# Grouping mechanisms for object-based vision and attention

by

Brian H. Hu

A dissertation submitted to The Johns Hopkins University in conformity with the requirements for the degree of Doctor of Philosophy.

Baltimore, Maryland

August, 2017

# Abstract

The visual brain faces the difficult task of reconstructing a three-dimensional (3D) world from two-dimensional (2D) images projected onto the two retinae. In doing so, visual information is organized in terms of objects in 3D space, and this organization is the basis for visual perception. In complex visual scenes, both the foreground and the background are rich in features of different types. The brain must find a way to group together the features that belong to objects on the foreground and distinguish them from features in the background.

The goal of this thesis is to understand how the neural circuits in primate cortex accomplish this task using grouping mechanisms for object-based vision and attention. Through computational modeling, I show that grouping mechanisms are fundamental for linking early feature representations to tentative perceptual objects known as proto-objects. Previous models on the neural coding of border ownership have identified a plausible network architecture for proto-object based perceptual organization. I extend these models to explain how the same grouping framework can be used to perform contour integration, border-ownership assignment, grouping of 3D

ABSTRACT

surfaces, and 3D visual saliency. My models offer several falsifiable predictions which can be tested in future experiments. My models also clarify how top-down attention interfaces with the neural circuits responsible for grouping together the features of an object. Overall, the models developed address the important question of how visual features are grouped into 2D and 3D object representations.

Primary Reader: Ernst Niebur

Secondary Reader: Rüdiger von der Heydt

# Acknowledgments

I had the privilege of sailing with Ernst to the neuroscience retreat one year. As I was sitting on the boat, I thought about how the trip was really a metaphor for my PhD. I have had opportunities to steer the boat (sometimes in the wrong direction!) and I have gone through both good and bad weather. Through it all, I am thankful to have had Ernst as my captain who guided me to the finish.

I would like to thank my family for their continued love and support. I thank my wife, Mingming, and my daughter, Katherine, for bearing with me in these final months. I thank my parents, Benjamin and Jennifer, for all the sacrifices they made for me and my siblings. I thank my brother Blair for his moral support and my sister Joy for being an inspiration to me.

I would also like to thank my thesis committee members, Rüdiger and Kechen, for their help and guidance during my PhD. Finally, I would like to thank my friend Matthew, for all the lunches and discussions that we shared together, and my labmates Danny and Grant, who made each day in lab an interesting one.

**Soli Deo Gloria**

# Dedication

This thesis is dedicated to Joy, who first got me interested in the study of vision. Your perseverance in the face of adversity has been my inspiration.

# Contents

# List of Tables

# List of Figures

# Chapter 1

# Introduction

## 1.1 Segmentation and figure-ground organization

The task of partitioning an image into regions bounded by contours (segmentation) and the task of assigning border-ownership of these contours to either the foreground or the background (figure-ground organization) are important first steps in achieving image understanding. Gestalt psychologists were the first to recognize the importance of the whole in influencing perception of the parts, and with this observation, laid out several principles for figure-ground organization (Koffka, 1935; Wertheimer, 1923). For example, the rule of good continuation states that well-aligned contour elements should be grouped together. This is closely related to the concept of a "local association field," where collinear contour elements excite each other and non-collinear elements inhibit each other (Ullman et al., 1992; Field et al., 1993). Results from neuroanatomy lend support to these ideas, as the lateral connections within V1 predominantly link similar-orientation cortical columns. However, our understanding of the neural mechanisms of these processes remains surprisingly limited.

The brain must keep track of which regions and contours belong to which objects. This is known as the binding problem, as it is not clear how the features of an object are bound together (Treisman, 1996). One class of models relies on the fast temporal coding structures of spike trains (Singer, 1999), but experimental evidence is controversial (Thiele and Stoner, 2003; Roelfsema et al., 2004; Dong et al., 2008). Another solution involves differential neural activity, where the neurons responding to the features of an object show increased firing compared with neurons responding to the background. This response enhancement is known as figure-ground modulation (FGM), and was first observed in primary visual cortex (V1) for texture-defined figures (Lamme, 1995). Similar results have been found using other tasks and techniques, including more recent voltage-sensitive dye imaging of populations of neurons during a contour grouping task (Gilad et al., 2013). However, this solution only works if there is a single object in the foreground, as multiple objects each labeled with higher neural activity could be interpreted as parts of a single object. Furthermore, each neuron's firing rate is inherently ambiguous, as higher activity could be due to labeling with FGM or because the neuron's preferred feature falls within its receptive field. As a result, the binding problem cannot be solved with models that only represent object information in terms of enhanced neural activity in early visual areas (Niebur, 2000). This strongly points to neural circuits that employ populations of neurons which explicitly represent (*i.e.* in their firing rate) the organization of the visual scene in terms of perceptual objects.

## 1.2 The role of cortical feedback

Early computational models (Stemmler et al., 1995b) and experimental studies (Simonotto et al., 1997; Polat et al., 1998; Chatterjee et al., 2011; Xie et al., 2014) put forth the view that horizontal connections are essentially static, or varying over time scales given by ontogenetic development or neuronal plasticity. Such structures could thus implement overall statistics of natural scenes (like circular structures, e.g. Sigman et al., 2001) but they could not flexibly represent the myriad of instantaneously present and constantly changing visual shapes observed during perception of dynamic scenes. This view has been considerably enlarged over the last decade or so, and what is emerging is a view in which the lateral connections are in place but can be actively and rapidly modulated by top-down connectivity from higher areas.

The degree of collinear facilitation observed in V1 is strongly context-dependent, and can change with the behavioral task (Li et al., 2004, 2006) as well as perceptual learning (Li et al., 2008; Yan et al., 2014). As a result, feedback connections from higher areas may play an important role in shaping the responses of neurons in early visual areas. In fact, simultaneous neural recordings from areas V1 and V4 during two different figure-ground segregation tasks show that V4 is intimately involved in the FGM process (Poort et al., 2012; Chen et al., 2014). In these studies, the FGM signal appears first in V4 and is then fed back to V1, with a delay representing recurrent processing. Additional studies of curve-tracing (Roelfsema et al., 1998) and border-ownership (Zhou et al., 2000; Qiu et al., 2007; Zhang and von der Heydt, 2010) further demonstrate that feedback mechanisms are necessary for explaining FGM in the presence of multiple objects. However, essential questions still remain about the nature of the interactions between and within different cortical areas. How is early-level feature information about an object combined with global context information about the object in a synergistic way in order to generate FGM?

## 1.3 The role of attention

Behavioral studies have shown that attention can be directed to objects (Egly et al., 1994) and electrophysiological results demonstrate that attention can act as a top-down signal which influences FGM (Qiu et al., 2007; Poort et al., 2012). In an ambiguous figure-ground display, attending to one region increases the probability that this region is perceived as figure (Driver and Baylis, 1996; Vecera et al., 2004). Spatial attention, which has been extensively studied, acts like a ”spotlight” that enhances neural responses within the focus of attention and suppresses responses outside (Motter, 1993). Attention can also operate in a feature-based or object-based manner. Feature-based attention acts broadly across the visual scene and increases the responses of all components that share similar feature attributes (e.g. color, orientation, or direction of movement) with the attended component (Treue, 1999). Object-based attention highlights all the parts of an object, also encompassing all the features that belong to the object (Roelfsema et al., 1998; Schoenfeld et al., 2014). Attention has been found to modulate border-ownership in an object-based manner (Qiu et al., 2007). Border-ownership is a property of many neurons in V2 which encodes the side to which an object border belongs relative to their receptive field (Zhou et al., 2000). To explain these results, Craft et al. (2007) proposed a model in which populations of grouping neurons explicitly represent (in their firing rates) the perceptual organization of the visual scene. Grouping neurons are reciprocally connected to border ownership selective (BOS) neurons through feedforward and feedback connections. Attention broadly targets grouping neurons, which can then modulate the activity of BOS neurons through feedback (Mihalas et al., 2011). A feedforward version of this model has been applied to natural images, where it outperforms other models in predicting the location of eye fixations (Russell et al., 2014).

## 1.4 Grouping of 3D surfaces

Grouping mechanisms are important not only for piecing together object contours, but also for providing a structure for selectively attending to groups of objects (Treisman and Gelade, 1980). Supported by extensive psychophysical data, Nakayama, He, and Shimojo (1995) proposed that surface representations play a key role in intermediate-level vision. For example, by selectively attending to a surface in 3D space, subjects can perform efficient search for a conjunction target (Nakayama and Silverman, 1986). In a separate cueing experiment, attention was shown to spread automatically across surfaces (He and Nakayama, 1995). These abilities indicate powerful mechanisms for grouping objects into surfaces in 3D space, and suggest that structuring the world in terms of surfaces might be an ecologically important function (*e.g.* for locomotion along the ground plane, reaching for objects along a table top *etc.*). The use of basis functions provides a suitable theoretical framework for understanding how multiple surfaces can be efficiently represented in the same population of neurons. Modeling the grouping of 3D surfaces will provide insight into the neural circuits that represent surfaces, filling a critical gap in our understanding of intermediate-level vision.

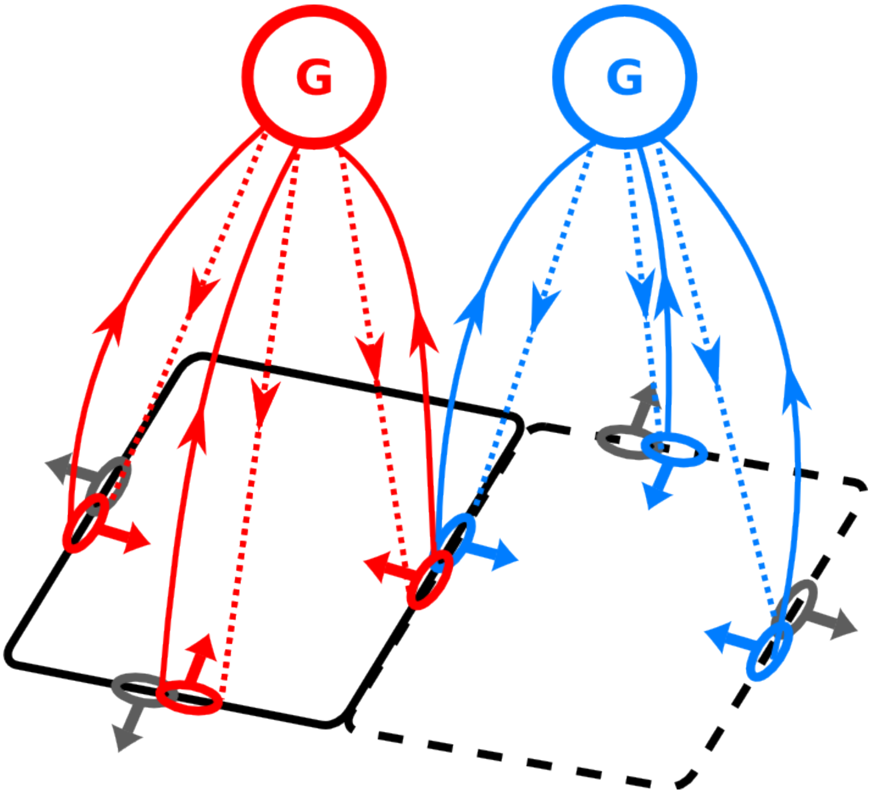

**Figure 1.1:** Overview of the grouping model

## 1.5 Overview of the grouping model

Several models have been proposed (Zhaoping, 2005; Sakai and Nishimura, 2006; Craft et al., 2007; Layton et al., 2012) to describe how a neuron's border ownership selectivity can be modulated by visual input far away from its classical receptive field (RF). In the grouping model (Craft et al., 2007), BOS neurons participate in neural circuits that define perceptual objects early-on in processing (Figure 1.1). An object border activates a pair of orientation-selective BOS neurons whose RFs are shown as ellipses. The arrows on the RFs point towards the preferred direction of the neuron, indicating where the object is relative to the neuron's RF. BOS neurons activated by the solid-line square (red ellipses) excite the appropriate grouping neuron (red G in circle) through feedforward connections (red solid lines). The grouping neuron in turn enhances the activity of the same BOS neurons through feedback connections (red dotted lines). This type of facilitatory feedback may be mediated by NMDAR channels, which allow gating of sensory input by top-down signals (Palmer et al., 2014). Neurons consistent with other objects (e.g. the dashed-line square) project to other grouping neurons (blue G in circle in this case). Presence of the solid-line square increases the firing rate of the red grouping neuron over that of the blue (and other) grouping neurons since the latter receive less feedforward input. As a consequence, the red grouping neuron provides more feedback to the red BOS neurons than the blue and gray BOS neurons would receive from their respective grouping neurons. Likewise, presence of the dashed-line square increases the firing rates of all blue neurons over those of gray and red neurons. Thus, an object border is represented by two BOS neurons whose relative activity codes for the side of ownership. The relative difference in firing rate is also known as the BOS signal (Zhou et al., 2000).

The BOS signal appears ˜25 ms after the visual response to an oriented edge, and the delay is essentially independent of object size (Zhou et al., 2000). This constant delay is consistent with a model in which grouping neurons of different sizes integrate local edge signals, and by feedback enhance the same edge signals (Craft et al., 2007). Attention enhances a BOS neuron's response when an object is on the neuron's preferred side, but has a suppressive effect if the object is on the non-preferred side (Qiu et al., 2007). This asymmetry is consistent with a model in which top-down attention targets grouping neurons, which then modulate the activity of BOS neurons through feedback (Mihalas et al., 2011). Additional support for the grouping model comes from observations of short-term memory of BOS signals (O'Herron and von der Heydt, 2009) and remapping of BOS signals across saccades and object movements (O'Herron and von der Heydt, 2013). These findings are difficult to explain with models that only represent object information in terms of neural activity in early visual areas. I believe that this strongly points to neural circuits that employ populations of neurons which explicitly represent (*i.e.* in their firing rate) the organization of the visual scene.

## 1.6 Summary of thesis

The work presented in this thesis deepens and extends our understanding of the neural mechanisms of FGM. In Chapter 1, we introduce background information needed to understand the physiology and previous modeling experiments. In Chapter 2, we propose a quantitative neural model of grouping constrained by physiological data. We validate the model by reproducing several experimental results related to contour integration and border-ownership assignment. In Chapter 3, we extend this model to natural scenes, and our model results are quantitatively compared with both experimental results and human-annotated figure-ground labels (Berkeley Segmentation Dataset). Beginning with Chapter 4, we shift our focus to the representation of 3D information in the visual system. First, we show that a grouping model can reproduce results from a set of psychophysical experiments where attention had to be directed to surfaces. We then show that 3D surfaces can be represented by a feedforward, linear combination of basis functions whose response properties are similar to those of disparity-selective neurons commonly found in early visual cortex. In Chapter 5, we propose a model of 3D visual saliency and show that the added depth information improves saliency prediction.

# Chapter 4

# 3D proto-object based saliency

## 4.1 Introduction

The brain receives large amounts of visual information that it must make sense of in real-time. Processing the entire visual field with the same level of detail present at the fovea would be an exceedingly complex and costly task requiring much greater computational resources than are available to the brain (Tsotsos, 1990). As a result, primates select only the most relevant information and discard the rest, a process known as selective visual attention. Many models of visual attention are constructed with a bottom-up architecture and rely on local contrast in low-level features such as intensity, color, orientation, or motion. Biologically-plausible center-surround differences across different feature channels of an input image can be used to compute a "saliency map" whose maxima indicate where selective attention is deployed (Koch and Ullman, 1985; Niebur and Koch, 1996; Itti et al., 1998).

However, there is both psychophysical (Einhäuser et al., 2008) and neurophysiological (Zhou et al., 2000; Qiu et al., 2007) evidence that attention relies not only on these simple image features, but also on the perceptual organization of the visual scene into tentative objects, or proto-objects (Rensink, 2000). Biologically-inspired models of proto-object based saliency have been proposed that take into account these recent findings (Craft et al., 2007; Mihalas et al., 2011; Russell et al., 2014). These models include border-ownership selective cells (referred to as border-ownership cells in the following) and grouping cells, which interact to achieve figure-ground segmentation of the image into proto-objects (figures) and the background (ground).

Border-ownership cells have been found in primate visual cortex, with the majority of neurons in area V2 having this property. These cells signal in their neural activity the one-sided assignment of an object border to the region perceived as figure (Zhou et al., 2000). Border-ownership cells are also modulated by attentional influences (Qiu et al., 2007). Grouping cells integrate global context information about proto-objects in the scene according to Gestalt principles such as closure, continuity, convexity *etc.* Importantly, grouping cells act at intermediate stages of vision and do not require higher-level information about object identity, semantic knowledge *etc.* They send feedback to border-ownership cells *via* fast white matter projections, which bias the activity of border-ownership cells to reflect the correct figure-ground segmentation of proto-objects. In this framework, visual saliency is a function of grouping cell activity, which represents the size and location of proto-objects within the image.

Border-ownership cells have been shown to respond to figure edges defined by a variety of image features, *e.g.* luminance edges, color edges *etc.* When no monocular edge information is present (*i.e.* when the figures are defined by random dot stereograms using only binocular disparity), border-ownership selectivity is also imparted by stereoscopic edges (Qiu and von der Heydt, 2005). Critically, their response to these different figural cues is typically the same in the two-dimensional (2D) and three-dimensional (3D) cases – the preferred side-of-figure of border-ownership cells is consistent for all cues that define the figure. The activity of border-ownership cells thus provides an interpretation of the visual scene in terms of depth-ordered surfaces that correspond to objects in 3D space. Despite these experimental observations, current models of border ownership do not explicitly use depth information and do not address how traditional 2D Gestalt cues interact with depth cues during the figure-ground segmentation process. An exception is a study by Mishra et al. (2012) who used computer vision methods to compute border ownership from low-level depth information and then performed object segmentation in natural images.

Even though in recent years stereoscopic 3D content has become increasingly prevalent, *e.g.* in the viewing of entertainment programs in cinemas and homes, little is known about how visual attention is deployed within 3D environments. It is thus important to understand how humans allocate their attention when viewing natural images and videos in 3D (Le Callet and Niebur, 2013). Binocular disparity cues, which can be used to generate strong depth percepts, have been shown to alter different aspects of eye movements when participants viewed 3D images (Jansen et al., 2009) and videos (Huynh-Thu and Schiatti, 2011). Only recently have 3D eye tracking datasets been made available which can be used to compare human eye movements with predictions of attentional models. The availability of these datasets and the recent explosion in new 3D content makes it possible to design computational models of 3D saliency and evaluate their performance objectively.

## 4.2 Related Work

### 4.2.1 Models of 3D visual attention

Compared to the number of models that have been proposed for 2D visual saliency, relatively few attempts have been made to study how visual attention is deployed within 3D environments. Existing models of 3D visual attention often compute a 2D saliency map which is then combined with the depth information to produce a new saliency map. These models fall into three categories (Wang et al., 2013) based on how the depth information is used: stereovision models, depth-weighting models, and depth-saliency models. For a comprehensive review of 3D visual attention models, see Wang et al. (2013); Ma and Hang (2015).

While the depth-weighting and depth-saliency models assume that a depth map has been computed, without specifying how, stereovision models explicitly implement the computation of depth information from the left and right views of the scene, thus replicating the human visual system's stereoscopic perception. An example of this is a study by Bruce and Tsotsos (2005), which extended a 2D selective tuning model of attention to also incorporate binocular information. However, no quantitative assessment of this model was performed.

Depth-weighting models use a base 2D saliency model (computed using one of the existing methods) and then multiplicatively weight the resulting saliency map with the depth information. Regions that are closer to the observer obtain higher weights, corresponding to greater combined saliency. In a model developed by Lang et al. (2012), novel depth priors are learned from a training portion of the data, and these are then combined with the output of a 2D saliency model either using pixel-wise addition or multiplication. With these depth priors, the authors find an increase of performance by 6-7% on their dataset compared to the base 2D model without depth information.

Depth-saliency models come in two flavors. In one, both a depth saliency map, obtained from depth alone, and a more traditional saliency map, obtained from 2D information alone, are computed. The two maps are then linearly combined to generate the final saliency map. Wang et al. (2013) determine depth saliency in a separate experiment involving synthetic stereoscopic stimuli, which allows them to reduce the influence of monocular depth cues, as well as control for the depth of objects and the background. With their experimental results, they propose a probabilistic model of depth saliency, where the probability of a point being fixated in 3D space is related to the magnitude of center-surround differences in depth contrast. Linearly combining these two saliency maps in a 1:1 ratio (50% weight each for 2D features and depth information) results in better performance on their dataset. In the second type of depth-saliency models, depth information is treated as an additional feature channel, on the same footing as intensity, color, orientation *etc.* The final saliency map is then a function of depth as well as of these other features (Ouerhani and Hügli, 2000; Jost et al., 2004; Hügli et al., 2005).

Our approach falls in the latter class of depth-saliency models, where all image features, including depth, interact through linear combination resulting in the final saliency map. Our model is completely integrated – depth information is treated as another cue which interacts with 2D Gestalt cues to influence figure-ground assignment of proto-objects within the scene. This agrees with anatomical and neurophysiological data that show that disparity selective cells, which are important for encoding stereoscopic depth information, are found in the same early cortical areas as neurons representing other features used in typical saliency models, like color and orientation (Hubel and Wiesel, 1962; Poggio et al., 1988). Different from previous models (Ouerhani and Hügli, 2000; Jost et al., 2004; Hügli et al., 2005), our model is not only based on basic image features (like color, intensity *etc.*) but it includes elements of perceptual organization, in particular proto-objects. The model is an extension of a previously described 2D model (Russell et al., 2014) and is constructed by including depth information as an additional feature. All features are used to determine proto-object based saliency. We tested our model on the three 3D eye tracking datasets listed in the next section, and we compared results with and without the added depth information.

### 4.2.2 3D eyetracking datasets

A common method for evaluating the quality of computational models of visual attention is to compare their performance in the prediction of human eye movements. Since its introduction by Parkhurst et al. (2002), this method has been used in a large number of studies, both for static and dynamic scenes (video) and both for human and non-human primates (for a recent review see Borji and Itti, 2013). Nearly all of this work has been limited, however, to 2D scenes.

In order to evaluate 3D attention models, eye tracking data on a variety of visual scenes have to be collected. We use datasets of natural images consisting of color images and associated depth maps along with human fixations for each image. Predictions of saliency maps for eye movements can then be compared to the ground truth fixation data using various metrics. Below, three such publicly available datasets are described. Figure 4.1 shows one example of the data available from each of them.

The NUS-3D dataset (Lang et al., 2012) contains 600 RGB and depth image pairs, each with a resolution of 640 × 480 pixels. The images show various scenes around the National University of Singapore (NUS) campus and were collected with a

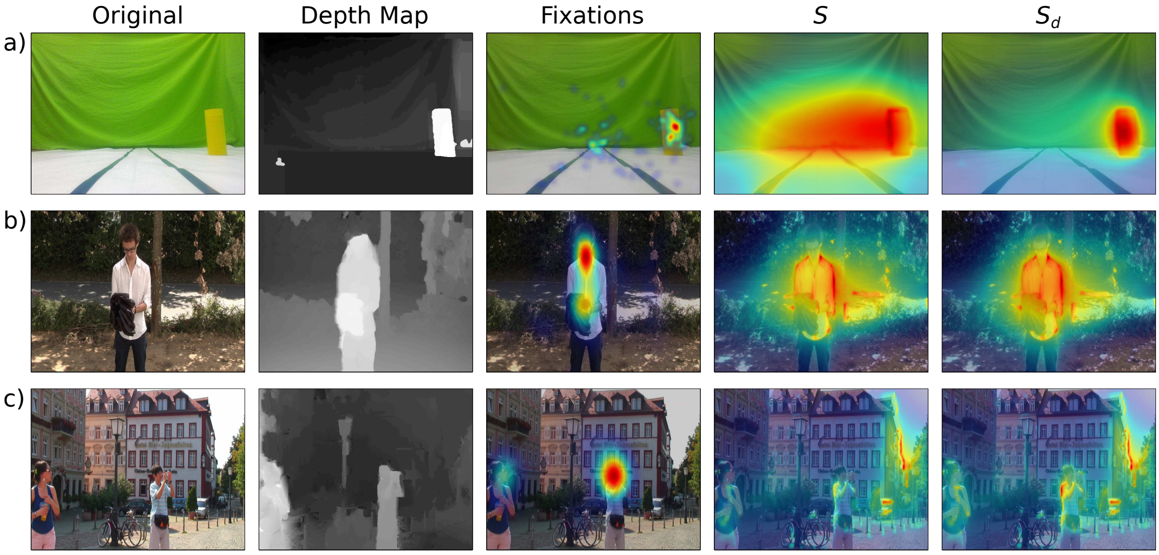

**Figure 4.1:** Examples of data used and results obtained. Columns (left to right) show one example each of the original image with its corresponding depth map, fixation map, our saliency model without ($S$) and with ($S_d$) depth information for the three 3D eye tracking datasets: a) NUS-3D. b) Gaze-3D. c) NCTU-3D.

Microsoft Kinect camera, which is capable of recording both RGB and depth images. The Kinect depth sensor is affected by ambient lighting and has a depth range of only about 4 m, which restricts the types of scenes it can accurately capture. The images were presented to 80 participants and each participant's eye tracking data was captured in both 2D and 3D free-viewing experiments. The 3D stimuli were generated by virtual view synthesis (see Lang et al., 2012, for details) but the synthesized 3D images are not available to the public. Only the raw and smoothed depth maps from the Kinect as well as the fixation density maps for the 2D and 3D viewing conditions are available.

The Gaze-3D dataset (Wang et al., 2013) consists of 18 stereoscopic images, along with their associated disparity maps and perceived depth maps (perceived depth is computed from raw disparity by taking into account viewing distance and display properties; see Wang et al., 2013, for details). The ground truth disparity maps were calculated from separate left and right image views using an optical flow method (Werlberger et al., 2009). The images come from the Middlebury 2005/2006 stereo image dataset (Scharstein and Pal, 2007) and the IVC 3D image dataset (Urvoy et al., 2012). The dataset also contains raw eye tracking data for both the left and right eyes, as well as processed fixation density maps from a total of 35 participants. The images are high resolution (1300 × 1080 pixels for the Middlebury subset and 1920 × 1080 pixels for the IVC subset) and have relatively accurate depth information. A limitation is the small number of images in the dataset.

The NCTU-3D dataset (Ma and Hang, 2015) consists of 475 2D images with a resolution of 1920 × 1080 pixels along with their corresponding depth maps and eye tracking data. The eye tracking data is in the form of fixation density maps and binary fixation maps. The images in the dataset were collected from randomly selected frames extracted from 11 different sequences of 3D videos from either Youtube (youtube.com) or 3dtv (3dtv.at). The depth maps were generated from left and right eye views using Depth Estimation Reference Software (DERS; version 5.0). A total of 16 participants freely viewed the videos in 3D.

## 4.3 Model

Our approach is based on the proto-object based saliency model proposed by Russell et al. (2014). In the model, grouping cells group visual features into proto-objects that are characterized by their locations and spatial scales. The large annular receptive fields of the grouping cells enforce the Gestalt principles of closure and convexity, which in turn biases the activity of proto-objects towards the center of objects. Proto-objects are then a means to organize the scene into separate figures as well as the background. The grouping mechanism operates on multiple feature channels and incorporates competition between proto-objects of similar size and feature type. The model explains the development of border-ownership findings in primate cortex (Craft et al., 2007; Zhou et al., 2000). Given that human eye movements tend to fall predominantly on objects (Einhäuser et al., 2008, but see Borji et al. (2013) for a different view), which are often closed and convex, the locations of proto-objects are also assumed to correlate with the salient points within the image. For a full description of the original model, we refer the reader to Russell et al. (2014).

We extend this model to include depth information. Since some images have border artifacts, all images are cropped to avoid spurious model responses at the borders. To achieve scale invariance, we create an image pyramid spanning 8-10 octaves (depending on the size of the image) by successively down-sampling the input image in steps of 2. We use a minimum cut-off image size of $3 \times 3$ pixels, such that pyramid levels that would reduce the image size below $3 \times 3$ pixels are not included

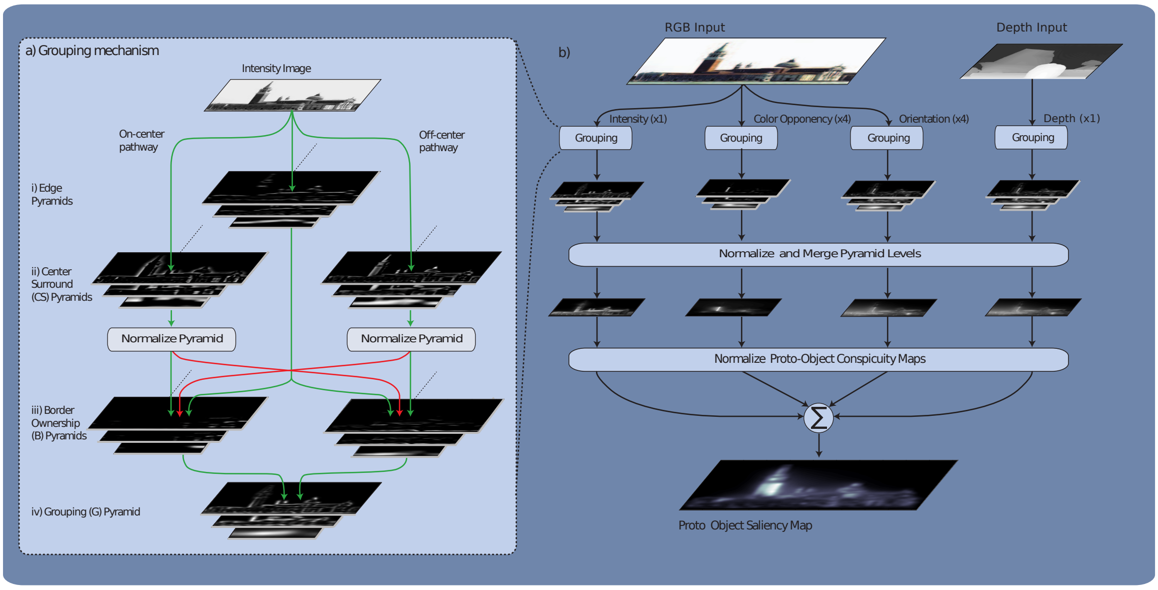

**Figure 4.2:** Proto-object saliency model with added depth information. The depth map is represented by the image at the top, far right, and the 2D image is to its left. Based on Figure 5 of Russell et al. (2014).

in our model. All operations are applied independently to each feature and at each level of the feature pyramids, except for when the scales and features are combined to obtain the final saliency map. Each layer of the network represents the neural activity of an array of neurons which tile the visual scene and which are propagated to other layers in the network through feedforward connections. The receptive fields of neurons are described by different correlation kernels, and the image input to each neuron in the model is calculated using the correlation operation. The model was implemented using MATLAB (Mathworks, Natick, MA). A model overview is shown in Figure 4.2.

In this chapter, we will refer to the saliency map generated by the original 2D model without depth information as $S$, and to the saliency model generated by our new model, which incorporates depth information as $S_d$. To compute $S$, the model accepts an input RGB image and decomposes it into different feature channels: one intensity channel, four color-opponency channels, and four orientation channels. Rather than providing "raw" stereoscopic disparity information in the form of different images to the two eyes (or retinae), we assume that the transformation from disparity information to depth has been performed at the input level to our model. There is well-known neuronal circuitry that transforms stereoscopic disparity into depth information (*e.g.* Poggio and Poggio, 1984) and the explicit representation of depth is more appropriate for the intermediate-vision conceptual level of our proto-object based model than the "raw" representation in terms of binocular disparity. Therefore, to compute $S_d$, we add an additional depth feature channel obtained from the input depth image. Within each feature channel, we perform edge filtering (using oriented Gabor filters at 4 orientations) to obtain the location of proto-object borders. In order to perform feedforward computation of figure-ground segregation, we employ a center-surround mechanism, similar to that used by Itti et al. (1998); such mechanisms have been observed at multiple stages in the brain, including retina, lateral geniculate nucleus, and cortex (*ibid*). This center-surround mechanism provides context information about proto-objects, and biases the activity of border-ownership cells with preferred directions that match the location of the figure.

For the 2D features used in our model, the center-surround mechanism is symmetric with respect to figure-ground contrast polarity (*e.g.* light figures on dark backgrounds or dark figures on light background result in the same net salience contribution). In contrast, for the depth channel we compute the center-surround differences in an asymmetrical manner, consistent with the kind of information provided by stereoscopic depth. While most feature differences across a contour are not predictive about which of its sides is the foreground[1], stereoscopic depth (disparity) provides nearly unambiguous information about which side of the border is closer to the observer. This side "owns" the object border when considered in a depth ordering sense and is part of the foreground. Physiological data show that the responses of border-ownership cells to disparity differences across a figure edge are in agreement with this observation (Zhou et al., 2000; Qiu and von der Heydt, 2005). Therefore, in our model near *vs.* far depth differences bias the activity of border-ownership cells such that the near side is more likely to be classified as the foreground object. Integrating depth information into the representation of an object by its contours is a critical difference between our model and other depth saliency models which directly combine depth feature information with 2D information, without taking into account perceptual grouping effects (*e.g.* Ouerhani and Hügli, 2000; Jost et al., 2004; Hügli et al., 2005). By modeling the response of border-ownership cells to depth edges, we enforce an additional constraint on how depth information is to be combined with 2D information in order to produce proto-objects and the resulting salient points of the image.

[1] Exceptions are T-junctions (Heitger and von der Heydt, 1993) and extremal edges (Palmer and Ghose, 2008; Ramenahalli et al., 2011, 2012, 2014) both of which are local cues that provide information about edge polarity.

Each channel is processed independently of the others by the same grouping mechanism, and then combined through a series of normalization operators, which allow for competition between proto-objects of similar scale and feature. The final feature map is obtained by scaling each map to a common scale (approximately the middle level of the image pyramid), and then performing a pixel-wise addition across scales. The different feature conspicuity maps are then normalized again and linearly combined with equal weights to form the proto-object saliency map $S$. When depth information is used, we use a linear combination of 80% 2D features, split equally among intensity, color, and orientation, and 20% depth features to form the depth-added proto-object saliency map $S_d$. Obviously, this fraction can be modified but we found that results do not critically depend on its choice (see also the Discussion).

## 4.4 Results

We evaluate the performance of our model by comparing our generated saliency maps with ground truth data in the form of human eye fixations. Different from many other studies, we tested our model with a whole battery of metrics that were *not* necessarily chosen to give good results. We did this in order to provide a more diverse picture of the model's overall performance. Riche et al. (2013) have suggested that at least three different metrics are needed to fairly evaluate a given model. To ensure that our results do not depend on the use of a specific evaluation method, we use a battery of commonly used saliency metrics: area under the curve (AUC), Pearson's linear cross-correlation (PLCC), normalized scanpath saliency (NSS), similarity (SIM), earth-mover's distance (EMD), and the Kullback-Leibler divergence (KLD). For a recent review see Riche et al. (2013), the following is a brief description of the metrics.

KLD, EMD, PLCC, and SIM are distribution-based metrics that measure the similarity/dissimilarity between two distributions (in our case, between the distribution of human eye fixations and of the salient points as predicted by the model). Larger values of KLD and EMD indicate a larger overall difference between the two distributions, while a value of zero indicates that the two distributions are not systematically different from each other. PLCC and SIM are bounded values, where a value of unity indicates that the two distributions are identical, while a value of zero indicates that the distributions are completely uncorrelated (PLCC can also be negative, indicating a negative correlation between the two distributions). AUC is a location-based metric, a measure borrowed from signal-detection theory. An equal number of fixated and random pixels are first chosen from the saliency map. A threshold is then applied to the saliency map, which acts as a classifier, with all saliency points above threshold considered "fixated," and all saliency points below threshold considered "background." For each threshold value, we can then determine a true positive rate and a false positive rate based on the ground truth eye fixation map, which allows us to generate a Receiver-Operator Characteristic (ROC) curve and calculate the corresponding Area Under the Curve (AUC) metric. An ideal score is unity while a random classification gives a score of 0.5 and systematic mis-classifications result in values between 0 and 0.5. NSS is a value-based metric, which compares predicted saliency values with the corresponding eye fixation maps. NSS effectively measures the average number of standard deviations that the predicted salient points are above the global mean of the saliency map, with larger values indicating fixated points having a higher saliency as predicted by the model.

With the exception of the KLD metric, the code for all evaluation metrics can be found online on the MIT Saliency Benchmark webpage (Judd et al., 2012). The metrics compare the saliency map with either the binary fixation maps that contain the locations of all eye fixations of all participants without smoothing, or the continuous fixation density maps (smoothed averages of fixations). For the datasets we used, both continuous and binary fixation data were either included with the dataset, could be generated from the raw eye tracking data, or were obtained through correspondence with the authors that collected the data. Fixation density maps were used with the PLCC, SIM, KLD, and EMD metrics and binary fixation maps with the AUC and NSS metrics.

To determine whether the addition of depth information improves performance of the base 2D saliency model, we performed two-tailed, paired Student t-tests, with a significance level of $\alpha = 0.05$. To adjust for multiple comparisons and the dependence between saliency metrics, we applied a Benjamini-Hochberg correction (Benjamini

| Eyetracking Dataset | | Saliency Metrics | | | | | |
|---|---|---|---|---|---|---|---|
| | | PLCC | SIM | AUC | NSS | EMD | KLD |
| NUS-3D | 2D model only (2D Fixations) | 0.349 | 0.305 | 0.769 | 1.036 | 2.917 | 1.485 |
| | 2D model + depth (2D Fixations) | **0.359**** | **0.307**** | **0.774**** | **1.068**** | 2.913 | 1.485 |
| | t(df) | 6.58(599) | 3.75(599) | 4.99(599) | 6.75(599) | -0.40(599) | 0.02(599) |
| | p-Value | $3.00 \times 10^{-10}$ | $2.87 \times 10^{-4}$ | $7.10 \times 10^{-6}$ | $2.04 \times 10^{-10}$ | 0.831 | 0.988 |
| | 2D model only (3D Fixations) | 0.336 | 0.289 | 0.772 | 1.046 | 2.987 | 1.559 |
| | 2D model + depth (3D Fixations) | **0.347**** | **0.291**** | **0.777**** | **1.080**** | 2.980 | 1.559 |
| | t(df) | 6.75(599) | 4.20(599) | 5.27(599) | 7.08(599) | -0.68(599) | -0.33(599) |
| | p-Value | $1.04 \times 10^{-10}$ | $4.65 \times 10^{-5}$ | $2.05 \times 10^{-7}$ | $2.50 \times 10^{-11}$ | 0.600 | 0.742 |
| Gaze-3D | 2D model only | 0.535 | 0.682 | 0.699 | 0.720 | 2.108 | 0.327 |
| | 2D model + depth | 0.552* | 0.688* | 0.705 | 0.743 | 2.060 | 0.312* |
| | t(df) | 2.18(17) | 2.74(17) | 1.79(17) | 1.53(17) | -0.98(17) | -2.36(17) |
| | p-Value | 0.086 | 0.084 | 0.137 | 0.174 | 0.342 | 0.086 |
| NCTU-3D | 2D model only | 0.473 | 0.513 | 0.760 | 1.052 | 3.751 | 0.755 |
| | 2D model + depth | **0.479**** | 0.514 | **0.764**** | **1.071**** | 3.761 | 0.755 |
| | t(df) | 2.35(474) | 1.04(474) | 2.84(474) | 3.1304(474) | 0.76(474) | $<10^{-3}$(474) |
| | p-Value | 0.038 | 0.446 | 0.014 | 0.011 | 0.540 | 0.999 |

**Table 4.1:** Depth information improves saliency prediction on 3D eye tracking datasets. A double asterisk (**) and **boldface type** indicate that the performance of the model with depth information differs significantly from that of the corresponding 2D model, in the row immediately above it (paired t-test with Benjamini-Hochberg correction for multiple comparisons, $p < 0.05$). Similarly, a asterisk (*) indicates that the performance of the model with and without depth information differs significantly, but at a higher alpha level ($p < 0.10$). The value of the t-test statistic (t), the degrees of freedom (df), and the adjusted p-values (p-Value) are also reported here.

and Hochberg, 1995) to control the false discovery rate ($q = 0.05$). Table 4.1 summarizes the results of our model with and without depth information for the different 3D eye tracking datasets. We also report adjusted p-values in the table.

For all three datasets, adding depth information ("2D model + depth" compared to "2D model only" in Table 4.1) improved the model's prediction of perceptual saliency in terms of eye fixations. At least three, and sometimes more of the six metrics with and without depth information differed in a statistically significant manner ($p < 0.05$ for the NUS-3D and NCTU-3D datasets, and $p < 0.10$ for the Gaze-3D dataset), although which of the metrics reached significance varied between datasets.

For the NUS-3D dataset, adding depth information improved the PLCC, SIM, AUC, and NSS metrics for both the 2D and 3D viewing conditions ($p < 0.05$, see Table 4.1 for the associated test statistics and p-Values). The EMD and KLD metrics showed improvement that was not statistically significant or no improvement, respectively. For the Gaze-3D dataset, adding depth information improved each of the metrics, but this improvement was not statistically significant at the chosen alpha level ($p > 0.05$). We note here that our model outperforms a previous model (in terms of the PLCC, AUC, and KLD metrics) that was evaluated using the same dataset (Wang et al., 2013). We also note that at the higher significance level used in that study, the improvement in the PLCC, SIM, and KLD metrics are statistically significant ($p < 0.10$). For the NCTU-3D dataset, adding depth information significantly improved the PLCC, AUC, and NSS metrics ($p < 0.05$), and also improved the SIM and KLD metrics but, again, these did not reach statistical significance. The EMD metric increased with depth information, indicating a greater difference between the distribution of salient points predicted by the model and the eye fixations, but this difference was not significant.

A special case is the NUS-3D dataset for which eye tracking data from a 2D viewing condition is also available. In this case, participants viewed the images binocularly without stereoscopic depth cues (identical input presented to both eyes) but monocular cues (like occlusion, shading, extremal edges, T-junctions *etc.*) remained available. While depth information plays an important role in the computation of proto-object representations in our model, it does not take into account other monocular depth cues. By comparing the fixation prediction performance of the model when it has access to depth information compared to when it does not (the first and second line of Table 4.1, respectively), we can assess the importance of monocular cues not included in the model. Results reveal significant differences for four of the six metrics (PLCC, SIM, AUC, and NSS, $p < 0.05$). As a result, we conclude that monocular depth cues play an important role for saliency prediction and that future models will likely benefit from including their influence.

Overall, incorporating depth into our proto-object based saliency model improved performance across all three tested datasets, as measured by different metrics that are sensitive to different components of the data. It should be noted that the effect of adding depth information is relatively small, which may point to the relative importance of traditional 2D features in visual saliency. In our model, depth information generally helps with perceptual saliency prediction, although the degree to which it does may vary greatly based on image content. We provide additional reasons for why the absolute size of the effect is small in the Discussion.

## 4.5 Discussion

For all 3D datasets, we found that incorporating binocular depth information in our model resulted in a small, but statistically significant improvement in perceptual saliency prediction on most of the evaluation metrics.

For the NUS-3D saliency dataset, adding depth information improved performance of the proto-object based saliency model for both 2D and 3D viewing conditions. The results for the 2D viewing condition agree with the previous finding that incorporating depth information gained from monocular depth cues can improve 2D saliency prediction (Ramenahalli and Niebur, 2013). In that study, however, depth information was inferred from the 2D image using the Make3D algorithm (Saxena et al., 2009), which computes a depth map from a 2D image, while in the current work, depth information is directly collected with the Kinect sensor.

Although our model performance does not exceed that of previously reported results on the NUS-3D dataset (Lang et al., 2012) or the NCTU-3D dataset (Ma and Hang, 2015), our model has the advantage of being a straightforward extension of an existing 2D model (Russell et al., 2014) which is based on biologically-realistic features of early and intermediate primate vision. Importantly, different from previous work, our model does not rely on learning novel depth priors or, for that matter, learning *anything* from a training set of images. This has at least two advantages. First, we eliminate the time and computational effort needed for training, which typically scales with the number of images and/or the number of features chosen to be learned.

Second, using depth information in a way that combines 2D Gestalt cues with depth cues is a mechanism of general validity, and therefore we believe that our model is applicable to a wide range of natural images, not just those included in the training datasets, or images similar to those. We also note that our model does extremely well on the Gaze-3D dataset, significantly outperforming the best previously reported results. This indicates that the proto-object based saliency model may be able to capture perceptual saliency more successfully than other 2D saliency models that are purely feature-based. However, we note that model performance even without the depth information is very good. While depth information does help in the prediction of eye fixations, its contribution is relatively small compared to that from 2D features and not statistically significant at the chosen alpha level ($\alpha = 0.05$).

We combined the 2D and 3D features with a 4:1 ratio of 2D features to 3D features, meaning a weight of 20% on the depth information and of 80% on the traditional 2D features. In contrast, Wang et al. (2013) used a ratio of 1:1 for weighting 2D features and depth information, giving depth information the same importance as the combination of all 2D features. In our experience, at least for conditions under which the three data sets that we have access to were collected, the contribution of depth information is comparable to some of the 2D submodalities, but substantially smaller than the combination of *all* 2D submodalities. Informal parametric studies showed that although the assignment of detailed relative weights is not critical, a clear dominance of 2D over depth information gave the best results.

The performance enhancement due to adding the depth channel results is significant but its absolute value is small. One reason for the small size of the effect could be the long viewing times which were, for the three datasets used, in the range of 4-15 seconds. With these relatively long viewing times, 2D features may play a more critical role in directing the participants's gaze, compared to the role of 3D features, which may be more important early in the viewing period. Indeed, others have shown time-dependent influences of the 2D and 3D features on saliency prediction (Gautier and Le Meur, 2012).

Secondly, we did not divide our images based on the depth range or depth-of-field, which have been shown to be important factors in determining to what extent depth information can influence visual attention (Lang et al., 2012). It is possible that large depth differences are disproportionately salient, but large differences can only occur in images with a large depth of field. Similarly, depth information may be particularly advantageous in highly-textured scenes, where 2D cues are not enough to perform segmentation of proto-objects. Ma and Hang (2015) present examples of images where depth information may help participants to segment objects among highly-textured backgrounds, or among objects that are not located at the center of the image.

A third reason why the absolute size of the 3D effect is small may be the presence of common surfaces in the images, which can influence the perception of metric depth. Several previous psychophysical studies have shown that perceived binocular depth can be affected by a common surface (???). Importantly, a recent result shows that perceived absolute distance to objects on the ground surface is not different between monocular and binocular viewing conditions (?). Binocular disparity may only play a critical role in perceiving absolute distance of a target in midair. Given that many objects rest on surfaces, binocular disparity information may not be needed – the relative depth of objects along a perceived common surface under 2D viewing conditions may be sufficient. It is then possible that the small absolute size of the 3D effect is due to the fact that many of the images in the datasets are natural scenes that carry common surfaces, such as the ground, floor, walls *etc.* (see Figure 4.1). This provides further evidence for the important role of surfaces defined either monocularly or binocularly for the perceptual organization of visual scenes (He and Nakayama, 1992, 1995; Hu et al., 2015).

Our proto-object based model operates at intermediate stages of vision, where perceptual organization of the visual scene is thought to occur. In contrast, other 3D saliency models use only low-level visual features, or incorporate additional semantically important cues that may only be found in higher visual areas. These other models either treat depth as an early feature which can be combined multiplicatively or additively with the 2D saliency map, or they include other features (such as human face and body detection, Cerf et al., 2008) as a means to improve performance. We believe that while adding these other features can improve model performance in many cases, intermediate stages of vision are critical for transforming low-level visual features into higher-level object representations that form the basis for further visual processing and allocation of attention. We show that by incorporating depth information as an additional cue into the grouping mechanism, we can more accurately predict where participants will fixate within a scene (*i.e.* as a marker of perceptual saliency). This is because binocular depth provides unambiguous information about the location and border ownership of object edges, which can be used for the perceptual organization of the scene in terms of proto-objects.

There has been some debate as to whether the computation of visual saliency is feature-based or object-based. Feature-based models rely on low-level feature contrast to generate a saliency map (*e.g.* Itti et al., 1998; Walther and Koch, 2006). Object-based saliency models, instead, start from the assumption that objects, and not necessarily their constituent features, are what is needed for determining the salient regions of an image and are the primary driver of fixations (Einhäuser et al., 2008; **?**; **?**). Support for object-based models comes from the analysis of fixation locations within objects. Fixations are well described by a two-dimensional Gaussian distribution with a mean biased towards the center of the object, which represents the preferred viewing location (PVL) of the object (**?**). Critically, proto-objects computed solely in terms of low-level features without the influence of Gestalt cues (Walther and Koch, 2006) do not exhibit a central PVL. However, the proto-objects in by our model integrate low-level feature information from different spatial locations and scales, such that their final activity should be biased towards the center of closed, convex objects. As a result, previous experimental results that cast doubt on the role of feature-based saliency (Einhäuser et al., 2008; ?; ?) do not rule out our proto-object based model, but rather support it. A direct comparison of the proto-objects in our model with real objects is still lacking, so it remains to be seen whether these proto-objects exhibit a central PVL, which is an area of future research. We believe our model fits most closely with the definition of proto-objects by Rensink (2000), as providing both a feedforward measure of objecthood and a "handle" for top-down processes. Saliency is then a function of proto-objects, and proto-objects may also causally drive attention. For a more detailed discussion of this issue, we refer the reader to Russell et al. (2014).

# Conclusion

We introduce a new proto-object based saliency model which makes use of information about 3D depth to segment natural scenes. Our model is an extension of previous models in 2D and it is constructed from first principles, without relying on learning of depth priors or depth features; it does not require any training. The model is biologically-inspired, with the computations needed being directly mapped to neural mechanisms that have been found in the brain. Using data from three separate 3D eye tracking datasets, we show that depth information improves performance in a robust manner using a number of evaluation metrics. Athough we find that proto-objects are largely formed based on 2D features, the added depth information has clear benefits in improving performance of the model in terms of predicting the location of human eye fixations.

# Bibliography

Y Benjamini and Y Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B*, 1995. URL `http://www.jstor.org/stable/2346101`.

Ali Borji and Laurent Itti. State-of-the-art in visual attention modeling. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2013.

Ali Borji, Dicky N Sihite, and Laurent Itti. Objects do not predict fixations better than early saliency: A re-analysis of Einhäuser et al.'s data. *Journal of vision*, 13 (10):18, 2013.

William H Bosking, Ying Zhang, Brett Schofield, and David Fitzpatrick. Orientation selectivity and the arrangement of horizontal connections in tree shrew striate cortex. *The Journal of Neuroscience*, 17(6):2112–2127, 1997.

S.L. Brincat and C.E. Connor. Underlying principles of visual shape selectivity in posterior inferotemporal cortex. *Nature Neuroscience*, 7:880–886, 2004.

Neil DB Bruce and John K Tsotsos. An attentional framework for stereo vision. In *Computer and Robot Vision, 2005. Proceedings. The 2nd Canadian Conference on*, pages 88–95. IEEE, 2005.

M. Cerf, J. Harel, W. Einhäuser, and C. Koch. Predicting human gaze using low-level saliency combined with face detection. *Advances in Neural Information Processing Systems*, 20:241–248, 2008.

Garga Chatterjee, Daw-An Wu, and Bhavin R Sheth. Phantom flashes caused by interactions across visual space. *Journal of vision*, 11(2):14, 2011.

Minggui Chen, Yin Yan, Xiajing Gong, Charles D Gilbert, Hualou Liang, and Wu Li. Incremental integration of global contours through interplay between visual cortical areas. *Neuron*, 82(3):682–694, 2014.

Michele A Cox, Michael C Schmid, Andrew J Peters, Richard C Saunders, David A Leopold, and Alexander Maier. Receptive field focus of visual area V4 neurons determines responses to illusory surfaces. *Proceedings of the National Academy of Sciences*, 110(42):17095–17100, 2013.

E. Craft, H. Schütze, E. Niebur, and R. von der Heydt. A neural model of figure-ground organization. *Journal of Neurophysiology*, 97(6):4310–26, 2007. PMID: 17442769.

Y. Dong, S. Mihalas, F. Qiu, R. von der Heydt, and E. Niebur. Synchrony and the binding problem in macaque visual cortex. *Journal of Vision*, 8(7):1–16, 2008. URL `http://journalofvision.org/8/7/30/,doi:10.1167/8.7.30`. PMC2647779.

J. Driver and G.C. Baylis. Edge-assignment and figure-ground segmentation in short-term visual matching. *Cognit Psychol*, 31(3):248–306, 1996.

J. Duncan. Selective attention and the organization of visual information. *J Exp Psychol Gen*, 113:501–517, Dec 1984.

R. Egly, J. Driver, and R. Rafal. Shifting visual attention between objects and locations: evidence for normal and parietal lesion subjects. *Journal of Experimental Psychology: General*, 123:161–77, 1994.

W. Einhäuser, M. Spain, and P. Perona. Objects predict fixations better than early saliency. *J. Vision*, 8(14):1–26, 2008.

Boris Epshtein, Ita Lifshitz, and Shimon Ullman. Image interpretation by a single bottom-up top-down cycle. *Proceedings of the National Academy of Sciences*, 105 (38):14298–14303, 2008.

D. J. Field, A. Hayes, and R. F. Hess. Contour integration by the human visual system: evidence for a local association field. *Vision Research*, 33(2):173–193, 1993.

Josselin Gautier and Olivier Le Meur. A time-dependent saliency model combining center and depth biases for 2D and 3D viewing conditions. *Cognitive Computation*, 4(2):141–156, 2012.

Ariel Gilad, Elhanan Meirovithz, and Hamutal Slovin. Population responses to contour integration: early encoding of discrete elements and late perceptual grouping. *Neuron*, 78(2):389–402, 2013.

C.D. Gilbert and T.N. Wiesel. Columnar specificity of intrinsic horizontal and cortico-cortical connections in cat visual cortex. *J. Neurosci.*, 9:2432–2442, 1989.

P. Girard, J.M. Hupé, and J. Bullier. Feedforward and feedback connections between areas V1 and V2 of the monkey have similar rapid conduction velocities. *J. Neurophysiol.*, 85:1328–1331, 2001.

D. M. Green and J. A. Swets. *Signal Detection Theory and Psychophysics.* Wiley, New York, N.Y., 1966.

S. Grossberg. 3-D vision and figure-ground separation by visual cortex. *Perception & Psychophysics*, 55:48–120, 1994.

S. Grossberg. Cortical dynamics of three-dimensional figure–ground perception of two-dimensional pictures. *Psychological review*, 104(3):618, 1997. PMID: 9243966.

Z. J. He and K. Nakayama. Surfaces versus features in visual search. *Nature*, 359: 231–233, 1992. PMID: 1528263.

Z. J. He and K. Nakayama. Visual attention to surfaces in three-dimensional space. *Proc. Natl. Acad. Sci. U. S. A.*, 9(24):11155–11159, 1995. PMID: 7479956.

Jay Hegdé and David C Van Essen. A comparative study of shape representation in macaque visual areas V2 and V4. *Cerebral Cortex*, 17(5):1100–1116, 2007.

F. Heitger and R. von der Heydt. A computational model of neural contour processing: figure-ground segregation and illusory contours. In *Proc. 4th Int. Conf. Computer Vision*, Proc. 4th Int. Conf. Computer Vision, pages 32–40. IEEE Computer Society Press, 1993.

Ming-Chou Ho and Su-Ling Yeh. Effects of instantaneous object input and past experience on object-based attention. *Acta psychologica*, 132(1):31–39, 2009.

Shaul Hochstein and Merav Ahissar. View from the top: Hierarchies and reverse hierarchies in the visual system. *Neuron*, 36(5):791–804, 2002.

B. Hu, R. von der Heydt, and E. Niebur. A neural model for perceptual organization of 3D surfaces. In *IEEE CISS-2015 49th Annual Conference on Information Sciences and Systems*, pages 1–6, Baltimore, MD, 2015. IEEE Information Theory Society. doi: 10.1109/CISS.2015.7086906.

D.H. Hubel and T.N. Wiesel. Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. *J. Physiol.*, 160:106–154, 1962.

Heinz Hügli, Timothée Jost, and Nabil Ouerhani. Model performance for visual attention in real 3D color scenes. In *Artificial Intelligence and Knowledge Engineering Applications: A Bioinspired Approach*, pages 469–478. Springer, 2005.

Quan Huynh-Thu and Luca Schiatti. Examination of 3D visual attention in stereoscopic video content. In *IS&T/SPIE Electronic Imaging*, pages 78650J–78650J. International Society for Optics and Photonics, 2011.

J. Intriligator and P. Cavanagh. The spatial resolution of visual attention. *Cognit. Psychol.*, 43:171–216, 2001.

L. Itti, C. Koch, and E. Niebur. A model of saliency-based fast visual attention for rapid scene analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(11):1254–1259, November 1998.

Lina Jansen, Selim Onat, and Peter König. Influence of disparity on fixation and saccades in free viewing of natural scenes. *Journal of Vision*, 9(1):29, 2009.

Ana Karla Jansen-Amorim, Mario Fiorani, and Ricardo Gattass. GABA inactivation of area V4 changes receptive-field properties of V2 neurons in Cebus monkeys. *Experimental Neurology*, 235(2):553–562, 2012.

Timothée Jost, Nabil Ouerhani, Roman von Wartburg, René Müri, and Heinz Hügli. Contribution of depth to visual attention: comparison of a computer model and human. In *Proceedings. Early cognitive vision workshop*, pages 1–4, 2004.

Tilke Judd, Frédo Durand, and Antonio Torralba. A Benchmark of Computational Models of Saliency to Predict Human Fixations. In *MIT Technical Report*, 2012.

R. Kimchi, Y. Yeshurun, and A. Cohen-Savransky. Automatic, stimulus-driven attentional capture by objecthood. *Psychon Bull Rev*, 14(1):166–172, Feb 2007.

C. Koch and S. Ullman. Shifts in selective visual attention: towards the underlying neural circuitry. *Human Neurobiol.*, 4:219–227, 1985.

K. Koffka. *Principles of Gestalt psychology.* Harcourt-Brace, New York, 1935.

V. A. F. Lamme. The neurophysiology of figure-ground segregation in primary visual cortex. *J Neurosci*, 15:1605–1615, 1995.

V. A. F. Lamme, K. Zipser, and H. Spekreijse. Figure-ground activity in primary visual cortex is suppressed by anesthesia. *Proc. Natl. Acad. Sci. U. S. A.*, 9(6): 3263–3268, 1998.

Congyan Lang, Tam V Nguyen, Harish Katti, Karthik Yadati, Mohan Kankanhalli, and Shuicheng Yan. Depth matters: Influence of depth cues on visual saliency. In *Computer Vision–ECCV 2012*, pages 101–115. Springer, 2012.

Oliver W Layton, Ennio Mingolla, and Arash Yazdanbakhsh. Dynamic coding of border-ownership in visual cortex. *Journal of vision*, 12(13):8, 2012.

P. Le Callet and E. Niebur. Visual Attention and Applications in Multimedia Technologies. *IEEE Proceedings*, 101(9):2058–67, 2013. NIHMS539064.

Tai Sing Lee, David Mumford, Richard Romero, and Victor A. F. Lamme. The Role of the Primary Visual Cortex in Higher Level Vision. *Vision Research*, 38:2429–2452, 1998.

Wu Li, Valentin Piëch, and Charles D Gilbert. Perceptual learning and top-down influences in primary visual cortex. *Nature neuroscience*, 7(6):651–657, 2004.

Wu Li, Valentin Piëch, and Charles D Gilbert. Contour saliency in primary visual cortex. *Neuron*, 50(6):951–962, 2006.

Wu Li, Valentin Piëch, and Charles D Gilbert. Learning to link visual contours. *Neuron*, 57(3):442–451, 2008.

Z. Li. A Neural Model of Contour Integration in the Primary Visual Cortex. *Neural Computation*, 10(903-940), 1998.

Chih-Yao Ma and Hsueh-Ming Hang. Learning-based saliency model with depth information. *Journal of vision*, 15(6):19–19, 2015.

D. Marr and T. Poggio. Cooperative Computation of Stereo Disparity. *Science*, 194, 1976. PMID: 968482.

Jonathan A Marshall, George J Kalarickal, and Elizabeth B Graves. Neural model of visual stereomatching: slant, transparency and clouds. *Network: Computation in Neural Systems*, 7(4):635–669, 1996.

Anne B Martin and Rüdiger von der Heydt. Spike Synchrony Reveals Emergence of Proto-Objects in Visual Cortex. *The Journal of Neuroscience*, 35(17):6860–6870, 2015.

C. J. McAdams and J. H. R. Maunsell. Effects of attention on orientation-tuning functions of single neurons in macaque cortical area V4. *J. Neurosci.*, 19:431–441, 1999.

S. Mihalas, Y. Dong, R. von der Heydt, and E. Niebur. Mechanisms of perceptual organization provide auto-zoom and auto-localization for attention to objects. *Proceedings of the National Academy of Sciences*, 108(18):7583–8, 2011. PMC3088583.

Ajay K Mishra, Ashish Shrivastava, and Yiannis Aloimonos. Segmenting ”simple”??? objects using RGB-D. In *Robotics and Automation (ICRA), 2012 IEEE International Conference on*, pages 4406–4413. IEEE, 2012.

B. C. Motter. Neural correlates of attentive selection for color or luminance in extrastriate area V4. *J. Neurosci.*, 14:2178–2189, 1994.

B.C. Motter. Focal attention produces spatially selective processing in visual cortical areas V1, V2, and V4 in the presence of competing stimuli. *J. Neurophysiology*, 70 (3):909–919, 1993.

K. Nakayama and G. H. Silverman. Serial and parallel processing of visual feature conjunctions. *Nature*, 320:264–265, 1986. PMID: 3960106.

K. Nakayama, Z. J. He, and S. Shimojo. Visual surface representation: a critical link between lower-level and higher-level vision. In S. Kosslyn and D. Osherson, editors, *Visual Cognition: An Invitation to Cognitive Science*, volume 2, chapter 1, pages 1–70. The MIT Press, 2nd edition, 1995.

E. Niebur. Separate but equal: Different kinds of information require different neural representations. In H. Bothe, editor, *Proceedings of the International Congress on Intelligent Systems and Applications (ISA-BIS)*, pages 1544–9, Wetaskiwin, Canada, December 2000. ICSC Academic Press. ISBN 3-906454-24-X.

E. Niebur and C. Koch. Control of Selective Visual Attention: Modeling the "Where" Pathway. In D. S Touretzky, M. C. Mozer, and M. E. Hasselmo, editors, *Advances in Neural Information Processing Systems*, volume 8, pages 802–808. MIT Press, Cambridge, MA, 1996.

P.J. O'Herron and R. von der Heydt. Short-term memory for figure-ground organization in the visual cortex. *Neuron*, 61(5):801–809, 2009. PMC2707495.

P.J. O'Herron and R. von der Heydt. Remapping of Border Ownership in the Visual Cortex. *The Journal of Neuroscience*, 33(5):1964–1974, 2013. PMID: 23365235 [PubMed - in process].

I. Ohzawa, G. C. DeAngelis, and R. D. Freeman. Stereoscopic depth discrimination in the visual cortex: neurons ideally suited as disparity detectors. *Science*, 249: 1037–1041, 1990.

Nabil Ouerhani and Heinz Hügli. Computing visual attention from scene depth. In *Pattern Recognition, 2000. Proceedings. 15th International Conference on*, volume 1, pages 375–378. IEEE, 2000.

Lucy M Palmer, Adam S Shai, James E Reeve, Harry L Anderson, Ole Paulsen, and Matthew E Larkum. NMDA spikes enhance action potential generation during sensory input. *Nature neuroscience*, 17(3):383–390, 2014.

S.E. Palmer and T. Ghose. Extremal Edge– A Powerful Cue to Depth Perception and Figure-Ground Organization. *Psychological Science*, 19(1):77, 2008. ISSN 0956-7976.

D. Parkhurst, K. Law, and E. Niebur. Modelling the role of salience in the allocation of visual selective attention. *Vision Research*, 42(1):107–123, 2002.

Anitha Pasupathy and Charles E Connor. Population coding of shape in area V4. *Nature neuroscience*, 5(12):1332–1338, 2002.

Valentin Piëch, Wu Li, George N Reeke, and Charles D Gilbert. Network model of top-down influences on local gain and contextual interactions in visual cortex. *Proceedings of the National Academy of Sciences*, 110(43):E4108–E4117, 2013. PMC3808648.

G. F. Poggio and B. Fischer. Binocular interaction and depth sensitivity in striate and prestriate cortex of behaving rhesus monkey. *J. Neurophysiol.*, 40:1392–1405, Nov 1977. PMID: 411898.

G.F. Poggio and T. Poggio. The analysis of stereopsis. *Ann. Rev. Neurosci.*, 7: 379–412, 1984.

Gian F Poggio, Francisco Gonzalez, and F Krause. Stereoscopic mechanisms in monkey visual cortex: binocular correlation and disparity selectivity. *The Journal of neuroscience*, 8(12):4531–4550, 1988.

Uri Polat, Keiko Mizobe, Mark W. Pettet, Takuji Kasamatsu, and Anthony M. Norcia. Collinear Stimuli Regulate Visual Responses Depending on Cell's Contrast Threshold. *Nature*, 391:580–584, February 1998.

Jasper Poort, Florian Raudies, Aurel Wannig, Victor AF Lamme, Heiko Neumann, and Pieter R Roelfsema. The role of attention in figure-ground segregation in areas V1 and V4 of the visual cortex. *Neuron*, 75(1):143–156, 2012.

F. T. Qiu and R. von der Heydt. Figure and ground in the visual cortex: V2 combines stereoscopic cues with Gestalt rules. *Neuron*, 47:155–166, 2005.

F. T. Qiu and R. von der Heydt. Neural representation of transparent overlay. *Nat. Neurosci.*, 10(3):283–284, 2007.

F. T. Qiu, T. Sugihara, and R. von der Heydt. Figure-ground mechanisms provide structure for selective attention. *Nat. Neurosci.*, 10(11):1492–9, October 2007.

S. Ramenahalli and E. Niebur. Computing 3D saliency from a 2D image. In *Information Sciences and Systems (CISS), 2013 47th Annual Conference on*, pages 1–5, 2013. doi: 10.1109/CISS.2013.6552297. URL `http://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=6552297`.

S. Ramenahalli, S. Mihalas, and E. Niebur. Extremal Edges: Evidence in Natural Images. In *IEEE CISS-2011 45th Annual Conference on Information Sciences and Systems*, pages 1–6, Baltimore, MD, 2011. IEEE Information Theory Society.

S. Ramenahalli, S. Mihalas, and E. Niebur. Figure-ground classification based on spectral properties of boundary image patches. In *IEEE CISS-2012 46th Annual Conference on Information Sciences and Systems*, pages 1–6, Princeton, NJ, 2012. IEEE Information Theory Society.

Sudarshan Ramenahalli, Stefan Mihalas, and Ernst Niebur. Local spectral anisotropy is a valid cue for figure-ground organization in natural scenes. *Vision Research*, 103: 116–126, Oct 2014. doi: 10.1016/j.visres.2014.08.012. URL `http://dx.doi.org/10.1016/j.visres.2014.08.012`. NIHMSID 631573.

R. A. Rensink. The dynamic representation of scenes. *Visual Cognition*, 7(1/2/3): 17–42, 2000.

Nicolas Riche, Matthieu Duvinage, Matei Mancas, Bernard Gosselin, and Thierry Dutoit. Saliency and human fixations: state-of-the-art and study of comparison metrics. In *Computer Vision (ICCV), 2013 IEEE International Conference on*, pages 1153–1160. IEEE, 2013.

P. R. Roelfsema, V. A. F. Lamme, and H. Spekreijse. Object-based attention in the primary visual cortex of the macaque monkey. *Nature*, 395(6700):376–381, 1998.

Pieter R Roelfsema. Cortical algorithms for perceptual grouping. *Annu. Rev. Neurosci.*, 29:203–227, 2006.

Pieter R Roelfsema, Victor A F Lamme, and Henk Spekreijse. Synchrony and covariation of firing rates in the primary visual cortex during contour grouping. *Nat Neurosci*, 7(9):982–991, Sep 2004. doi: 10.1038/nn1304. URL `http://dx.doi.org/10.1038/nn1304`.

Ari Rosenberg, Noah J Cowan, and Dora E Angelaki. The Visual Representation of 3D Object Orientation in Parietal Cortex. *The Journal of Neuroscience*, 33(49): 19352–19361, 2013.

A. F. Russell, S Mihalas, R. von der Heydt, E. Niebur, and R. Etienne-Cummings. A model of proto-object based saliency. *Vision Research*, 94:1–15, 2014.

P Sajda and L.H. Finkel. Intermediate-Level Visual Representations and the Construction of Surface Perception. *J Cogn Neurosci*, 7:267–291, 1995.

K. Sakai and H. Nishimura. Surrounding suppression and facilitation in the determination of border ownership. *Journal of Cognitive Neuroscience*, 18(4):562–579, 2006.

Ashutosh Saxena, Min Sun, and Andrew Y Ng. Make3d: Learning 3d scene structure from a single still image. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 31(5):824–840, 2009.

Daniel Scharstein and Chris Pal. Learning conditional random fields for stereo. In *Computer Vision and Pattern Recognition, 2007. CVPR'07. IEEE Conference on*, pages 1–8. IEEE, 2007.

Mircea A Schoenfeld, Jens-Max Hopf, Christian Merkel, Hans-Jochen Heinze, and Steven A Hillyard. Object-based attention involves the sequential activation of feature-specific cortical modules. *Nature neuroscience*, 17(4):619–624, 2014.

B. J. Scholl. Objects and attention: the state of the art. *Cognition*, 80(1-2):1–46, 2001.

M. Sigman, G. A. Cecchi, C. D. Gilbert, and M. O. Magnasco. On a common circle: natural scenes and Gestalt rules. *Proc Natl Acad Sci U S A*, 98(4):1935–1940, Feb 2001. doi: 10.1073/pnas.031571498. URL `http://dx.doi.org/10.1073/pnas.031571498`.

Enrico Simonotto, Massimo Riani, Charles Seife, Mark Roberts, Jennifer Twitty, and

Frank Moss. Visual perception of stochastic resonance. *Physical Review Letters*, 78(6):1186, 1997.

W. Singer. Neuronal synchrony: a versatile code for the definition of relations? *Neuron*, 24:49–65, 1999.

M. Stemmler, M. Usher, and E. Niebur. Lateral cortical connections may contribute to both contour completion and redundancy reduction in visual processing. *Soc. Neurosci. Abstr.*, 21(1):510, 1995a.

M. Stemmler, M. Usher, and E. Niebur. Lateral Interactions in Primary Visual Cortex: A Model Bridging Physiology and Psychophysics. *Science*, 269:1877–1880, 1995b.

Dan D Stettler, Aniruddha Das, Jean Bennett, and Charles D Gilbert. Lateral connectivity and contextual interactions in macaque primary visual cortex. *Neuron*, 36(4):739–750, 2002.

Tadashi Sugihara, Fangtu T Qiu, and Rüdiger von der Heydt. The speed of context integration in the visual cortex. *Journal of neurophysiology*, 106(1):374–385, 2011. PMC3129740.

Hans Supèr and Victor AF Lamme. Altered figure-ground perception in monkeys with an extra-striate lesion. *Neuropsychologia*, 45(14):3329–3334, 2007.

A. Thiele and G. Stoner. Neuronal synchrony does not correlate with motion coherence in cortical area MT. *Nature*, 421(6921):366–370, 2003.

A. Treisman. The binding problem. *Curr Opin Neurobiol*, 6(2):171–178, April 1996.

A. Treisman and G. Gelade. A feature-integration theory of attention. *Cognitive Psychology*, 12:97–136, 1980. PMID: 7351125.

J. C. M. Treue, S.and Trujillo. Feature-based attention influences motion processing gain in macaque visual cortex. *Nature*, 399:575–9, 1999.

S. Treue. Neural correlates of attention in primate visual cortex. *Trends in Neurosciences*, 24:295–300, May 2001.

J. K. Tsotsos. Analyzing vision at the complexity level. *Behav. Brain Sci.*, 13: 423–469, 1990.

S. Ullman. Visual routines. *Cognition*, 18:97–159, 1984.

Shimon Ullman, RL Gregory, and J Atkinson. Low-Level Aspects of Segmentation and Recognition [and Discussion]. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*, 337(1281):371–379, 1992.

Matthieu Urvoy, Marcus Barkowsky, Romain Cousseau, Yao Koudota, Vincent Ricorde, Patrick Le Callet, Jesus Gutierrez, and Narciso Garcia. NAMA3DS1-COSPAD1: Subjective video quality assessment database on coding conditions introducing freely available high quality 3D stereoscopic sequences. In *Quality of*

*Multimedia Experience (QoMEX), 2012 Fourth International Workshop on*, pages 109–114. IEEE, 2012.

Shaun P Vecera, Anastasia V Flevaris, and Joseph C Filapek. Exogenous spatial attention influences figure-ground assignment. *Psychological Science*, 15(1):20–26, 2004.

R. von der Heydt, F. T. Qiu, and Z. J. He. Neural mechanisms in border ownership assignment: motion parallax and gestalt cues. *J. Vision*, 3(9):666a, 2003.

D. Walther and C. Koch. Modeling attention to salient proto-objects. *Neural Networks*, 19:1395–1407, Nov 2006.

Junle Wang, Matthieu Perreira DaSilva, Patrick LeCallet, and Vincent Ricordel. Computational model of stereoscopic 3D visual saliency. *Image Processing, IEEE Transactions on*, 22(6):2151–2165, 2013.

Manuel Werlberger, Werner Trobin, Thomas Pock, Andreas Wedel, Daniel Cremers, and Horst Bischof. Anisotropic Huber-L1 Optical Flow. In *Proceedings of the British Machine Vision Conference (BMVC)*, London, UK, September 2009.

M. Wertheimer. Untersuchungen zur Lehre von der Gestalt II. *Psychol. Forsch.*, 4: 301–350, 1923.

Jonathan R Williford and Rudiger von der Heydt. Border-ownership coding. *Scholarpedia*, 8(10):30040, 2013.

Jonathan R Williford and Rudiger von der Heydt. Early Visual Cortex Assigns Border Ownership in Natural Scenes According to Image Context. *Journal of Vision*, 14 (10):588–588, 2014.

Jun Xie, Guanghua Xu, Jing Wang, Sicong Zhang, Feng Zhang, Yeping Li, Chengcheng Han, and Lili Li. Addition of visual noise boosts evoked potential-based brain-computer interface. *Scientific reports*, 4, 2014.

Yin Yan, Malte J Rasch, Minggui Chen, Xiaoping Xiang, Min Huang, Si Wu, and Wu Li. Perceptual training continuously refines neuronal population codes in primary visual cortex. *Nature neuroscience*, 17(10):1380–1387, 2014.

Shih-Cheng Yen and Leif H Finkel. Extraction of perceptually salient contours by striate cortical networks. *Vision research*, 38(5):719–741, 1998.

N.R. Zhang and R. von der Heydt. Analysis of the context integration mechanisms underlying figure–ground organization in the visual cortex. *The Journal of Neuroscience*, 30(19):6482–6496, 2010. PMC2910339.

Li Zhaoping. Border ownership from intracortical interactions in visual area V2. *Neuron*, 47:143–153, 2005. PMID: 15996554.

H. Zhou, H. S. Friedman, and R. von der Heydt. Coding of border ownership in monkey visual cortex. *J. Neurosci.*, 20(17):6594–6611, 2000. PMID: 10964965.