ORIGINAL PAPER

Saliency prediction based on multi-channel models of visual processing

Qiang Li[1,2] (iD)

Abstract
Visual attention is one of the most significant characteristics for selecting and understanding the outside redundancy world. The human vision system cannot process all information simultaneously due to the visual information bottleneck. In order to reduce the redundant input of visual information, the human visual system mainly focuses on dominant parts of scenes. This is commonly known as visual saliency map prediction. This paper proposed a new psychophysical oriented saliency prediction architecture, which inspired by *multi-channel model* of visual cortex functioning in humans. The model consists of opponent color channels, wavelet transform, wavelet energy map, and contrast sensitivity function for extracting low-level image features and providing a maximum approximation to the low-level human visual system. The proposed model is evaluated using several datasets, including the MIT1003, MIT300, TORONTO, SID4VAM, and UCF Sports datasets. We also quantitatively and qualitatively compare the saliency prediction performance with that of other state-of-the-art models. Our model achieved strongly stable and better performance with different metrics on natural images, psychophysical synthetic images and dynamic videos. Additionally, we suggested that Fourier and spectral-inspired saliency prediction models outperformed other state-of-the-art non-neural network and even deep neural network models on psychophysical synthetic images. In the meantime, we suggest that deep neural networks need specific architectures and goals to be able to predict salient performance on psychophysical synthetic images better and more reliably. Finally, the proposed model could be used as a computational model of primate low-level vision system and help us understand mechanism of primate low-level vision system. The project page can be available at: https://sinodanishspain.github.io/HVS_SaliencyModel/.

Abbreviations

HVS	Human Vision System
V1	Primary Visual Cortex
ICL	Incremental Coding Length
CNN	Convolution Neural Network
DNN	Deep Neural Network
WT	Wavelet Transform
IWT	Inverse Wavelet Transform
AUC	Area Under Curve
NSS	Normalized Scanpath Saliency
CC	Pearson's Correlation Coefficient
SIM	Similarity or Histogram Intersection
IG	Information Gain
KL	Kullback–Leibler Divergence
CSFs	Contrast Sensitivity Functions
FT	Fourier Transform
DWT	Discrete Wavelet Transform
IDWT	Inverse Discrete Wavelet Transform
LGN	Lateral Geniculate Nucleus
GT	Ground Truth

✉ Qiang Li
qiang.li@uv.es

1 Image Processing Laboratory, University of Valencia, Valencia, Spain

2 Tri-Institutional Center for Translational Research in Neuroimaging and Data Science (TReNDS), Atlanta, USA

1 Introduction

The human vision system (HVS) uses visual attention to extract information from the redundancy of the natural world [1]. Redundancy in natural scene images is generally ineffective for scene classification or recognition. Focused visual

attention can be used to identify and remove irrelevant information from a cluttered natural environment, according to Barlow's *efficiency-coding hypothesis* [2]. Bottom-up and top-down visual attention mechanisms are the most common. Models built from the bottom up are primarily motivated by external stimuli. A variety of visual information (such as color, frequency, texture, orientation, and motion) is processed to extract image features using [3]. Contrast this with bottom-up approaches that aim to achieve a specific goal, which typically involve high-level information feedback and modulate lower-level vision functions [4]. Saliency map prediction has been successfully applied to both of the above computational neural models, and long-term research has been done on the visual attention mechanism [5].

Many studies have tried to carry out bottom-up or top-down computational modeling to predict saliency maps. In the following, I will briefly review some saliency prediction models that have achieved remarkable performance in saliency prediction. One of the earliest computational models, proposed by Itti et al. [6], was based on the bottom-up mechanisms of low-level vision systems. The model structure contains linear filtering, center-surround differences, across-scale combinations, and linear combinations. Achanta [7] devised a model for area segmentation that generates saliency maps to identify standout objects with clearly defined boundaries. Bruce and Tsotsos [8] proposed a model based on Shannon's self-information assessment to identity saliency map. Based on an investigation of the amplitude spectrum of natural images, Li [9] presented a novel bottom-up computational model for determining visual saliency. Hou and Zhang [10] proposed a spatial-temporal visual attention model based on feature rarity. Using the incremental coding length (ICL) method, they figured out each feature's perspective entropy gain and then made a saliency map. Murray et al. [11] showed a color appearance model could be used for producing saliency maps because it involves parameter selection and spatial pooling function. Boolean map-based saliency (BMS) is a novel Boolean map-based saliency model created by Zhang and Sclaroff [12]. According to Gestalt's theory of *figure-ground segregation*, the BMS model computes saliency maps by examining Boolean maps' topological structure [13]. A simple image descriptor known as the image signature was established by Hou et al. [14]. They developed a saliency algorithm based on the image signature. Goferman [15] described a context-aware saliency approach that seeks to recognize image regions that represent the scene. According to this approach, instead of identifying fixation locations, the dominating item is detected. Hou and Zhang [16] demonstrated a straightforward approach for generating the corresponding saliency map in the spatial domain by analyzing the log-spectrum of an input image. Guo [17] proposed a fast method that uses the spectral residual of the amplitude spectrum to build the saliency map. Schauerte and Stiefelhagen [18] firstly proposed employing Eigenaxes and Eigenangles for models of spectral saliency dependent on the Fourier transform. Murray [19] proposed a saliency model based on a low-level spatial-chromatic function of HVS, which successfully predicted chromatic induction phenomena and generated a saliency map [20]. To detect both static and spatial-temporal saliency, Seo and Milanfar [21] provided an innovative, unifying computational framework for detecting saliency. Wavelet transforms have also begun to be widely used to estimate computational vision saliency maps [11]. Compared to the Fourier transform, it has a high resolution in both the frequency and time domains. Wavelet can decompose signals at different scales, also known as multi-resolution/multi-scale analysis or sub-band coding, which can capture more low-level information from the original signal [22]. On the other hand, the wavelet transform approach can explain the primary visual cortex (V1) properties and produce multi-scale and multi-orientation features when provided with stimuli. The final estimated saliency map could sum up all the processed wavelet coefficients through an inverse wavelet transform [23]. When using the wavelet transform to create saliency maps, however, global contrast is lost rather than local information. Spratling [24] proposed a saliency prediction model based on *predictive coding theory* [25].

In the last few years, some studies have attempted to estimate saliency maps with deep convolutional neural networks (CNNs), which have achieved impressive performances compared to conventional methods [26–31]. Cornia et al. [32] introduced a novel deep architecture for saliency prediction. To predict saliency maps, current state-of-the-art models for saliency prediction use fully convolution networks, which utilize a non-linear combination of features extracted from the last convolution layer. DeepGazeII [33] is a model which forecasts where people will look in images. The model employs features from the VGG-19 deep neural network, which has been trained to recognize objects in images. Compared to conventional methods, deep learning implementations are also easier to transfer to real-life applications such as object detection, video understanding, and image compression. In this study, our main contributions are four-fold:

1. A psychophysically oriented saliency prediction model was proposed, which was inspired by the *multi-channel model* of human vision system function. The model has a contrast sensitivity function, an opponent color channel, a wavelet transform, a wavelet energy map, and a wavelet transform energy map. The proposed model is a bottom-up model, and it was tested on the different datasets using certain metrics.
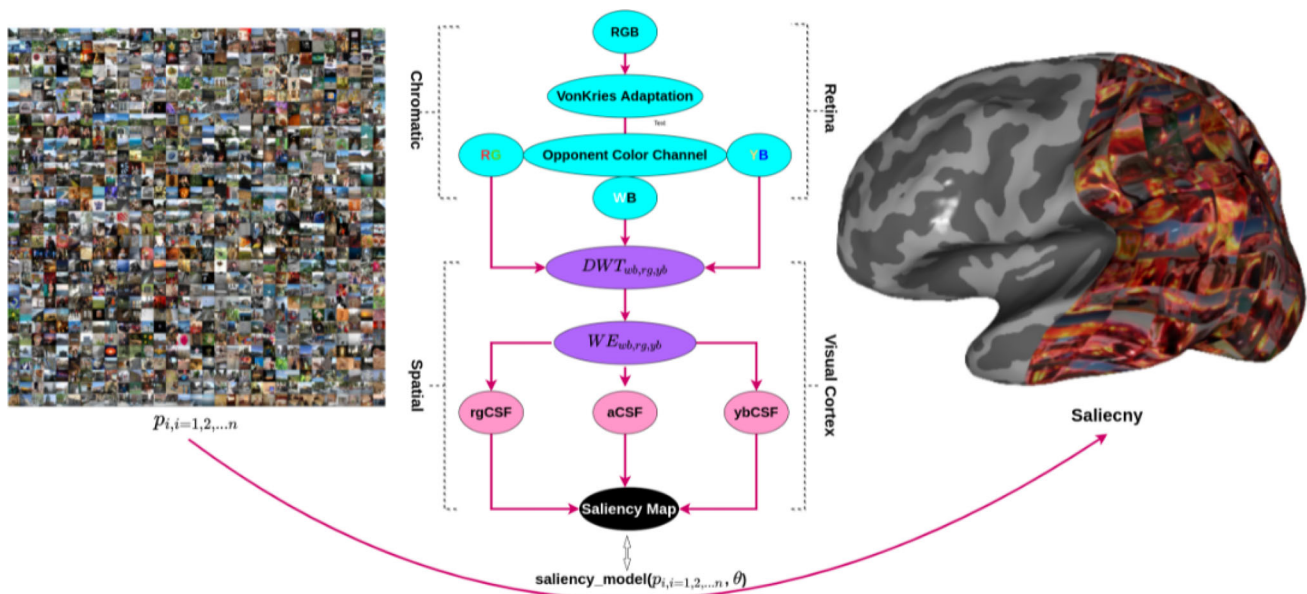
Fig. 1 Architecture of the proposed saliency prediction model. The left panel image was selected from the MIT1003 dataset. The flow chart shows the framework of the proposed model, containing the chromatic response in the retina and spatial feature processing in the visual cortex. The natural image is first adapted, before decomposing it into white-black, red-green, and yellow-blue opponent neural channels. In the spatial component, a discrete wavelet transform is applied to each opponent color channel, then the wavelet energy map is measured. In the last step of the proposed model, the CSF is applied to each opponent wavelet energy channel and combined with each opponent's feature. i and θ indicate image and model parameters, respectively. The details of each component are described in the following section. The graph on the right refers to the map of the left panel image's saliency on the inflated visual cortex using the proposed model

2. The spatial chromatic contrast sensitivity function was implemented by Python,[1] which is available at https://github.com/sinodanishspain/CSFpy.
3. The proposed model achieved strongly stable and better performance with different metrics on natural images, psychophysical synthetic images, and dynamic scenes. Beyond the accuracy of saliency prediction, we take more neuroscience concepts into account rather than statistical concepts, and it is also another computational goal in the study.
4. We suggested that Fourier and spectral-inspired saliency prediction models outperformed other state-of-the-art non-neural network and even deep neural network models on psychophysical synthetic images. The proposed model can be successfully applied to explain the "pop-out" effects in the visual search and attention mechanisms of primate vision systems and inspire the development of better deep learning models. Furthermore, we also suggest that deep neural networks require distinct architectures and goals in order to reliably predict salient performance on psychophysical synthetic images.

The rest of this paper is organized as follows: Sect. 2 introduces the concepts of opponent color space, wavelet decomposition, wavelet energy map estimation, and CSF. Section 3 introduces the saliency map prediction model, along with different datasets and evaluation metrics. Section 4 presents the experimental results. The final section provides discussions and conclusions for the paper.

2 The proposed saliency prediction model

2.1 Saliency prediction model

In this paper, we propose a biologically inspired visual saliency prediction map, based on the human low-level visual system. The extraction of information from the retina, LGN, and V1 is a critical component of visual neural networks. The color opponent channel, wavelet transform, wavelet energy map, and contrast sensitivity function are the main components of the proposed model architecture. The color opponent channel simulates the response of retinal cells to different spectral wavelengths, and the wavelet transform presents the multi-scale and multi-orientation properties of the V1. The CSF is used to describe the human brain's susceptibility to spatial frequencies. The details of each component are described in the following sections. Figure 1 depicts the computational saliency prediction model architecture.
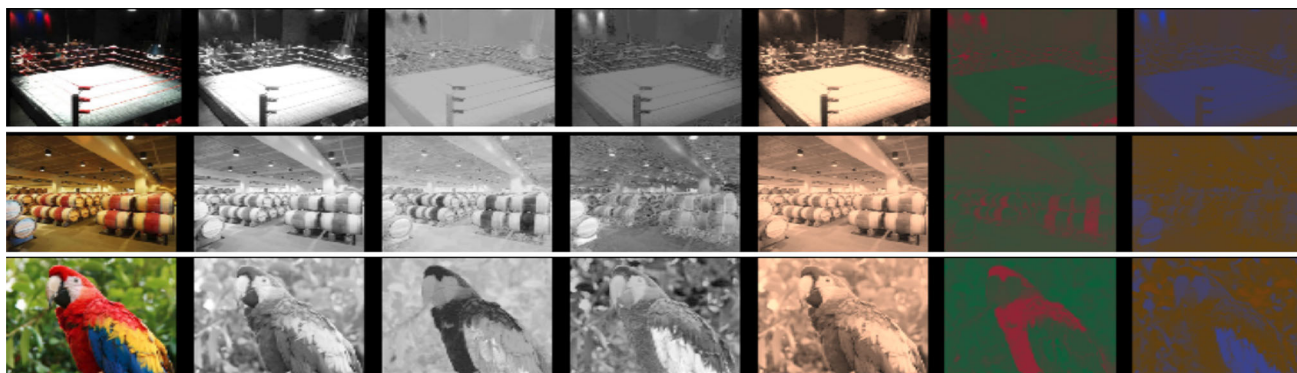
---

[1] https://www.python.org/.

Fig. 2 Opponent color processing. The first column represents the raw RGB color space, followed by the white–black (WB) channel, red–green (RG) channel, and yellow–blue (YB) channel, each with a gray colormap. The final three columns depict the WB, RG, and YB channels in artificial color, in order to better visualize the opponent color processing in the visual system

2.2 Gain control with von Kries chromatic adaptation model

Gain control exists in the visual information processing pipeline in the retina and cortex. In other words, gain control influences both top-down and bottom-up visual information flows, as well as attention-related cognitive functioning [34]. Meanwhile, gain control always strives to maintain a steady-state brain and self-regulation condition between the brain and the natural environment. In the von Kries model, we multiply each channel of the image with the gain value after normalizing its intensity [35–37]. However, there are some implications to this approach. The first is that the channels are considered independent signals, which is why we use independent gains. Second, this gain is added not in the RGB space but, instead, in the tristimulus LMS space. Assuming that the LMS is the same as our image's tristimulus values, the von Kries model can be written in math as:

$$L_2 = \frac{L_1}{L_{\max}} L_{\max 2},$$
$$M_2 = \frac{M_1}{M\max} M_{\max 2},$$
$$S_2 = \frac{S_1}{S\max} S_{\max 2}, \tag{1}$$

$$\begin{bmatrix} L_{post} \\ M_{post} \\ S_{post} \end{bmatrix} = \begin{bmatrix} \frac{1}{L_{\max}} & 0 & 0 \\ 0 & \frac{1}{M_{\max}} & 0 \\ 0 & 0 & \frac{1}{\sin ax} \end{bmatrix} \begin{bmatrix} L_1 \\ M_1 \\ S_2 \end{bmatrix}, \tag{2}$$

$$\begin{bmatrix} L_2 \\ M_2 \\ S_2 \end{bmatrix} = \begin{bmatrix} L_{\max 2} & 0 & 0 \\ 0 & M_{\max 2} & 0 \\ 0 & 0 & S_{\max 2} \end{bmatrix}, \tag{3}$$

where L_1 corresponds to the original image's L values; L_{Max}, M_{Max}, and S_{Max}, respectively, correspond to the maximum value of each channel in the LMS image; L_{Max2}, M_{Max2}, and S_{Max2} are the gain values with a set value of 0.6 in the proposed model; and L_2 is the corrected L channel after adaptation.

2.3 Color appearance model

Representation of color in the brain can improve object recognition and identity. Trichromatic theory [38] and the color appearance model proposed based on the functioning of the sensors encode color information and have been widely used in low-level image processing. Two functional types of chromatic sensitivity or selectivity sensors were found—single-opponent and double-opponent neurons—based on the responses of long (L), mediate (M), and short (S) cones in the physical world [39]. Most saliency prediction models use CIELAB and YUV color spaces for the opponent color spaces. In our case, we use another opponent's color space [40, 41], and the color space transform matrix from RGB to $O_1 O_2 O_3$ can be expressed as:

$$\begin{bmatrix} O_1 \\ O_2 \\ O_3 \end{bmatrix} = \begin{bmatrix} 0.2814 & 0.6938 & 0.0638 \\ -0.0971 & 0.1458 & -0.0250 \\ -0.0930 & -0.2529 & 0.4665 \end{bmatrix} \begin{bmatrix} R \\ G \\ B \end{bmatrix}. \tag{4}$$

The test natural scene images (of sizes 256 × 256 and 512 × 512) were selected from the Signal and Image Processing Institute, University of Southern California, [2] and the Kodak lossless true-color image database [3] (of sizes 512 × 768, 768 × 512, and 768 × 512). The total natural color images were resized into the same size (8 bits, 256 × 256) as test images. All natural chromatic images were converted from RGB space to the $O_1 O_2 O_3$ domain, based on the above conversion matrix. As can be seen in Fig. 2, the chromatic information
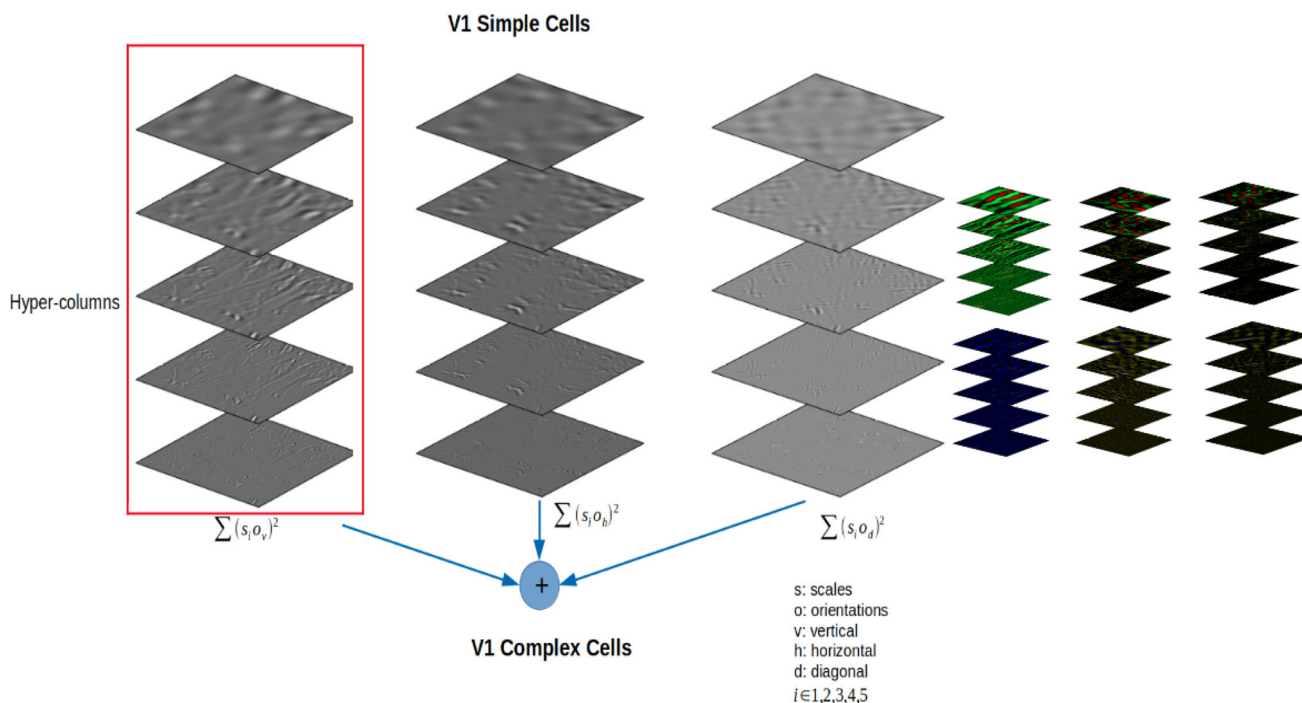

V1 Simple Cells

Fig. 3 The modeling of V1 simple and complex cells in each opponent channel. The red rectangle indicates hypercolumns in the visual cortex. From left to right, the graph depicts WB opponent neurons with different orientations and scales. The following zoomed out top/bottom graphs with artificial color, for better visualization of features, in each hypercolumn indicate RG/YB opponent neurons across different orientations and scales. The V1 complex cells can be obtained from the sum of squares of wavelet transform features across scales and orientations in the simple cells

(white-black, red-green, and yellow-blue) was decomposed into each channel.

2.4 Wavelet energy map

2.4.1 Visual cortex receptive fields with wavelet filters

The primary visual cortex contains neurons that reflect the structure of the retinal image in terms of a wavelet basis, and the visual simple and complex cells can be modeled with wavelet filters. In our case, we did not consider the in-depth details of each hypercolumn neuron's interaction mechanisms (e.g., Li's model [42]). The simulated V1 complex receptive fields sum all the squares of different scales and orientations after the wavelet transform (see Fig. 3). The V1 simple receptive fields in each opponent channel are mathematically defined as:

$$V_{iv} = s_i o_v \tag{5}$$
$$V_{ih} = s_i o_h \tag{6}$$
$$V_{id} = s_i o_d, \tag{7}$$

where s indicates receptive filed scales, o refers to orientation—

that is, vertical (v), horizontal (h), and diagonal (d)—and i indicates the number of neurons/features. The V1 complex cells can be formulated as:

$$V_{\text{complex}} = \sum_1^i (s_i o_v)^2 + \sum_1^i (s_i o_h)^2 + \sum_1^i (s_i o_d)^2. \tag{8}$$

2.4.2 Wavelet transform and wavelet energy map

The wavelet image analysis can decompose an image into multi-scale and multi-orientation features, similar to the visual cortex representation. Compared to the Fourier transform (FT), a wavelet transform can represent spatial and frequency information simultaneously. Alfred Haar first proposed the wavelet transform approach, and it has already been widely used in signal analysis [43]; for example, for image compression, image denoising, and classification. Wavelet transforms have already been applied in visual saliency map prediction, and achieved good performance [44]. However, wavelet energy maps remain barely used in visual saliency map prediction, and they can be used to enhance local contrast information in the decomposition sub-bands. In our
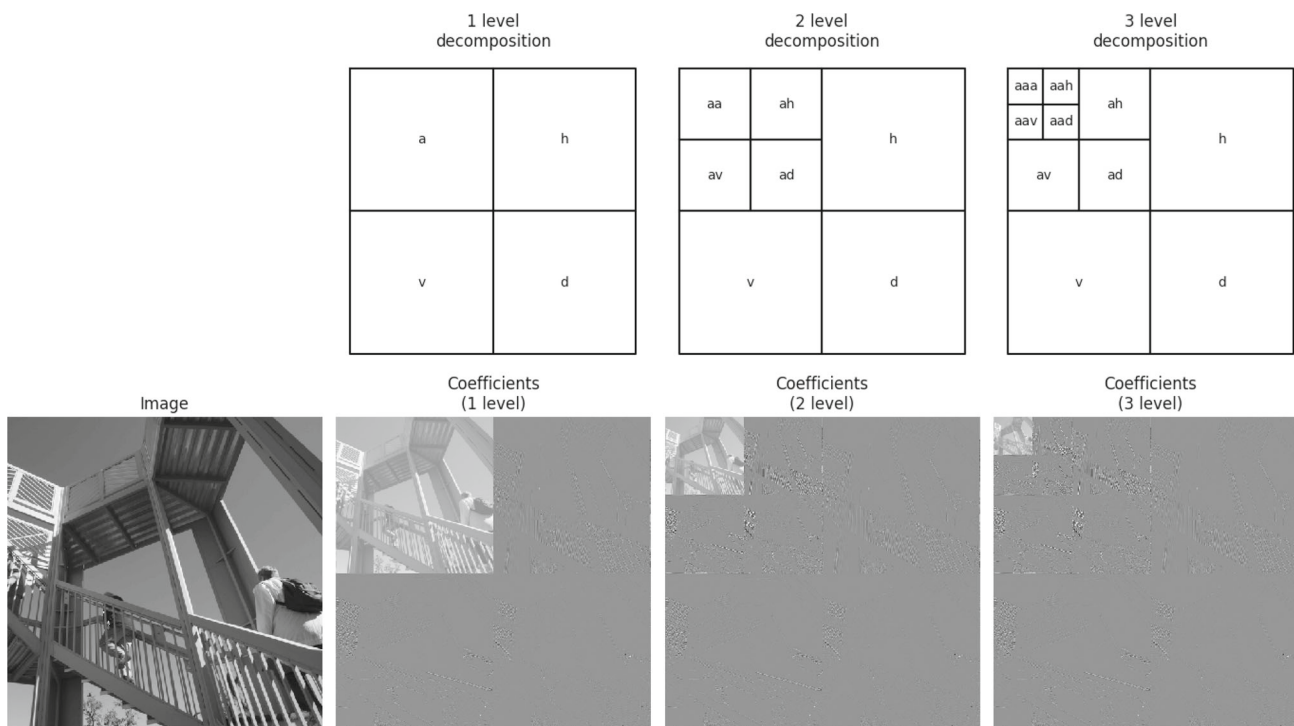
Fig. 4 The different decomposition levels of the DWT (e.g., first, second, and third levels): "a" indicates the original image, "h" indicates the horizontal feature, "v" refers to the vertical feature, and "d" represents the diagonal feature. The bottom-left image is the original image, and the following images represent the first-, second-, and third-level decomposition features from the original image

proposed model, we use a discrete wavelet transform (DWT), which can be written in math as:

$$r[n] = ((I * f)[n]) \downarrow 2 = \left(\sum_{k=-\infty}^{\infty} I[k]f[n-k] \right) \downarrow 2 \quad (9)$$

where I indicates the input images, f represents a series of filter banks (low-pass and high-pass), and $\downarrow 2$ indicates down-sampling until the next layer's signal cannot be decomposed any more (see Fig. 4). A series of sub-band images are produced after convolution with the DWT; then, the wavelet energy map can be calculated from each sub-band feature (see Fig. 5).

The wavelet energy map can be expressed as:

$$\mathcal{WE}(i, j) = \|I(i, j)\|^2 = \sum_{k=1}^{3ind+1} |I_k(i, j)|^2, \quad (10)$$

where $3ind$ indicates the maximum level of an image that can be decomposed in the last level, e.g., 3 level decomposition, and $I_k(i, j)^2$ represents the energy map of each sub-band feature.

2.5 Contrast sensitivity function

The human visual system is sensitive to contrast changes in natural environments. The visual cortex function can be decomposed into subset compositions, where one of the significant features is the CSF, which can be divided into achromatic and chromatic spatial CSFs [45]. In this proposed computational model, an achromatic CSF (aCSF) and chromatic CSFs (rgCSF and ybCSF) were implemented, which was first proposed by Mannos and Sakrison in 1974 [46], and further improved later [47, 48] (see Fig. 6). The achromatic CSF mathematics is as follows:

$$CSF(f_x, f_y) = Q(f) * L(f_x, f_y), \quad (11)$$
$$Q(f) = g * \left(\exp(-(f/f_m)) - l * \exp\left(-\left(f^2/s^2\right)\right)\right), \quad (12)$$

$$L(f_x, f_y) =$$
$$1 - w * \left(4(1 - \exp(-(f/os))) * f_x^2 * f_y^2 \right)/f^4, \quad (13)$$

where (f_x, f_y) indicates a 2D spatial frequency vector (in cycle/deg), f represents the modulus of the spatial frequency (cycle/deg), g represents the overall gain ($g = 330.74$), f_m is a parameter that controls the exponential decay of the CSF Tyler ($f_m = 7.28$), l represents the loss at low frequencies

Fig. 5 Each channel's DWT map and the wavelet energy maps corresponding to it. The first column shows the DWT maps for achromatic (WB) and chromatic (RG, YB) channels. The second column is the wavelet energy map, obtained by summing across scales and orientation features for WB, RG, and YB opponent channels, respectively. The last column shows the sum of squares energy maps in each opponent channel
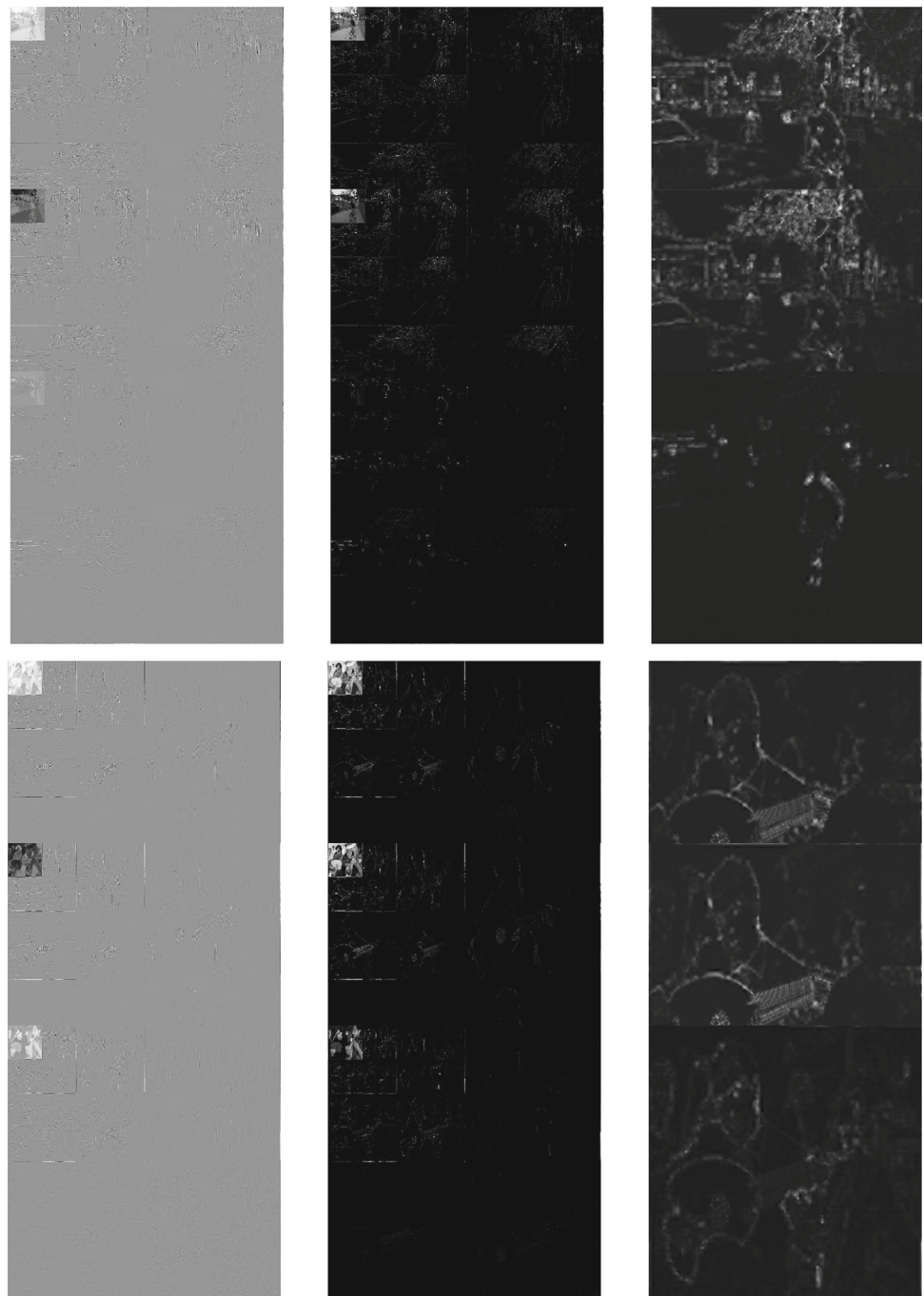
($l = 0.837$), s is a parameter that controls the attenuation of the loss factor at high frequencies ($s = 1.809$), w indicates the weighting of the oblique effect ($w = 1$), and os indicates the oblique effect scale ($os = 6.664$). The CSFs were applied to the wavelet energy image in the Fourier domains. It can be described by the following formula:

$$CSF_{WE} = real(\mathcal{F}(\mathcal{I}(\mathrm{I}(\mathrm{F}(WE.real)) \odot CSF))), \quad (14)$$

where F indicates the 2D Fourier transform, $\mathcal{F}$ indicates the 2D inverse Fourier transform, I indicates *fftshift*, which rearranges a Fourier transform by shifting the zero-frequency component to the center of the image, and $\mathcal{I}$ indicates *ifftshift*, which rearranges a zero-frequency-shifted Fourier transform back to the original transform output. In other words, *ifftshift* undoes the result of *fftshift*. The Python implementations of the above CSFs (aCSF, rgCSF, and ybCSF) are available at https://github.com/sinodanishspain/CSFpy.
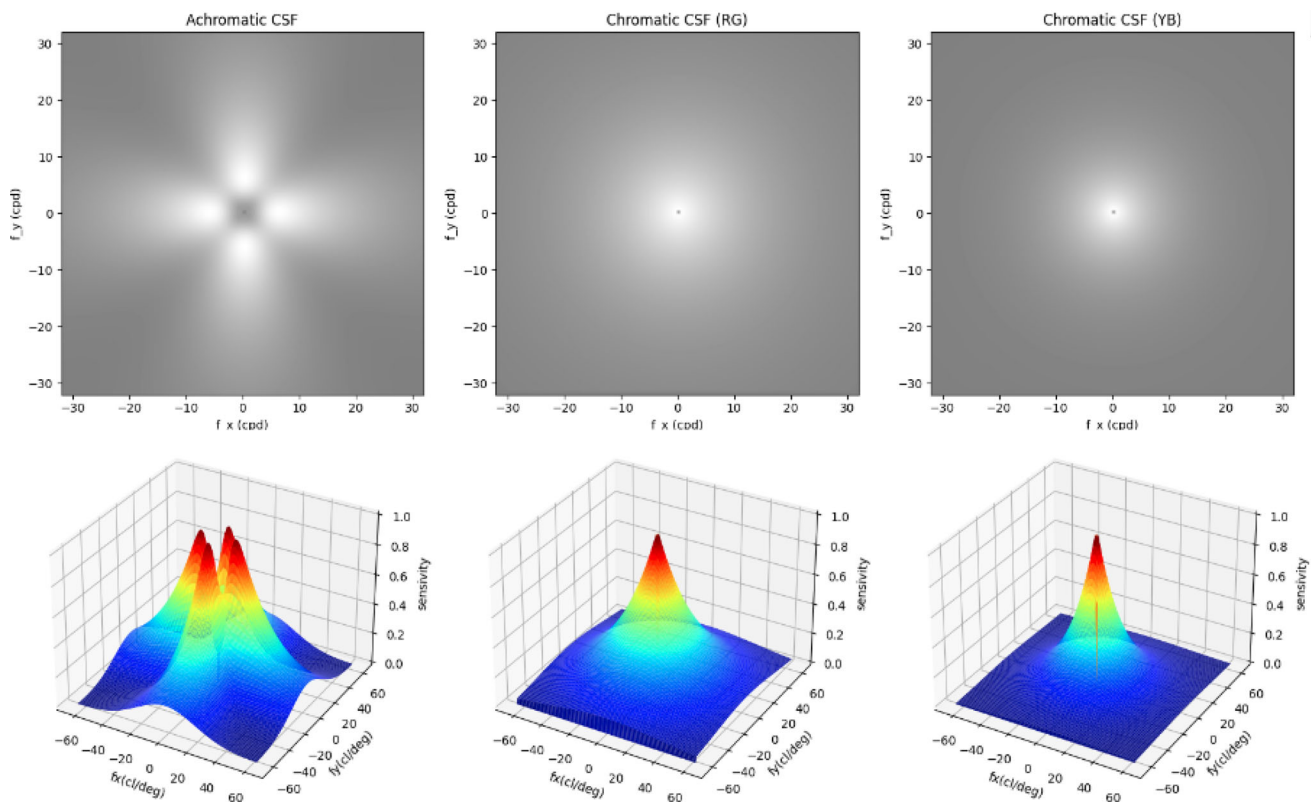
Fig. 6 Achromatic and chromatic CSFs. The images in the top row are 2D CSFs, and the bottom row shows 3D CSFs

3 Materials and methods

3.1 datasets

The proposed model was tested on several well-known datasets, including MIT1003, MIT300, TORONTO, and SID4VAM. The following sections introduce the basic information of each dataset.

- MIT1003 is an image dataset that includes 1003 images from the Flickr and LabelMe collections. The fixation map was generated by recording the eye-tracking data of 15 participants. It is the largest eye-tracking dataset. The dataset includes 779 landscape and 228 portrait images with sizes spanning from 405×405 to 1024×1024 pixels [49].
- MIT300 is a benchmark saliency test dataset that includes 300 images obtained by recoding a 39-observer eye-tracking dataset. The MIT300 dataset categories are highly varied and natural. The dataset can be used for model evaluation [49].
- TORONTO includes 120 chromatic images free-viewed by 20 subjects. The dataset contains both outdoor and indoor scenes with a fixed resolution of 511×681 pixels [8].

- SID4VAM is a synthetic image database that is mainly used to psychophysically evaluate the V1 properties. This database is composed of 230 synthetic images, including 15 distinct types of low-level features (e.g., brightness, size, color, and orientation) with different target-distractor pop-out-type synthetic images [50].

3.2 Evaluation metrics

As mentioned before, there are several approaches to evaluate metrics between visual saliency and model prediction. In general, saliency evaluation can be divided into two branches: location-based and distribution-based. The former mainly focuses on the district located in the saliency map, and the latter considers both the predicted saliency and human eye fixation maps as continuous distributions [51]. In this research, we used AUC, NSS, CC, SIM, IG, and KL to evaluate the methods and details of each evaluation metric, as described in the following.[4]

3.2.1 Area under the ROC curve (AUC)

The AUC metric is a popular approach for the evaluation of saliency model performance. The saliency map can be

---

[4] https://github.com/cvzoya/saliency.

treated as a binary classifier to split positive samples from negative samples by setting different thresholds. The true positive (TP) is the proportion of salient map values beyond a specific threshold at the fixation locations. In contrast, the false positive (FP) is the proportion of salient map values beyond a specific threshold at the non-fixation locations. In our case, the thresholds were set from the saliency map values and the AUC-Judd, AUC-Borji, and sAUC measures [52].

3.2.2 Normalized scanpath saliency (NSS)

The NSS metric usually measures the relationship between human eye fixation maps and model-predicted saliency maps [53]. The NSS can be formally defined as follows given a binary fixation map F and a saliency map S:

$$NSS = \frac{1}{N} \sum_{i=1}^{N} \bar{S}(i) \times F(i), \tag{15}$$

$$N = \sum_i F(i) \text{ and } \bar{S} = \frac{S - \mu(S)}{\sigma(S)}, \tag{16}$$

where N is the total number of human eye positions, $\mu(s)$ is the mean value of saliency maps, and $\sigma(S)$ is the standard deviation.

3.2.3 Similarity metric (SIM)

The similarity metric (SIM) is a very famous algorithm for measuring image structure similarity, which has already been widely used in image quality and image processing disciplines [54]. The SIM mainly measures the normalized probability distributions of eye fixation and model-predicted saliency maps. The SIM can be mathematically described as:

$$SIM = \sum_{i=1}^{N} \min(P(i), Q(i)), \tag{17}$$

where $P(i)$ and $Q(i)$ are the normalized saliency map and the fixation map, respectively. A similarity score should be in the range between zero and one.

3.2.4 Information gain (IG)

Information gain is an approach to measuring saliency map prediction accuracy from an information-theoretic view. It mainly measures the critical information contained in the predicted saliency map, compared with a ground-truth map [55]. The mathematical formula for the IG can be expressed as:

$$IG\left(P, Q^B\right) = \frac{1}{N} \sum_i Q_i^B$$

$$\left[\log_2 (\epsilon + P_i) - \log_2 (\epsilon + B_i)\right], \tag{18}$$

where P indicates the predicted saliency map, Q^B is the baseline map, and ϵ represents a regularity parameter.

3.2.5 Pearson's correlation coefficient (CC)

Pearson's correlation coefficient (CC) is a linear approach that measures how many similarities there are between the predicted saliency map and the baseline map [56].

$$CC\left(P, Q^D\right) = \frac{\sigma\left(P, Q^D\right)}{\sigma(P) \times \sigma\left(Q^D\right)}, \tag{19}$$

where P indicates the predicted saliency map and Q^D is the ground-truth saliency map.

3.2.6 Kullback–Leibler divergence (KL)

The Kullback–Leibler divergence (KL) is used to measure the distance between the samples of two distributions from an information-theoretic perspective [55]. It can be formally defined as:

$$KL\left(P, Q^D\right) = \sum_i Q_i^D \log\left(\epsilon + \frac{Q_i^D}{\epsilon + P_i}\right), \tag{20}$$

where P indicates the predicted saliency map, Q^D is the ground-truth saliency map, and ϵ represents a regularity parameter.

3.2.7 Other metrics

We also evaluated the performance of different salient prediction models through two main metrics: precision-recall curves (PR curves) and the F-measure.[5] By binarizing the predicted saliency map with thresholds in [0,255], a series of precision and recall score pairs were calculated for each dataset image. The PR curve was plotted using the average precision and recall of the dataset under different thresholds [57].

4 Experimental results

4.1 Quantitative comparison of the proposed model with other state-of-the-art models

To evaluate the performance of the proposed model, we compared it with eight other state-of-the-art models. For

---

Table 1 Quantitative scores of several models for the MIT1003 dataset. The baseline ITT model is shaded in light blue and the proposed model is shown in green

Methods	DNN	AUC_Judd	AUC_Borji	sAUC	NSS	SIM
ITT	*B*	*0.674*	*0.655*	*0.610*	*0.629*	*0.291*
Ours	**N**	**0.705**	**0.692**	**0.653**	**0.849**	**0.362**
SR	N	0.708	0.683	0.638	0.791	0.329
AIM	N	0.706	0.696	0.639	0.780	0.282
BMS	N	0.684	0.637	0.576	0.729	0.346
CASD	N	0.747	0.731	0.651	0.977	0.350
DCTS	N	0.746	0.732	0.650	1.000	0.322
HFT	N	0.797	0.764	0.619	1.258	0.416
ICL	N	0.769	0.713	0.617	1.048	0.420
PFT	N	0.708	0.683	0.636	0.787	0.326
PQFT	N	0.643	0.530	0.519	0.459	0.292
QDCT	N	0.736	0.714	0.647	0.920	0.338
RARE	N	0.777	0.755	0.665	1.198	0.380
SIM	N	0.701	0.693	0.653	0.743	0.283
SUN	N	0.665	0.647	0.601	0.629	0.287
Achanta	N	0.534	0.526	0.526	0.174	0.240
Simpsal	N	0.735	0.721	0.610	0.892	0.337
Spratling	N	0.512	0.508	0.510	0.039	0.234
SIMgrouping	N	0.724	0.716	0.668	0.873	0.308
SeoMilanfar	N	0.710	0.688	0.633	0.808	0.351
ML_Net	*Y*	*0.836*	*0.743*	*0.689*	*1.928*	*0.565*
DeepGazeII	*Y*	*0.886*	*0.837*	*0.779*	*2.483*	*0.527*

The black-bold scores demonstrate the performance of our model's saliency prediction. "N" indicates NO, "Y" indicates YES, and "B" indicates Baseline. The results of ML_Net and DeepGazeII models for the MIT1003 dataset are shown in pink, as this dataset was used to train ML_Net and DeepGazeII and, so, their results could not be compared with those of the other models. Table 5 includes an appendix with model details

comparison of the quantitative results, we selected the MIT1003 and SID4VAM benchmarks. These results are reported in Tables. 1, 2, and 3. The superior performance, in terms of saliency prediction, was achieved by models based on biological/cognitive and Fourier/spectral foundations. Our model achieved stable and superior performance, in terms of different evaluation metrics compared to other biological/cognitive- and Fourier/spectral-inspired models. However, saliency map prediction based on a convolutional neural network outperformed other models in natural scene images, as more images were used to train the neural network. Consequently, these cannot be compared to other models, as they are based more on statistical than neuroscientific principles. In this paper, we emphasize understanding saliency prediction from a neuroscience perspective, in order to further help us understand the mechanism of visual attention cognitive function. Furthermore, biological/cognitive- and Fourier/spectral-inspired saliency detection models were outperformed by deep learning approaches (ML_Net and DeepGazeII) in the SID4VAM dataset (see Tab. 3 and Fig. 10). As previously said, SID4VAM is a synthetic image database that is primarily used to psychophysically test the V1 properties, which is also why we stated that deep learn-

ing models refer more to statistics than neuroscience in the explanation of human visual attention mechanisms.

4.2 Qualitative comparison of the proposed model with other state-of-the-art models

We qualitatively tested the proposed model using the MIT1003, MIT300, TORONTO, SID4VAM, and UCF Sports datasets.[6] We also compared the model's performance with that of other state-of-the-art saliency prediction models on the MIT1003, TORONTO, and SID4VAM datasets. Figures 7, 8, 9, 10, and 12 show the saliency map results when the proposed model and other state-of-the-art models were applied to sample images from the studied datasets. The performance each of saliency prediction model was evaluated through AUC and PR curves, as shown in Fig. 11. We can see that the proposed model could predict most of the salient objects in the given images. Furthermore, the proposed model could successfully detect the orientation, boundary, and pop-out functions when the model was applied to the SID4VAM dataset. In summary, our

---

[6] https://www.crcv.ucf.edu/data/UCF_Sports_Action.php

Table 2 Quantitative scores of several models for the TORONTO dataset. The baseline ITT model is shown in light blue and the proposed model is shown in green

Methods	DNN	AUC_Judd	AUC_Borji	sAUC	NSS	SIM
ITT	*B*	*0.700*	*0.679*	*0.641*	*0.816*	*0.317*
Ours	**N**	**0.701**	**0.686**	**0.674**	**0.844**	**0.365**
SR	N	0.744	0.722	0.683	1.019	0.343
AIM	N	0.727	0.718	0.664	0.885	0.356
SIM	N	0.754	0.744	0.707	0.951	0.361
SUN	N	0.674	0.653	0.613	0.656	0.285
HFT	N	0.820	0.792	0.659	1.548	0.522
ICL	N	0.792	0.737	0.652	1.245	0.532
PFT	N	0.742	0.717	0.684	1.001	0.339
CASD	N	0.780	0.764	0.688	1.237	0.364
PQFT	N	0.650	0.524	0.517	0.482	0.263
QDCT	N	0.769	0.748	0.691	1.174	0.354
RARE	N	0.806	0.774	0.693	1.514	0.402
Achanta	N	0.551	0.541	0.539	0.249	0.305
Simpsal	N	0.769	0.754	0.648	1.121	0.355
Spratling	N	0.508	0.503	0.509	0.015	0.242
SIMgrouping	N	0.769	0.760	0.710	1.090	0.326
SeoMilanfar	N	0.769	0.744	0.695	1.185	0.382
ML_Net	*Y*	*0.823*	*0.803*	*0.750*	*1.824*	*0.536*
DeepGazeII	*Y*	*0.846*	*0.827*	*0.756*	*2.199*	*0.620*

The black-bold scores demonstrate the performance of our model's saliency prediction. The results of ML_Net and DeepGazeII for the TORONTO dataset are shown in pink, as this dataset was used to train ML_Net and DeepGazeII and, so, their results could not be compared with those of the other models. Table 5 includes an appendix with model details

proposed biological/Fourier/spectral-inspired saliency prediction model achieved superior and stable performance on natural images, psychophysical synthetic images, and dynamic scenes, compared with other existing models.

4.3 Ablation study

Here, we will explore the efficacy of the main components in this ablation study, and we will only measure the model performance with AUC_Judd metric on the MIT1003 dataset. The color appearance model, wavelet transform, wavelet energy map, and contrast sensitivity functions used for image feature extraction are firstly evaluated for their efficiency. Our method's performance has been stable and consistent. However, without using any of the modules, for example, remove color opponent channels, wavelet transform, wavelet energy map, and contrast sensitivity functions, performance degradation occurs (see Tab. 4).

5 Discussion

Saliency modeling has become a well-known area of study in computer vision and neuroscience. As a result, academics have attempted to employ different architectures that have

not significantly increased model performance. Instead, we should look into how humans perceive scenes; what draws their attention is crucial. In this study, we addressed numerous methods for going beyond the capabilities of such models.

First, it is vital to comprehend the cognitive attention mechanism. Visual attention is a selective cognitive process that helps us deal with this issue successfully by focusing on key information while disregarding unnecessary information. Spatial attention is important in discrimination and appearance tasks in saliency prediction studies. Investigating the brain underpinnings of visual attention can help us design better and more accurate saliency prediction models. On the other hand, better saliency prediction models can assist us in comprehending the cognitive process of visual attention in the brain. Second, we require multi-model and multi-label datasets to assess the effectiveness of saliency prediction models. The MIT Saliency benchmark datasets were used to evaluate the saliency prediction model. As previously stated, the saliency prediction model should perform better on natural scene photographs and psychophysical synthetic images (e.g., SID4VAM). This could aid in the improvement of the model architecture and our understanding of the cognitive attention process. Third, we offered extensive experimental findings demonstrating that our method consistently

Table 3 Quantitative scores of several models for the SID4VAM dataset. The baseline ITT model is shown in light blue and the proposed model is shown in green

Models	DNN	AUC_Judd	AUC_Borji	sAUC	CC	NSS	KL	SIM	IG
GT	B	0.943	0.882	0.860	1.000	4.204	0.000	1.000	2.802
Baseline-CG	B	0.703	0.697	0.525	0.281	1.577	0.722	0.372	-0.189
ITT	B	0.645	0.592	0.591	0.165	0.801	1.778	0.315	-0.104
Ours	**N**	**0.667**	**0.647**	**0.646**	**0.378**	**1.019**	**1.769**	**0.459**	**-1.508**
AIM	N	0.572	0.566	0.560	0.118	0.515	14.96	0.216	-18.742
Achanta	N	0.505	0.513	0.513	0.060	0.334	8.431	0.117	-9.634
BMS	N	0.811	0.517	0.515	0.129	0.702	19.161	0.624	-24.258
CASD	N	0.737	0.673	0.663	0.429	2.124	2.659	0.397	-1.263
DCTS	N	0.732	0.725	0.715	0.440	2.191	1.442	0.377	0.363
HFT	N	0.780	0.753	0.697	0.561	2.364	1.424	0.458	0.406
ICL	N	0.742	0.716	0.630	0.344	1.189	1.974	0.391	-0.443
PFT	N	0.704	0.692	0.688	0.407	2.074	2.556	0.361	-1.230
PQFT	N	0.581	0.518	0.517	0.112	0.611	12.262	0.175	-15.074
QDCT	N	0.717	0.705	0.699	0.419	2.139	1.850	0.368	-0.213
SIM	N	0.648	0.638	0.628	0.181	0.744	1.799	0.340	-0.152
SR	N	0.747	0.695	0.689	0.423	2.087	1.508	0.414	0.354
SUN	N	0.545	0.536	0.535	0.086	0.403	17.387	0.155	-22.268
Simpsal	N	0.753	0.740	0.661	0.370	1.380	1.461	0.393	0.323
Spratling	N	0.535	0.527	0.534	0.030	0.176	17.801	0.147	-22.890
SIMgrouping	N	0.649	0.638	0.638	0.191	0.866	1.925	0.350	-0.338
SeoMilanfar	N	0.703	0.692	0.698	0.246	1.181	2.696	0.337	-1.400
ML_Net	Y	0.600	0.593	0.501	0.118	0.341	7.529	0.370	-2.153
DeepGazeII	Y	0.645	0.581	0.475	0.103	0.555	2.450	0.274	-1.128

The black-bold scores demonstrate the performance of our model's saliency prediction. The Fourier/spectral-inspired models had the best prediction scores, compared to the other start-of-the-art non-neural network (and even deep neural network) models on the SID4VAM dataset. The results of ML_Net and DeepGazeII on the SID4VAM dataset are shown in pink, as this dataset was used to train them, and their results could not be compared with those of the other models. Table 5 includes an appendix with model details
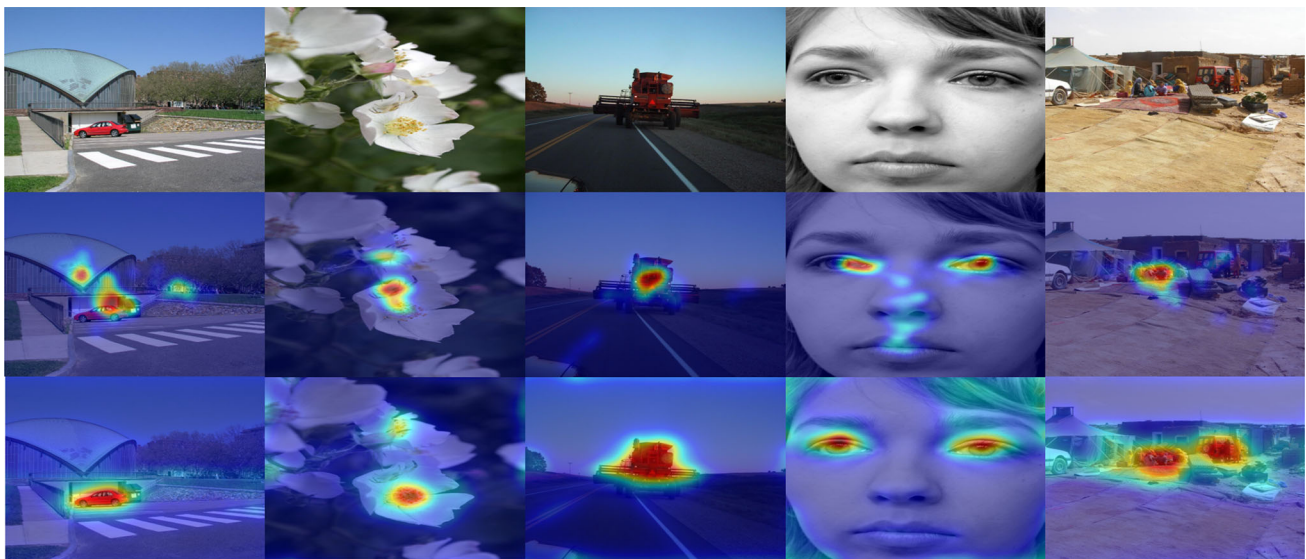

Fig. 7 Performance evaluation on the MIT1003 dataset. The first row shows color images, the second row shows ground-truth saliency maps, and the last row shows the proposed model's predicted saliency maps, respectively
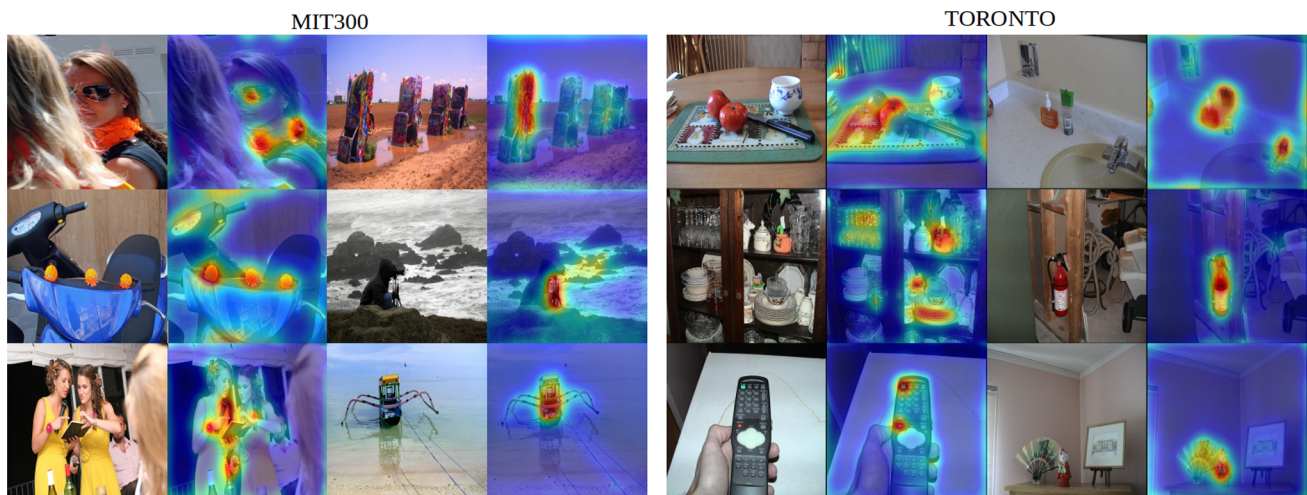
Fig. 8 Left: Performance evaluation on the MIT300 dataset. The first and third columns are color images. The second and fourth columns are the proposed model's predicted saliency maps. **Right: Performance** **evaluation on the TORONTO dataset.** The first and third columns are color images. The second and fourth columns are the proposed model's predicted saliency maps

achieved steady and better results when compared to other state-of-the-art technologies. It is worth noting that we used biologically inspired visual model estimation to determine saliency. Our proposed saliency prediction model incorporates more neuroscience notions than statistical concepts.

The proposed model incorporates opponent color channels, the wavelet transform, the wavelet energy map, and the contrast sensitivity function, but our model ignores the fact that natural images have been preprocessed multiple times, from retinal response to LGN response, for better saliency prediction accuracy and a close approximation of the human visual system. In addition, other wavelet transform families [22](e.g., Symlets, Morlet, Mexican Hat, Meyer, and Steerable pyramid etc.) may be worth investigating in the future. Saliency prediction in dynamic videos may benefit from converting the spatio-chromatic contrast sensitivity functions to a spatio-temporal one, for instance, white-black, red-green, and yellow-blue channel. Most importantly, we must understand how to properly optimize model parameters. This is one of the most crucial things we can do to improve a model's ability to perform multiple tasks. Furthermore, the suggested model contains a small number of parameters that were essentially set and fixed for all tests. One of the most important things that contributed to the suggested method's efficiency was the use of wavelet energy maps and opponent CSFs as features. Furthermore, our extensive experimental results showed that the proposed saliency prediction measure generated from a local image energy estimator is far more successful and straightforward to implement than existing methods. Even though our method was built solely on biologically inspired computational principles, the resulting model structure showed significant agreement with the fixation behavior of the human visual system.

The contribution of this study is limited, and there are some limitations to this study. First, the proposed model is inspired by the human low-level vision system, and each component of the model is already used separately in various saliency models. The proposed model architecture is based primarily on the simulated multi-channel coding principle and integrates and separates image features at different components; additionally, in the final stage of the model, we applied spatio-chromatic CSFs to each channel, which is the first time CSFs have been applied to each channel rather than an achromatic channel. Moreover, as we stated before, the traditional saliency prediction model includes both non-neural network and deep neural network-based saliency prediction models; all of these models only focus on natural image saliency prediction and completely ignore psychophysical synthetic images. As demonstrated in the results, deep neural networks outperform traditional (non-neural network) saliency prediction models in natural image saliency prediction, but they have some limitations in psychophysical synthetic images, where, surprisingly, traditional (non-neural network) saliency prediction models outperform deep neural networks. Furthermore, deep neural networks became more unstable in the presence of distracting noise in natural images, while non-neural network saliency prediction remained more reliable [58]. Bowers et al. [59] showed more problems with using deep neural networks to model the human vision system, and all of these suggested that deep neural networks have to be tested with more broadly defined tasks to improve their generality.

The extension study can be examined in the future following this study. First, the noise or other degradation processing is applied to natural images or psychophysical synthetic images to completely check how they affect the saliency pre-
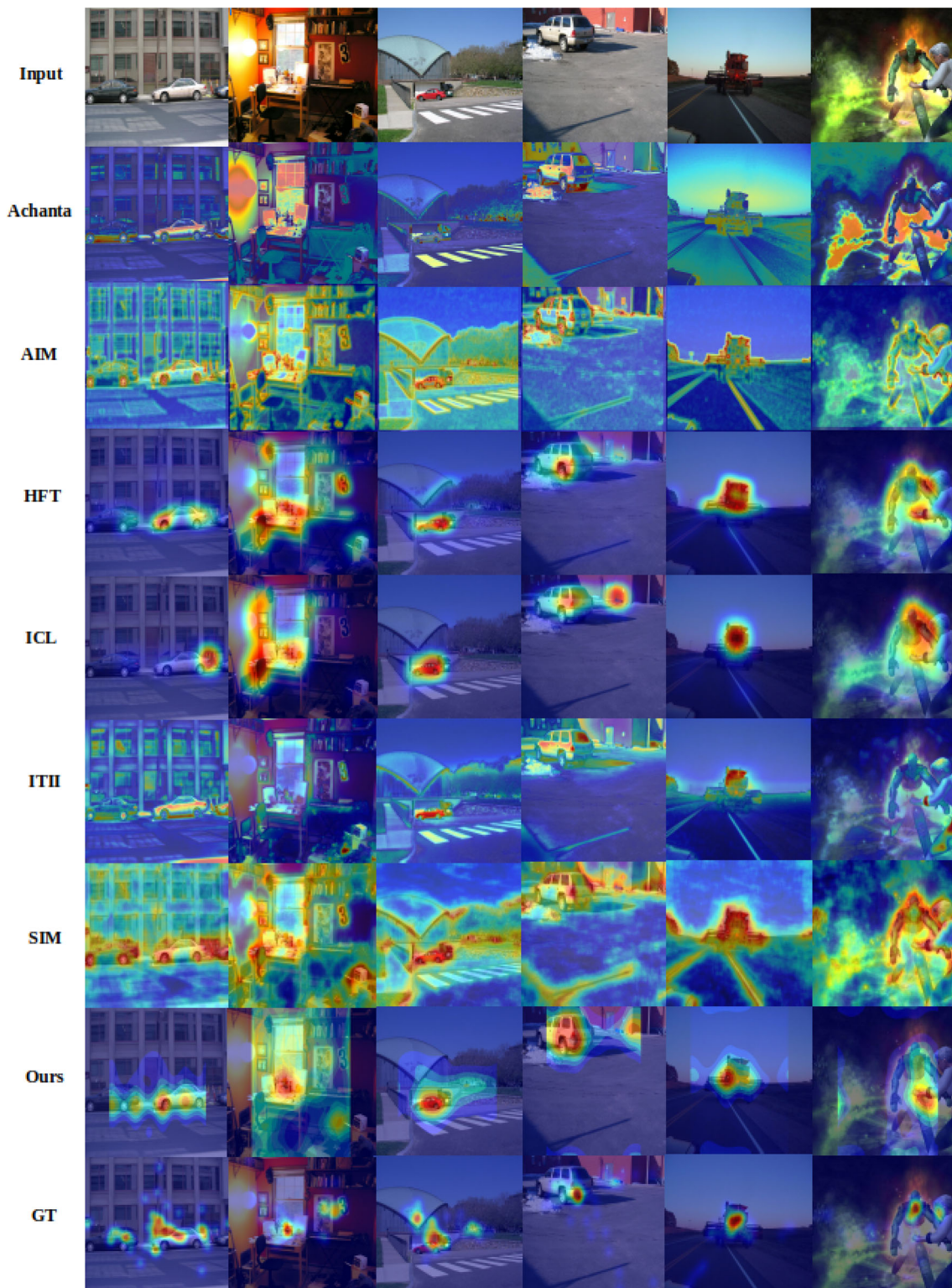
Fig. 9 Qualitative saliency prediction results on the MIT1003 dataset with different models. The first row shows six stimuli images selected from the MIT1003 dataset. The rows below show the predicted saliency maps obtained with Achanta, AIM, HFT, ICL, ITII, SIM, and the proposed model, as well as the ground-truth (GT) saliency, with artificial color for better visualization
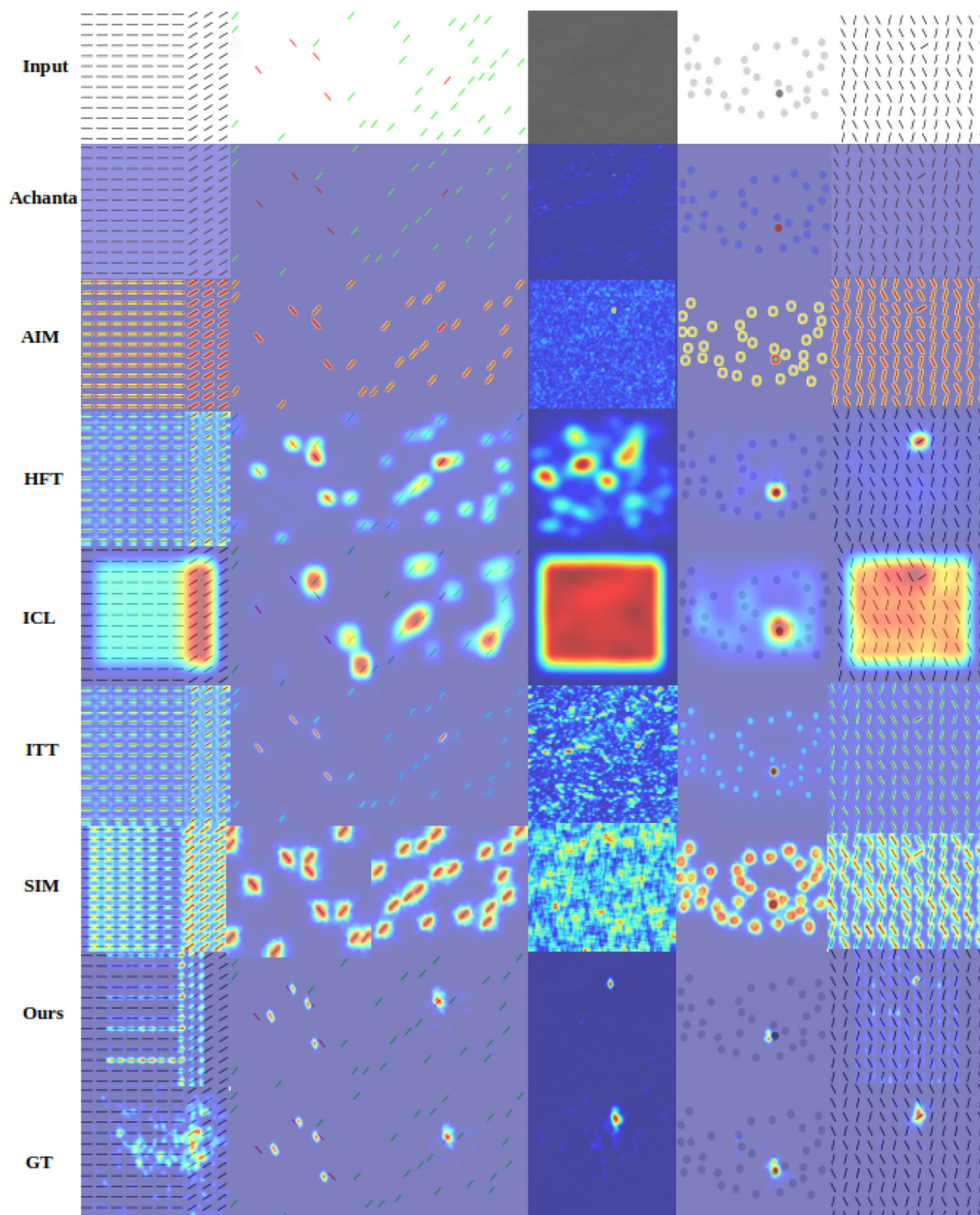
Fig. 10 Qualitative saliency prediction results on the SID4VAM dataset with different models. The first row shows six stimuli images selected from the SID4VAM dataset. The rows beneath show the salience prediction results obtained with Achanta, AIM, HFT, ICL, ITII, SIM, and the proposed model, as well as the ground truth (GT) salience, with artificial color for better visualization. The proposed model can be successfully applied to explain the "pop-out" effects in the visual search

diction results, and it is very important for us to improve the performance of saliency prediction models in real complex environments, for example, in a fog or rain situation. Second, the saliency performance of both natural images and psychophysical synthetic images can be checked with vision transforms to see whether they have some similar behavior compared to deep nets. Third, concerning training psychophysical synthetic images with deep neural nets, we need large psychophysical synthetic images, which will be more fair for deep nets and can better help us check their
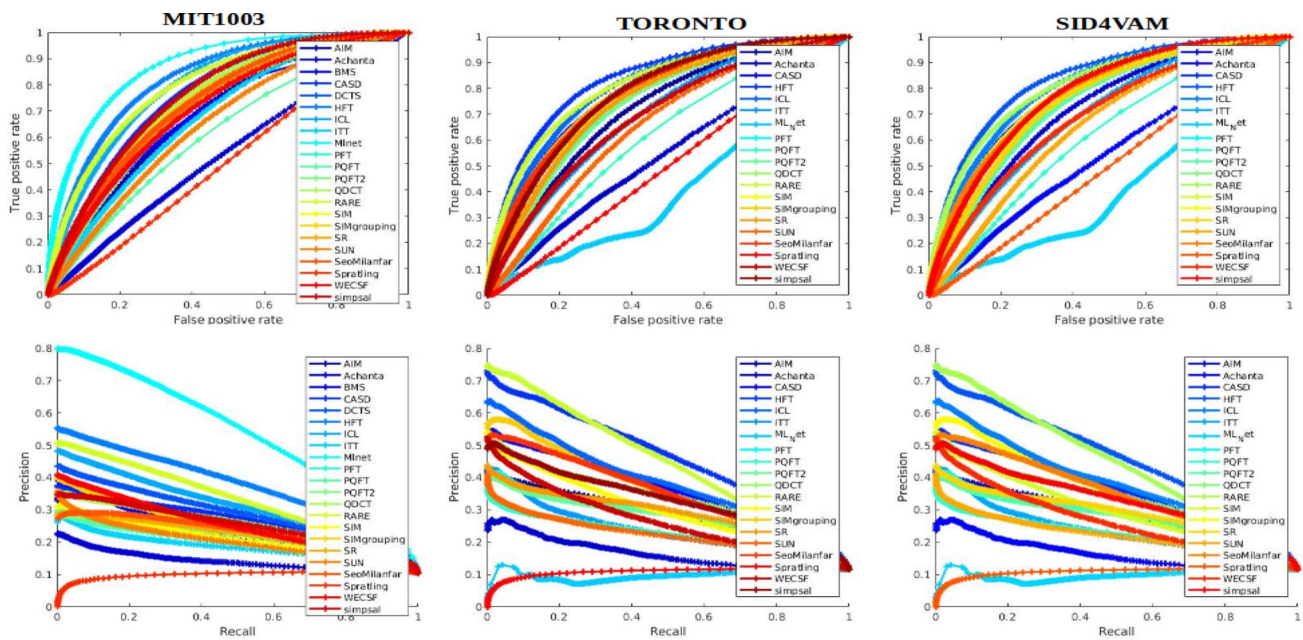
Fig. 11 ROC curve (AUC) and PR curves. Comparison of the area under the ROC curve (AUC) and PR curves, with different thresholds between our method and other state-of-the-art methods on three benchmark datasets

Fig. 12 Dynamic saliency prediction. For these sample frames from the UCF Sports Action dataset, the model clearly produced better results and perfectly captured the text information on the bottom-left
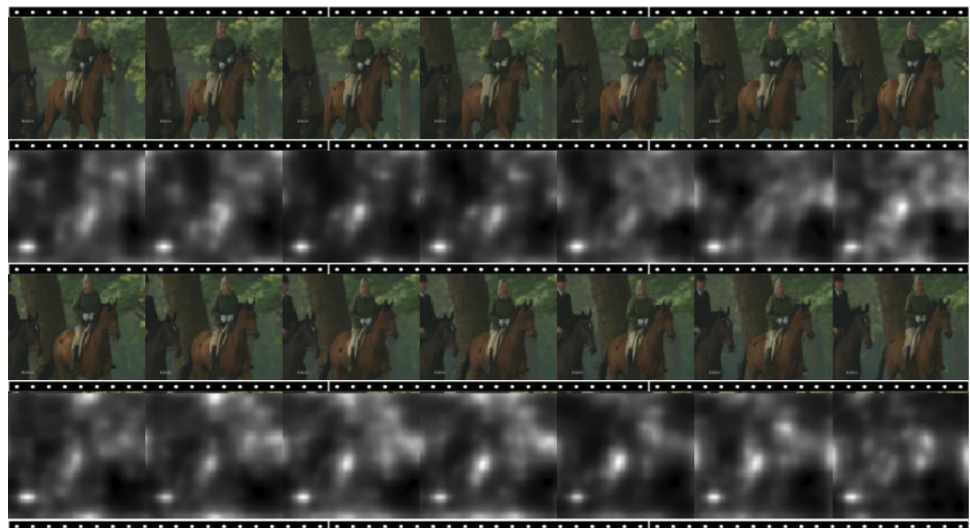

Table 4 Ablation study of the proposed model. The module's role in the saliency prediction model on the MIT1003 dataset

Modules	GT	$O_1 O_2 O_3$	WT	WE	CSFs
AUC_Judd	0.705	0.603	0.502	0.563	0.456

performance in psychophysical synthetic images and give us more evidence of the similarities and differences between deep nets and human vision.

6 Conclusions

In this study, a computational psychophysical visual saliency prediction model inspired by a low-level human visual path-way was proposed. The model includes color opponent channels, wavelet transform, wavelet energy map, and contrast sensitivity functions in order to predict the saliency map. The model was evaluated by classical benchmark datasets and achieved strongly stable and better performance in terms of visual saliency prediction, compared with the baseline model. Furthermore, we found that models based on deep neural networks outperformed ours in terms of natural image salience prediction but underperformed for psychophysical synthetic images. In contrast, Fourier/spectral-inspired models had the opposite effect, as Fourier/spectral-inspired models simulate optical neural processing from the retina to the V1. However, deep neural networks take statistics

into account more than low-level vision system functioning, and we argue that deep neural networks cannot reliably predict salient performance on psychophysical synthetic images without using specialized architectures and goals. Lastly, we added spatial-temporal saliency prediction to our model, and it was able to pick out the most important thing in the videos.

Data and Code availability The code performs main part of experiments described in this article are available at project page: https://sinodanishspain.github.io/HVS_SaliencyModel/.

Declarations

Conflict of interest The author declares no conflicts of interest.

Appendix

See Fig. 5.

Table 5 Saliency prediction models. The model function categories that inspired the developed corresponding models are shown on the right side of the table. Most saliency prediction models were inspired by biological/cognitive, Fourier/spectral, information-theoretic, and probabilistic principles

Models	Year	Inspiration
ITII [6]	1998	Biological
Achanta [7]	2009	Fourier/Spectral
AIM [8]	2005	Biological/Information-Theoretic
HFT [9]	2013	Fourier/Spectral
ICL [10]	2008	Information-Theoretic
SIM [11]	2011	Biological
BMS [12]	2013	Probabilistic
DCTS [14]	2011	Fourier/Spectral
CASD [15]	2010	Biological/Probabilistic
PFT [16]	2007	Fourier/Spectral
PQFT [17]	2008	Fourier/Spectral
QDCT [18]	2012	Fourier/Spectral
SIMgrouping [19]	2013	Biological/Cognitive
RARE [60]	2012	Information-Theoretic
SUN [61]	2008	Probabilistic
SeoMilanfar [21]	2009	Biological/Cognitive
Spratling [24]	2011	Biological/Cognitive
Simpsal [62]	2012	Biological/Cognitive
ML_Net [32]	2016	Deep Neural Network
DeepGazeII [33]	2016	Deep Neural Network

References

1. Treisman, A., Gelade, G.: A feature-integration theory of attention. Cogn. Psychol. **12**, 97–136 (1980)
2. Barlow, H.: Sensory mechanisms, the reduction of redundancy, and intelligence. Mech. Thought Proc. **10**, 535–539 (1959)
3. Wang, D., Kristjansson, A., Nakayama, K.: Efficient visual search without top-down or bottom-up guidance. Percept. Psychophys. **67**, 239–53 (2005)
4. Itti, L.: Models of bottom-up and top-down visual attention, Ph.D. dissertation, Pasadena, California, Jan (2000)
5. Sun, Y., Fisher, R.: Object-based visual attention for computer vision. Artif. Intell. **146**, 77–123 (2003)
6. Itti, L., Koch, C., Niebur, E.: A model of saliency-based visual attention for rapid scene analysis. IEEE Trans. Pattern Anal. Mach. Intell. **20**, 1254–1259 (1998)
7. Achanta, R., Hemami, S., Estrada, F., Süsstrunk, S.: Frequency-tuned salient region detection. In: Proceedings of the IEEE Conference on Computer Vision Pattern Recognition (CVPR), Jun. (2009)
8. Bruce, N.D.B., Tsotsos, J.K.: Saliency based on information maximization, In: Proceedings of the 18th International Conference on Neural Information Processing Systems, ser. NIPS'05, Vancouver, British Columbia, Canada: MIT Press, pp. 155–162 (2005)
9. Li, J., Levine, M., An, X., Xu, X., He, H.: Visual saliency based on scale-space analysis in the frequency domain. IEEE Trans. Pattern Anal. Mach. Intell. **35**, 996–1010 (2012)
10. Hou, X., Zhang, L.: Dynamic visual attention: searching for coding length increments. Adv. Neural Inf. Process. Syst **21**, 681–688 (2008)
11. Murray, N., Vanrell, M., Otazu, X., Párraga, C.A.: Saliency estimation using a non-parametric low-level vision model, In: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp. 433–440 (2011)
12. Zhang, J., Sclaroff, S.: Saliency detection: A boolean map approach, In: Proceedings of the IEEE International Conference on Computer Vision (ICCV), Dec. pp. 153–160 (2013)
13. Pinna, B., Reeves, A., Koenderink, J., Doorn, A., Deiana, K.: A new principle of figure-ground segregation: the accentuation. Vis. Res. **143**, 9–25 (2017)
14. Hou, X., Harel, J., Koch, C.: Image signature: highlighting sparse salient regions. IEEE Trans. Pattern Anal. Mach. Intell. **34**, 194–201 (2011)
15. Goferman, S., Zelnik-Manor, L., Tal, A.: Context-aware saliency detection, In: Proceedings IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp. 2376–2383 (2010)
16. Hou, X., Zhang, L.: Saliency detection: a spectral residual approach, In: IEEE Conference in Computer Vision and Pattern Recognition, vol. 2007, (2007)
17. Guo, C., Ma, Q.: Zhang, L.: Spatio-temporal saliency detection using phase spectrum of quaternion fourier transform, In: IEEE Conference in Computer Vision and Pattern Recognition, (2008)
18. Schauerte, B., Stiefelhagen, R.: Quaternion-based spectral saliency detection for eye fixation prediction, In: European Conference on Computer Vision (ECCV), Oct. pp. 116–129, ISBN: 978-3-642-33708-6 (2012)
19. Murray, N., Vanrell, M., Otazu, X., Párraga, C.A.: Low-level spatiochromatic grouping for saliency estimation. IEEE Trans. Pattern Anal. Mach. Intell. **35**, 2810–6 (2013)
20. Otazu, X., Pàrraga, C.A., Vanrell, M.: Toward a unified chromatic induction model. J. Vis. **10**, 5 (2010)
21. Seo, H., Milanfar, P.: Static and space-time visual saliency detection by self-resemblance. J. Vis. **9**, 15–27 (2009)
22. Louis, A., Maass, P., Rieder, A.: Wavelets: theory and Applications. Jan. ISBN: 978-0-471-96792-7 (1997)

23. Selvaraj, A., Shebiah, N.: Object recognition using wavelet based salient points. Open Signal Process. J. **2**, 14–20 (2009)

24. Spratling, M.: Predictive coding as a model of the v1 saliency map hypothesis. Neural Netwo. Offi. J. Int. Neural Netw. Soc. **26**, 7–28 (2011)

25. Rao, R., Ballard, D.: Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects. Nat. Neurosci. **2**, 79–87 (1999)

26. Borji, A.: Saliency prediction in the deep learning era: successes and limitations. IEEE Trans. Pattern Anal. Mach. Intell. **43**(2), 679–700 (2019)

27. Kruthiventi, S., Gudisa, V., Dholakiya, J., Babu, R.: Saliency unified: A deep architecture for simultaneous eye fixation prediction and salient object segmentation, Jun. pp. 5781–5790 (2016)

28. Cong, R., Lei, J., Fu, H., Cheng, M.-M., Lin, W., Huang, Q.: Review of visual saliency detection with comprehensive information. IEEE Trans. Circuits Syst. Video Technol. **29**(10), 2941–2959 (2019)

29. Cong, R., Lei, J., Fu, H., Porikli, F., Huang, Q., Hou, C.: Video saliency detection via sparsity-based reconstruction and propagation. IEEE Trans. Image Process. **28**(10), 4819–4831 (2019)

30. Fang, C., Tian, H., Zhang, D., Zhang, Q., Han, J., Han, J.: Densely nested top-down flows for salient object detection. Sci. China Inf. Sci. **65**(8), 1–14 (2022)

31. Zhang, D., Han, J., Zhang, Y., Xu, D.: Synthesizing supervision for learning deep saliency network without human annotation. IEEE Trans. Pattern Anal. Mach. Intell. **42**(7), 1755–1769 (2020)

32. Cornia, M., Baraldi, L., Serra, G., Cucchiara, R.: A deep multi-level network for saliency prediction. In: 23rd International Conference on Pattern Recognition (ICPR), pp. 3488–3493 (2016)

33. Kümmerer, M., Wallis, T. S. A., Bethge, M.: Deepgaze II: reading fixations from deep features trained on object recognition, CoRR, vol. abs/1610.01563, arXiv: 1610.01563. [Online]. Available: (2016)

34. Butz, M.: Toward a cognitive sequence learner: hierarchy, self-organization, and top-down bottom-up interaction, Apr. [Online]. Available: https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.177.6739&rep=rep1&type=pdf (2004)

35. Finlayson, G., Drew, M., Funt, B.: Color constancy: Enhancing von Kries adaptation via sensor transformations, Proc SPIE, vol. 1913, Sep (1993)

36. Finlayson, G.D., Alsam, A., Hordley, S.D.: Local linear models for improved von kries adaptation, In: The Tenth Color Imaging Conference: Color Science and Engineering Systems, Technologies, Applications, CIC,: Scottsdale, Arizona, USA, November 12–15, 2002, IS&T - The Society for Imaging. Science and Technology **2002**, pp. 139–144 (2002)

37. Krauskopf, J., Gegenfurtner, K.: Color discrimination and adaptation. Vis. Res. **32**, 2165–75 (1992)

38. Brill, M.: Trichromatic theory, pp. 827–829, ISBN: 978-0-387-30771-8 (2014)

39. Shapley, R., Hawken, M.: Color in the cortex: single- and double-opponent cells. Vis. Res. **51**, 701–17 (2011)

40. Hering, E.: Outlines of a Theory of the Light Sense. Harvard University Press, Cambridge (1920)

41. Hurvich, L., Jameson, D.: An opponent-process theory of color vison. Psychol. Rev. **64**, 384–404 (1957)

42. Zhaoping, L.: A neural model of contour integration in the primary visual cortex. Neural Comput. **10**, 903–40 (1998)

43. Haar, A.: Zur theorie der orthogonalen funktionensysteme. (zweite mitteilung)., Mathematischen Annalen, vol. 71, pp. 38–53, [Online]. Available: http://eudml.org/doc/158516 (1912)

44. Imamoğlu, N., Lin, W., Fang, Y.: A saliency detection model using low-level features based on wavelet transform. IEEE Trans. Multimedia **15**, 96–105 (2013)

45. Mullen, K.: The contrast sensitivity of human color vision to red-green and blue-yellow chromatic gratings. J. Physiol. **359**, 381–400 (1985)

46. Mannos, J., Sakrison, D.: The effects of a visual fidelity criterion of the encoding of images. IEEE Trans. Inf. Theory **20**(4), 525–536 (1974)

47. Watson, A., Malo, J.: Video quality measures based on the standard spatial observer, In: Proceedings International Conference on Image Processing (ICIP), vol. 3, Feb. pp. III–41, ISBN: 0-7803-7622-6 (2002)

48. Watson, A., Ahumada, A.: The spatial standard observer. J. Vis. **4**, 51–51 (2010)

49. Judd, T., Durand, F., Torralba, A.: A benchmark of computational models of saliency to predict human fixations, In: MIT Technical Report (2012)

50. Berga, D., Fdez-Vidal, X. R., Otazu, X., Pardo, X. M.: Sid4vam: A benchmark dataset with synthetic images for visual attention modeling, 2019 IEEE/CVF International Conference on Computer Vision (ICCV), pp. 8788– 8797 (2019)

51. Bylinskii, Z., Judd, T., Oliva, A., Torralba, A., Durand, F.: What do different evaluation metrics tell us about saliency models? arXiv preprint arXiv:1604.03605 (2016)

52. Borji, A., Rezazadegan Tavakoli, H., Sihite, D., Itti, L.: Analysis of scores, datasets, and models in visual saliency prediction, In: Proceedings of the IEEE International Conference on Computer Vision (ICCV), Dec, pp. 921–928 (2013)

53. Emami, M., Hoberock, L.: Selection of a best metric and evaluation of bottom-up visual saliency models. Image Vis. Comput. **31**, 796–808 (2013)

54. Riche, N., Duvinage, M., Mancas, M., Gosselin, B., Dutoit, T.: Saliency and human fixations: State-of-the-art and study of comparison metrics, In: Proceedings of the IEEE International Conference on Computer Vision, (2013)

55. Kümmerer, M., Wallis, T., Bethge, M.: Information-theoretic model comparison unifies saliency metrics. Proc. Natl. Acad. Sci. **112**, 201 510 393 (2015)

56. Jost, T., Ouerhani, N., Wartburg, R., Müri, R., Hügli, H.: Assessing the contribution of color in visual attention. Comput. Vis. Image Underst. **100**, 107–123 (2005)

57. Feng, M.: Evaluation toolbox for salient object detection., https://github.com/ArcherFMY/sal_eval_toolbox (2018)

58. Li, Q.: Understanding saliency prediction with deep convolutional neural networks and psychophysical models (2022)

59. Bowers, J.S., et al.: Deep problems with neural network models of human vision. Behav. Brain Sci., 1–74 (2022)

60. Riche, N., Mancas, M., Gosselin, B., Dutoit, T.: Rare: A new bottom-up saliency model, In: Image Processing, 19th IEEE Conference on (IEEE), (2012)

61. Zhang, L., Tong, M., Marks, T., Shan, H., Cottrell, G.: Sun: a Bayesian framework for saliency using nature statistics. J. Vis. **8**, 32 (2008)

62. Harel, J.: A saliency implementation in matlab, http://www.klab.caltech.edu/~harel/share/gbvs.php (2012)

Qiang Li I am a researcher working on computational neuroscience, computational neuroimaging, deep learning, information theory, and image and video processing. I got my PhD from the University of Valencia and am now a postdoc in the TReNDs.