Vision Research 212 (2023) 108304

Contents lists available at ScienceDirect

Vision Research

journal homepage: www.elsevier.com/locate/visres

The influence of stereopsis on visual saliency in a proto-object based model of selective attention

Takeshi Uejima [a,*], Elena Mancinelli [a], Ernst Niebur [b], Ralph Etienne-Cummings [a]

[a] The Department of Electrical and Computer Engineering, The Johns Hopkins University, Baltimore, MD, USA
[b] The Solomon Snyder Department of Neuroscience and the Zanvyl Krieger Mind/Brain Institute, The Johns Hopkins University, Baltimore, MD, USA

ARTICLE INFO

Keywords:
Saliency
Visual attention
Stereopsis
Proto-object
Disparity
Eye fixation

ABSTRACT

Some animals including humans use stereoscopic vision which reconstructs spatial information about the environment from the disparity between images captured by eyes in two separate adjacent locations. Like other sensory information, such stereoscopic information is expected to influence attentional selection. We develop a biologically plausible model of binocular vision to study its effect on bottom-up visual attention, i.e., visual saliency. In our model, the scene is organized in terms of proto-objects on which attention acts, rather than on unbound sets of elementary features. We show that taking into account the stereoscopic information improves the performance of the model in the prediction of human eye movements with statistically significant differences.

1. Introduction

We are surrounded by three-dimensional space; however, each retina captures only a two-dimensional image. Each retinal image individually can contain clues for depth information such as shading, looming sizes, and occlusion, with the latter including the presence of T-junctions (Nakayama et al., 1995; von der Heydt, 2015; Welchman, 2016). In addition, binocular vision which uses triangulation by two eyeballs, provides a reliable cue for depth and results in vivid three-dimensional (3D) perception. Not only is this capability used for range finding, some animals with front-facing eyes, including humans, also exploit binocular stereopsis for camouflage breaking (Nityananda & Read, 2017); in the words of Bela Julesz, "with stereoscopic vision there is no camouflage" (Julesz, 1989). This is likely due to the organization of visual scenes into objects: camouflage exploits erroneous assignment of perceptual edges which is made more difficult by the existence of explicit depth differences at object borders, and their absence within object borders (Adams et al., 2019; Poggio & Poggio, 1984).

Binocular vision has been studied both in neuroscience and in engineering, the former primarily focusing on revealing how nervous systems achieve stereovision and the latter on finding efficient and precise algorithms. In both fields, a major difficulty is the stereo correspondence problem: to find out which features in two retinal images originate from the same point in 3D space. The problem may seem trivial because we effortlessly and quickly solve it in daily life. Nonetheless, it is not simple, and the brain devotes multiple cortical areas to solve it (Cumming & DeAngelis, 2001; Kumano et al., 2008; Tanabe et al., 2004). The mechanisms employed in the primary visual cortex have been extensively studied, and the "disparity energy model" is widely accepted as it agrees well with data from neurophysiological experiments (Ohzawa et al., 1990, 1997).

Binocular information, together with other visual features, not only underlies functions like object recognition but presumably also provides inputs for the determination of which parts of the visual scene are the most relevant, i.e. which require detailed processing. Identifying these regions is the task solved by visual selective attention. In general, this is a highly complex function which involves perceptual and cognitive processes at many levels. An important part of this function is data-driven, or bottom-up attention, that finds the most relevant image regions based on low-level visual features and their combinations. These regions are usually called the most "salient" areas of the scene. The seminal work by Koch and Ullman (Koch & Ullman, 1985) established a systematic way to find these regions in the form of a saliency map which ranks the level of saliency at different locations in the visual scene. Predictions of this theory need to be compared with behavioral observations. Two considerations are of relevance here. First, the saliency map was originally proposed as a mechanism for covert attention which is correlated with, but not identical to, overt attention, i.e. eye movements. There are methods to measure behavioral consequences of covert visual attention, e.g. (Posner, 1980) but in practice, it is much easier to

* Corresponding author.
E-mail addresses: tuejima1@alumni.jh.edu, uejima.takeshi@gmail.com (T. Uejima).

https://doi.org/10.1016/j.visres.2023.108304
Received 9 March 2022; Received in revised form 18 July 2023; Accepted 18 July 2023
Available online 3 August 2023

measure the state of overt attention(Parkhurst et al., 2002) which, as mentioned, is known to correlate with covert attention (Deubel & Schneider, 1996; Hoffman & Subramaniam, 1995; Moore & Fallah, 2001; van der Stigchel & Theeuwes, 2007). For this reason, (Parkhurst et al., 2002) proposed to use overt attention as an approximation to covert attention for the purposes of testing models of the latter. Second, the saliency map only takes account of bottom-up information. To minimize (though not eliminate) the effects of top-down attention, which is goal-directed and depends on the internal state of the observer in addition to the visual input (Parkhurst et al., 2002) and later studies use a free-viewing paradigm.

Many models of visual saliency rely on local contrast in low-level features such as intensity, color, and orientation. However, a plethora of studies in psychology and neurophysiology have shown that visual attention is also influenced by the rapid perceptual organization of the visual scene into tentative objects rather than the basic features themselves (Egly et al., 1994; Einhäuser et al., 2008; Nuthmann & Henderson, 2010; Qiu et al., 2007; Stoll et al., 2015; Zhou et al., 2000). By "tentative objects" we mean that even relatively low-level visual processes can capture the structure (e.g., foreground and background) of the input scene. These tentative objects, or areas that possess "objectness," are called proto-objects (Rensink, 2000). A proto-object based saliency model was shown to predict eye fixations with good accuracy (Russell et al., 2014). In that model, the combination of edge detection, center-surround mechanism, and grouping processes extracts the tentative objects based on closure and proximity. While originally this model used information from maps of intensity, color, and orientation, it was later extended to additionally utilize motion, depth, and texture features (Hu et al., 2016; Mancinelli et al., 2018; Molin et al., 2015; Uejima et al., 2020), and it was also implemented in biofidelic neuromorphic hardware (Ghosh et al., 2022; Iacono et al., 2019; Molin et al., 2021; Ramenahalli et al., 2013).

In this paper, we propose a model of biological stereopsis and incorporate it into that proto-object based saliency model. While depth and disparity features have been integrated into the model previously (Hu et al., 2016; Mancinelli et al., 2018), we discuss below why our approach employs a different mechanism to exploit binocular disparity that is biologically plausible. As in previous work, the output of our model is a saliency map, and we compare it with published human fixation data obtained while participants freely viewed stereogram images. As we show below, our model shows better predictive performance than the original two-dimensional (2D) proto-object based saliency model.

The main novel contributions of our study are: 1) building a biofidelic 3D visual saliency model that includes disparity-tuned neurons and border-ownership coding neurons in areas V1, V2, and V4, achieving proto-object based perception; and 2) applying the model to natural 3D scenes and evaluating 3D effects on saliency with fixation data collected from humans viewing natural scenes.

2. Related studies

2.1. Stereopsis and eye fixations

Many studies have sought an understanding of how the brain achieves stereoscopic vision. Since Julesz introduced random dot stereograms (Julesz, 1971), which do not include any 2D depth cues and provide only disparity information, neuroscientists have used these stimuli to study brain activity corresponding to depth perception solely generated by binocular disparity (Poggio et al., 1985). These experiments have also shown that no prior knowledge about objects is needed for stereo correspondence because an observer can perceive nontrivial image contents only after fusing the stereogram.

One of the major difficulties of stereoscopic vision is the correspondence problem, i.e. to find corresponding features in the two 2D images. A cooperative process is an early model to solve this problem, by finding matching points using iterative computations which minimize an error-measure(Marr, 1982; Marr & Poggio, 1979). While this could potentially be realized in the nervous system, the biological system seems to process stereovision more rapidly than is expected from an iterative process, at least in its early stages. Neurophysiological studies have revealed that stereoscopic vision can be explained by the so-called disparity energy model, in which binocular simple cells sum the activity of monocular simple cells linearly. Subsequently, binocular complex cells sum the squared responses of quadrature pairs of the simple cells (Ohzawa, 1998; Ohzawa et al., 1990, 1997). Marr and Poggio pointed out that combining multi-spatial-frequency filters (i.e., neuronal receptive fields) aids to prevent false matches between parts of one image to non-corresponding portions of the other image (Marr & Poggio, 1979). Their original idea employs a sequential coarse-to-fine structure that first computes coarse (low spatial frequencies) disparities and then proceeds to finer scales. Later, pooling multi-spatial-frequency features to find correspondence points based on information from all scales at once (rather than sequentially at different scales) was proposed (Fleet et al., 1996). While it is not clear whether the biological system employs a sequential mechanism or a simultaneous pooling algorithm, integration over multiple spatial-frequencies has been observed in primate area V4 (Kumano et al., 2008) and in primary visual cortex of cat (Baba et al., 2015), and we adopt it in our model.

Visual saliency has been widely studied in 2D (Bruce & Tsotsos, 2005; Hou et al., 2012; Itti et al., 1998; Itti & Koch, 2000; Judd et al., 2009; Koch & Ullman, 1985; Li, 2002; Niebur & Koch, 1996), see (Borji et al., 2013) for a comparative study. While these models approach the problem from a mechanistic point of view, deep learning based models have also been used, and showed remarkably high performance in predicting human fixations (Cornia et al., 2016; Huang et al., 2015; Kruthiventi et al., 2015; Kümmerer et al., 2014, 2016; Vig et al., 2014). Recently, some studies attempted to incorporate psychological concepts, based on Gestalt principles, into saliency models (Russell et al., 2014; Zhang & Sclaroff, 2013, 2016). Gestalt psychology argues that the whole of an object is more important than individual features for perception. This assertion has been supported by neurophysiological studies of figure-ground organization coding in visual cortex (Qiu et al., 2007; Qiu & von der Heydt, 2005; von der Heydt, 2015; Williford & von der Heydt, 2016; Zhou et al., 2000) which link perception and neural responses.

Saliency regarding stereoscopic images has been also studied, although not to the extent of 2D saliency. Reports investigating how depth information affects human eye movements (Gautier & Le Meur, 2012; Huynh-Thu & Schiatti, 2011; Jansen et al., 2009; Khaustova et al., 2013; Lang et al., 2012) showed that, overall, humans tend to fixate similar locations in situations with or without binocular information. More specifically, the fixation locations are almost the same for 3D and 2D images in long observation windows (20 s) but different in short observation windows (about four or five seconds) (Gautier & Le Meur, 2012; Jansen et al., 2009; Khaustova et al., 2013). Notably, researchers reported a tendency for humans to look at closer points soon after they look at an image (Gautier & Le Meur, 2012; Jansen et al., 2009; Lang et al., 2012). The effect of 3D cues for visual perception was also observed on shorter time scales in a texture segmentation task (Zhaoping et al., 2009). The study showed that the 3D process shortens the reaction time to segment two textures if and only if the task is difficult for the 2D process, which implies that V1 plays a dominant role during the initial attentional process, and that extrastriate cortex later provides additional information. More specifically, when texture segmentation is sufficiently easy, human observers typically require a reaction time of half a second to one second to report the location of the boundary between two neighboring textures. Their experiments indicate that this reaction time is not shortened by adding depth information to visual inputs unless the segmentation is so difficult that the reaction time is longer than one second. We note that these influences of binocular vision on saliency should not be confounded with the effect of ocularity, predicted by (Li, 2002) and observed experimentally by (Zhaoping, 2008, 2012, 2018);

see also the discussion below.

Based on these behavioral results, researchers have proposed visual saliency models for 3D still scenes. Lang et al. calculated "depth priors" that indicates how eye fixations of human observers differed between 2D images and the corresponding 3D scenes, and they proposed to include these priors in existing 2D saliency models (Lang et al., 2012). Wang et al. took a Bayesian approach to incorporate depth effects on fixations (Wang et al., 2013), in which the parameters for the probability distribution were tuned by observed fixation data. Ma and Hang (Ma & Hang, 2015) proposed another learning-based model that employed a similar method for the Judd et al. model (Judd et al., 2009) for 2D static images, which includes various features including a face detection mechanism. They extended the model to incorporate features from a depth map. In these data-driven approaches, the parameters are determined by human behavioral data and did not explicitly implement biological mechanisms. An alternative biologically plausible binocular segmentation model has been proposed, which employs disparity selective V2 neurons (Zhaoping, 2002). In that study, intracortical interaction generate the stereo correspondence and pre-attentive stereo segmentation on random-dot stereograms. The study focused on early visual processing involving area V1 and V2 and did not include later processing areas such as V4.

In the context of binocular vision, it is known that dichoptic features also affect saliency, which was an important prediction generated by a saliency map model implemented in area V1 (Li, 2002). In a series of studies, Zhaoping (Zhaoping, 2008, 2012, 2018) showed that these features include ocularity. Specifically, she used dichoptic viewing in which a center stimulus was presented to one eye and its surround to the other. She showed that ocular singletons, in which the center stimulus differed from the surround in some feature, e.g. orientation, elicit behavior that was consistent with strong saliency at the stimulus location. Remarkably, this was the case even though under such viewing conditions humans are typically not aware to which eye a stimulus is presented, i.e. their perception is identical to that of a stimulus in a surround field in a monocularly presentation. Although these results are highly interesting, we here do not study dichoptic vision but focus on complex natural scenes.

In this study, we implement algorithms inspired by the information processing principles employed in the primate brain. More specifically, we use the framework of a proto-object based model of perceptual organization and attentional control (Russell et al., 2014) and integrate a biologically-plausible stereovision mechanism in this model.

Previously, Hu and collaborators published a proto-object based saliency model that includes depth features (Hu et al., 2016). This model takes a depth map as input along with a 2D image, which means depth information must be calculated or measured before it is used in the model. The depth map is then treated similar to any other feature map, e.g., the intensity map. Another model, proposed by Mancinelli et al. (Mancinelli et al., 2018) takes two images from the right and left cameras rather than a depth map. This approach is more biofidelic because the visual cortex is not provided with an explicit depth map as input, but instead, with the output from two retinae. However, the images in the Mancinelli et al model need to be rectified which either requires precise knowledge of the optical geometry, which often is not available, or additional knowledge of the depth at several locations in the scene. This is the main disadvantage of the model since it leaves open where this information comes from. We therefore take a different approach for modelling visual saliency based on Gestalt principles with a stereopsis mechanism that does not suffer from these limitations, by only requiring input from two cameras.

3. 3D eye fixations datasets

As mentioned previously, a widely accepted method to evaluate the quality of saliency models is to compare how well they can predict human eye fixations. Although many datasets of human fixations for 2D scenes have been published, only few are available for 3D stimuli.

The Gaze-3D dataset is a publicly-available 3D fixation dataset (Wang et al., 2013). It consists of 18 stereoscopic images and corresponding disparity and depth maps calculated by an optical flow method. The fixation data were collected from 35 participants sitting at 93 cm distance from a 26-inch display with a resolution of 1920 × 1200 pixels for 15 s after stimulus onset. Their eye tracking data were recorded from the left eye, meaning the fixation locations correspond to the left image.

The NCTU-3D dataset consists of 475 stereoscopic scenes and corresponding depth maps (Ma & Hang, 2015). The eye-tracking data were collected from 16 subjects. The 3D images were displayed on a 23-inch monitor with 1920 × 1080 pixels resolution, placed at 78.5 cm from the observers for 4 s after stimulus onset. Fig. 1 (a) shows an example of stimuli, fixation map, and depth map from the NCTU-3D dataset. The provided fixation data are based on right-eye tracking.

4. Proto-object based saliency model

4.1. Model framework

The disparity channels to be described in Section 3.2 provide input to a variation of the proto-object based saliency model which was originally introduced by Russell et al. (Russell et al., 2014). We use an improved algorithm developed by Uejima et al. (Uejima et al., 2020) but omit the texture features introduced in that model. Since the model framework used in this paper is the same as that in those prior studies, we briefly explain it in Section 3.3.

The source code of the proposed model is available online (https://github.com/csmslab/proto-object-saliency-stereopsis).

4.2. Disparity channels

We start by modeling the retina under photopic conditions, i.e., under light conditions in which rods are saturated and cones play the main role. Retinal output is generated by three types of retinal ganglion cells: parasol, midget, and bistratified (Nassi & Callaway, 2009). Simplified, the parasol cells mainly represent intensity (luminance) while the other two represent chromatic information: the midget cells red-green colors, and the bistratified cells yellow-blue colors. We model the intensity channel, I, as:

$$I = \frac{r+g+b}{3} \tag{1}$$

where r, g, and b are the red, green, and blue components of the image (Itti et al., 1998).

The color channels are modeled as below:

$$\boldsymbol{R} = \left\lfloor \boldsymbol{r} - \frac{\boldsymbol{g}+\boldsymbol{b}}{2} \right\rfloor, \boldsymbol{G} = \left\lfloor \boldsymbol{g} - \frac{\boldsymbol{r}+\boldsymbol{b}}{2} \right\rfloor$$

$$B = \left\lfloor b - \frac{\boldsymbol{r}+\boldsymbol{g}}{2} \right\rfloor, Y = \left\lfloor \frac{\boldsymbol{r}+\boldsymbol{g}}{2} - \frac{|\boldsymbol{r}-\boldsymbol{g}|}{2} - b \right\rfloor \tag{2}$$

$$RG = \lfloor R-G \rfloor., GR = \lfloor G-R \rfloor$$

$$BY = \lfloor B-Y \rfloor., YB = \lfloor Y-B \rfloor \tag{3}$$

where $\lfloor \cdot \rfloor$ is half-wave rectification, and RG, GR, BY, and YB are color opponency channels. The color signals are only computed for pixels whose intensity value exceeds 10% of the maximum intensity of the input image since hue variations are not perceivable at very low luminance. It is still unclear what role chromatic information plays for stereopsis. It has been reported (Gregory, 1977; Jordan et al., 1990; Lu & Fender, 1972) that random-dot stereograms need luminance cues to cause depth perception although isoluminant figural stimuli can be

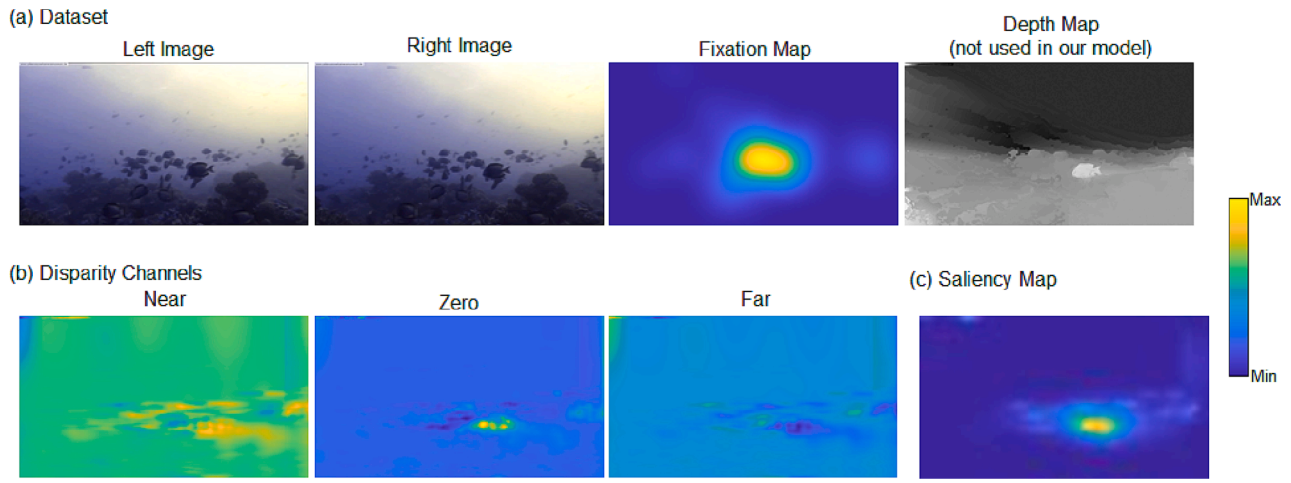

Fig. 1. Examples of a 3D eye tracking dataset and the saliency map generated by the proposed model. (a) Stereo image, fixation map and depth map from the NCTU-3D dataset. The depth map is not used in our model. (b) Disparity channels of Near, Zero, and Far calculated by our model. A fish in the lower-right is present in the Near and Zero channels. (c) Saliency map generated from the disparity channels. The color bar applies to (b), (c) and the fixation map in (a).

perceived as stereoscopic. In our models, we implement versions with and without contributions from color channels, see below.

We use the disparity energy model as a biologically-plausible method to extract depth (Ohzawa et al., 1990, 1997). The brain, as well as our model, combines multi-spatial-frequency filters to accurately detect disparity information (Baba et al., 2015; Fleet et al., 1996; Kumano et al., 2008). The calculation of disparity energy starts by computing receptive field properties of monocular simple cells. Simple cells in V1 are modeled by Gabor filters (Kulikowski et al., 1982), in our model,

$$g_{e,\theta}(x,y) = \exp\left(\frac{-x^{'2}+\gamma^2 y^{'2}}{2\sigma^2}\right)\cos(\omega x^{'}) \tag{4}$$

$$g_{o,\theta}(x,y) = \exp\left(\frac{-x^{'2}+\gamma^2 y^{'2}}{2\sigma^2}\right)\sin(\omega x^{'}) \tag{5}$$

$$x^{'} = x\cos\theta + y\sin\theta, y^{'} = -x\sin\theta + y\cos\theta \tag{6}$$

where $\theta \in \{0, \frac{\pi}{4}, \frac{\pi}{2}, \frac{3\pi}{4}\}$, and $g_{e,\theta}(x,y)$ and $g_{o,\theta}(x,y)$ are the even- and odd-Gabor filters with spatial aspect ratio γ, width $\sigma = 2.24$, and spatial frequency $\omega = 1.57$. As in many other models of early visual processing, the value of γ is chosen to be below unity, resulting in elongated filters as shown in Fig. 2(a). For instance, Russell et al. used $\gamma = 0.5$ and $\gamma = 0.8$ for edge detection and orientation channels, respectively, and we use the same values for those purposes (edge detection and orientation channels). However, for the simple cells of the disparity features we employ shortened Gabor filters with $\gamma = 5$ because such filters showed better results than elongated filters. Although most orientation-selective cells in early cortex have elongated receptive fields, which low spatial aspect ratios ($\gamma < 1$), a fraction of them have high spatial aspect ratios ($\gamma > 1$) (Xu et al., 2016). This is rare for simple cells but more common for complex cells (ibid.). An example of a shortened (non-elongated) filter is shown in Fig. 2 (b). Our modeling results predict that the subpopulation of cells with high spatial aspect ratio, which are ill-suited for functionalities like orientation filters, is preferentially involved with the computation of binocular disparity which is greatly improved by the presence of these cells.

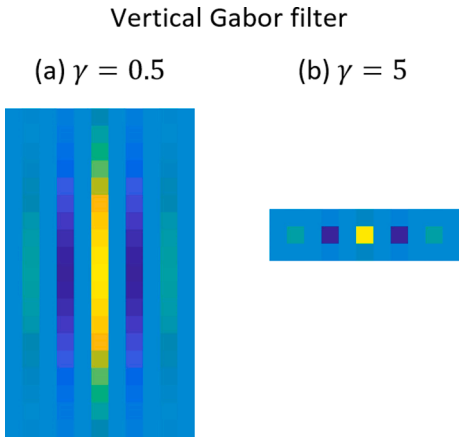

Fig. 2. Vertical Gabor filters. (a) A low spatial aspect ratio Gabor filter. (b) A high spatial aspect ratio Gabor filter.

Model receptive fields vary in size to make responses tolerant to changes in scale. For the sake of computational efficiency, we scale the input image, by full-octave steps, rather than the filters. The image of the k-th scale level is written as X^k, $\mathrm{X} \in \{I, RG, GR, BY, YB\}$. Monocular simple cells are represented as:

$$S^k_{\mathrm{X},R,e,\theta}(x,y) = \mathrm{X}^k_R(x,y) * g_{e,\theta}(x,y) \tag{7}$$

$$S^k_{\mathrm{X},R,o,\theta}(x,y) = \mathrm{X}^k_R(x,y) * g_{o,\theta}(x,y) \tag{8}$$

where $S^k_{R,e,\theta}$ is a k-th level simple cell activation function with even-symmetric Gabor filters from the right image which has a preferred angle of θ. $S^k_{R,o,\theta}$ is the same but for an odd-symmetric Gabor filter. The asterisk symbol * indicates convolution. The simple cells from the left image, $S^k_{L,e,\theta}$ and $S^k_{L,o,\theta}$, are obtained in the same way:

$$S^k_{\mathrm{X},L,e,\theta}(x,y) = \mathrm{X}^k_L(x,y) * g_{e,\theta}(x,y) \tag{9}$$

$$S^k_{\mathrm{X},L,o,\theta}(x,y) = \mathrm{X}^k_L(x,y) * g_{o,\theta}(x,y) \tag{10}$$

$$\text{Model A}: \theta = \frac{\pi}{2}, \mathrm{X} = I \tag{11}$$

$$\text{Model B}: \theta = \left\{0, \frac{\pi}{4}, \frac{\pi}{2}, \frac{3\pi}{4}\right\}, \mathrm{X} = I \tag{12}$$

$$\text{Model C}: \theta = \frac{\pi}{2}, \mathrm{X} = \{I, RG, GR, BY, YB\} \tag{13}$$

$$\text{Model D}: \theta = \left\{0, \frac{\pi}{4}, \frac{\pi}{2}, \frac{3\pi}{4}\right\}, \mathrm{X} = \{I, RG, GR, BY, YB\} \tag{14}$$

We define four 3D models which differ in the combinations of features and orientations. To compute disparity information, Model A uses intensity at one orientation only (vertical), model B uses intensity at four orientations, model C uses intensity plus color features at one orientation, and model D uses intensity plus color features at four orientations. This means that model A, the simplest, uses $\theta = \frac{\pi}{2}$ in equations (4) and (5), i.e., a vertical Gabor filter and only the I channel as X in equations (7–10). Model D, the most complex, utilizes $\theta \in \{0, \frac{\pi}{4}, \frac{\pi}{2}, \frac{3\pi}{4}\}$ for the Gabor filters and I, RG, GR, BY, and YB channel as X. All models include the 2D features described in Section 3.3, which means that all of them include color and orientation information in the computation of 2D saliency.

Responses of binocular complex cells are calculated from the simple cells with displaced images,

$$C^k_{X,\theta,d}(x,y) = \left(S^k_{X,R,e,\theta}(x,y) + S^k_{X,L,e,\theta}(x+d,y)\right)^2 + \left(S^k_{X,R,o,\theta}(x,y) + S^k_{X,L,o,\theta}(x+d,y)\right)^2 \quad (15)$$

where d is the disparity between the right and left images. The range of the disparity d is arbitrary. In this paper, d takes on the range of ± 8% of the input image width. The simple cell activities S^k are rescaled to the original image size before the calculation to make the disparity d cover the same displacement at all levels.

As described, this is called a position-based model since the displacement between two images is represented as a position difference. indicated by d in the equations. The displacement can also be represented by a phase difference, resulting in phase-based models (Fleet et al., 1991, 1996; Ohzawa et al., 1997). Physiological experiments show that the brain uses both approaches (DeAngelis et al., 1991; Ohzawa et al., 1990), but their roles in stereopsis are controversial. One possibility is that the phase disparity tuned cells are used because they provide higher accuracy (Qian & Zhu, 1997). However, pure phase-based disparity can only capture disparities smaller than the receptive field's wavelength. Furthermore, Read and Cumming pointed out that phase disparity does not exist in natural images and that cells responding with phase disparity characteristics may function as "lie-detectors" to eliminate false matches (Read & Cumming, 2007). In this study, for the sake of simplicity we use only the position-based model.

The output of a binocular complex cell is enhanced when the two images match at the cell's preferred disparity. Focusing on a specific location (x,y), the value of $C^k_{\beta,\theta,d}$ represents the "confidence" that the location belongs to the specific disparity d, which depends on X, θ, and k. However, its response is also enhanced where false matches or high monocular contrast exist. To compensate for such unreliable responses, we employ a softmax function to compute normalized complex cell responses C',

$$C'^k_{X,\theta,d}(x,y) = \frac{\exp\left(C^k_{X,\theta,d}(x,y)\right)}{\sum_{d'}\exp\left(C^k_{X,\theta,d'}(x,y)\right)} \quad (16)$$

This can be interpreted as the "normalized confidence" of the disparities of each location (x,y) and for each of the parameters (scale, angle, color, and intensity). This computation is similar to the divisive normalization mechanism that is found in many cortical circuits (Heeger, 1992).

Then, C' is linearly summed up over scales, intensity and color maps, and orientations to compute disparity confidence maps, D'. This is written as:

$$D'_d(x,y) = \sum_\theta \sum_\beta \sum_k C'^k_{X,\theta,d}(x,y) \quad (17)$$

Integration of multiple spatially frequency maps was reported in visual cortex (Baba et al., 2015; Kumano et al., 2008) (although we use a broader range of frequencies in our model) as was integration of color and orientation (Garg et al., 2019; Ghose and Ts'o, 2017).

In the primate brain, disparity information is sent to both dorsal and ventral areas (Preston et al., 2008). Here, we focus on the ventral stream which encodes depth information categorically while the dorsal stream represents it in a parametric manner. Thus, we collapse the normalized disparity map into three categories: near-, zero-, and far-positions. This can be written as:

$$Near(x,y) = \sum_d^{d\in near} D'_d(x,y)$$

$$Zero(x,y) = \sum_d^{d\in zero} D'_d(x,y)$$

$$Far(x,y) = \sum_d^{d\in far} D'_d(x,y) \quad (18)$$

where *near*, *zero*, and *far* mean ranges of disparities for each position in the scene. We set these ranges based on Panum's fusional area which is defined as encompassing any point where binocular fusion can be achieved (i.e. absence of diplopia) and which spans approximately 10 to 20 min of arc disparity (Qin et al., 2004). The channel of *zero* is set to approximate Panum's fusional area, and *near* and *far* include all nearer and farther disparities, respectively. We used ± 5 pixels (which corresponds to approximately ± 10 min of arc for our validation setup described in Section 3.4) as the range for the *zero* channel.

Fig. 3 shows a schematic of the disparity channel computation. The calculated disparity channels form one set of inputs to the proto-object based saliency model, on the same footing with the 2D features intensity, color, and orientation (Russell et al., 2014; Uejima et al., 2020). Fig. 4 explains how the depth information is processed in the proposed disparity process.

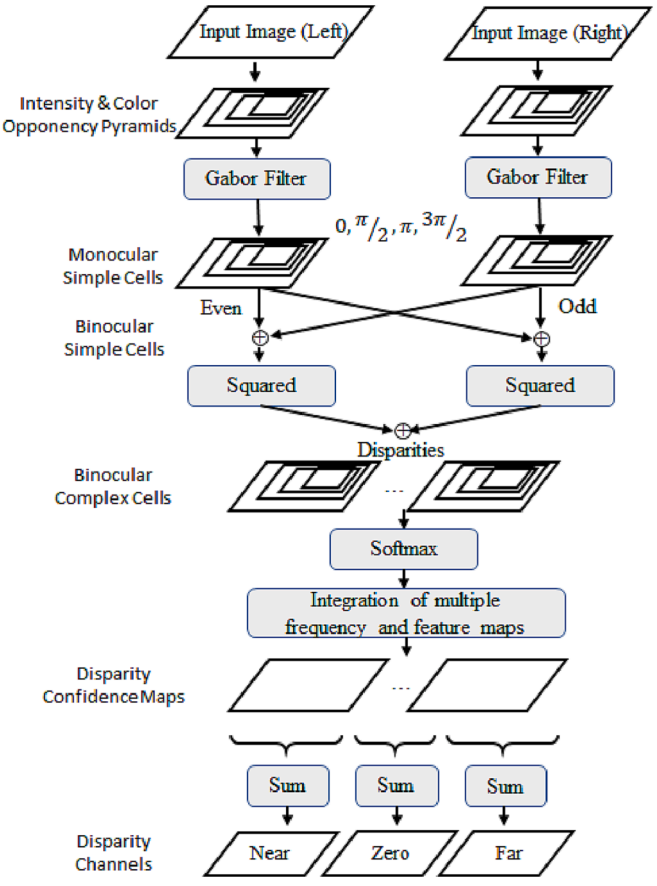

Fig. 3. Schematic of disparity channel computation.

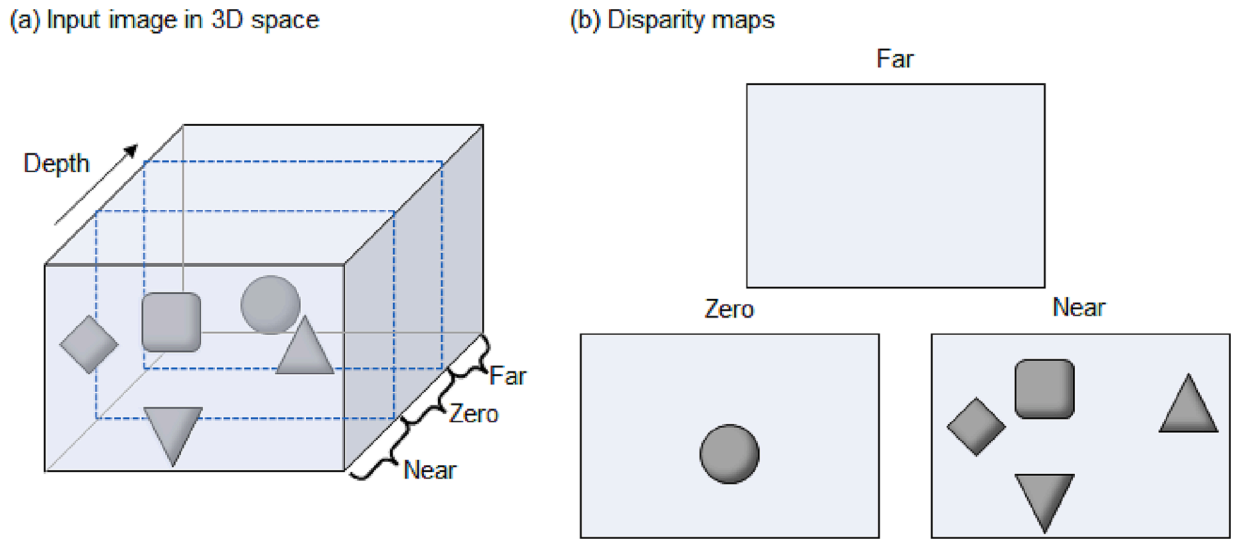

Fig. 4. Conceptual images of the disparity maps. (a) An input image represented in 3D space (actual stimulus is provided as stereo images). The depth is divided into three category: Near, Zero, and Far. (b) Extracted disparity maps.

4.3. Proto-object based model

2D feature channels are generated by extracting intensity, color, and orientation feature maps from the input, and then the 2D channels and disparity channels are processed by the proto-object based model. These algorithms except disparity channels are basically the same as the supplementary materials in (Uejima et al., 2020) and in the remainder of this subsection we follow the description in that paper closely.

The intensity and color channels are computed by eq. (1–3), respectively.

The orientation channels with four angles, O_0, $\mathrm{O}_{\frac{1}{4}\pi}$, $\mathrm{O}_{\frac{1}{2}\pi}$, $\mathrm{O}_{\frac{3}{4}\pi}$, are calculated from the intensity features, and they are the same as I at this point.

$$\mathrm{O}_\theta = I = \frac{r+g+b}{3}\left(\theta \in \left\{0, \frac{1}{4}\pi, \frac{1}{2}\pi, \frac{3}{4}\pi\right\}\right) \tag{19}$$

10-level pyramid images are created by scaling the intensity, color, and disparity channels. The scaling is done by half-octave for each level. In a similar way to using the *meta*-variable X in Section 3.2, we refer to the image features at level k as β^k, $\beta \in \left\{I, RG, GR, BY, YB, \mathrm{O}_0, \mathrm{O}_{\frac{1}{4}\pi}, \mathrm{O}_{\frac{1}{2}\pi}, \mathrm{O}_{\frac{3}{4}\pi}, Near, Zero, Far\right\}$.

The pyramid images are processed by surface detectors employing center-surround mechanism with receptive fields, cs_{on} and cs_{off}, which are modeled as two 2D Gaussian filters represented as:

$$\begin{aligned} cs_{on}(x,y) &= \frac{1}{2\pi\sigma_i^2} e^{-\frac{x^2+y^2}{2\sigma_i^2}} - \frac{1}{2\pi\sigma_o^2} e^{-\frac{x^2+y^2}{2\sigma_o^2}} \\ cs_{off}(x,y) &= -\frac{1}{2\pi\sigma_i^2} e^{-\frac{x^2+y^2}{2\sigma_i^2}} + \frac{1}{2\pi\sigma_o^2} e^{-\frac{x^2+y^2}{2\sigma_o^2}} \end{aligned} \tag{20}$$

where σ_i is the standard deviation of the center (inner) Gaussian, and σ_o is the standard deviation of the surrounding (outer) Gaussian. Here, these are set as $\sigma_i = 0.9$ and $\sigma_o = 2.7$. These kernels are replaced by the even Gabor filters to calculate the orientation channel so that the kernels, $cs_{on,\mathrm{O},\theta}$ and $cs_{off,\mathrm{O},\theta}$, are written as;

$$cs_{on,\mathrm{O},\theta}(x,y) = \exp\left(-\frac{x^{'2} + \gamma_1^2 y^{'2}}{2\sigma_1^2}\right)\cos(\omega_1 x') \tag{21}$$

$$cs_{off,\mathrm{O},\theta}(x,y) = -cs_{on,\mathrm{O},\theta}(x,y) \tag{22}$$

$$x' = x\cos\theta + y\sin\theta, y' = -x\sin\theta + y\cos\theta \tag{23}$$

Where $\gamma_1 = 0.8$, $\sigma_1 = 3.2$, and $\omega_1 = 0.7854$ in the same manner as the original proto-object based model (Russell et al., 2014). The cs_{on} detects light objects on dark backgrounds and cs_{off} does dark objects on light backgrounds.

The center-surround activities, $\mathscr{CS}$, are calculated as products of the center-surround kernels and each feature, which can be written as:

For $\beta \in \{I, RG, GR, BY, YB, Near, Zero, Far\}$

$$\begin{aligned} \mathscr{CS}^k_{\beta,D}(x,y) &= \mathscr{N}_1\left(\left\lfloor \beta^k(x,y) * cs_{off}(x,y)\right\rfloor\right) \\ \mathscr{CS}^k_{\beta,L}(x,y) &= \mathscr{N}_1\left(\left\lfloor \beta^k(x,y) * cs_{on}(x,y)\right\rfloor\right) \end{aligned}$$

For $\beta \in \left\{\mathrm{O}_0, \mathrm{O}_{\frac{1}{4}\pi}, \mathrm{O}_{\frac{1}{2}\pi}, \mathrm{O}_{\frac{3}{4}\pi}\right\}$

$$\begin{aligned} \mathscr{CS}^k_{\beta,D}(x,y) &= \mathscr{N}_1\left(\left\lfloor \beta^k(x,y) * cs_{off,O,\theta}(x,y)\right\rfloor\right) \\ \mathscr{CS}^k_{\beta,L}(x,y) &= \mathscr{N}_1\left(\left\lfloor \beta^k(x,y) * cs_{on,O,\theta}(x,y)\right\rfloor\right) \end{aligned} \tag{24}$$

where $\mathscr{CS}^k_D$ and $\mathscr{CS}^k_L$ form the dark and light object pyramids with k-th level scaled images, and $\mathscr{N}_1(\cdot)$ is a normalization operator used in the same way as by (Russell et al., 2014). In the $\mathscr{N}_1$ normalization process, $\mathscr{CS}_D$ and $\mathscr{CS}_L$ are simultaneously normalized to the range of 0 to 10. Then the average of all local maxima, $\overline{\mathrm{m}}$, is computed across both maps, and each map is multiplied by $(10-\overline{\mathrm{m}})^2$. It emphasizes the global maximum center-surround response and suppresses maps with multiple local maxima. Similar normalization procedures are used in many model implementations of saliency maps, including the original (Itti et al., 1998) study. Simple cells activities, which work as edge detectors

$$S^k_{\beta,e,\theta}(x,y) = \beta^k(x,y) * g_{e,\theta}(x,y) \tag{25}$$

$$S^k_{\beta,o,\theta}(x,y) = \beta^k(x,y) * g_{o,\theta}(x,y) \tag{26}$$

where $g_{e,\theta}$ and $g_{o,\theta}$ are defined in eq. (4–6) with $\gamma = 0.5$.

The activity of complex cells are calculated from the simple cells:

$$C^k_{\beta,\theta}(x,y) = \sqrt{S^k_{\beta,e,\theta}(x,y)^2 + S^k_{\beta,o,\theta}(x,y)^2} \tag{27}$$

The surface (center-surround activities) and edge (complex cells) maps are used to calculate border-ownership coding which is physiologically observed mainly in cortical area V2 (Zhou et al., 2000). The

firing activity of some of these cells is independent of contrast polarity. To simulate this function, we first compute $\mathscr{B}^k_{\theta,L}$, the border ownership activity for a light object on a dark background and $\mathscr{B}^k_{\theta,D}$, the border ownership activity for a dark object on a light background:

$$\mathscr{B}^k_{\beta,\theta,L}(x,y) = \left[\mathscr{C}^k_{\beta,\theta}(x,y) \times \left(1 + \sum_{j\geq k} \frac{1}{2^j} v_{\theta+\pi}(x,y) * \mathscr{CS}^j_{\beta,L}(x,y) - \sum_{j\geq k} \frac{1}{2^j} v_{\theta}(x,y) * \mathscr{CS}^j_{\beta,D}(x,y) \right) \right]$$

$$\mathscr{B}^k_{\beta,\theta,D}(x,y) = \left[\mathscr{C}^k_{\beta,\theta}(x,y) \left(1 + \sum_{j\geq k} \frac{1}{2^j} v_{\theta+\pi}(x,y) * \mathscr{CS}^j_{\beta,D}(x,y) - \sum_{j\geq k} \frac{1}{2^j} v_{\theta}(x,y) * \mathscr{CS}^j_{\beta,L}(x,y) \right) \right] \quad (28)$$

where v is the von Mises distribution (Russell et al., 2014) calculated as:

$$v_\theta(x,y) = -\frac{\exp\left[\left(\sqrt{x^2+y^2} - R_0 \right) \sin\left(\tan^{-1}\frac{y}{x} - \theta \right) \right]}{2\pi I_0 \left(\sqrt{x^2+y^2} - R_0 \right)} \quad (29)$$

R_0 is the zero-crossing radius of the center surround masks, and I_0 is the modified Bessel function of the first kind. v_θ is then normalized as:

$$v_\theta(x,y) = \frac{v_\theta(x,y)}{\max(v_\theta(x,y))} \quad (30)$$

The responses of the border-ownership cells to light and dark objects are combined to make them independent of figure-ground contrast polarity.

$$\mathscr{B}^k_{\beta,\theta}(x,y) = \mathscr{B}^k_{\beta,\theta,L}(x,y) + \mathscr{B}^k_{\beta,\theta,D}(x,y) \quad (31)$$

At each pixel, multiple border ownership cells exist for each direction of ownership, organized in pairs with mutually opposing spatial preferences. For instance, at a pixel located on a vertical border, there is one border ownership with a higher rate when the foreground object is to its right, and its partner which has identical feature preferences but which fires with a higher rate when the foreground object is to its left. To determine which border a pixel belongs to, the model uses a winner-take-all algorithm between the response of a border ownership selective neuron $\mathscr{B}^k_\theta$ and its partner $\mathscr{B}^k_{\theta+\pi}$,

$$\widehat{\mathscr{B}}^k_\beta(x,y) = \mathscr{B}^k_{\beta,\hat{\theta}}(x,y) \quad (32)$$

where

$$\hat{\theta} = \underset{\theta}{\operatorname{argmax}} \left(\mathscr{B}^k_{\beta,\theta}(x,y) - \mathscr{B}^k_{\beta,\theta+\pi}(x,y) \right) \quad (33)$$

Then, the grouping cell responses are calculated by summing the winning border ownership activity in an annular fashion (Craft et al., 2007; Russell et al., 2014)

$$\mathscr{G}^k_\beta(x,y) = \sum_\theta \left[\delta\left(\mathscr{B}^k_{\beta,\theta}(x,y), \widehat{\mathscr{B}}^k_\beta \right) \times \left(\mathscr{B}^k_{\beta,\hat{\theta}}(x,y) - \mathscr{B}^k_{\beta,\hat{\theta}+\pi}(x,y) \right) * v_{\hat{\theta}}(x,y) \right] \quad (34)$$

where $\delta\left(\mathscr{B}^k_{\beta,\theta}(x,y), \widehat{\mathscr{B}}^k_\beta \right) = 1$ if $\mathscr{B}^k_{\beta,\theta}(x,y) = \widehat{\mathscr{B}}^k_\beta$ and zero otherwise.

A final saliency map is computed by normalizing and combining each grouping cell response from each channel. The combined channels for intensity, color, orientation, and disparity (namely, $\overline{\mathscr{I}}$, $\overline{\mathscr{C}}$, $\overline{\mathscr{O}}$, and $\overline{\mathscr{D}}$) are calculated by:

$$\overline{\mathscr{I}} = \oplus_{k=1}^{k=10} \mathscr{N}_2\left(\mathscr{G}^k_I \right)$$

$$\overline{\mathscr{C}} = \oplus_{k=1}^{k=10} \left(\mathscr{N}_2\left(\mathscr{G}^k_{RG} \right) + \mathscr{N}_2\left(\mathscr{G}^k_{GR} \right) + \mathscr{N}_2\left(\mathscr{G}^k_{BY} \right) + \mathscr{N}_2\left(\mathscr{G}^k_{YB} \right) \right)$$

$$\overline{\mathscr{O}} = \sum_{\alpha\in\{0,\pi/4,\pi/2,3\pi/4\}} \left(\oplus_{k=1}^{k=10} \mathscr{N}_2\left(\mathscr{G}^k_{O\alpha} \right) \right)$$

$$\overline{\mathscr{D}} = \oplus_{k=1}^{k=10} \left(\mathscr{N}_2\left(\mathscr{G}^k_{Near} \right) + \mathscr{N}_2\left(\mathscr{G}^k_{Zero} \right) + \mathscr{N}_2\left(\mathscr{G}^k_{Far} \right) \right) \quad (35)$$

In these equations, $\mathscr{N}_2(\cdot)$ is almost identical to $\mathscr{N}_1(\cdot)$ which was defined after eq. (24), but rather than normalizing two maps ($\mathscr{CS}_D$ and $\mathscr{CS}_L$), normalization only is performed on one map (the argument of $\mathscr{N}_2(\cdot)$).

The sum over normalized proto-object maps of all features constitutes the final saliency map, $\mathscr{S}$, which is represented as:

$$\mathscr{S} = \left(\mathscr{N}_2(\overline{\mathscr{I}}) + \mathscr{N}_2(\overline{\mathscr{C}}) + \mathscr{N}_2(\overline{\mathscr{O}}) + \mathscr{N}_2(\overline{\mathscr{D}}) \right) \quad (36)$$

The overall view of the model is shown in Fig. 5. In this paper, we call the model without the disparity channels (i.e., calculated from only intensity, color, and orientation features) the 2D model, and models that include the disparity channels 3D models (Model A – D, described in Section 3.2).

For the comparison of model results with ground truth, we use eye movement data from two empirical studies, the Gaze-3D and the NCTU-3D data sets (Section 2.2). The former reports eye movements of the left eye and the latter from the right eye. In both cases, we compute the saliency map from the image projected to that eye whose movements were recorded, i.e. from the left eye for the Gaze-3D dataset and the right eye for the NCTU-3D dataset.

4.4. Validation

To evaluate the quality of the proposed model, we used publicly-available eye fixation datasets to compare our saliency maps with human eye movements which are taken as ground truth for the deployment of selective attention. We used the Gaze-3D dataset (Wang et al., 2013) which includes 18 images and the NCTU-3D dataset (Ma & Hang, 2015) comprising 475 images. We reduced the image size of the datasets by a factor of two before using them as input to our model to decrease computation time.

For quantitative validation, we employed five metrics to assess the predictive performance of the generated saliency maps for human fixations. The metrics are normalized scanpath saliency (NSS), Pearson's correlation coefficient (CC), similarity (SIM), Kullback-Leibler divergence (KLD), and shuffled area under the ROC curve (sAUC). These metrics were calculated using published codes (Bylinskii et al., 2019).

As a short overview, NSS is the mean value of the normalized saliency map at the fixation locations. The normalized saliency map is calculated by transforming the map values to zero mean and unit standard deviation. CC takes on zero value for two uncorrelated variables and unit value for identical ones. The SIM measure of two maps is zero when the maps have no overlap and unity if the two maps are identical. KLD quantifies the dissimilarity between two probability distributions, and smaller KLD indicates higher similarity. The sAUC is a modified version of the area under the ROC curve. The Receiver Operating Characteristic (ROC) measures the ratio of true positives and false positives at various thresholds. The sAUC samples negative points to calculate the false positive rate from equivalent fixation locations of *other* images, rather than uniformly random locations from the *same* image that standard AUC uses. This compensates for systematic biases present in all images, such as the well-known center-bias, see. e.g. Parkhurst et al 2002 (Parkhurst et al., 2002). For a discussion how center bias is corrected for the other four metrics see below.

It is known that blurring saliency maps can affect metrics (Borji & Itti, 2012; Hou et al., 2012). Basically, blurring approximates the sampling error of the eye tracker used for recording fixations. We applied 2D Gaussian kernels with various widths and determined the optimal blurring kernel for each model and metric. The kernel width was varied between 1% and 20% of the image widths for NCTU-3D, and between

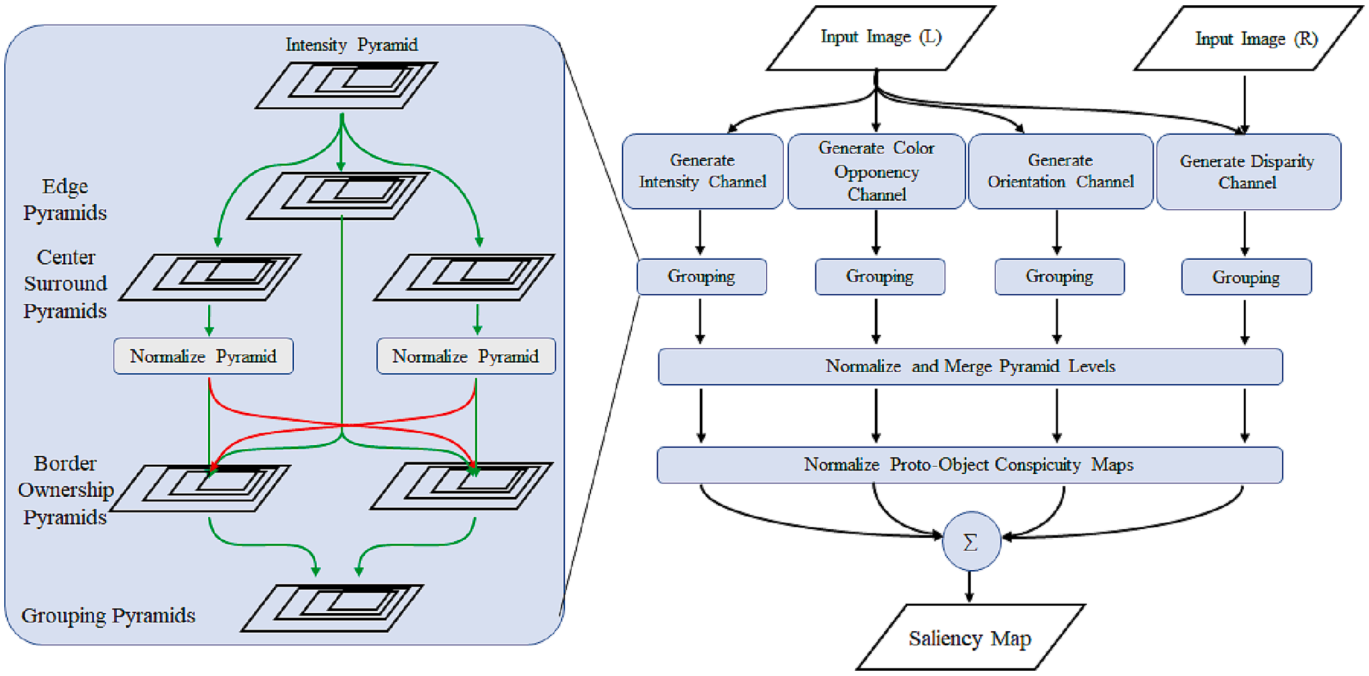

Fig. 5. Overall view of the model. In this overview, the non-depth related components of the saliency map (intensity, color, orientation) are computed from the input to the left eye. During validation, this is the case for the Gaze-3D dataset. For the NCTU-3D dataset, these components are computed from the input to the right eye.

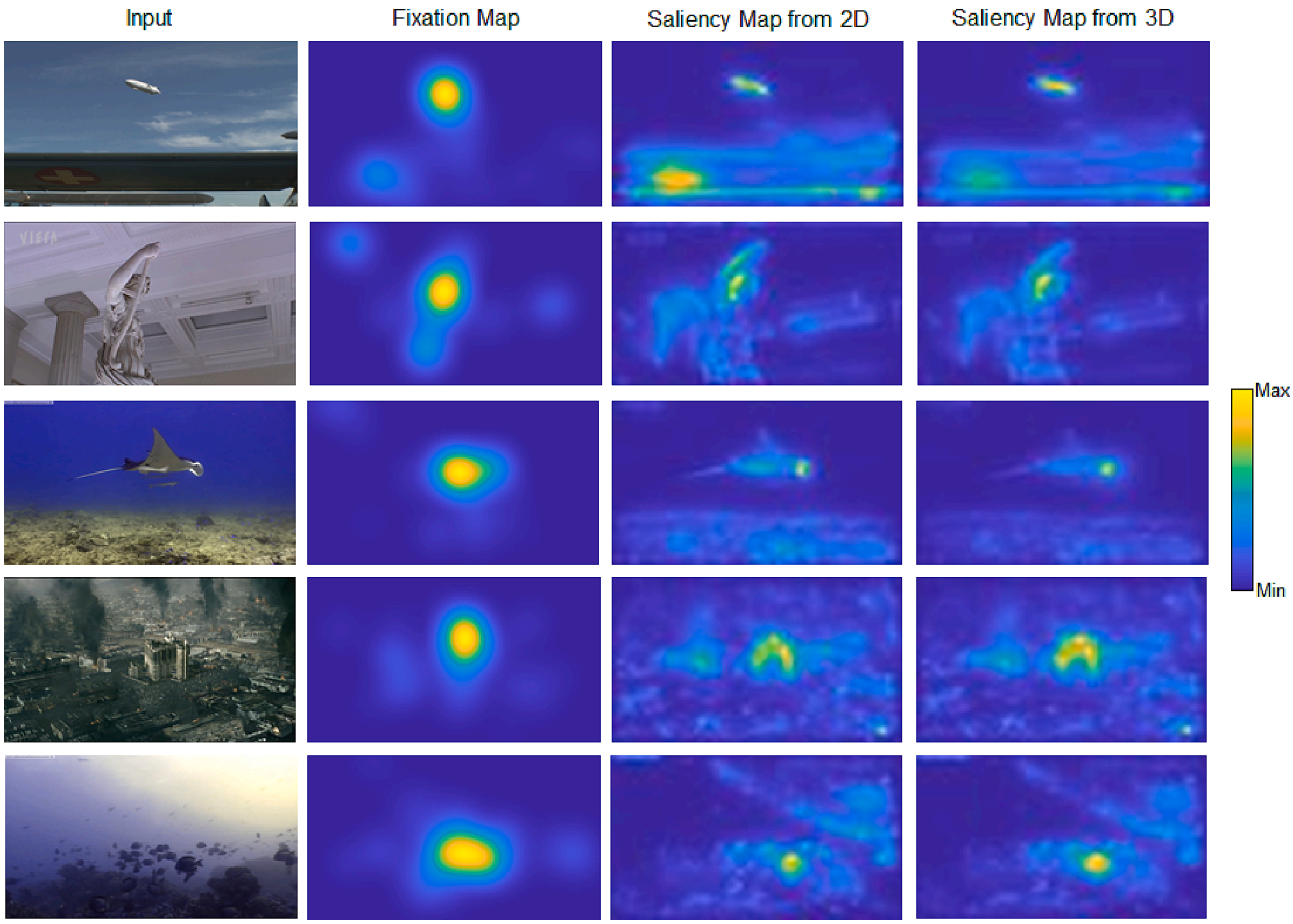

Fig. 6. Examples of saliency maps generated by the 2D model and a 3D model (model A). The input image and fixation map are from the NCTU-3D dataset. Our 3D model suppresses background saliency compared to the 2D model. Color scale applies to all saliency and fixation maps.

1% and 40% for Gaze-3D in steps of 1%. An exception was the sAUC metric which was calculated under the condition of the blurring kernel width between 1% and 8%, because the sAUC's metric produced higher values for smaller kernels than the other metrics.

All metrics except sAUC are affected by the center-bias: human observers tend to fixate preferentially at locations in the vicinity of the centers of images. Parkhurst et al (Parkhurst et al., 2002) showed that weighting saliency with a Gaussian at the image center resulted in better fixation prediction, and that it could be improved even more by centering a Gaussian on the location of the instantaneous fixation, to take into account the fall-off of visual acuity in the periphery. Following (Zhang & Sclaroff, 2016), here we use a simpler approach of using a fixed parabolic distance-to-center (DTC) re-weighting. This is computed as:

$$DTC(i,j) = 1 - \frac{\sqrt{\left(i - \frac{H}{2}\right)^2 + \left(j - \frac{W}{2}\right)^2}}{\sqrt{\left(\frac{H}{2}\right)^2 + \left(\frac{W}{2}\right)^2}} \tag{37}$$

where i and j are the row and column indice and H and W are the height and width of the input image. The generated saliency map blurred by the 2D Gaussian kernel with the best sigma is pixel-wise multiplied with DTC. The DTC re-weighting procedure was not applied for the sAUC metric because sAUC automatically compensates for the center-bias.

5. Results

Examples of the proposed disparity channels and the resulting saliency map are shown in Fig. 1 (b) and (c), respectively. The fixation map in this example indicates that participants tend to fixate a fish in the lower half of the image that is closer to the observer than the other fish. The computed disparity maps in Fig. 1 (b) show that the location of the fish is captured by the near and zero channels. The saliency map is calculated based on the disparity channels and shows high value at the approximate location of this fish as shown in Fig. 1 (c).

Fig. 6 shows comparisons between fixation maps and saliency maps generated by the 2D and 3D models. We here use model A described in Section 3.2 which employs vertical orientation and intensity. The 2D model saliency maps were calculated from intensity, color, and orientation features as in (Uejima et al., 2020), but without the texture features in that model. The 3D saliency maps were computed by adding depth maps generated from the disparity channels. In these examples, the 2D model predicted the human fixations to some extent, and the depth information improved the predictions. In the example of the first row, for instance, the 3D model shows higher saliency value at the location of the airship than the 2D model, in agreement with human fixations. In all examples shown, the 3D model suppresses saliency to background patterns, relative to foreground objects. This shows that depth plays a role in the prediction of overt attention.

To quantify the performance enhancement due to adding depth channels, we calculate the metrics of the saliency maps generated with and without the depth information by comparing them against human fixations. These metrics are shown in Table 1. As described in Section 3.2, we used four variations of the 3D model which include different combinations of color features and orientations. Our results indicate that the depth channels improve the prediction of human fixations. Because of the small size of the Gaze-3D dataset (only 18 images), in the following we focus our analysis on the NCTU-3D data.

We first look at the models that only use intensity information (not color), i.e., models A and B. For all but two of the ten comparisons between the 2D model and the two corresponding models with 3D information (Models A and B), the models with 3D information performed equally well or better than the 2D model. The increase in performance was significant for 5 of these comparisons (two-tailed paired t-test, $p < 0.05$).

If color information is added, models incorporating 3D cues (Models C and D) are equal or better than the corresponding 2D model for all ten comparisons. However, the differences are small and don't reach significance. A surprising result is that the simplest of the 3D models, Model A with only one orientation (vertical) in which neither color nor the other orientations contribute, overall performs best. It has the best scores by all metrics on the Gaze-3D dataset, and the highest by all metrics bar-one on the NCTU-3D dataset. Even in case of the one exception in the NCTU-3D data, the best performance occurs in the second model with only one orientation, Model C. Given that human eyes are typically at the same height, and binocular disparity thus occurring between locations symmetric to the vertical, it seems intuitive that the vertical orientation is the most important. It appears that taking into account other orientations is not only unnecessary but, in fact, interfering with optimal performance. As for the lacking contribution of color information, we will come back to this question in our Discussion section.

Primates express strong interest in faces and bodies which attract attention even when they are task-irrelevant (Landman et al., 2014). Indeed, detection of faces and body parts is supported by anatomical structures in monkeys (Desimone, 1991; Gross, 2008; Tsao et al., 2003) and humans (Downing et al., 2001). We expect that the human fixation locations that we use as ground truth in this study show a similar bias. Since none of our models has corresponding explicit detection mechanisms for faces or body parts, we expected that models predict fixation

Table 1
Models incorporating depth features predict human eye fixations equally well or better than the 2D model. In column 1, "1 orientation" indicates a model with only one (vertical) Gabor filter and "4 orientations" models with Gabor filter with four orientations. Labels "I" and "I&C" indicate models with only the intensity feature and a combination of intensity and color features, respectively. Underscore denotes the best score for each metric. A parenthesis next to the performance value of a model indicates that this model performs significantly better than any of the models listed in the parenthesis, where B, C, D, indicate the different 3D models, and "2" the 2D model. For instance, by the CC metric Model A performs better than models 2, B, and C. Significance was evaluated by two-tailed paired t-tests ($p < 0.05$). Larger values are better for all metrics but KLD.

Model	Metrics				
	NSS	CC	SIM	KLD	sAUC
NCTU-3D dataset					
2D model	1.468	0.633	0.612	0.518	0.657
Model A (3D, 1 orientation, I)	1.473 (B)	0.638 (2,C,D)	0.614 (2,B,C)	0.513	0.655
Model B (3D, 4 orientations, I)	1.463	0.638 (2,C,D)	0.613 (2)	0.515 (2)	0.654
Model C (3D, 1 orientation, I&C)	1.468	0.633	0.612	0.517	0.660 (A,B)
Model D (3D, 4 orientations, I&C)	1.470	0.633	0.613	0.516	0.657
Gaze-3D dataset					
2D model	0.958	0.685	0.731	0.254	0.609
Model A (3D, 1 orientation, I)	0.968 (C,D)	0.687 (C,D)	0.732	0.252	0.611
Model B (3D, 4 orientations, I)	0.959 (D)	0.684 (D)	0.731	0.255	0.610
Model C (3D, 1 orientation, I&C)	0.949	0.678	0.729	0.256	0.608
Model D (3D, 4 orientations, I&C)	0.947	0.680	0.729	0.255	0.605

locations better for images that have no persons in the scene than for images with persons. We therefore divided the images into two sets: one set with humans visible (fully or partially) and the other set without. The latter could, however, include animals or statues of humans. In the NCTU data set, we found 303 images in which humans were visible, at least partially, and 172 images in which that was not the case. In the Gaze-3D data, 7 images included visible humans and 11 did not.

The results shown in Table 2 confirm our expectation: for all five models, and for all five metrics, fixation prediction performance decreased with the presence of faces, bodies, or body parts in the NCTU-3D dataset. In the majority of the tests (15/25), this decrease was significant (one-tailed Welch test, $p < 0.05$). Similarly for the Gaze-3D data, the majority (17/25) of tests showed higher performance for images without humans and the majority of these differences (10/17) were significant. Overall, the effect was somewhat weaker than for the NCTU-3D data, but the Gaze-3D dataset is very small which limits the statistical power that can be achieved. We also note that a separate channel for face detection, for instance using the standard Viola-Jones algorithm (Viola & Jones, 2001), can be easily added to the models and would most likely increase fixation prediction performance substantially, as it did in a previously-developed class of models from the same pedigree as ours (Cerf et al., 2008).

6. Discussion

Binocular information processing in our model is based on physiological and psychological evidence. Its basic mechanism is the binocular energy model which combines information from two monocular input sources into a binocular signal, akin to the generation of complex cell responses from the activity of monocular simple cells in two eyes. Binocular complex cells are tuned to specific disparity ranges and their activity represents a "confidence" measure of the disparity difference within their receptive fields. Our model of cells with binocular receptive fields uses Gabor filters with high spatial aspect ratio ($\gamma > 1$). While the spatial aspect ratio of the majority of orientation selective cortical cells is $\gamma > 1$, a fraction of cells in early visual cortex have a spatial aspect ratio larger than unity. We hypothesize that their role is primarily in disparity computations, rather than in the representation of oriented edges for which elongated filters are better suited.

Despite decades of neurophysiological research, it is still not entirely clear how the brain deals with depth information. Early studies of stereopsis proposed that the disparity processing is achieved categorically, by cells tuned near, far, and zero (in the focal plane) (Poggio & Poggio, 1984; Richards, 1971). Later, this three-channel model was replaced by a continuous representation, similar to that for orientation or motion direction (Poggio, 1995). A recent imaging study indicates that, in fact, both ideas may be correct but used in different pathways: dorsal cortical areas, including V3A and V7 encode parametric disparity while areas in the ventral pathway, in particular the lateral occipital area, represent the categorical responses "near" and "far (Preston et al., 2008). The former may be more relevant for visual control of action while the latter may be more useful for tasks like object recognition. In our model, we adopt the latter, i.e. a categorical model akin to what is found in the ventral pathway, with the responses of the binocularly tuned cells organized in three channels: near, far, and (close-to-)zero.

The boundaries between the zero channel and the other two are defined by Panum's area. This gives rise to a limitation of our approach because the model assumes that the focal plane is always in the zero disparity range of the input image. Because subjects can move their eyes, they can fixate objects outside the focal plane in which case Panum's fusion area changes. Our model does not include such dynamic change of subjective disparity perception.

The output of these disparity channels is given as input to a proto-object based saliency model (Russell et al., 2014) with V4 playing a major role. This is clearly different from previous studies of biofidelic 2D saliency (Itti et al., 1998; Itti & Koch, 2000; Koch & Ullman, 1985; Li, 2002; Niebur & Koch, 1996) and 3D saliency (Zhaoping, 2002). It is known that latencies of V2 responses are longer than V1 responses by approximately a dozen of milliseconds in the primate brain (Gawne & Martin, 2002; Nowak et al., 1995; Schmolesky et al., 1998). Reaction time studies (Zhaoping et al., 2009) indicate that depth features, presumably processed in V2, influence attentional guidance in complex stimulus stimuli (those requiring more than one second of human manual reaction times to report their visual perception), in contrast to the low-level features involving V1, which direct attention in tasks where reaction times are much faster. This would imply that depth-derived attentional guidance in our proto-object based mechanism would also be rather slow, with reaction times on the order of a second or more. We can not address directly whether this is the case because in the datasets we use in our study, eye fixations last for 4 to 15 s.

In our model, proto-objects are calculated separately in the spaces of near, far, and zero channels. As in earlier saliency map models, interactions between features and spatial scales emphasize the influence of those maps with a small number of local maxima and suppress those with many peaks. In typical scenes, a small number of foreground objects tend to be in the near or zero depth zones while broad areas of cluttered background are in the "far" channel. For such scene structures, the model emphasizes the foreground objects in the near and zero channels which only have a few peaks. This agrees with the observation that humans tend to fixate objects in foreground (near) locations (Gautier & Le Meur, 2012; Jansen et al., 2009; Lang et al., 2012). We find that by most metrics our intensity-based models (Models A and B)

Table 2
Effect of presence of persons in images. The left and right columns of each metrics show model performance on images without and with persons, respectively. Underscore denotes the best score for each metric. For the NCTU-3D data, model performance is always higher on images without persons. This is not the case for Gaze-3D but this dataset has only 7 images with persons. An asterisk (*) indicates that the performance is significantly better on images without persons (one-tailed Welch test, $p < 0.05$).

Model	Metrics									
	NSS		CC		SIM		KLD		sAUC	
NCTU-3D dataset	without	with	without	with	without	with	without	with	without	with
2D model	1.540	1.427	0.685*	0.603	0.637*	0.597	0.478*	0.541	0.657	0.657
Model A (3D, 1 orientation, I)	1.551*	1.430	0.693*	0.607	0.640*	0.599	0.465*	0.541	0.661	0.653
Model B (3D, 4 orientations, I)	1.545*	1.416	0.692*	0.607	0.640*	0.598	0.469*	0.542	0.655	0.653
Model C (3D, 1 orientation, I&C)	1.532	1.432	0.685*	0.604	0.635*	0.599	0.479*	0.539	0.665	0.657
Model D (3D, 4 orientations, I&C)	1.543	1.429	0.685*	0.604	0.637*	0.599	0.476*	0.538	0.660	0.654
Gaze-3D dataset	without	with	without	with	without	with	without	with	without	with
2D model	0.874	1.090	0.703	0.657	0.760*	0.686	0.217*	0.311	0.610	0.607
Model A (3D, 1 orientation, I)	0.878	1.110	0.704	0.660	0.760*	0.690	0.219*	0.304	0.618	0.599
Model B (3D, 4 orientations, I)	0.871	1.098	0.702	0.656	0.759*	0.687	0.222*	0.307	0.604	0.621
Model C (3D, 1 orientation, I&C)	0.855	1.096	0.693	0.655	0.757*	0.684	0.220*	0.312	0.606	0.610
Model D (3D, 4 orientations, I&C)	0.862	1.082	0.696	0.653	0.758*	0.684	0.216*	0.316	0.598	0.616

predict human fixation patterns better than the analogous model that uses monocular information only (Table 1).

This was not the case for the two models that, in addition to intensity, also use color information. While intensity-only based Models A and B showed nearly uniformly better performance than the 2D model, the models incorporating color information to calculate depth in addition (Models C and D) performed similarly to the 2D model on the NCTU-3D dataset, and even slightly worse on the Gaze-3D dataset (the small size of Gaze-3D makes it difficult to assign high importance to the latter result). Furthermore, comparing the intensity-only models directly with the models that use both intensity and color information, the former predict fixation clearly better. This result seems at first puzzling: why would a model that has access to more information perform worse than one with less information?

There are two considerations to take into account. First, parameters in our models are assigned fixed values, they are not selected by a learning algorithm that optimizes performance in a given task (here, optimal agreement with eye fixations). There is therefore no guarantee that providing additional information improves the performance. Second, and in addition, we define performance as agreement with human eye movements. By this measure, the expectation that making available additional source of types of information improves performance rests on the assumption that this information is also available to, and used by, the mechanisms that guide eye movements in humans. This is, however, not necessarily the case; the role of chromatic information in stereopsis is complicated. At the same time that Julesz invented random-dot stereograms (Julesz, 1971), he reported that anti-correlated dots (with reversed intensity polarity) did not give rise to stereoscopic vision, but that correlated colors of the dots aided binocular fusion. Shortly thereafter, Lu and Fender disputed the latter finding, concluding their paper with the sentence "luminance alone is used as the principal signal to determine their depth" (Lu & Fender, 1972). Later studies found evidence for stereoscopic vision in isoluminant conditions (Comerford, 1974; Grinberg & Williams, 1985) but methodological problems with their approach led Livingstone and Hubel to question these results (Livingstone & Hubel, 1987). Instead, they proposed that color, processed in the parvocellular stream, interacts minimally, at best, with orientation, motion and depth which are predominantly represented in the magnocellular stream. This interpretation has been questioned yet again by newer results that provide evidence for effective cross-communication between the magnocellular and parvocellular channels that support stereopsis driven by luminance, even though it is not clear that color information by itself is sufficient to induce depth perception (Scharff & Geisler, 1992; Simmons & Kingdom, 1997; Tyler & Cavanagh, 1991). While newer evidence supports that color information does modulate perception of depth (Den Ouden et al., 2005), it is not clear if the effect is strong enough in natural scenes, the stimuli that we use, to result in statistically significant differences. Overall, while it was not the focus of our work to study the role of color in stereopsis, our results provide some support to the notion that for this perceptual function and for this set of stimuli, chromatic information pays only a minor role if any.

Understanding stereoscopic vision is not only of interest for basic science but it also has practical implications. As mentioned in the Introduction, one is obviously the determination of the distance of objects from the observer. Another (related) one is camouflage breaking. The goal of camouflage is to disrupt the process of segmenting the to-be-camouflaged object from its background. In addition to the obvious method of trying to match the local visual properties of the object as closely as possible with those of the background, a (related) technique is to create strong contours inside the object boundaries, resulting in the creation of internal "false" edges. This process interferes with the grouping of the object's features into a coherent entity, and therefore its identification (Adams et al., 2019). Availability of depth information reduces the effect of monocular edges internal to the object and therefore aids the visual system in the formation of correct object borders, segregating the object from the background. It is known that visual scenes are organized in terms of such segregated objects, or more precisely, proto-objects (Rensink, 2000) which are entities towards attention is directed (Egeth & Yantis, 1997; Egly et al., 1994; Scholl, 2001). The fact that our 3D models predict human fixation patterns better than the 2D model supports the idea that human fixations are partially guided by depth cues. Furthermore, this result may have consequences for the application of our models in the context of camouflage. Given that depth cues contribute strongly to scene segmentation and, therefore, camouflage breaking (Adams et al., 2019), we may conjecture that this is a possible application where algorithms like our models A and B may be useful. We did not test this hypothesis specifically on scenes with camouflage patterns, but this is a topic for future research.

7. Conclusion

We incorporate a biologically plausible stereopsis mechanism into a proto-object based saliency model. The proposed model takes stereoscopic images as input and computes categorical depth information from a disparity energy model. We combine the resulting depth information with monocular information (intensity, color, and orientation) to form a representation of the visual scene in terms of proto-objects. The resulting saliency map generates predictions for the allocation of overt attention that agree significantly better with human behavior than those from the corresponding maps using monocular cues only. We also note that, different from typical machine learning approaches, all stages of this process are derived from first principles. No training is required other than the setting of a small number of general parameters.

CRediT authorship contribution statement

Takeshi Uejima: Methodology, Software, Visualization, Writing – original draft. **Elena Mancinelli:** Methodology, Software. **Ernst Niebur:** Conceptualization, Writing – review & editing, Supervision, Funding acquisition. **Ralph Etienne-Cummings:** Conceptualization, Methodology, Resources, Writing – review & editing, Supervision, Project administration, Funding acquisition.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

The code is available at our Github page. https://github.com/csmslab

Acknowledgments

We thank the National Science Foundation (Grant 167522 and 1835202) and the National Institutes of Health (Grant 5R01EY027544 and R01DC020123) which support this work through the CRCNS mechanisms. We also acknowledge support from the Office of Naval Research (Grant N00014-22-1-2699). T. Uejima's PhD training was supported by the Japanese Acquisition, Technology & Logistic Agency, Government of Japan.

References

Adams, W. J., Graf, E. W., & Anderson, M. (2019). Disruptive coloration and binocular disparity: Breaking camouflage. *Proceedings of the Royal Society B: Biological Sciences, 286*(1896), 20182045. https://doi.org/10.1098/rspb.2018.2045

Baba, M., Sasaki, K. S., & Ohzawa, I. (2015). Integration of Multiple Spatial Frequency Channels in Disparity-Sensitive Neurons in the Primary Visual Cortex. *The Journal of*

Neuroscience, 35(27), 10025–10038. https://doi.org/10.1523/JNEUROSCI.0790-15.2015
Borji, A., & Itti, L. (2012). Exploiting local and global patch rarities for saliency detection. *IEEE Conference on Computer Vision and Pattern Recognition, 2012*, 478–485. https://doi.org/10.1109/CVPR.2012.6247711
Borji, A., Sihite, D. N., & Itti, L. (2013). Quantitative Analysis of Human: Model Agreement in Visual Saliency Modeling-A Comparative Study. *IEEE Transactions on Image Processing, 22*(1), 55–69. https://doi.org/10.1109/TIP.2012.2210727
Bruce, N. D. B., & Tsotsos, J. K. (2005). In *Saliency Based on Information Maximization* (pp. 155–162). MIT Press.
Bylinskii, Z., Judd, T., Oliva, A., Torralba, A., & Durand, F. (2019). What Do Different Evaluation Metrics Tell Us about Saliency Models? *IEEE Transactions on Pattern Analysis and Machine Intelligence, 41*(3), 740–757. https://doi.org/10.1109/TPAMI.2018.2815601
Cerf, M., Harel, J., Einhauser, W., & Koch, C. (2008). Predicting human gaze using low-level saliency combined with face detection. In J. C. Platt, D. Koller, Y. Singer, & S. T. Roweis (Eds.), *Advances in Neural Information Processing Systems 20* (pp. 241–248). Curran Associates Inc.
Comerford, J. P. (1974). Stereopsis with chromatic contours. *Vision Research, 14*(10), 975–982. https://doi.org/10.1016/0042-6989(74)90166-7
Cornia, M., Baraldi, L., Serra, G., & Cucchiara, R. (2016). Predicting Human Eye Fixations via an LSTM-based Saliency Attentive Model. *ArXiv Preprint, 1611*, 09571. http://arxiv.org/abs/1611.09571.
Craft, E., Schutze, H., Niebur, E., & von der Heydt, R. (2007). A neural model of figure–ground organization. *Journal of Neurophysiology, 97*(6), 4310–4326.
Cumming, B. G., & DeAngelis, G. C. (2001). The Physiology of Stereopsis. *Annual Review of Neuroscience, 24*(1), 203–238. https://doi.org/10.1146/annurev.neuro.24.1.203
DeAngelis, G. C., Ohzawa, I., & Freeman, R. D. (1991). Depth is encoded in the visual cortex by a specialized receptive field structure. *Nature, 352*(6331), 156–159. https://doi.org/10.1038/352156a0
Den Ouden, H. E. M., Van Ee, R., & De Haan, E. H. F. (2005). Colour helps to solve the binocular matching problem. *The Journal of Physiology, 567*(2), 665–671. https://doi.org/10.1113/jphysiol.2005.089516
Desimone, R. (1991). Face-Selective Cells in the Temporal Cortex of Monkeys. *Journal of Cognitive Neuroscience, 3*(1), 1–8. https://doi.org/10.1162/jocn.1991.3.1.1
Deubel, H., & Schneider, W. X. (1996). Saccade target selection and object recognition: Evidence for a common attentional mechanism. *Vision Research, 36*(12), 1827–1837. https://doi.org/10.1016/0042-6989(95)00294-4
Downing, E. P., Yuhong, J., Miles, S., & Nancy, K. (2001). A Cortical Area Selective for Visual Processing of the Human Body. *Science, 293*(5539), 2470–2473. https://doi.org/10.1126/science.1063414
Egeth, H. E., & Yantis, S. (1997). Visual attention: Control, representation, and time course. *Annual Review of Psychology, 48*(1), 269–297. https://doi.org/10.1146/annurev.psych.48.1.269
Egly, R., Driver, J., & Rafal, R. D. (1994). Shifting visual attention between objects and locations: Evidence from normal and parietal lesion subjects. *Journal of Experimental Psychology: General, 123*(2), 161–177. https://doi.org/10.1037/0096-3445.123.2.161
Einhäuser, W., Spain, M., & Perona, P. (2008). Objects predict fixations better than early saliency. *Journal of Vision, 8*(14), 1–26. https://doi.org/10.1167/8.14.18
Fleet, D. J., Jepson, A. D., & Jenkin, M. R. M. (1991). Phase-based disparity measurement. *CVGIP: Image Understanding, 53*(2), 198–210. https://doi.org/10.1016/1049-9660(91)90027-M
Fleet, D. J., Wagner, H., & Heeger, D. J. (1996). Neural encoding of binocular disparity: Energy models, position shifts and phase shifts. *Vision Research, 36*(12), 1839–1857. https://doi.org/10.1016/0042-6989(95)00313-4
Garg, A. K., Li, P., Rashid, M. S., & Callaway, E. M. (2019). Color and orientation are jointly coded and spatially organized in primate primary visual cortex. *Science, 364*(6447), 1275–1279. https://doi.org/10.1126/science.aaw5868
Gautier, J., & Le Meur, O. (2012). A Time-Dependent Saliency Model Combining Center and Depth Biases for 2D and 3D Viewing Conditions. *Cognitive Computation, 4*(2), 141–156. https://doi.org/10.1007/s12559-012-9138-3
Gawne, T. J., & Martin, J. M. (2002). Responses of Primate Visual Cortical Neurons to Stimuli Presented by Flash, Saccade, Blink, and External Darkening. *Journal of Neurophysiology, 88*(5), 2178–2186. https://doi.org/10.1152/jn.00151.200
Ghose, G. M., & Ts'o, D. Y. (2017). Integration of color, orientation, and size functional domains in the ventral pathway. *Neurophotonics, 4*(3), 1–18. https://doi.org/10.1117/1.NPh.4.3.031216
Ghosh, S., D'Angelo, G., Glover, A., Iacono, M., Niebur, E., & Bartolozzi, C. (2022). Event-driven proto-object based saliency in 3D space to attract a robot's attention. *Scientific Reports, 12*(1), 7645. https://doi.org/10.1038/s41598-022-11723-6
Gregory, R. L. (1977). Vision with isoluminant colour contrast: 1.A projection technique and observations. *Perception, 6*(1), 113–119. https://doi.org/10.1068/p060113
Grinberg, D. L., & Williams, D. R. (1985). Stereopsis with chromatic signals from the blue-sensitive mechanism. *Vision Research, 25*(4), 531–537. https://doi.org/10.1016/0042-6989(85)90156-7
Gross, C. G. (2008). Single neuron studies of inferior temporal cortex. *Neuropsychologia, 46*(3), 841–852. https://doi.org/10.1016/j.neuropsychologia.2007.11.009
Heeger, D. J. (1992). Normalization of cell responses in cat striate cortex. *Visual Neuroscience, 9*(2), 181–197. https://doi.org/10.1017/S0952523800009640
Hoffman, J. E., & Subramaniam, B. (1995). The role of visual attention in saccadic eye movements. *Perception & Psychophysics, 57*(6), 787–795. https://doi.org/10.3758/BF03206794
Hou, X., Harel, J., & Koch, C. (2012). Image Signature: Highlighting Sparse Salient Regions. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 34*(1), 194–201. https://doi.org/10.1109/TPAMI.2011.146
Hu, B., Kane-Jackson, R., & Niebur, E. (2016). A proto-object based saliency model in three-dimensional space. *Vision Research, 119*, 42–49. https://doi.org/10.1016/j.visres.2015.12.004
Huang, X., Shen, C., Boix, X., & Zhao, Q. (2015). SALICON: Reducing the Semantic Gap in Saliency Prediction by Adapting Deep Neural Networks. *The IEEE International Conference on Computer Vision (ICCV)*.
Huynh-Thu, Q., & Schiatti, L. (2011). Examination of 3D visual attention in stereoscopic video content. *Proc.SPIE, 7865*. https://doi.org/10.1117/12.872382
Iacono, M., D'Angelo, G., Glover, A., Tikhanoff, V., Niebur, E., & Bartolozzi, C. (2019). Proto-object based saliency for event-driven cameras. *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), 2019*, 805–812. https://doi.org/10.1109/IROS40897.2019.8967943
Itti, L., & Koch, C. (2000). A saliency-based search mechanism for overt and covert shifts of visual attention. *Vision Research, 40*(10–12), 1489–1506. https://doi.org/10.1016/S0042-6989(99)00163-7
Itti, L., Koch, C., & Niebur, E. (1998). A model of saliency-based visual attention for rapid scene analysis. *IEEE Transactions on Pattern Analysis & Machine Intelligence, 11*, 1254–1259.
Jansen, L., Onat, S., & König, P. (2009). Influence of disparity on fixation and saccades in free viewing of natural scenes. *Journal of Vision, 9*(1), 29. https://doi.org/10.1167/9.1.29
Jordan, J. R., Geisler, W. S., & Bovik, A. C. (1990). Color as a source of information in the stereo correspondence process. *Vision Research, 30*(12), 1955–1970. https://doi.org/10.1016/0042-6989(90)90015-D
Judd, T., Ehinger, K., Durand, F., & Torralba, A. (2009). Learning to predict where humans look. In *2009 IEEE 12th International Conference on Computer Vision* (pp. 2106–2113). https://doi.org/10.1109/ICCV.2009.5459462
Julesz, B. (1971). *Foundations of cyclopean perception* (U.). Chicago Press.
Julesz, B. (1989). AI And Early Vision - Part II. *Proc.SPIE, 1077*. 10.1117/12.952723.
Khaustova, D., Fournier, J., Wyckens, E., & le Meur, O. (2013). How visual attention is modified by disparities and textures changes? *Proc. SPIE - Human Vision and Electronic Imaging XVIII, 8651*, 276–290. https://doi.org/10.1117/12.2003587
Koch, C., & Ullman, S. (1985). Shifts in selective visual attention: Towards the underlying neural circuitry. *Human Neurobiology, 4*(4), 219–227.
Kruthiventi, S. S. S., Ayush, K., & Babu, R. V. (2015). DeepFix: A Fully Convolutional Neural Network for predicting Human Eye Fixations. *IEEE Transactions on Image Processing, 26*(9), 4446–4456. https://doi.org/10.1007/s00138-014-0621-6
Kulikowski, J. J., Marčelja, S., & Bishop, P. O. (1982). Theory of spatial position and spatial frequency relations in the receptive fields of simple cells in the visual cortex. *Biological Cybernetics, 43*(3), 187–198. https://doi.org/10.1007/BF00319978
Kumano, H., Tanabe, S., & Fujita, I. (2008). Spatial Frequency Integration for Binocular Correspondence in Macaque Area V4. *Journal of Neurophysiology, 99*(1), 402–408. https://doi.org/10.1152/jn.00096.2007
Kümmerer, M., Theis, L., & Bethge, M. (2014). Deep Gaze I: Boosting Saliency Prediction with Feature Maps Trained on ImageNet. In *arXiv e-prints* (p. arXiv:1411.1045). https://ui.adsabs.harvard.edu/abs/2014arXiv1411.1045K.
Kümmerer, M., Wallis, T. S. A., & Bethge, M. (2016). DeepGaze II: Reading fixations from deep features trained on object recognition. In *eprint arXiv:1610.01563* (p. arXiv:1610.01563). https://ui.adsabs.harvard.edu/abs/2016arXiv161001563K.
Landman, R., Sharma, J., Sur, M., & Desimone, R. (2014). Effect of distracting faces on visual selective attention in the monkey. In *Proceedings of the National Academy of Sciences*. https://doi.org/10.1073/pnas.1420167111
Lang, C., Nguyen, T. V., Katti, H., Yadati, K., Kankanhalli, M., & Yan, S. (2012). Depth Matters: Influence of Depth Cues on Visual Saliency. *European Conference on Computer Vision, 2012*, 101–115.
Li, Z. (2002). A saliency map in primary visual cortex. *Trends in Cognitive Sciences, 6*(1), 9–16. https://doi.org/10.1016/S1364-6613(00)01817-9
Livingstone, M. S., & Hubel, D. H. (1987). Psychophysical evidence for separate channels for the perception of form, color, movement, and depth. *The Journal of Neuroscience, 7*(11), 3416–3468. https://doi.org/10.1523/JNEUROSCI.07-11-03416.1987
Lu, C., & Fender, D. H. (1972). The interaction of color and luminance in stereoscopic vision. *Investigative Ophthalmology & Visual Science, 11*(6), 482–490.
Ma, C.-Y., & Hang, H.-M. (2015). Learning-based saliency model with depth information. *Journal of Vision, 15*(6), 19. https://doi.org/10.1167/15.6.19
Mancinelli, E., Niebur, E., & Etienne-Cummings, R. (2018). Computational stereo-vision model of proto-object based saliency in three-dimensional space. *IEEE Biomedical Circuits and Systems Conference (BioCAS), 2018*, 1–4. https://doi.org/10.1109/BIOCAS.2018.8584679
Marr, D. (1982). *Vision: A computational investigation into the human representation and processing of visual information*. MIT Press.
Marr, D., & Poggio, T. (1979). A computational theory of human stereo vision. *Proceedings of the Royal Society of London. Series B. Biological Sciences, 204*(1156), 301–328. https://doi.org/10.1098/rspb.1979.0029
Molin, J. L., Etienne-Cummings, R., & Niebur, E. (2015). How is motion integrated into a proto-object based visual saliency model? *2015 49th Annual Conference on Information Sciences and Systems, CISS 2015*. 10.1109/CISS.2015.7086902.
Molin, J. L., Thakur, C. S., Niebur, E., & Etienne-Cummings, R. (2021). A Neuromorphic Proto-Object Based Dynamic Visual Saliency Model With a Hybrid FPGA Implementation. *IEEE Transactions on Biomedical Circuits and Systems, 15*(3), 580–594. https://doi.org/10.1109/TBCAS.2021.3089622
Moore, T., & Fallah, M. (2001). Control of eye movements and spatial attention. *Proceedings of the National Academy of Sciences, 98*(3), 1273 LP – 1276. 10.1073/pnas.98.3.1273.
Nakayama, K., He, Z. J., & Shimojo, S. (1995). Visual surface representation. In S. M. Kosslyn, & D. N. Osherson (Eds.), *An Invitation to Cognitive Science: Visual Cognition* ((2nd ed.,, pp. 1–70). The MIT Press.

Nassi, J. J., & Callaway, E. M. (2009). Parallel processing strategies of the primate visual system. *Nature Reviews. Neuroscience, 10*(5), 360–372. https://doi.org/10.1038/nrn2619

Niebur, E., & Koch, C. (1996). Control of Selective Visual Attention: Modeling the "Where" Pathway. In D. S. Touretzky, M. C. Mozer, & M. E. Hasselmo (Eds.), *Advances in Neural Information Processing Systems 8* (pp. 802–808). MIT Press.

Nityananda, V., & Read, J. C. A. (2017). Stereopsis in animals: Evolution, function and mechanisms. *The Journal of Experimental Biology, 220*(14), 2502–2512. https://doi.org/10.1242/jeb.143883

Nowak, L. G., Munk, M. H. J., Girard, P., & Bullier, J. (1995). Visual latencies in areas V1 and V2 of the macaque monkey. *Visual Neuroscience, 12*(2), 371–384. https://doi.org/10.1017/S095252380000804X

Nuthmann, A., & Henderson, J. M. (2010). Object-based attentional selection in scene viewing. *Journal of Vision, 10*(8), 20. https://doi.org/10.1167/10.8.20

Ohzawa, I. (1998). Mechanisms of stereoscopic vision: The disparity energy model. *Current Opinion in Neurobiology, 8*(4), 509–515. https://doi.org/10.1016/S0959-4388(98)80039-1

Ohzawa, I., DeAngelis, G. C., & Freeman, R. D. (1990). Stereoscopic depth discrimination in the visual cortex: Neurons ideally suited as disparity detectors. *Science, 249*(4972), 1037–1041. https://doi.org/10.1126/science.2396096

Ohzawa, I., Deangelis, G. C., & Freeman, R. D. (1997). Encoding of Binocular Disparity by Complex Cells in the Cat's Visual Cortex. *Journal of Neurophysiology, 77*(6), 2879–2909. https://doi.org/10.1152/jn.1997.77.6.2879

Parkhurst, D., Law, K., & Niebur, E. (2002). Modeling the role of salience in the allocation of overt visual attention. *Vision Research, 42*(1), 107–123. https://doi.org/10.1016/S0042-6989(01)00250-4

Poggio, G. F. (1995). Mechanisms of Stereopsis in Monkey Visual Cortex. *Cerebral Cortex, 5*(3), 193–204. https://doi.org/10.1093/cercor/5.3.193

Poggio, G. F., Motter, B. C., Squatrito, S., & Trotter, Y. (1985). Responses of neurons in visual cortex (V1 and V2) of the alert macaque to dynamic random-dot stereograms. *Vision Research, 25*(3), 397–406. https://doi.org/10.1016/0042-6989(85)90065-3

Poggio, G. F., & Poggio, T. (1984). The Analysis of Stereopsis. *Annual Review of Neuroscience, 7*(1), 379–412. https://doi.org/10.1146/annurev.ne.07.030184.002115

Posner, M. I. (1980). Orienting of Attention. *Quarterly Journal of Experimental Psychology, 32*(1), 3–25. https://doi.org/10.1080/00335558008248231

Preston, T. J., Li, S., Kourtzi, Z., & Welchman, A. E. (2008). Multivoxel Pattern Selectivity for Perceptually Relevant Binocular Disparities in the Human Brain. *The Journal of Neuroscience, 28*(44), 11315–11327. https://doi.org/10.1523/JNEUROSCI.2728-08.2008

Qian, N., & Zhu, Y. (1997). Physiological computation of binocular disparity. *Vision Research, 37*(13), 1811–1827. https://doi.org/10.1016/S0042-6989(96)00331-8

Qin, D., Takamatsu, M., & Nakashima, Y. (2004). Measurement for the Panum's Fusional Area in Retinal Fovea Using a Three-Dimention Display Device. *Journal of Light & Visual Environment, 28*(3), 126–131. https://doi.org/10.2150/jlve.28.126

Qiu, F. T., Sugihara, T., & von der Heydt, R. (2007). Figure-ground mechanisms provide structure for selective attention. *Nature Neuroscience, 10*(11), 1492–1499. https://doi.org/10.1038/nn1989

Qiu, F. T., & von der Heydt, R. (2005). Figure and Ground in the Visual Cortex: V2 Combines Stereoscopic Cues with Gestalt Rules. *Neuron, 47*(1), 155–166. https://doi.org/10.1016/j.neuron.2005.05.028

Ramenahalli, S., Mendat, D. R., Dura-Bernal, S., Culurciello, E., Niebur, E., & Andreou, A. (2013). Audio-visual saliency map: Overview, basic models and hardware implementation. *2013 47th Annual Conference on Information Sciences and Systems (CISS)*, 1–6. 10.1109/CISS.2013.6552285.

Read, J. C. A., & Cumming, B. G. (2007). Sensors for impossible stimuli may solve the stereo correspondence problem. *Nature Neuroscience, 10*(10), 1322–1328. https://doi.org/10.1038/nn1951

Rensink, R. A. (2000). The Dynamic Representation of Scenes. *Visual Cognition, 7*(1–3), 17–42. https://doi.org/10.1080/135062800394667

Richards, W. (1971). Anomalous Stereoscopic Depth Perception. *Journal of the Optical Society of America, 61*(3), 410–414. https://doi.org/10.1364/JOSA.61.000410

Russell, A. F., Mihalaş, S., von der Heydt, R., Niebur, E., & Etienne-Cummings, R. (2014). A model of proto-object based saliency. *Vision Research, 94*, 1–15. https://doi.org/10.1016/j.visres.2013.10.005

Scharff, L. V., & Geisler, W. S. (1992). Stereopsis at isoluminance in the absence of chromatic aberrations. *Journal of the Optical Society of America A. Optics and Image Science, 9*(6), 868–876. https://doi.org/10.1364/JOSAA.9.000868

Schmolesky, M. T., Wang, Y., Hanes, D. P., Thompson, K. G., Leutgeb, S., Schall, J. D., & Leventhal, A. G. (1998). Signal Timing Across the Macaque Visual System. *Journal of Neurophysiology, 79*(6), 3272–3278. https://doi.org/10.1152/jn.1998.79.6.3272

Scholl, B. J. (2001). Objects and attention: The state of the art. *Cognition, 80*(1), 1–46. https://doi.org/10.1016/S0010-0277(00)00152-9

Simmons, D. R., & Kingdom, F. A. A. (1997). On the independence of chromatic and achromatic stereopsis mechanisms. *Vision Research, 37*(10), 1271–1280. https://doi.org/10.1016/S0042-6989(96)00273-8

Stoll, J., Thrun, M., Nuthmann, A., & Einhäuser, W. (2015). Overt attention in natural scenes: Objects dominate features. *Vision Research, 107*, 36–48. https://doi.org/10.1016/j.visres.2014.11.006

Tanabe, S., Umeda, K., & Fujita, I. (2004). Rejection of False Matches for Binocular Correspondence in Macaque Visual Cortical Area V4. *The Journal of Neuroscience, 24*(37), 8170–8180. https://doi.org/10.1523/JNEUROSCI.5292-03.2004

Tsao, D. Y., Freiwald, W. A., Knutsen, T. A., Mandeville, J. B., & Tootell, R. B. H. (2003). Faces and objects in macaque cerebral cortex. *Nature Neuroscience, 6*(9), 989–995. https://doi.org/10.1038/nn1111

Tyler, C. W., & Cavanagh, P. (1991). Purely chromatic perception of motion in depth: Two eyes as sensitive as one. *Perception & Psychophysics, 49*(1), 53–61. https://doi.org/10.3758/BF03211616

Uejima, T., Niebur, E., & Etienne-Cummings, R. (2020). Proto-Object Based Saliency Model With Texture Detection Channel. *Frontiers in Computational Neuroscience, 14*, 84. https://www.frontiersin.org/article/10.3389/fncom.2020.541581.

van der Stigchel, S., & Theeuwes, J. (2007). The relationship between covert and overt attention in endogenous cuing. *Perception & Psychophysics, 69*(5), 719–731. https://doi.org/10.3758/BF03193774

Vig, E., Dorr, M., & Cox, D. (2014). Large-Scale Optimization of Hierarchical Features for Saliency Prediction in Natural Images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 2798–2805).

Viola, P., & Jones, M. (2001). Rapid object detection using a boosted cascade of simple features. *Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR 2001, 1*, I–I. 10.1109/CVPR.2001.990517.

von der Heydt, R. (2015). Figure–ground organization and the emergence of proto-objects in the visual cortex. *Frontiers in Psychology, 6*, 1695. https://www.frontiersin.org/article/10.3389/fpsyg.2015.01695.

Wang, J., Da Silva, M. P., Callet, P. L., & Ricordel, V. (2013). Computational Model of Stereoscopic 3D Visual Saliency. *IEEE Transactions on Image Processing, 22*(6), 2151–2165. https://doi.org/10.1109/TIP.2013.2246176

Welchman, A. E. (2016). The Human Brain in Depth: How We See in 3D. *Annual Review of Vision Science, 2*(1), 345–376. https://doi.org/10.1146/annurev-vision-111815-114605

Williford, J. R., & von der Heydt, R. (2016). Figure-Ground Organization in Visual Cortex for Natural Scenes. ENEURO.0127-16.2016 *ENeuro, 3*(6). https://doi.org/10.1523/ENEURO.0127-16.2016.

Xu, T., Li, M., Chen, K., Wang, L., & Yan, H.-M. (2016). Aspect Ratio of the Receptive Field Makes a Major Contribution to the Bandwidth of Orientation Selectivity in Cat V1. *Advances in Cognitive Neurodynamics*, 133–142.

Zhang, J., & Sclaroff, S. (2013). Saliency Detection: A Boolean Map Approach. *IEEE International Conference on Computer Vision, 2013*, 153–160. https://doi.org/10.1109/ICCV.2013.26

Zhang, J., & Sclaroff, S. (2016). Exploiting Surroundedness for Saliency Detection: A Boolean Map Approach. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 38*(5), 889–902. https://doi.org/10.1109/TPAMI.2015.2473844

Zhaoping, L. (2002). Pre–attentive segmentation and correspondence in stereo. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences, 357*(1428), 1877–1883. https://doi.org/10.1098/rstb.2002.1158

Zhaoping, L. (2008). Attention capture by eye of origin singletons even without awareness—A hallmark of a bottom-up saliency map in the primary visual cortex. *Journal of Vision, 8*(5), 1. https://doi.org/10.1167/8.5.1

Zhaoping, L. (2012). Gaze capture by eye-of-origin singletons: Interdependence with awareness. *Journal of Vision, 12*(2), 17. https://doi.org/10.1167/12.2.17

Zhaoping, L. (2018). *Ocularity Feature Contrast Attracts Attention Exogenously. In Vision* (Vol. 2, Issue 1). https://doi.org/10.3390/vision2010012

Zhaoping, L., Guyader, N., & Lewis, A. (2009). Relative contributions of 2D and 3D cues in a texture segmentation task, implications for the roles of striate and extrastriate cortex in attentional selection. *Journal of Vision, 9*(11), 20. https://doi.org/10.1167/9.11.20

Zhou, H., Friedman, H. S., & von der Heydt, R. (2000). Coding of border ownership in monkey visual cortex. *The Journal of Neuroscience, 20*(17), 6594–6611. https://doi.org/10.1523/JNEUROSCI.2797-12.2013