of saliency maps (NSWAM|NSWAM-CM).

In Fig. 19 we illustrated results of PFI and SI in relation to relative target-distractor feature contrast for cases of Brigthness, Color, Size and Orientation differences. After computing SI for each distinct psychophysical stimuli, we can see in Fig. 17 that our model performs best for searching objects in stimuli where there are clear differences in brightness, color, size and/or angle, rather than for the case of different combination of features, specially with heterogeneous, nonlinear or categorical angle configurations.

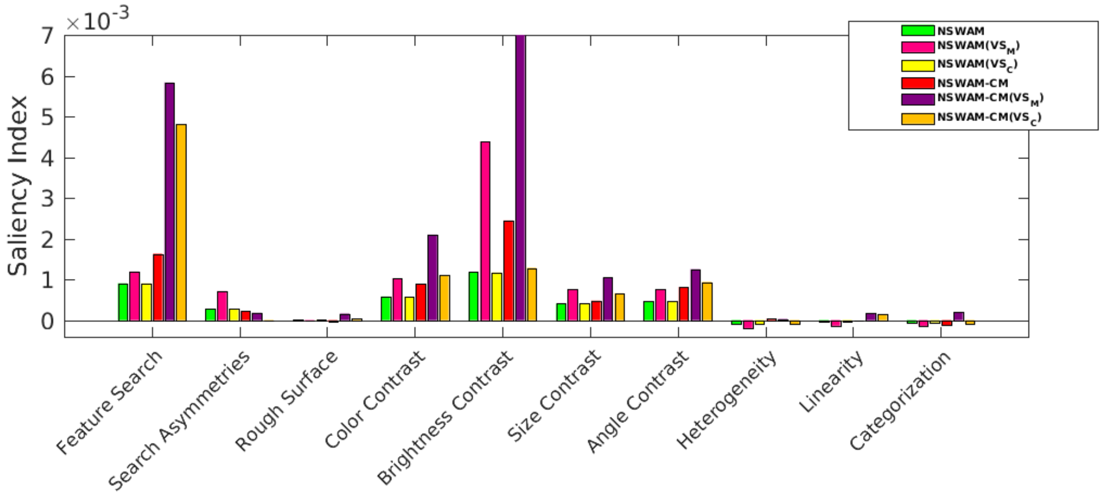


Figure 17: Performance on visual search evaluated on each distinct low-level feature, stimulus instances are from SID4VAM's dataset [17, 75].

4.3.1. Discussion

Overall results show that features computed by the top-down approach seemingly performs better in visual search than saliency, both considering features with maximal cortical activity (VS$_M$) and average statistics of low-level features (VS$_C$). Search in both objects and psychophysical image patterns is significantly more efficient in SI and PFI when selecting maximal feature activations (VS$_M$). Our model is able to localize objects in real scenes, specially when objects are distinct enough from others (in these low-level feature computations). However, the model fails when there are sparse regions of the image that interfere with the selected object (being too salient, such as in Fig. 18-"Telephone") and when characteristics of some parts of these objects (comprised in the mask) do not significantly pop-out or either coincide with other non-relevant objects (see Fig. 18-"Car"). This could be improved by computing a higher number of features [85, 86] (which would represent in more detail each cortical cell sensitivity at higher visual areas of cortex). We can observe that when using both cortical magnification transform and top-down selection (-CM+VS), some non-relevant parts of the image are discriminated easier than using top-down selection alone (see non-relevant artifacts caused by repetitive patterns or wrap-around filtering effects Fig. 19-Bottom). This suggests that using foveation not only can improve performance on localizing objects (Fig. 16) but also that provides biologically-plausible perceptual characteristics not considered in most artificial models. Even if our computations of top-down selection are fed to the model as a constant factor (according to the activity from exemplars), our model's lateral interactions leverage at each saccade the activity from both bottom-up and top-down attention.

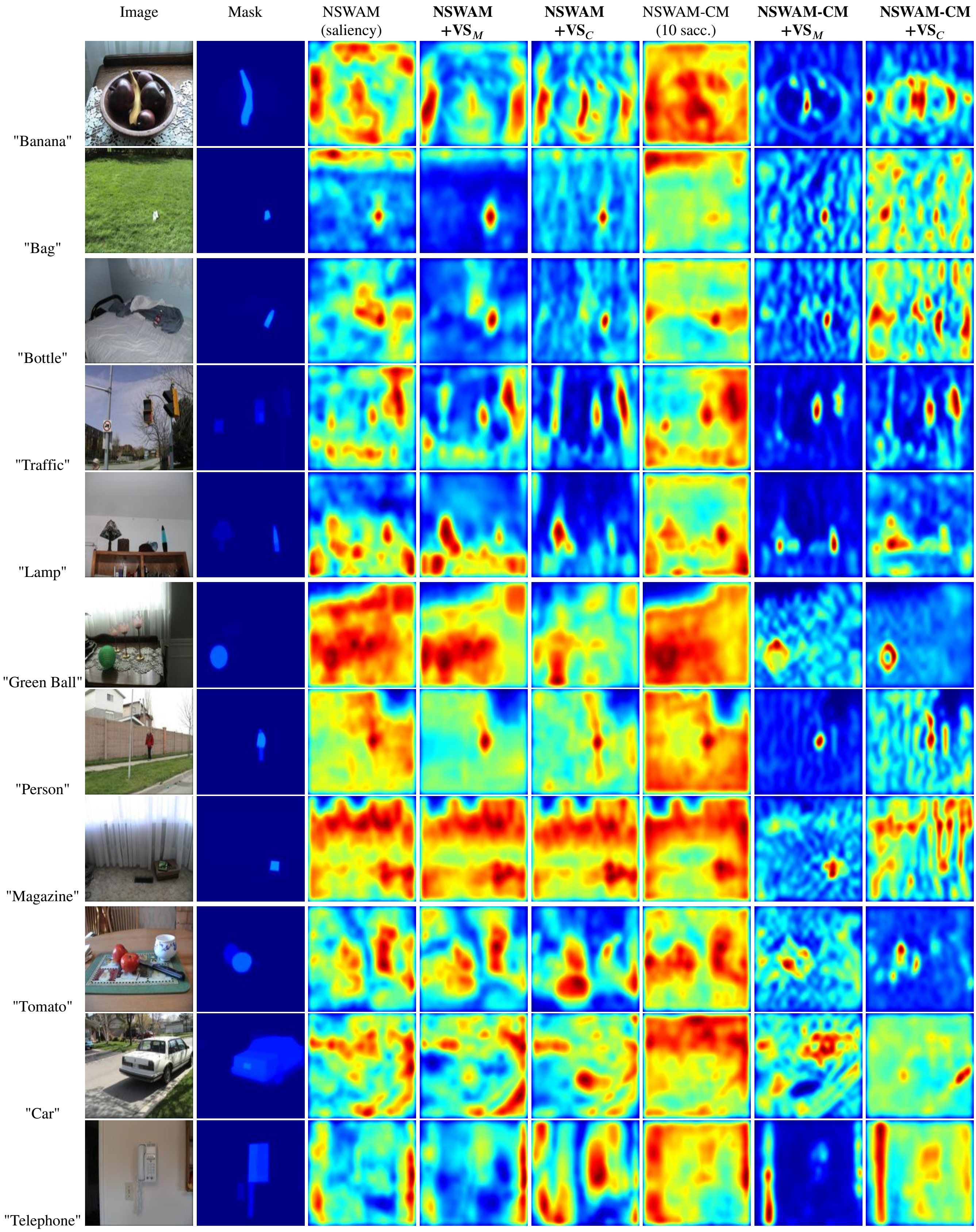


Figure 18: Search instances with a specific ROI (Mask) based on a category/word exemplar.

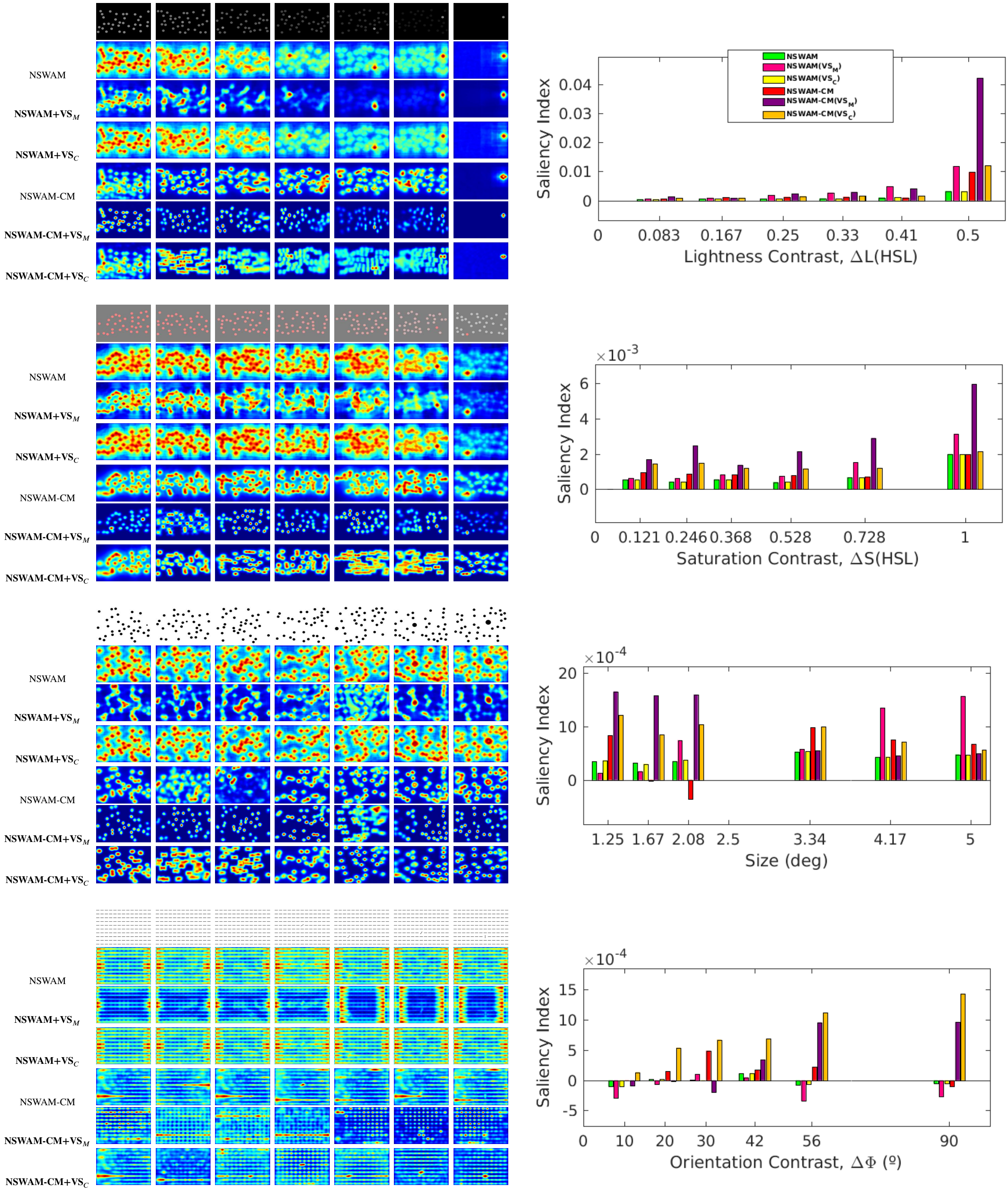


Figure 19: Performance on visual search examples with a specific low-level feature contrast (for Brightness, Color, Size and Orientation). We represented 7 instances ordered by search difficulty of each feature sample.

5. General Discussion

Current implementation of our V1 model is based on Li's excitatory-inhibitory firing rate network [26], following previous hypotheses of pyramidal and interneuron connectivity for orientation selectivity in V1 [55, 56]. To support and extend this hypothesis, distinct connectivity schemas (following up V1 cell subtype characterization) [87, 88] could be tested (e.g. adding dysynaptic connections between inhibitory interneurons) to better understand V1 intra-cortical computations. Furthermore, modeling intra-layer interactions of V1 cells [44] could explain how visual information is parallely processed and integrated by simple and complex cells [85], how distinct chromatic opponencies (P-,K- and M-) are computed at each layer [89], and how V1 responses affect SC activity (i.e. from layer 5) [90]. Testing contributions of each of these chromatic pathways (at distinct single/double opponencies and polarities), as well as distinct fusion mechanisms regarding feature integration, would define a more detailed description of how visual features affect saliency map predictions.

Previous and current scanpath model predictions could be considered to be insufficient due to the scene complexity and numerous factors (such as the task specificity, scene semantics, etc.) simultaneously involved in saccade programming. These factors increase overall errors on scanpath predictions, as systematic tendencies increase over time [17, 19, 80, 84], making late saccades difficult to predict. In that aspect, in free-viewing tasks (when there is no task definition), top-down attention is likely to be dependent on the internal state of the subject. Further understanding of high level attentional processes have only been approximated through statistical and optimization techniques uniquely with fixation data (yet participant decisions on fixations are not accounted and usually have high variability). It has also been later observed that fixations during free-viewing and visual search have distinct temporal properties. This could explain that saliency and relevance are elicited differently during viewing time. Latest literature on that aspect, discern two distinct patterns of fixations (either ambient or focal) where subjects first observe the scene (possibly towards salient regions), then focus their attention on regions that are relevant to them [70], and these influences are mainly temporal. Its modelization for eye movements in combination with memory processing is still under discussion. Current return mechanisms have long been computed by inhibiting the regions of previous fixations (spatially-based), nonetheless, IoR could also have feature-selective properties [91] to consider.

We suggest that not all fixations should have the same importance when evaluating saliency predictions. Nature and synthetic scene images lack of semantic (man-made) information, which might contribute to the aforementioned voluntary (top-down guided) eye movements [92]. Acknowledging that objects are usually composed by the combination of several features (either in shape, color, etc.), we should analyze if low-level features are sufficient to perform complex categorical search tasks. Extrastriate computations could allow the usage of object representations at higher-level processing, introducing semantically-relevant information and several image samples per category. Cortical processing of extrastriate areas (from V2 and V3) towards temporal (V4 & IT) and dorsal (V5 & MT) pathways [93, Section II][44] could represent cortical activity at these distinct levels of processing, modeling in more detail the computations within the two-stream hypothesis (what & where pathways). Color, shape and motion processing in each of these areas could generate more accurate representations of SC activity [20], producing more complex predictions such as microsaccadic and smooth pursuit eye movements with dynamic scenes.

6. Future Work

Current and future implementations of the model are able to process dynamic stimuli as to represent attention using videos. By simulating motion energy from V1 cells and MT direction selective cells [25, Section 2.3.5], would allow our model to reproduce object motion and flicker mechanisms found in the HVS. Moreover, foveation through more plausible cortical mapping algorithms [94] could provide better spatial detail of the cortical field organization of foveal and peripheral retinotopic regions and lateralization, currently seen to reproduce V1/V2/V3 physiological responses. Adding to that, hypercolumnar feature computations of geniculocortical pathways could be extended with a higher number of orientation and scale sensitivities with self-invertible 2D Log-Gabor filters [95]. In that regard, angle configuration pop-out effects and contour detection computations [96, 97] can be done by changing neuron connectivity and orientation tuning modulations. Spatiotemporal convolutions shown for center-surround RF [98] could be integrated for mimicking the dynamics and feature tuning at each pre-cortical pathway.

We aim in future implementations to model the impact of feedback in cortico-cortical interactions with respect striate and extrastriate areas in the HVS. Some of these regions project directly to SC, including the intermediate areas (pulvinar and medial dorsal) and basal ganglia [20, 64, 68]. Our current implementation can be extended with a large scale network of spiking neurons [99, 100], also being able to learn certain image patterns through spike-timing dependent plasticity (STDP) mechanisms [101]. With such a network, the same model would be able to perform both psychophysical and electrophysiological evaluations while providing novel biologically-plausible computations with large scale image datasets.

7. Conclusion

In this study we have presented a biologically-plausible model of visual attention by mimicking visual mechanisms from retina to V1 using real images. From such, computations at early visual areas of the HVS (i.e. RP, RGC, LGN and V1) are performed by following physiological and psychophysical characteristics. Here we state that lateral interactions of V1 cells are able to obtain real scene saliency maps and to predict locations of visual fixations. We have also proposed novel scanpath computations of scene visualization using a cortical magnification function. Our model outperforms other biologically inspired saliency models in saliency predictions (specifically with nature and synthetic images) and has a trend to acquire similar scanpath prediction performance with respect other artificial models, outperforming them in saccade amplitude correlations. The aim of this study, besides from acquiring state-of-the-art results, is to explain how lateral connections can predict visual fixations and how these can explain the role of V1 in this and other visual effects. In addition, we formulated projections of recurrent and selective attention using the same model (simulating frontoparietal top-down inhibition mechanisms). Our implementation of these, included top-down projections from DLPFC, FEF and LIP (regarding visual selection and inhibition of return mechanisms). We have shown how scanpath predictions improve by parametrizing the inhibition of return, with highest performance at a size of 2 deg and a decay time between 1 and 5 fixations. By processing low-level feature representations of real images (considering statistics of wavelet coefficients for each object or feature exemplar) and using them as top-down cues, we have been able to perform feature and object search using the same computational architecture. Two search strategies are presented, and we show that both the probability to gaze inside a ROI and the amount of fixations inside that ROI increase with respect saliency. In previous studies, the same model has been able to reproduce brightness [28] and chromatic [30] induction, as well as explaining V1 cortical hyperexcitability as a indicator of visual discomfort [31]. With the same parameters and without any type of training or optimization, NSWAM is also able predict bottom-up and top-down attention for free-viewing and visual search tasks. Model characteristics has been constrained (in both architecture and parametrization) with human physiology and visual psychophysics, and can be considered as a simplified and unified simulation of how low-level visual processes occur in the HVS.

Acknowledgments

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.

References

[1] C. E. Shannon. A mathematical theory of communication. *Bell System Technical Journal*, 27(3):379–423, jul 1948.
[2] H. B. Barlow. Redundancy reduction revisited. *Network*, 12 3:241–53, 2001.
[3] Li Zhaoping. From the optic tectum to the primary visual cortex: migration through evolution of the saliency map for exogenous attentional guidance. *Current Opinion in Neurobiology*, 40:94–102, oct 2016.
[4] Anne M. Treisman and Garry Gelade. A feature-integration theory of attention. *Cognitive Psychology*, 12(1):97–136, January 1980.
[5] Jeremy M. Wolfe, Kyle R. Cave, and Susan L. Franzel. Guided search: An alternative to the feature integration model for visual search. *Journal of Experimental Psychology: Human Perception and Performance*, 15(3):419–433, 1989.
[6] Christof Koch and Shimon Ullman. Shifts in selective visual attention: Towards the underlying neural circuitry. In *Matters of Intelligence*, pages 115–141. Springer Netherlands, 1987.
[7] 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.
[8] 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.
[9] Liming Zhang and Weisi Lin. *Selective Visual Attention*. John Wiley & Sons (Asia) Pte Ltd, mar 2013.
[10] Neil D. B. Bruce and John 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.
[11] Bruno A. Olshausen and David J. Field. Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature*, 381(6583):607–609, jun 1996.

[12] D. Brockmann. Are human scanpaths levy flights? In *9th International Conference on Artificial Neural Networks: ICANN '99*. IEE, 1999.

[13] Giuseppe Boccignone and Mario Ferraro. Modelling gaze shift as a constrained random walk. *Physica A: Statistical Mechanics and its Applications*, 331(1-2):207–218, jan 2004.

[14] Olivier Le Meur and Zhi Liu. Saccadic model of eye movements for free-viewing condition. *Vision Research*, 116:152–164, nov 2015.

[15] Jonathan Harel, Christof Koch, and Pietro Perona. Graph-based visual saliency. *Proc. Advances in Neural Information Processing Systems (NIPS 2007)*, 19:545–552, 2007.

[16] 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.

[17] David Berga, Xosé R. Fdez-Vidal, Xavier Otazu, Víctor Leborán, and Xosé M. Pardo. Psychophysical evaluation of individual low-level feature influences on visual attention. *Vision Research*, 154:60–79, jan 2019.

[18] Robert Desimone and John Duncan. Neural mechanisms of selective visual attention. *Annual Review of Neuroscience*, 18(1):193–222, mar 1995.

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

[20] Brian White and Douglas P. Munoz. *The Oxford Handbook of Eye Movements*. Oxford University Press, aug 2011.

[21] Edmund Rolls. *Memory, attention, and decision-making: a unifying computational neuroscience approach*. Oxford University Press, Oxford New York, 2008.

[22] 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.

[23] Liqiang Huang and Harold Pashler. A boolean map theory of visual attention. *Psychological Review*, 114(3):599–631, 2007.

[24] Zhaoping Li. A saliency map in primary visual cortex. *Trends in Cognitive Sciences*, 6(1):9–16, jan 2002.

[25] Li Zhaoping. *Understanding vision : theory, models, and data*. Oxford University Press, Oxford, United Kingdom, 2014.

[26] Zhaoping Li. A neural model of contour integration in the primary visual cortex. *Neural Computation*, 10(4):903–940, may 1998.

[27] 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.

[28] Olivier Penacchio, Xavier Otazu, and Laura Dempere-Marco. A neurodynamical model of brightness induction in v1. *PLoS ONE*, 8(5):e64086, may 2013.

[29] David Berga and Xavier Otazu. A neurodynamic model of saliency prediction in v1. *arXiv preprint arXiv:1811.06308*, 2018.

[30] Xim Cerda and Xavier Otazu. A Multi-Task Neurodynamical Model of Lateral Interactions in V1: Chromatic Induction. *39th European Conference of Visual Perception, PERCEPTION*, 45(2):51, 2016.

[31] Olivier Penacchio, Arnold J. Wilkins, Xavier Otazu, and Julie 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.

[32] Olivier Penacchio and Arnold J. Wilkins. Visual discomfort and the spatial distribution of fourier energy. *Vision Research*, 108:1–7, mar 2015.

[33] An T.D. Le, Jasmine Payne, Charlotte Clarke, Murphy A. Kelly, Francesca Prudenziati, Elise Armsby, Olivier Penacchio, and Arnold J. Wilkins. Discomfort from urban scenes: Metabolic consequences. *Landscape and Urban Planning*, 160:61–68, apr 2017.

[34] Gustavo Deco and Edmund T. Rolls. A neurodynamical cortical model of visual attention and invariant object recognition. *Vision Research*, 44(6):621–642, mar 2004.

[35] Yuqiao Gu and Hans Liljenström. A neural network model of attention-modulated neurodynamics. *Cognitive Neurodynamics*, 1(4):275–285, oct 2007.

[36] Sylvain Chevallier, Nicolas Cuperlier, and Philippe Gaussier. Efficient neural models for visual attention. In *Computer Vision and Graphics*, pages 257–264. Springer Berlin Heidelberg, 2010.

[37] Ruben Coen-Cagli, Peter Dayan, and Odelia Schwartz. Cortical surround interactions and perceptual salience via natural scene statistics. *PLoS Computational Biology*, 8(3):e1002405, March 2012.

[38] Hung-Cheng Chang, Stephen Grossberg, and Yongqiang Cao. Whereâ€™s waldo? how perceptual, cognitive, and emotional brain processes cooperate during learning to categorize and find desired objects in a cluttered scene. *Frontiers in Integrative Neuroscience*, 8, jun 2014.

[39] Mateja Marić and Dražen Domijan. A neurodynamic model of feature-based spatial selection. *Frontiers in Psychology*, 9, mar 2018.

[40] 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.

[41] Samuel G. Solomon and Peter Lennie. The machinery of colour vision. *Nature Reviews Neuroscience*, 8(4):276–286, apr 2007.

[42] Yasushi Imamoto and Yoshinori Shichida. Cone visual pigments. *Biochimica et Biophysica Acta (BBA) - Bioenergetics*, 1837(5):664–673, may 2014.

[43] Andrew Stockman, Donald I. A. MacLeod, and Nancy E. Johnson. Spectral sensitivities of the human cones. *Journal of the Optical Society of America A*, 10(12):2491, dec 1993.

[44] Lawrence C. Sincich and Jonathan C. Horton. THE CIRCUITRY OF v1 AND v2: Integration of color, form, and motion. *Annual Review of Neuroscience*, 28(1):303–326, jul 2005.

[45] P Lennie, J Krauskopf, and G Sclar. Chromatic mechanisms in striate cortex of macaque. *The Journal of Neuroscience*, 10(2):649–669, feb 1990.

[46] 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.

[47] H. Strasburger, I. Rentschler, and M. Juttner. Peripheral vision and pattern recognition: A review. *Journal of Vision*, 11(5):13–13, dec 2011.

[48] Marisa Carrasco. Covert attention increases contrast sensitivity: psychophysical, neurophysiological and neuroimaging studies. In *Visual Perception - Fundamentals of Vision: Low and Mid-Level Processes in Perception*, pages 33–70. Elsevier, 2006.

[49] Brian A. Wandell, Serge O. Dumoulin, and Alyssa A. Brewer. Visual field maps in human cortex. *Neuron*, 56(2):366–383, oct 2007.
[50] E. L. Schwartz. Spatial mapping in the primate sensory projection: Analytic structure and relevance to perception. *Biological Cybernetics*, 25(4):181–194, dec 1977.
[51] Zhaoping Li. Pre-attentive segmentation in the primary visual cortex. *Spatial Vision*, 13(1):25–50, 2000.
[52] Li Zhaoping. V1 mechanisms and some figure–ground and border effects. *Journal of Physiology-Paris*, 97(4-6):503–515, jul 2003.
[53] 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.
[54] Li Zhaoping and Li Zhe. Primary visual cortex as a saliency map: A parameter-free prediction and its test by behavioral data. *PLOS Computational Biology*, 11(10):e1004375, oct 2015.
[55] Charles D. Gilbert. Horizontal integration and cortical dynamics. *Neuron*, 9(1):1–13, jul 1992.
[56] Michael Weliky, Karl Kandler, David Fitzpatrick, and Lawrence C. Katz. Patterns of excitation and inhibition evoked by horizontal connections in visual cortex share a common relationship to orientation columns. *Neuron*, 15(3):541–552, sep 1995.
[57] Eilen Nordlie, Marc-Oliver Gewaltig, and Hans Ekkehard Plesser. Towards reproducible descriptions of neuronal network models. *PLoS Computational Biology*, 5(8):e1000456, aug 2009.
[58] 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.
[59] Brian J. White, Janis Y. Kan, Ron Levy, Laurent Itti, and Douglas 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.
[60] Peter H. Schiller and Edward J. Tehovnik. Chapter 9 look and see: how the brain moves your eyes about. In *Progress in Brain Research*, pages 127–142. Elsevier, 2001.
[61] Anja K.E. Horn and Christopher Adamczyk. Reticular formation. In *The Human Nervous System*, pages 328–366. Elsevier, 2012.
[62] Naila Murray, Maria Vanrell, Xavier Otazu, and C. Alejandro Parraga. Saliency estimation using a non-parametric low-level vision model. In *CVPR 2011*. IEEE, jun 2011.
[63] Olivier LeMeur and Thierry Baccino. Methods for comparing scanpaths and saliency maps: strengths and weaknesses. *Behavior Research Methods*, 45(1):251–266, jul 2012.
[64] Charles Pierrot-Deseilligny, Dan Milea, and René Müri. Eye movement control by the cerebral cortex. *Current Opinion in Neurology*, 17(1):17–25, feb 2004.
[65] Wloka, Calden; York University, Canada, Kotseruba, Iuliia; York University, Canada, and Tsotsos, John; York University, Canada. A focus on selection for fixation, 2016.
[66] Mehran Ahmadlou, Larry S. Zweifel, and J. Alexander Heimel. Functional modulation of primary visual cortex by the superior colliculus in the mouse. *Nature Communications*, 9(1), sep 2018.
[67] Yin Yan, Li Zhaoping, and Wu 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.
[68] C. Pierrot-Deseilligny, R.M. Müri, C.J. Ploner, B. Gaymard, and S. Rivaud-Péchoux. Cortical control of ocular saccades in humans: a model for motricity. In *Progress in Brain Research*, pages 3–17. Elsevier, 2003.
[69] J.D. Schall. Frontal eye fields. In *Encyclopedia of Neuroscience*, pages 367–374. Elsevier, 2009.
[70] Michelle L. Eisenberg and Jeffrey M. Zacks. Ambient and focal visual processing of naturalistic activity. *Journal of Vision*, 16(2):5, mar 2016.
[71] Richard Godijn and Jan Theeuwes. Oculomotor capture and inhibition of return: Evidence for an oculomotor suppression account of IOR. *Psychological Research*, 66(4):234–246, nov 2002.
[72] James W. Bisley and Michael E. Goldberg. Neural correlates of attention and distractibility in the lateral intraparietal area. *Journal of Neurophysiology*, 95(3):1696–1717, mar 2006.
[73] 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.
[74] Ali Borji and Laurent Itti. Cat2000: A large scale fixation dataset for boosting saliency research. *CVPR 2015 workshop on "Future of Datasets"*, 2015. arXiv preprint arXiv:1505.03581.
[75] David Berga, Xose R. Fdez-Vidal, Xavier Otazu, and Xose M. Pardo. Sid4vam: A benchmark dataset with synthetic images for visual attention modeling. In *The IEEE International Conference on Computer Vision (ICCV)*, October 2019.
[76] Stefan Winkler and Ramanathan Subramanian. Overview of eye tracking datasets. In *2013 Fifth International Workshop on Quality of Multimedia Experience (QoMEX)*. IEEE, jul 2013.
[77] Yin Li, Xiaodi Hou, Christof Koch, James M. Rehg, and Alan L. Yuille. The secrets of salient object segmentation. In *2014 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, jun 2014.
[78] 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.
[79] M.W. Spratling. Predictive coding as a model of the v1 saliency map hypothesis. *Neural Networks*, 26:7–28, feb 2012.
[80] Benjamin W. Tatler, Roland J. Baddeley, and Iain D. Gilchrist. Visual correlates of fixation selection: effects of scale and time. *Vision Research*, 45(5):643–659, mar 2005.
[81] James R. Antes. The time course of picture viewing. *Journal of Experimental Psychology*, 103(1):62–70, 1974.
[82] L. Zhaoping. Gaze capture by eye-of-origin singletons: Interdependence with awareness. *Journal of Vision*, 12(2):17–17, feb 2012.
[83] Arthur G. Samuel and Donna Kat. Inhibition of return: A graphical meta-analysis of its time course and an empirical test of its temporal and spatial properties. *Psychonomic Bulletin & Review*, 10(4):897–906, dec 2003.
[84] Lars O. M. Rothkegel, Hans A. Trukenbrod, Heiko H. Schütt, Felix A. Wichmann, and Ralf Engbert. Temporal evolution of the central fixation bias in scene viewing. *Journal of Vision*, 17(13):3, nov 2017.
[85] David A. Mély and Thomas Serre. Towards a theory of computation in the visual cortex. In *Computational and Cognitive Neuroscience of*

Vision, pages 59–84. Springer Singapore, oct 2016.

[86] David Berga, Calden Wloka, and John K Tsotsos. Modeling task influences for saccade sequence and visual relevance prediction. *Journal of Vision*, 19(10):106c, September 2019.

[87] Stephen Grossberg, Jesse Palma, and Massimiliano Versace. Resonant cholinergic dynamics in cognitive and motor decision-making: Attention, category learning, and choice in neocortex, superior colliculus, and optic tectum. *Frontiers in Neuroscience*, 9, jan 2016.

[88] Jung H. Lee, Christof Koch, and Stefan Mihalas. A computational analysis of the function of three inhibitory cell types in contextual visual processing. *Frontiers in Computational Neuroscience*, 11, apr 2017.

[89] E. N. Johnson, M. J. Hawken, and R. Shapley. The orientation selectivity of color-responsive neurons in macaque v1. *Journal of Neuroscience*, 28(32):8096–8106, aug 2008.

[90] Hoang L. Nhan and Edward M. Callaway. Morphology of superior colliculus- and middle temporal area-projecting neurons in primate primary visual cortex. *The Journal of Comparative Neurology*, 520(1):52–80, Nov 2011.

[91] Kesong Hu, Junya Zhan, Bingzhao Li, Shuchang He, and Arthur G. Samuel. Multiple cueing dissociates location- and feature-based repetition effects. *Vision Research*, 101:73–81, aug 2014.

[92] 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.

[93] John Werner and Leo M. Chalupa. *The new visual neurosciences*. The MIT Press, Cambridge, Massachusetts, 2014.

[94] Mark M. Schira, Christopher W. Tyler, Branka Spehar, and Michael Breakspear. Modeling magnification and anisotropy in the primate foveal confluence. *PLoS Computational Biology*, 6(1):e1000651, jan 2010.

[95] Sylvain Fischer, Filip Šroubek, Laurent Perrinet, Rafael Redondo, and Gabriel Cristóbal. Self-invertible 2d log-gabor wavelets. *International Journal of Computer Vision*, 75(2):231–246, jan 2007.

[96] Martin A. Asenov. Dynamic model of interactions between orientation selective neurons in primary visual cortex. Master's thesis, University of Edinburg, Edinburgh, UK, 2016.

[97] Akiyuki Anzai, Xinmiao Peng, and David C Van Essen. Neurons in monkey visual area v2 encode combinations of orientations. *Nature Neuroscience*, 10(10):1313–1321, sep 2007.

[98] DC Somers, SB Nelson, and M Sur. An emergent model of orientation selectivity in cat visual cortical simple cells. *The Journal of Neuroscience*, 15(8):5448–5465, August 1995.

[99] E.M. Izhikevich. Which model to use for cortical spiking neurons? *IEEE Transactions on Neural Networks*, 15(5):1063–1070, sep 2004.

[100] Amirhossein Tavanaei, Masoud Ghodrati, Saeed Reza Kheradpisheh, Timothée Masquelier, and Anthony Maida. Deep learning in spiking neural networks. *Neural Networks*, 2018.

[101] Timothée Masquelier and Simon J Thorpe. Unsupervised learning of visual features through spike timing dependent plasticity. *PLoS computational biology*, 3(2):e31, 2007.