Vision Research 160 (2019) 60–71

Contents lists available at ScienceDirect

Vision Research

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

Unique objects attract attention even when faint

Daniel M. Jeck[a,b], Michael Qin[c], Howard Egeth[d], Ernst Niebur[a,d,e,*]

[a] Zanvyl Krieger Mind/Brain Institute, Johns Hopkins University, Baltimore, MD, USA
[b] Department of Biomedical Engineering, Johns Hopkins University, Baltimore, MD, USA
[c] Department of Biomedical Engineering, University of Connecticut at Storrs, USA
[d] Department of Psychological and Brain Sciences, Johns Hopkins University, Baltimore, MD, USA
[e] Solomon Snyder Department of Neuroscience, Johns Hopkins University, Baltimore, MD, USA

ARTICLE INFO

Keywords:
Attention
Saliency
Uniqueness
Weak signals
Top-down
Bottom-up

ABSTRACT

Locally contrasting objects, *e.g.* a red apple surrounded by green apples, attract attention. Does this generalize to differences in feature space? That is, do unique objects-regardless of their location-stand out from a collection of objects that are similar to one another, even when the unique object has lower local contrast with the background than the other objects? Behavioral data show indeed a preference for unique items but previous experiments enabled viewers to anticipate what response they were "supposed" to give. We developed a new experimental paradigm that minimizes such top-down effects. Pitting local contrast against global uniqueness, we show that unique stimuli attract attention even in not-anticipated, never-seen images, and even when the unique stimuli are faint (low contrast). A computational model explains how competition between objects in feature space favors dissimilar objects over those with similar features. The model explains how humans select unique objects, without a loss of performance on natural scenes.

1. Introduction

Human behavior depends on input from all sensory modalities. Most of the information collected by the sensory apparatus is, however, quickly discarded by a filtering process commonly referred to as selective attention. It is this process that allows organisms with limited information processing capabilities to operate quickly and efficiently in a highly complex world (Tsotsos, 1990). Over the last decades, much progress has been made on understanding attentive selection. In particular, computational "saliency map" models (Koch & Ullman, 1985), which are based on local differences of visual (Itti, Koch, & Niebur, 1998) and auditory (Kayser, Petkov, Lippert, & Logothetis, 2005) sensory features, are quite successful in explaining human behavior in both controlled laboratory and real world situations (Borji & Itti, 2013). These models only use image information (saliency) but not observers' goals, memory, or other internal states.

It has long been suspected that the idea that perceptual saliency, which is derived from stimulus contrast and drives attention, can be generalized to more abstract spaces. That is, that a stimulus that differs from others *in feature space* (rather than in geometrical space) stands out by itself, without needing help from observer-inherent biases like anticipation, goals *etc*. In the simplest case, which we test here, this would imply that a stimulus that is weak by itself but unique because it differs in one of its features from all other stimuli present *anywhere* in a scene is inherently salient, and that it therefore attracts attention. As an example, consider the image in Fig. 1A showing a number of black squares and one gray square on a white background. By our hypothesis, uniqueness in color (intensity) space should make the gray square the most salient stimulus even though locally the black squares have higher contrast with the background, which generally would drive up their salience.

There have been several behavioral tests of this hypothesis, as discussed in Section 1.2, which confirm that unique stimuli are indeed preferentially attended. As discussed in that Section, it is not clear, however, whether attentional allocation in these studies is controlled by bottom-up cues, our focus of interest, or by non-specified top-down information that participants gleaned implicitly or explicitly from the task instructions.

The situation is further complicated by the fact that the saliency map models which predict eye movements and other indicators of attentional selection quite well (Borji & Itti, 2013; Jeck, Qin, Egeth, & Niebur, 2017; Masciocchi, Mihalas, Parkhurst, & Niebur, 2009; Parkhurst, Law, & Niebur, 2002) predict that uniqueness by itself does not enhance attentional deployment, Fig. 1B, C. We found that the same is the case for all computational models of attentional control that we applied to stimuli like those in Fig. 1A, see Section 1.1. It is thus not clear whether the intuition that unique stimuli attract attention is rooted in reality.

* Corresponding author.

https://doi.org/10.1016/j.visres.2019.04.004
Received 15 September 2018; Received in revised form 11 April 2019; Accepted 14 April 2019
Available online 21 May 2019

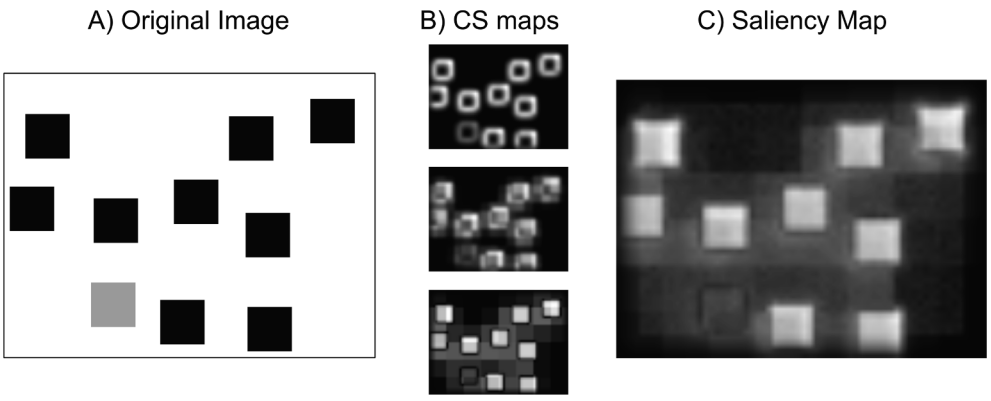

Fig. 1. Visual scene with unique stimulus and model results. (A) Simple scene with one unique stimulus. (B) Three example scales (fine to coarse, top to bottom) of center-surround (CS) responses of the saliency map model (Itti et al., 1998) to the stimulus in (A). At all scales, the gray square has the weakest response. (C) Final output of the model. Intensity represents saliency. Color and orientation channels are included in the computation but they do not make a substantial contribution for this image.

We therefore decided to find out whether our intuition, that a unique object attracts attention of human observers, is true even in the absence of any expectation or anticipation. We developed an experimental paradigm specifically designed to avoid biasing observers. The basic ideas are that (a) task instructions are kept to a bare minimum, (b) spontaneity of responses is encouraged over long deliberations, and (c) every participant performs the task only a very limited number of times. These features are designed to minimize expectations about which stimulus will be delivered next. It is impractical to measure (overt) attention by fitting subjects with an eye tracker and then have them perform a task for only a few seconds or a few trials. Therefore participants were shown a short sequence of visual scenes on a tablet computer and asked to "tap the first place you look when the image appears" (Materials and Methods). Our recent work (Jeck et al., 2017) has shown that these tapping responses are significantly correlated with other measures of attention.

1.1. Model predictions for unique objects

Consider the image in Fig. 1A, a number of black squares and one gray square on a white background. As long as the unique gray square is easily discriminated from both the background and the black squares, our intuition suggested that its uniqueness makes it the most salient stimulus. However, saliency models predict the opposite. First, we observe that only intensity contributes to saliency in this simple scene (no color *etc*). Second, the center-surround contrast of the gray square is smaller than that of the black squares since its intensity is closer to the background than that of the black squares. Therefore, the models predict that the black-on-white squares have higher saliency than the gray-on-white square. This is illustrated in Fig. 1B which shows center-surround responses to the intensity channel of the image in Fig. 1A at different spatial scales for the center and surround. For each of these center-surround computations, the gray square produces a weaker response than the black squares. Because of the lowered center-surround responses, the resulting saliency map [Fig. 1C, computed from the model in Itti et al. (1998)] assigns a lower saliency level to the unique gray square than to the black squares. Indeed, no linear combination of these center surround maps can generate a saliency map in which the gray square has a higher value than both the black squares and the background (see Appendix A).

One might expect that at large spatial scales the center-surround operation would compare the intensity of the squares with that of their neighbors, enhancing the gray square. Salience of a unique stimulus is, indeed, enhanced if the distance between this stimulus and the other stimuli is small enough that the latter are located in the surround (as defined by the model) of the former [see *e.g.* Niebur, Itti, and Koch, 2002, Fig. 4]. However, for the stimulus in Fig. 1A, even though the surround of each square at larger spatial scales includes the black squares, it also includes much of the white background. The latter dominates in all cases, resulting in a mostly-white surround for all squares. When the center-surround operation computes the difference between the center (black or gray) and the surround (mostly-white), the black squares produce a larger difference than the gray one. We hypothesize that what makes the gray square unique, and therefore salient, is the difference of intensity between it and the set of black squares. Thus, saliency is still determined by a center-surround difference but this difference is computed in "feature space", with comparisons between *objects* rather than between spatially defined *regions* of the visual field. This raises a conflict between predictions of saliency map models [e.g. Itti et al., 1998] and our intuition that the gray square is salient. The goal of this paper is to test whether human behavior agrees with our intuition, or with this and other models of salience. In Section 3.3 we propose a novel model of inter-object comparison that assigns high salience to unique objects.

Is the failure to assign higher saliency to a unique object limited to the original saliency map studies (Itti et al., 1998; Niebur & Koch, 1996), or does it affect a larger class of models? Since we believe that higher salience is assigned to unique objects because visual scenes are processed in terms of objects rather than of elementary visual features (Discussion), we were first particularly interested in models that involve the formation of perceptual objects (or proto-objects). discussed in Section 2.4. This is the case for the models developed by Russell, Mihalas, von der Heydt, Niebur, and Etienne-Cummings (2014) and Walther and Koch (2006). We therefore ran these models on the input shown in Fig. 1A, with results shown in Fig. 2A and B, respectively. Both models assign lower salience to the gray square compared to the

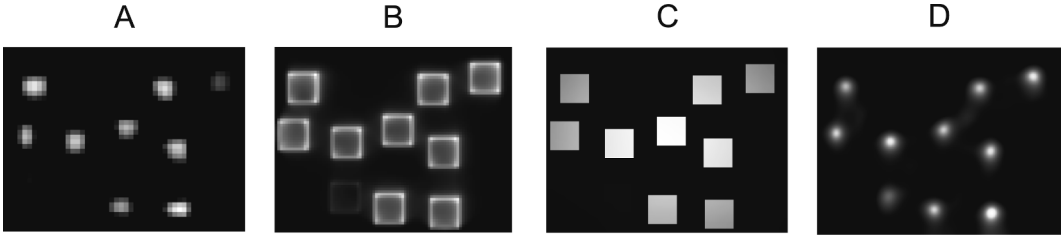

Fig. 2. Output of models of salience for the input shown in Fig. 1A with white regions indicating high salience. (A) Walther and Koch (2006) (B) Russell et al. (2014), (C) Perazzi et al. (2012), (D) Kümmerer et al. (2016).

black squares, in disagreement with our intuition.

Given that we believe it is the uniqueness of the color of the gray square that makes it stand out (or salient) in perception, we then searched for a model that is specifically designed to detect unique elements. We narrowed the choice down further by demanding that the model takes into account the one feature, namely color (of which intensity is a special case) that distinguishes the unique stimulus from all others in our images. Both of these conditions are fulfilled by a study by Perazzi, Krahenbuhl, Pritch, and Hornung (2012). In their model, an image is decomposed into compact regions of roughly similar colors, and the salience of a region is driven by the relation of two factors, the "uniqueness" and the "distribution" of a color. Colors that are far away in color space from others present in the image (such as gray in Fig. 1A) are considered "unique" and therefore salient. Colors that are spatially distributed in the image (black and white in Fig. 1A) are given a high "distribution" level, resulting in lowered salience.

The output of the Perazzi et al. (2012) model for the scene in Fig. 1A is shown in Fig. 2C. Given its design, we were surprised that it does not label the gray square as salient. We ran the Perazzi et al. (2012) model on all of the ten stimulus arrays with faint objects that we used in our behavioral experiments, described below and shown in the left panels of Figs. S3 and S4. For eight of them, the unique square was not labeled as salient. We analyzed the internal model function and found that the distribution term (*i.e.* the spatial extent of the color) was estimated to be high because the color of the unique square is close enough to the background color to cause the model to group them together somewhat. By fine-tuning one of its model parameters (σ_c), we managed to have the unique square labeled as salient for each of the ten images. However, we do not know how the modified model performs on images other than those ten for which its parameters were specifically tuned. We are also uncertain how this parameter change affects the behavior of the model on natural scenes. A further consideration is that the Perazzi et al. (2012) model is defined in purely functional terms and its relationship with what we know about early visual processing in biological systems is remote at best.

Finally, we tested our hypothesis on a model that predicts salience which was developed by Kümmerer, Wallis, and Bethge (2016). The model uses features from VGG-16, a deep network trained for object recognition. This model achieves high performance on natural scenes and is currently ranked highly on the MIT300 saliency benchmark which was introduced by Judd, Durand, and Torralba (2012). For compatibility with other tested models, we used a version with no explicit preference for objects in the center of the image. As seen in Fig. 2D, this model also assigns a lower salience to the unique object, indicating that the training of the model does not generalize well from object recognition to saliency computation.

1.2. Attention to unique stimuli: behavioral studies

Empirical data from the visual attention literature are consistent with our intuition that the gray square in Fig. 1A is salient. The gray square in Fig. 1A can be described as a singleton, which is related, in a general way to ongoing research on the attention-capturing properties of salient singletons [*e.g.* Bacon and Egeth, 1994; Ernst and Horstmann, 2018; Folk, Remington, and Johnston, 1992; Godijn & Theeuwes, 2002; Leber and Egeth, 2006; Theeuwes, 1992, 1994, 2010; Theeuwes and Van der Burg, 2011]. In these studies, subjects are given a visual task (such as reporting the orientation of a single element) while a single element in the display, which is irrelevant to the task, is different from the rest in a single dimension. A decrease in task performance when the irrelevant singleton is present is interpreted as attentional capture by the singleton, drawing attentional resources away from the assigned task. However, unlike Fig. 1A, in the singleton literature it is usually the case that the feature contrast for the distracting irrelevant singleton is stronger than the feature contrast for the relevant singleton. In order to determine whether the gray square in Fig. 1A is actually more salient than the black squares, we need to find studies where the feature contrast is actually weaker for the singleton.

In the realm of visual search experiments, Treisman and Gormican (1988) showed that search for a light gray target among dark gray distractors (on a white background) is efficient (independent of the number of distractors) when the distractor, backgroand and target colors are all easily discriminated (their Table 2). Efficient search can be explained most easily by assuming that attention is directed to the target. Bauer, Jolicoeur, and Cowan, 1996 support a similar intuition for a case with distractors of two colors and the target falling between them in color space. Nothdurft (2006) also reports on several psychophysical experiments in which the participants assess the salience of different stimuli, and report the salience of a single low-contrast target among multiple high-contrast ones.

A substantial concern with all of these experiments is, however, that all of them informed participants about the upcoming visual scene either directly, through the task instructions, or indirectly, by repeatedly presenting similar scenes (*e.g.* circular displays of dots where one or a few differ from the rest). Such "top-down" influences (which we define as being dependent on the internal state of the observer) of the task to be performed strongly affect participants' eye movements (Yarbus, 1967; DeAngelus & Pelz, 2009) and thus likely their attentional state (for more discussion see Section 4.2). Therefore, we cannot conclude from the aforementioned evidence that the measured attentional effects are driven by "bottom-up" cues (dependent on the image only). An experiment not affected by such anticipatory effects was performed by Pashler and Harris (2001) who showed that subjects selected a unique flashing stimulus over several static ones, and a unique static stimulus over several flashing ones. Most scenes do not have flashing stimuli, though, and it remains unclear whether uniqueness attracts attention in more general environments, in particular in static scenes.

2. Materials and methods

We first describe the methods used for the behavioral experiments (Sections 2.1,2.2,2.3) and then the computational models (Sections 2.4 and 2.5).

2.1. Experimental paradigm

All methods were approved by the Johns Hopkins Institutional Review Board and carriedout in accordance with the Code of Ethics of the World Medical Association (Declaration of Helsinki). In previous work (Jeck et al., 2017) we have addressed the difficulty of minimizing the contamination of behavioral measurements by top-down effects. The essential ideas from that work are that (a) task instructions are kept to a bare minimum, (b) spontaneity of responses is encouraged over long deliberations, and (c) every participant performs the task only a very limited number of times. These features are designed to minimize expectation which stimuli will be delivered next and anticipation of responses that participants might believe they are expected to provide. More specifically, we described a method of obtaining attentional responses from participants who were only minimally informed about the upcoming stimulus.

Our approach was inspired by an experiment developed by Firestone and Scholl (2014). Participants were passers-by on the Johns Hopkins University Homewood Campus. They were approached and asked if they wanted to do a quick psychology experiment on a tablet computer. If they agreed, they were shown a short sequence of complex natural scenes on the tablet and asked to "tap the first place you look when the image appears". As shown by Jeck et al. (2017), tapping responses were found to be significantly correlated with other measures of attention, specifically eye movements (Parkhurst et al., 2002), conscious selections of interesting image parts (Masciocchi et al., 2009), and the computational model of Itti et al. (1998).

In the present study, we use these methods to empirically address

the question of whether unique faint stimuli appear salient among a set of stronger stimuli, like the gray square among black squares in Fig. 1A. We used simplified images, similar to that figure, that are designed specifically to address this question by minimizing other possible influences although we believe that the effect is also present in more complex images. Instructions were identical to that described for the Jeck et al. (2017) study. The scenes we use are described in detail in Section 2.2.

All stimuli were shown on an iPad tablet (Apple Computers, iOS 8.3 operating system, screen 9.7″ with 1024 × 768 resolution). The screen occupied approximately 15°–35° of visual angle depending on how far away from the face it was held by the participant. Participant were first shown a white screen with two small black squares (see Fig. 4), which we call the initialization screen. They were instructed to tap on either one of the squares to bring up a test image, and were told "When the image appears, tap the first place you look". The image then appeared and, after the participant had tapped his or her selected location on it, the position of the tap and the reaction time (time between this tap and the tap on the initialization screen) were recorded using the pixel and time given by the operating system. Test images strictly alternated between a natural scene and a scene of colored squares, see below. Each participant saw a total of 12 images of which the first always was a natural scene.

We recorded 1512 taps on simple scenes from 252 participants (101 male, 151 female). Population results are shown in Fig. 6. Reaction time (RT) was defined as the time from tapping on the initialization screen to tapping on the test image. Median RT was 1.3 s, the mean was 1.4 s. We did not analyze RTs in detail because our data collection system (iPad) did not allow control of the exact timing of image presentation.

2.2. Stimuli

The stimulus set consisted of 30 images, with each showing a set of colored squares on a white background. We refer to these as "simple scenes" to distinguish them from the natural scenes that were presented in alternation with them. As mentioned, responses to the natural scenes have been analyzed previously (Jeck et al., 2017); for the purpose of the current study they only serve to separate the simple scenes and to possibly reduce the predictability of the image sequence. On each of the simple scenes, the screen was divided evenly into a 5 × 3 grid. In ten of the grid locations (randomly chosen) a square (120 × 120 pixels) was placed. Each square was placed at a random location (uniform distribution) within the central 80% of the grid cell in the horizontal and vertical directions. This placement pattern spaced out the squares evenly on average without creating a percept of a predictable pattern. The color of the squares varied among the six square image types generated, Gray/Black, All-Black, Black/Gray, Blue/Yellow, Pink/Red, and All-Red; an example of each is shown in Fig. 5. Five images were generated for each image type. Each of the All-Black images was identical to one of the Gray/Black images except that the color of the single gray square was changed to black. This design allowed for a direct comparison between the gray square and the corresponding black square since the geometries of one All-Black and one Gray/Black image were identical. Likewise, each of the All-Red images was identical to one of the Pink/Red images except that the pink square was changed to red. Otherwise, all images were independent of each other. All computational models we tested predict (Section 1.1) that in a pair of Gray/Black and All-Black images with the squares at the same positions, the black square in the All-Black image at the same position of the single gray square is more salient (see Fig. 1C) and therefore should be tapped more than the gray square. The analogous argument applies to the All-Red and Pink/Red pairs of images.

The simple scenes were separated into five groups of six, with each group containing one image from each type. Each participant saw exactly one Gray/Black image, and one matched All-Black image was always shown to the same participant. Likewise, each participant saw one Pink/Red image and its matched All-Red image. Images were presented in randomized order, with the constraint that the first simple scene of a matched pair was always chosen such that each of the two members of a matched pairs of images had an equal number of participants see it first. For instance, the number of participants that saw the first Gray/Black image was the same as that of participants that saw the matched All-Black image as their first simple scene. This allowed us to perform a direct comparison of data gathered from the first time a participant saw a simple scene with matched sample sizes.

2.3. Statistics

Data from the first view were analyzed using a one-tailed Fisher's exact test. Data from all views were analyzed using paired one-tailed t-tests. Significance of the fits for center bias and intercepts in Fig. 6 was assessed using F-tests, the latter with False-Discovery Rate correction (Benjamini & Hochberg, 1995) to control for multiple comparisons. The significance level for all statistical tests was set to $\alpha = 0.05$.

2.4. Proto-object comparison model: overview

One strategy humans and other animals use to cope with the complexity of their environments is to transform raw sensory input into representations that match more closely the functional relationships in the world. In the visual and auditory modalities this process is called perceptual organization (Bregman, 1990; Kimchi, Behrmann, & Olson, 2003). In the visual system, the fundamental units of this representation are no longer activity levels of retinal ganglion cells but their correlated patterns that correspond to visually perceived objects. We and others have developed quantitative models to understand the underlying computations (Ardila, Mihalas, von der Heydt, & Niebur, 2012; Craft, Schütze, Niebur, & von der Heydt, 2007; Hu, von der Heydt, & Niebur, 2015; Brian, Kane-Jackson, & Niebur, 2016; Hu et al., 2017; Mihalas, Dong, von der Heydt, & Niebur, 2011; Pentland, 1986; Ramenahalli, Mihalas, & Niebur, 2014; Russell et al., 2014; Walther & Koch, 2006). As was observed by Rensink (2000) and Zhou, Friedman, and von der Heydt (2000), perceptual organization does not require the formation of fully-formed objects as would be needed for tasks like object recognition or discrimination. Instead, it is sufficient that the scene is segmented into entities that are characterized by a few elementary features, like their position, size *etc*. Following Rensink (2000), we call these entities proto-objects. For the sake of simplicity, we use the terms "object" and "proto-object" interchangeably.

Electrophysiological studies (Martin & von der Heydt, 2015; Qiu & von der Heydt, 2005, 2007; Williford & von der Heydt, 2016; Zhou et al., 2000) show that "ownership" of object borders is represented by the firing rate of individual neurons in (mainly) extrastriate visual cortex. Computational models (Craft et al., 2007; Hu et al., 2015; Hu et al., 2016; Mihalas et al., 2011; Russell et al., 2014; Wagatsuma, von der Heydt, & Niebur, 2016) show that these results can be explained by a population of "grouping" cells that represent in their firing rates the proto-objects underlying the model in the present study (their name stems from their role of binding, or grouping, the different features of proto-objects). Grouping neurons receive input from striate and early extrastriate cortex and, in turn, modulate activity of some neurons in these areas. It is this modulation which imparts border-ownership selectivity on these "border-ownership selective" neurons. Response properties of grouping neurons are quite simple, in the simplest form their activity represents the presence of "something" within a certain size range at an approximate position. Therefore, any single grouping neuron can not represent objects with complex shapes. Our working hypothesis is that such objects are represented in the population activity of grouping cells. A modeling study (Ardila et al., 2012) demonstrated that visual activation by complex geometrical shapes results in an activity pattern at the grouping cell level that reproduces the output of the medial axis transform, an abstract representation of

complex visual shapes commonly used in computational vision (Blum, 1973; Feldman & Singh, 2006). The purpose of this early proto-object representation is only to provide structure to the visual input. Qiu, Sugihara, and von der Heydt (2007) showed that perceptual organization, or at least the part that manifests itself in the form of border ownership selectivity, is a pre-attentive process on which other mechanisms can build, *e.g.* object recognition, attention to objects, or navigation.

Our behavioral results (below) suggest that humans compute relative saliency of simultaneously present (proto-) objects by comparing the features of these objects, rather than properties of regions that are defined by simple spatial relationships, as in center-surround contrast computations. To understand these behavioral results, we develop a computational model of visual saliency based on comparisons between objects. The model generalizes the idea that objects that differ from other objects are salient, while repeated objects are not salient. While in early models (Itti et al., 1998; Itti & Koch, 2001; Koch & Ullman, 1985; Niebur & Koch, 1996) the elements to be compared were defined purely spatially, newer approaches are based on proto-objects (Russell et al., 2014; Walther & Koch, 2006). However, as discussed previously (Section 1.1), these models cannot explain that humans assign higher salience to unique objects over repeated objects.

To obtain a representation of proto-objects in a visual scene, we compute grouping cell activity as in the Russell et al. (2014) model but we remove the normalization procedure that follows in that model and replace it by a normalization that takes into account other stimuli anywhere in the scene. Grouping cells tile the entire image with overlapping proto-objects of many different radii, and are computed on different submodalities (intensity, color, and orientation). A grouping cell in the Russell et al. (2014) model will have a strong response if it is at the center of a set of co-circular edges at the preferred radius of the grouping cell. Grouping cells in our model have a minimum preferred radius of 32 pixels and a maximum of 256 pixels. The model is illustrated in Fig. 3 and defined formally in Section 2.5. Here we propose a simple color-based model, as it is sufficient to explain the data collected. However, a number of other features (*e.g.* shape, orientation, etc.) could easily be added to our normalization procedure.

2.5. Proto-object comparison model: formal definition

Proto-objects in our model are defined by their position (X, Y) and radius (r). We do not, however, assume a binary distinction between the presence and absence of proto-objects at any location. Instead, the activity of grouping cell responses in the Russell et al., 2014 model provides a graded measure of the "belief" that a proto-object with a specific radius is present at a specific location. Let (X_i, Y_i), and r_i represent position and radius for the i-th proto-object. We define its strength S_i as the product of r_i^2 with the i-th grouping cell response. Since proto-objects are calculated by contrast-based mechanisms, the S values of proto-objects representing the gray square in Fig. 1A are lower than those of black squares. An example of a set of grouping cell responses with a radius of 32 pixels to the stimulus in Fig. 3A is shown in Fig. 3B.

In order to compare between proto-objects in our new normalization step below, we must first compute a set of features for each proto-object. For each proto-object, the proto-object comparison (POC) model computes features over the image region defined by the circle with center at (X_i, Y_i) with radius r_i (see 3C for an example). We compute histograms of L, a^*, and b^* values from the CIELAB color space (Ibraheem, Hasan, Khan, & Mishra, 2012). These color dimensions have been found to be represented in a number of visual areas (Brouwer & Heeger, 2009; Conway & Tsao, 2009; Li, Liu, Juusola, & Tang, 2014). Each of these histograms has nine bins and is normalized to sum to 1 so that patches of different radii can be compared appropriately. We also compute histograms with nine bins for the potential radii of proto-objects in the patch. Eight of the bins have the value zero and the one which corresponds to the actual object radius having the value unity. For the i-th proto-object, this gives us a feature vector F_i whose components are the values of 36 different bins (nine bins for each of the four features L_i, a_i^*, b_i^*, and r_i). We refer to the value for the i-th proto-object in the j-th bin as F_{ij}. In the brain, we presume that these features are computed simultaneously with the computation of proto-objects themselves. The entries in the histograms correspond to activity patterns of separate neuronal populations that are tuned to the features represented by the histogram bins. We chose those features in our model partly for reasons of computational efficiency rather than as

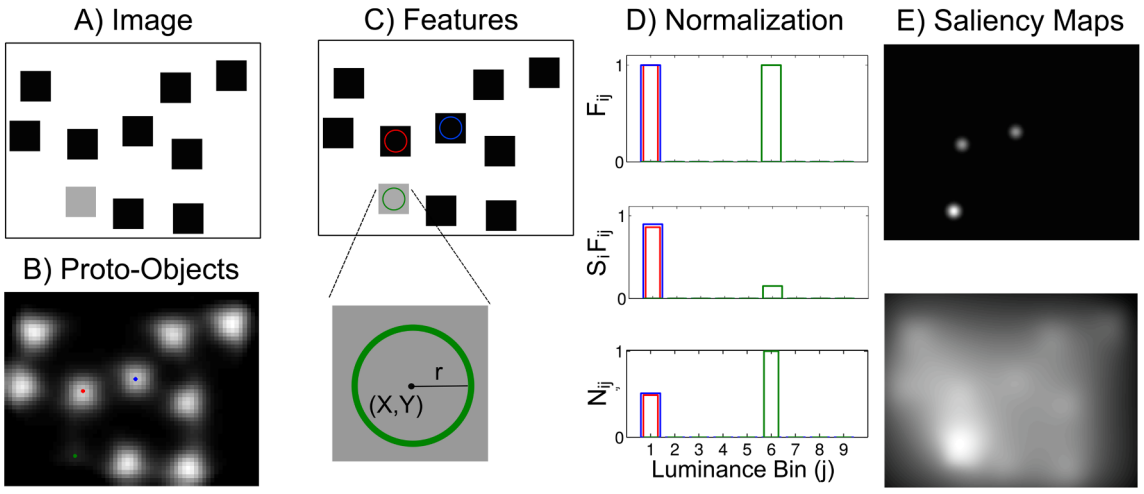

Fig. 3. Simplified illustration of the POC model. (A) Example input image. (B) Map of the strengths of proto-objects with a radius of $r = 32$ pixels only. Three colored dots correspond to proto-objects used in the following panels. Note the weak response to the gray square, centered on green dot. (C) Illustration of the feature computation for the proto-object highlighted by the green dot in B. Histograms are formed over the pixels within the circle of radius r around the center (X, Y) of the proto-object. Note that all pixels are gray, which corresponds to the sixth luminance bin in panel D. (D) Top panel, feature representations for three proto-objects. Two are on black regions (red and blue dots in B, overlapping red and green histogram bins in the 3 panels of D), these proto-objects have identical colors (black) and their values in the histogram are identical, unity in the first histogram bin and zero in all others. The third is on the gray region of the image (green dot in panel B, zoom-in in Panel C, and green histogram bin in D), it has unity value in bin 6 and zero elsewhere. Middle panel, feature values multiplied by proto-object strength S. Bottom panel, strength after normalization. (E) Top panel, saliency output for the three proto-objects processed in D. Output for all other proto-objects omitted for clarity. Bottom panel, output for all proto-objects. Note the enhanced saliency of the gray square in D, E. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

detailed implementations of biological processes. For instance, we do not claim that there are neuronal populations that one-by-one code L, a^*, and b^* values from the CIELAB color space but we do believe that color is represented explicitly in neuronal activity patterns.

Interaction between proto-objects is introduced through a normalization process,

$$N_{ij} = \frac{S_i F_{ij}}{\sum_k \sum_k F_{kj}} \tag{1}$$

if $\sum_k S_i F_{kj} \neq 0$, otherwise $N_{ij} = 0$. In this equation, the value in each histogram bin is multiplied by S_i so that strongly detected objects are given more weight in the normalization process, and strongly detected objects with the same feature values will suppress one another. This is illustrated in Fig. 3D for three proto-objects: two with high S_i that share the same feature values and one with a lower S_i value that is unique. The strength of each proto-object is then computed as

$$P_i = \sum_j N_{ij} \tag{2}$$

and a saliency map Q is defined as a sum of Gaussians with weight given by P_i and locations given by the proto-object.

$$Q(x, y) = \sum_i \sum_i \exp\left(-\frac{(x - X_i)^2 + (y - Y_i)^2}{2\sigma_i^2}\right) \tag{3}$$

where the spread $\sigma_i = \frac{r_i}{2}$ of the Gaussian ensures that most of a proto-object's activation is near its center. A saliency map for the three proto-objects and the full output using all proto-objects is shown in Fig. 3E.

We noted that the saliency maps generated by the POC model are blurrier than the Russell et al. (2014) model, which has in the past correlated with improved performance on natural scenes (Judd et al., 2012). To illustrate the importance of the normalization process rather than the other implementation details of the model (*e.g.* the creation of the saliency map using a sum of Gaussians), we also generate a saliency map using Eq. (3) but replacing P_i with S_i. Note that this is equivalent (up to a scaling factor) of computing Eq. (1) without the denominator, since $\sum_j F_{ij} = 1$. We found that the average R for this modified model was 0.464, similar to the values found for the non-modified model (Results).

3. Results

3.1. First view only

Participants were approached on a campus of Johns Hopkins University and asked whether they were willing to do a quick psychology experiment on a tablet computer. By tapping on the screen, they made appear an alternating sequence of natural scenes and "simple scenes" consisting of colored squares (Methods; see Fig. 4). They were instructed to tap with a finger of their choice on the first place they looked at in the scene. Each participant saw a total of 12 images of which the first always was a natural scene. For the purposes of this study, only responses to simple scenes are analyzed, natural scenes only served to minimize potential interactions between subsequent simple scenes. One class of simple scenes consisted of one unique gray square among several black squares, as in Fig. 1A. Each observer who saw one of these scenes also saw an identical one in which the unique gray square was replaced by a black square, Fig. 5A. In this case existing models predict that the black square should be more salient than the gray square due to its higher contrast with the background despite having the uniqueness property removed. Other patterns of uniqueness (or its absence) were also created with other colored squares, Fig. 5B, C.

We obtained the main result of this study by analyzing tap locations for the very first trial on which a subject was presented with a simple scene (this was always the second trial since the first was a natural scene). Example stimuli with all taps shown as overlaid green dots are in Fig. 5 (for first taps only see Suppl. Figs. S1 and S2 in Appendix B). Participants tapped the unique ("singleton") square (gray or pink) significantly more frequently than the matched square on a control image (black or red). This result holds both for a gray square among black squares (Fig. 6A) and for a pink square among red squares (Fig. 6C). In the former case, we observed 14 taps on the gray square taps *vs.* 6 taps on the black square, both out of 51 taps. A one-tailed Fisher's exact test gives $p = 0.039$. In the latter case, we observed 16 taps on the pink square *vs.* 8 on the red square, both out of 47 taps. A one-tailed Fisher's exact test gives $p = 0.048$. Note that in both cases, the local contrast of the unique object to the background is lower than that of the non-unique objects which is what leads the computational models astray. Also note that the black among gray and the blue among yellow trials do not speak to the issue of whether a faint object attracts attention.

This result confirms the intuition that a unique stimulus is more salient than a non-unique stimulus, even if the latter has higher contrast to the background.

3.2. All presentations

In the next analysis, we studied the relative tapping rates of all six simple scenes a given participant saw, averaged over all participants. Each image had a different singleton tap rate, defined as the fraction of times that participants tapped on the singleton square. We did not enforce that the same number of participants saw the same scene as their first, second, or n-th simple scene. Therefore, it is difficult to quantify whether the location of an image in the image sequence influences tap rates. For the remainder of our analysis, we therefore aggregate data over all six presentations of simple scenes.

Including all six views by each participant, gray squares are tapped significantly more than the black squares in corresponding positions,

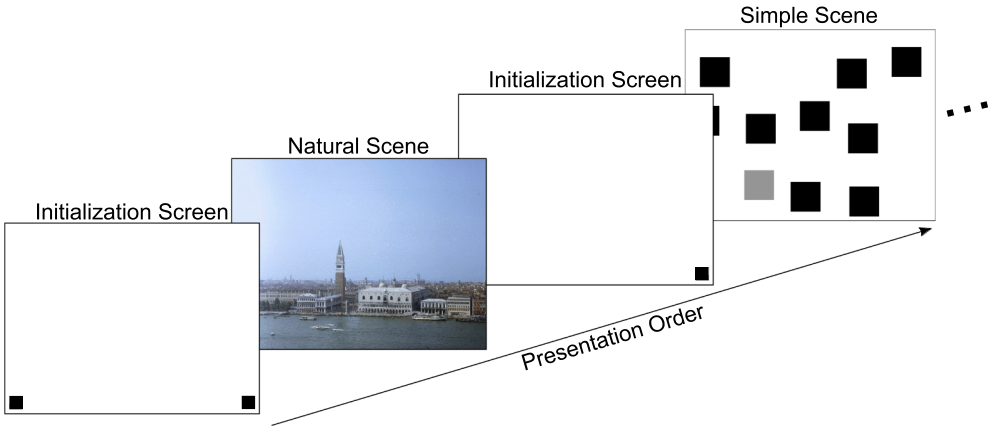

Fig. 4. Experimental procedure. Participants first saw an initialization screen and tapped on one of the small black squares at the bottom. This brought up a test image (alternating natural scenes and simple scene). They then tapped on it at a place of their choosing which was, according to instructions, the first place they looked at when the image had appeared. Tapping position and reaction time were collected, the initialization screen reappeared, and the cycle re-commenced. (Natural scenes were shown in full color, see the web version of this article.)

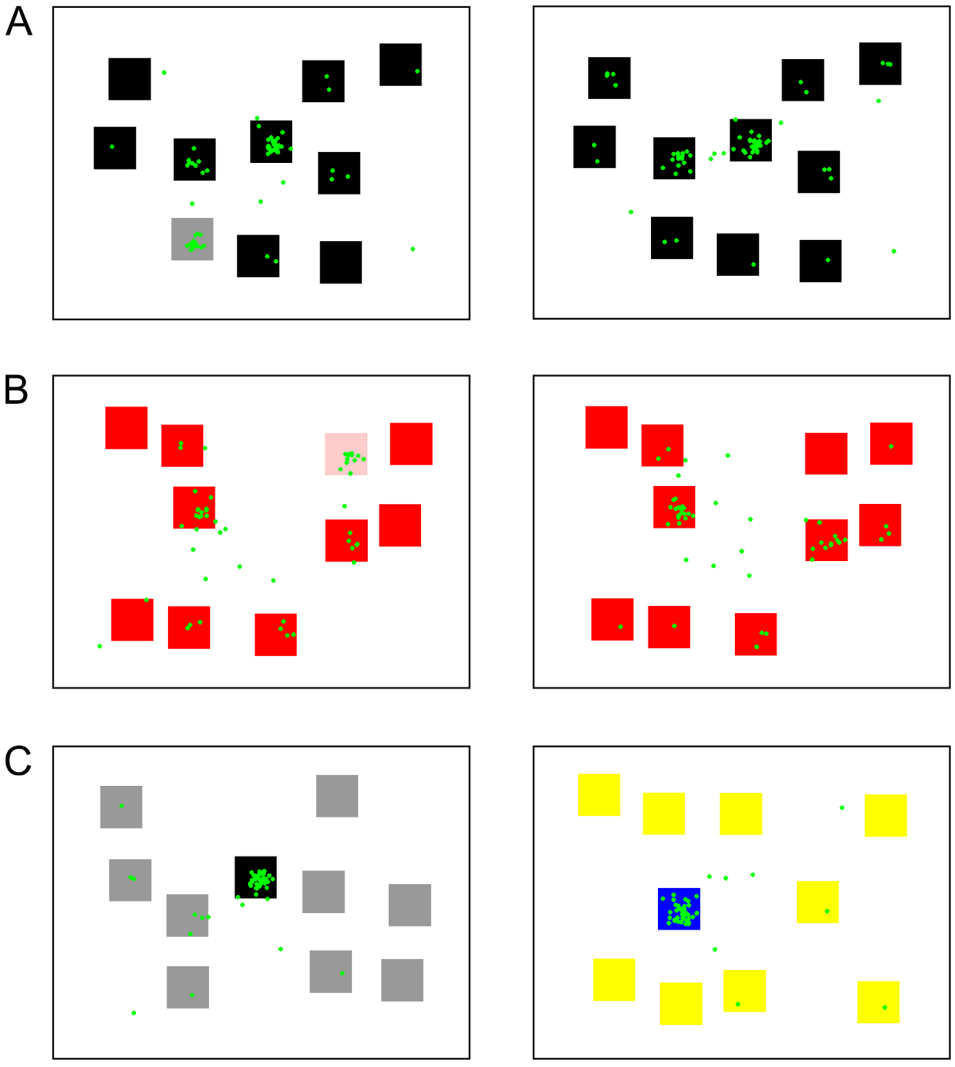

Fig. 5. Example set of simple scenes, overlaid (green dots) with tap locations of all participants who saw this set. All taps shown are in response to the first time participants saw these scenes. (A) Gray/Black (left) and corresponding All-Black (Right) images. Note that the corresponding image has an identical spatial arrangement of squares. (B) Pink/Red (left) and corresponding All-Red (right) images. Again, the spatial arrangement is identical. (C) Black/Gray (left) and Blue/Yellow (right) images. The Black/Gray and Blue/Yellow images had independent spatial arrangements. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

with 90 taps on gray squares *vs.* 21 on black squares, out of 252 taps in both cases ($p < 10^{-14}$ paired t-test; Fig. 6B). The same holds for pink squares *vs.* corresponding red squares, 101 taps *vs.* 59 out of 252 ($p < 10^{-5}$; Fig. 6D). Note that a paired t-test is appropriate in these cases because the same participants saw paired Gray/Black and All-Black images on different presentations (and the same for Pink/Red and All-Red images). More detailed analysis shows that direct comparisons between Gray/Black and All-Black images were significant individually for each of the five pairs (Suppl. Fig. S3), as well as for three out of the five pairs of Pink/Red and All-Red images (Suppl. Fig. S4). For the two images without significantly increased tap rates on the pink squares, the corresponding red square was the most tapped square on the All-Red image and in both images it was located close to the center of the scene. A ceiling effect, likely due to the geometrical arrangements of stimuli (center bias, see next paragraph), may thus be the reason why we did not find a significant effect in these cases.

As in previous studies (Buswell, 1935; Parkhurst et al., 2002; Tseng, Carmi, Cameron, Munoz, & Itti, 2009), we found a strong center bias in our results. Fig. 6E shows the rates at which participants tapped on a singleton square as a function of the square's Euclidean distance from the center of the image. The lines in the figure are generated from a linear regression model where each type of stimulus and the distance from the center are used to predict the tap rate. Also shown are the tap rates and linear fits for the non-singletons in the All-Black and All-Red images. A significant effect of distance from the center was found for each line (F-test, all $p < 10^{-5}$). The negative slope of all lines confirms the existence of a center bias in all conditions. Interaction terms between the distance from the center and the stimulus type were not found to be significant except in the case of the non-singletons.

By analyzing the intercepts of the fit lines (Fig. 6F) we can roughly gauge the salience for the different image types independently of the center bias. By performing pairwise comparisons between the intercepts, we found that the Blue/Yellow intercept was significantly higher than the Gray/Black and the Pink/Red intercepts (F-test, all $p < 0.05$), and intercept of the fit line for non-singletons was lower than for any of the images containing singletons (all $p < 10^{-11}$). These results held after performing a False-Discovery Rate correction (Benjamini & Hochberg, 1995) to control for multiple comparisons. We also found that the singletons in Black/Gray and Blue/Yellow images are generally more salient than either the singletons in Gray/Black or singletons in Pink/Red images. These results agree qualitatively with previous search asymmetry studies (Treisman & Gormican, 1988) since the Gray/Black singletons are less salient than the Black/Gray singletons, (Fig. 6E) while confirming that the singleton gray squares in the Gray/Black images can still be salient.

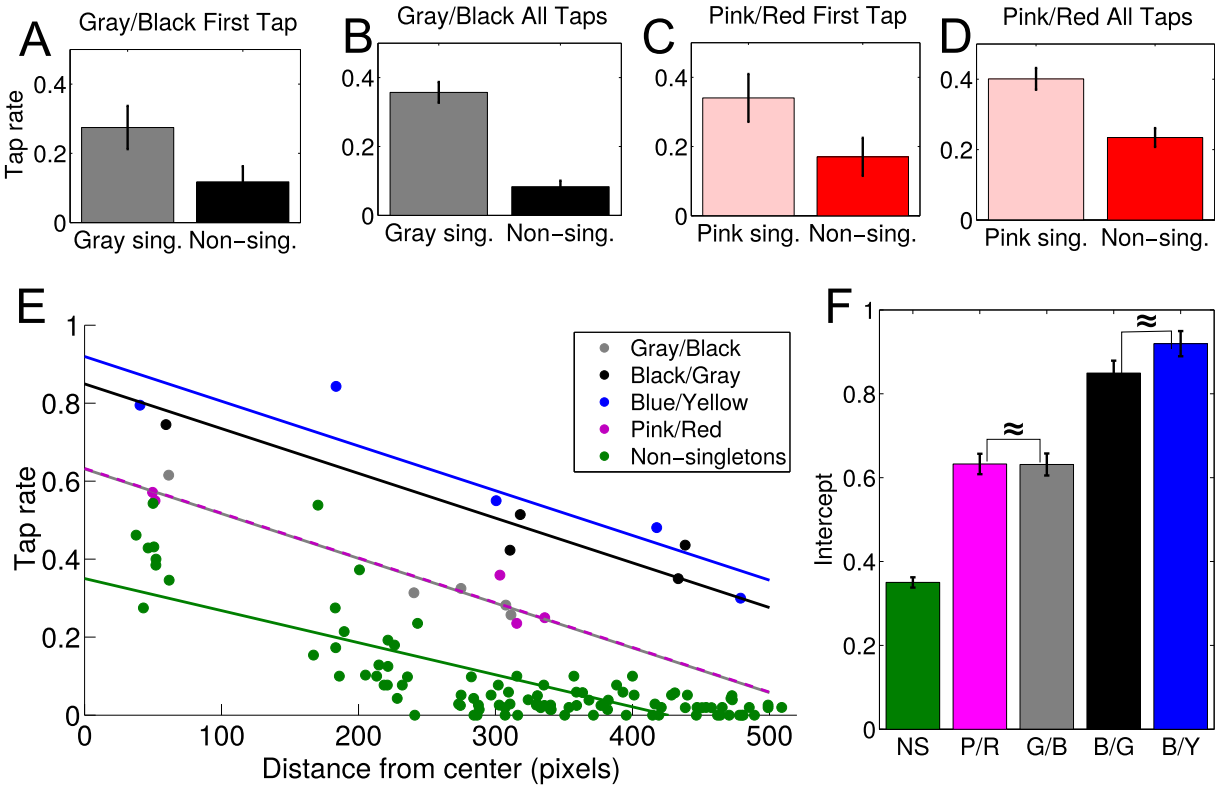

Fig. 6. (A–D) Rates at which participants tapped on singleton (sing.) squares *vs.* non-singleton corresponding squares (Non-sing.) in a control image. Error bars represent standard error. (A) Gray/Black *vs.* All-Black comparison for each participant's first tap. (B) Gray/Black *vs.* All-Black comparison for all taps. (C) Pink/Red *vs.* All-Red comparison for each participant's first tap. (D) Pink/Red *vs.* All-Red comparison for all taps. (E) Rates at which participants tapped on the singleton squares (colored circles, see legend), and each of the various non-singleton squares in the All-Red and All-Black images (green circles). The horizontal axis is the Euclidean distance from the center of the image. Fit lines were generated for each singleton image type individually and for Non-singletons combined, colors same as for the corresponding circle symbols. (F) The vertical intercept of each fit line from (E) with standard error bars (G/B = Gray/Black, P/R = Pink/Red, B/G = Black/Gray, B/Y = Blue/Yellow, NS = Non-singletons). The symbol ≈ indicates that no pairwise difference was found ($p \geqslant 0.05$). All other intercept pairs differed significantly ($p < 0.05$). (For a full colour version of this figure, the reader is referred to the web version of this article.)

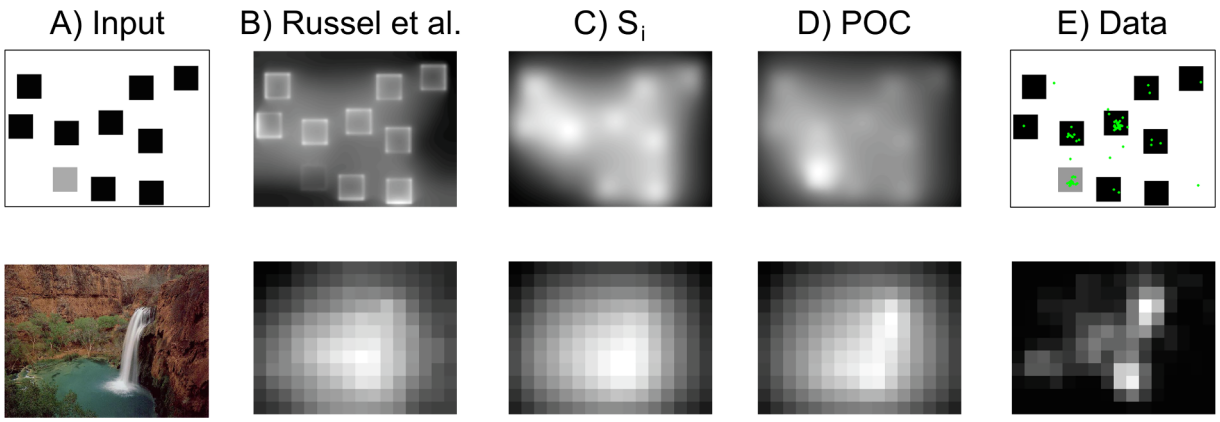

Fig. 7. Model behavior on an example simple scene (top row) and natural scene (bottom). (A) Input image. (B) Output of the Russell et al. (2014) model. (C) Output of the model without normalization, using S_i instead of P_i in Eq. (3). (D) Output of the POC model. (E) Image with tap data (top; green dots) and fixation map (bottom) overlaid. For the natural scene saliency and fixation maps are downsampled to the 12×16 resolution used by Jeck et al. (2017). (For interpretation of the references to colour in this figure legend and a full colour version of the lower panel of (A), the reader is referred to the web version of this article.)

3.3. Inter-object competition in feature spaces

Our behavioral results suggest that humans compute relative saliency of simultaneously present (proto-) objects[1] by comparing the features of these objects, rather than properties of regions that are defined by simple spatial relationships, as in center-surround contrast computations. To understand these behavioral results, we use the Proto-Object Comparison (POC) model defined in Sections 2.4 and 2.5. For each of the simple scene stimuli shown to participants, the POC model was able to predict that the unique object was the most salient object in the image (see Fig. 7, top row for an example). This held for all stimuli shown in Suppl. Figs. S3 and S4, even without accounting for center

[1] We refer to "proto-objects" rather than to "objects" because our focus is on low-level perception which does not require many of the properties of an object, see Rensink (2000) and Zhou et al. (2000); see Section 2.5 for more details.

bias. It would be simple to add such a bias (Parkhurst et al., 2002) but since the POC model already has perfect performance in this respect, it seems besides the point to add this modification.

We also wondered whether the POC model is able to maintain the same level of performance on natural scenes as the Russell et al. (2014) model on which it is based. We tested the output of the POC model on images where we had previously recorded fixations to generate saliency maps, and used the Pearson correlation R between fixation maps and these saliency maps as our measure of performance (see Jeck et al., 2017 for details). Average R over the 100 images we tested [the same as used by Russell et al. (2014)] was 0.484 for the POC model, actually slightly higher than the value of $R = 0.472$ for the Russell et al. (2014) model.

4. Discussion

4.1. Simple scenes and models of salience

Using the novel tapping paradigm we have shown that participants preferentially report that the first place they look is at a unique (singleton) object, even when that item is faint (low contrast) compared to other items in the display. This suggests that subjects have a default set in place when they view an unknown scene. Default sets have been discussed in general terms [see, *e.g.* Folk, Remington, and Johnston, 1993]. As mentioned in Section 1.2, our results are similar to those in the ongoing research on the attentional capture of singletons [*e.g.* Bacon and Egeth, 1994; Folk et al., 1992; Leber and Egeth, 2006; Theeuwes, 1992, 2010; Ernst and Horstmann, 2018]. However, those approaches differ in two important ways from the present study. First, in the capture literature it is typically the case that subjects have an explicit task such as finding a singleton on one dimension, while an irrelevant salient singleton is present or not. Thus competing attentional sets may be in play. Second, the feature contrast for the distracting irrelevant singleton (e.g., a green item when the others are all red) is typically stronger than the feature contrast for the relevant singleton (e.g., a diamond shape in an array of circles). Indeed, when the discriminabilities of the shape and color dimensions are adjusted appropriately such that form is easier to discriminate than color, then search for a form target is not slowed by the presence of a unique color, while search for color is affected by the presence of a unique form (Theeuwes, 1992).

It was to be expected, and is predicted by quantitative models of salience computation, that singletons that have high contrast relative to all other scene elements including the background are attended preferentially. We show that this is the case for intensity contrast (black square surrounded by white background and gray squares) and for color contrast (blue square surrounded by white background and yellow squares). In contrast, all computational models we tested predict the opposite outcome for a "faint" singleton object: As long as the singleton is well-isolated from other objects so that local center-surround differences incorporate substantial input from the background, models assign low salience to a singleton with lower contrast against the background than other objects in the image (gray or pink squares surrounded by white background and black/red squares). We show that humans select these singletons over otherwise identical stimuli that are not singletons. As long as the singleton, background, and other objects are sufficiently far apart from each other in color space, the singleton will be preferentially selected.

While previous research in the visual search and psychophysical literature has arrived at similar conclusions about salience (Bauer et al., 1996; Nothdurft, 2006; Treisman & Gormican, 1988), the participants involved in those studies were either explicitly informed about the nature of the images being presented, or they performed enough trials that they likely expected certain types of images. It is therefore not clear to what extent responses influenced by systematic top-down effects rather than controlled by the perceptual qualities of the visual stimuli. We therefore developed an experiment in which participants received minimal information on the stimuli. Our results show significantly increased salience of unique objects even for the very first time a participant sees a scene.

4.2. Top down influences and possible experimental confounds

Our experimental paradigm makes it feasible to measure behavioral responses from a large number of participants while minimizing expectations of stimulus contents they may have or develop during the experiment. Nevertheless, given the complexity of the human mind, it is extremely difficult to completely exclude top-down influences. In the following, we discuss potential remaining top-down effects that may have biased our results, from the more generic to the more specific.

Our experimental paradigm certainly does not remove all top-down influences on attention. Participants' behavior will naturally be affected by outdoor distractions, their internal state *etc*. However, we contend that removal of (explicitly or implicitly provided) information about the visual stimuli removes those top-down influences that are specific to the images they see, leaving only those of a generic nature that are independent of the images. In our experiment each image with a faint singleton square is paired with another one that is identical in all respects except that the singleton is replaced by a distractor square. There is no reason to assume that top-down influences due to generic distractions *etc* are different between the two images of a pair. Therefore, differences in internal state of a general nature cannot explain our results.

A different, rather pessimistic interpretation of our results is that the participants were priming themselves to look for unusual objects because they knew that they were participating in a psychology experiment, and that they responded in a way that they believed the scientist wanted them to respond ["demand bias;" (Firestone & Scholl, 2016)]. It has, indeed, been found that participants in psychology experiments will modulate their responses based on what they believe the purpose of the experiment is. For example, Durgin et al. (2009) showed that participants will give a higher estimate of a slope to be scaled while wearing a heavy backpack if they infer that experimenters expect that the weight will influence their judgment, compared to a situation where they carry the same backpack but believe its weight is irrelevant for estimating the slant (because they are told that it contains measurement equipment). See also Brown (1953) for an earlier account of experimental design affecting subjective assessments. We acknowledge the possibility that subjects tap the unique object because they think the experimenter wants that response, though it does not seem probable. Our experiment was designed specifically to minimize this effect which, if present, should be much more prevalent in the cited previous behavioral studies of the salience of faint objects. In our experiment, a significant effect is observed when participants respond to the very first singleton image they ever see (after one other image showing an unrelated natural scene), with the response given within about one second. It seems highly unlikely that participants come to a conclusion of what the experimenter expects from them in literally a split second without any additional information but the image itself. In addition, the fact that our previous results (Jeck et al., 2017) showed significant correlations between behavior in this task and several measures of saliency strongly suggests that taps do occur on salient stimuli. Furthermore, in informal debriefing of participants after the experiment, none of the participants asked if they were supposed to tap on the singletons. All this supports the interpretation that our results are not due to demand bias or similar effects.

Another criticism may be that the participants had enough time viewing the image to engage in a mixture of top-down and bottom-up processing. Under this view, the fact that the participants have a reaction time greater than a second is a serious design limitation. Rather than their attention being drawn immediately to the most salient stimulus and then reporting it, during that amount of time the participants

may saccade to multiple locations and modify their choice of where they report they first look based on a higher-level interpretation of the scene. While it is true that the median reaction time of 1.3 s would theoretically allow several fixations, this does not take into account the time needed for making a controlled hand-movement to a specific location in a task executed without any previous training, performed in a casual environment (standing on a walkway on campus), and without encouragement for a rapid response. We consider it likely that most of the 1.3 s long period between the time the image was presented and the finger reached the tablet surface was devoted to motor planning and actual limb motion. We also note that this criticism would likewise apply to fixation data which is typically gathered over several seconds of free viewing per scene (Borji & Itti, 2013).

Finally, we briefly address two concerns that are not related to top-down influences. The first concerns the scene gist hypothesis [for Review see Oliva, 2005], stating that the overall structure of the scene may begin to have an effect on attention immediately (within $\approx$ 100 ms). This is not a top-down effect by our definition since the gist is a property of the scene, rather than of the internal state of the observer. It would be extremely difficult to separate this effect from guidance of attention to salient stimuli in any experiment. In fact, if the gist of our singleton scenes can be described by "several similar objects and one dissimilar object on a homogeneous background" it would even conceptually be difficult to distinguish its effect from salience-driven guidance of attention to the singleton. Both explanations may simply be different descriptions of the same underlying process.

One possible methodological concern may be that participants may not follow our instructions to "tap the first place you look". Indeed, we do not control whether they do but we see this formulation rather as a non-technical way to instruct participants to indicate where they are attending than as an action that needs to be followed precisely. Our interest is to assess where attention is deployed, rather than finding a precise surrogate for eye movements. Pointing with a finger is a very natural and universal human behavior (Kita, 2003) which already appears during infancy, at about one year of age (Leavens, Hopkins, & Bard, 2005; Tomasello, Carpenter, & Liszkowski, 2007). The purpose of finger pointing is typically to direct attention (either that of another person or occasionally of the pointing person him or herself) towards a specific part of the world. This behavior is thus a direct, voluntary expression of attentional selection.

4.3. Object-based models

Regardless of interpretation, a model that would capture the observed behavior must rely on a computation more advanced than spatially defined center-surround operations. A natural step in this direction is the formation of proto-objects by grouping together low level features of the same type. A computational model by Perazzi et al. (2012) breaks an image into compact color patches of approximately equal size and assigns heightened salience to a patch of unique color. However, it does not take into account that most cells in early primate cortex are orientation selective and, more importantly, it fails to assign high salience to the unique objects in our stimuli (see Section 1.1). More biologically realistic models are based on explicit representations of proto-objects. Proto-objects offer a representation that has more fidelity to the physical world, with distinct objects of widely varying size occupying consistent locations in visual space. Such representations are more behaviorally useful than color patches, as predictable changes in the visual scene could be encoded for a proto-object but not for a color patch, which may arbitrarily break up an object over multiple patches. Russell et al. (2014) define proto-objects based on the strength and organization of edges in the image. Walther and Koch (2006) identify the submodality with the highest contribution at the peak of the saliency map (Itti et al., 1998) and define proto-objects by spreading activation in this submodality around the saliency peak. Both models are unable to capture the observed behavior because they do not perform any comparison between proto-objects (Section 1.1). Our new model implements this competition in feature space and we show that it is in agreement with human behavior.

4.4. Future work

This work opens the possibility for a number of avenues for further research. Despite the existence of a vibrant literation on the capture of attention by singletons (Bacon & Egeth, 1994; Ernst & Horstmann, 2018; Folk et al., 1992; Godijn & Theeuwes, 2002; Leber & Egeth, 2006; Theeuwes, 1992, 2010; Theeuwes & Van der Burg, 2011), the nature of the underlying mechanisms is still unresolved. Perceptual grouping, spatial organization and feature similarities all likely play a role by themselves or in combination (Belopolsky, Zwaan, Theeuwes, & Kramer, 2007). A question to answer is the spatial extent over which comparisons between objects influence their relative salience. Similarly, the details of the comparison in feature space have not been rigorously determined. Space and feature dimensions can also be combined, as in the normalization theory of attention (Reynolds & Heeger, 2009). Our model assigns color into bins rather than doing a continuous comparison across color space, which may also provide a plausible explanation of the data. Either of these would likely necessitate their own studies. Furthermore, the addition of features beyond color would open up the possibility that multiple features would conflict with each other.

5. Conclusion

We show that uniqueness makes objects perceptually salient, even when uniqueness is pitted against features that by themselves suppress their saliency. These results cannot be explained by top-down influences on behavior (Section 4.2). Existing computational models do not capture this fundamental behavior but a model that includes competition in (non-local) feature space agrees both with intuition and with observed human behavior. Our new model points to the importance of inter-object comparisons when predicting human behavior and our findings demonstrate the utility of the tapping paradigm in testing models of visual attention. These results are of wide-ranging importance for understanding human behavior, as they have theoretical, methodological, and practical implications. They may also be of interest to the human factors community in the design of visual interfaces.

Data and code availability

Data is made freely available at https://github.com/dannyjeck/Proto-Object-Comparison.

Computer code for the model described in the text is made freely available at https://github.com/dannyjeck/Proto-Object-Comparison.

Author contributions

DJ, MQ, HE and EN designed the experiment. DJ collected and analyzed the data and implemented and tested the model. DJ, MQ, HE and EN wrote the manuscript.

Competing interests

The authors declare that they have no competing interests.

Acknowledgement

This work is supported by the National Institutes of Health under grants R01EY027544 and R01DA040990. We thank Grant Gillary for helpful discussions.

Appendix A. Linear separability of the unique faint square in feature space

While Fig. 1C shows that the Itti et al. (1998) model of visual salience does not produce a strong response for the unique gray square, we wondered whether *any* linear combination of the intensity features in that model could produce the observed behavior where the unique square is selected more than an otherwise identical black square at the same location. If a linear combination of the intensity center-surround features could produce the desired behavior, then there would exist some set of weights on the center-surround maps that enhance the gray square while suppressing the black squares as well as the background.

We therefore set up a set of linear constraints. If they are impossible to satisfy then the feature space is unable to replicate human behavior. The center surround features c generated by the original code from Itti et al. (1998) are six 48×64 pixel maps, one at every spatial scale in the Laplace pyramid of the intensity submodality. For a given (x, y) location on the image, $c(x, y)$ is therefore a vector with six elements corresponding to each center surround feature value at (x, y). Our constraints are that a set of weights w must satisfy

$$w^T c(x, y) < K \tag{4}$$

for some constant K, when (x, y) is outside of the gray square. Additionally, for some point (x', y') inside the gray square,

$$w^T c(x', y') > K \tag{5}$$

must also be satisfied. For a given value of (x', y') this set of constraints can be reformulated as a linear programming problem which allows us to check the satisfiability of the constraints using off-the-shelf software (Matlab R2012b). We found that for the image in Fig. 1A the constraints could not be satisfied for any given (x', y') value on the gray square.

Appendix B. Supplementary data

Supplementary data associated with this article can be found, in the online version, at https://doi.org/10.1016/j.visres.2019.04.004.

References

Ardila, D., Mihalas, S., von der Heydt, R., & Niebur, E. (2012). Medial axis generation in a model of perceptual organization. *IEEE CISS-2012 46th Annual Conference on Information Sciences and Systems* (pp. 1–4). NJ: Princeton University IEEE.

Bacon, W. F., & Egeth, H. E. (1994). Overriding stimulus-driven attentional capture. *Perception & Psychophysics, 55*, 485–496.

Bauer, B., Jolicoeur, P., & Cowan, W. B. (1996). Visual search for colour targets that are or are not linearly separable from distractors. *Vision Research, 36*(10), 1439–1466. https://doi.org/10.1016/0042-6989(95)00207-3 ISSN 00426989,http://linkinghub.elsevier.com/retrieve/pii/0042698995002073.

Belopolsky, A. V., Zwaan, L., Theeuwes, J., & Kramer, A. F. (2007). The size of an attentional window modulates attentional capture by color singletons. *Psychonomic Bulletin & Review, 14*(5), 934–938.

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B, 57*(1), 289–300.

Blum, H. (1973). Biological shape and visual science (Part I). *Journal of Theoretical Biology, 38*(2), 205–287.

Borji, A., & Itti, L. (2013). State-of-the-art in visual attention modeling. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 35*(1), 185–207.

Bregman, A. S. (1990). *Auditory scene analysis: The perceptual organisation of sound.* Cambrigde, MA: MIT Press.

Brouwer, G. J., & Heeger, D. J. (2009). Decoding and reconstructing color from responses in human visual cortex. *Journal of Neuroscience, 29*(44), 13992–14003.

Brown, D. R. (1953). Stimulus similarity and the anchoring of subjective scales. *The American Journal of Psychology, 66*(2), 199–214.

Buswell, G. T. (1935). *How people look at pictures.* University of Chicago Press Chicago.

Conway, B. R., & Tsao, D. Y. (2009). Color-tuned neurons are spatially clustered according to color preference within alert macaque posterior inferior temporal cortex. *Proceedings of the National Academy of Sciences* pages pnas-0810943106.

Craft, E., Schütze, H., Niebur, E., & von der Heydt, R. (2007). A neural model of figure-ground organization. *Journal of Neurophysiology, 97*(6), 4310–4326 PMID: 17442769.

DeAngelus, M., & Pelz, J. B. (2009). Top-down control of eye movements: Yarbus revisited. *Visual Cognition, 17*(6–7), 790–811. https://doi.org/10.1080/13506280902793843 ISSN 1350–6285,http://www.tandfonline.com/doi/abs/10.1080/13506280902793843.

Durgin, F. H., Baird, J. A., Greenburg, M., Russell, R., Shaughnessy, K., & Waymouth, S. (2009). Who is being deceived? The experimental demands of wearing a backpack. *Psychonomic Bulletin & Review, 16*(5), 964–969.

Ernst, D., & Horstmann, G. (2018). Pure colour novelty captures the gaze. *Visual Cognition, 26*(5), 366–381. https://doi.org/10.1080/13506285.2018.1459997.

Feldman, J., & Singh, M. (2006). Bayesian estimation of the shape skeleton. *Proceedings of the National Academy of Sciences, 103*(47), 18014–18019.

Firestone, C., & Scholl, B. J. (2014). Please tap the shape, anywhere you like shape skeletons in human vision revealed by an exceedingly simple measure. *Psychological Science, 25*(2), 377–386. https://doi.org/10.1177/0956797613507584 0956797613507584, ISSN 1467-9280,http://www.ncbi.nlm.nih.gov/pubmed/24406395.

Firestone, C., & Scholl, B. J. (2016). Cognition does not affect perception: Evaluating the evidence for top-down effects. *Behavioral and Brain Sciences, 39*.

Folk, C. L., Remington, R. W., & Johnston, J. C. (1992). Involuntary covert orienting is contingent on attentional control setting. *Journal of Experimental Psychology: Human Perception and Performance, 18*, 1030–1044.

Folk, C. L., Remington, R. W., & Johnston, J. C. (1993). Contingent attentional capture: A reply to Yantis (1993). *Journal of Experimental Psychology: Human Perception and Performance, 19*(3), 682–685.

Godijn, R., & Theeuwes, J. (2002). Programming of endogenous and exogenous saccades: Evidence for a competitive integration model. *Journal of Experimental Psychology: Human Perception and Performance, 28*(5), 1039.

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

Hu, B., von der Heydt, R., Niebur E. (2017). Proto-object based contour detection and figure-ground segmentation. In *Proceedings of CoSyNe* (page 123). Salt Lake City, UT.

Brian, H., 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 PMID: 26739278.

Ibraheem, N. A., Hasan, M. M., Khan, R. Z., & Mishra, P. K. (2012). Understanding color models: A review. *ARPN Journal of Science and Technology, 2*(3), 265–275.

Itti, L., & Koch, C. (2001). Computational modelling of visual attention. *Nature Neuroscience, 2*, 194–203.

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

Jeck, D. M., Qin, M., Egeth, H., & Niebur, E. (2017). Attentive pointing in natural scenes correlates with other measures of attention. *Vision Research, 135*, 54–64 PMID: 28427890.

Judd, T., Durand, F., Torralba, A. (2012). A Benchmark of Computational Models of Saliency to Predict Human Fixations. In MIT Technical Report.

Kayser, C., Petkov, C. I., Lippert, M., & Logothetis, N. K. (2005). Mechanisms for allocating auditory attention: an auditory saliency map. *Current Biology, 15*, 1943–1947.

Kimchi, R., Behrmann, M., & Olson, C. R. (2003). *Perceptual organization in vision: Behavioral and neural perspectives.* Mahwah, NJ: Lawrence Erlbaum Associates.

Kita, S. (2003). *Pointing: Where language, culture, and cognition meet.* Mahwah, NJ: Lawrence Erlbaum Associates.

Koch, C., & Ullman, S. (1985). Shifts in selective visual attention: Towards the underlying neural circuitry. *Human Neurobiology, 4*, 219–227.

Kümmerer, M., Wallis, T. S. A., Bethge, M. (2016). Deepgaze II: Reading fixations from deep features trained on object recognition. arXiv:1610.01563 [cs.CV].

Leavens, D. A., Hopkins, W. D., & Bard, K. A. (2005). Understanding the point of chimpanzee pointing epigenesis and ecological validity. *Current Directions in Psychological Science, 14*(4), 185–189.

Leber, A. B., & Egeth, H. E. (2006). It's under control: Top-down search strategies can override attentional capture. *Psychonomic Bulletin & Review, 13*, 132–138.

Li, M., Liu, F., Juusola, M., & Tang, S. (2014). Perceptual color map in macaque visual area v4. *Journal of Neuroscience, 34*(1), 202–217.

Martin, A. B., & von der Heydt, R. (2015). Spike synchrony reveals emergence of proto-objects in visual cortex. *The Journal of Neuroscience, 35*(17), 6860–6870.

Masciocchi, C., Mihalas, S., Parkhurst, D., & Niebur, E. (2009). Everyone knows what is interesting: Salient locations which should be fixated. *Journal of Vision, 9*(11), 1–22 PMC 2915572.

Mihalas, S., Dong, Y., von der Heydt, R., & Niebur, E. (2011). Mechanisms of perceptual organization provide auto-zoom and auto-localization for attention to objects.

Proceedings of the National Academy of Sciences, 108(18), 7583–7588 PMC3088583.
Niebur, E., & Koch, C. (1996). Control of selective visual attention: Modeling the "Where" pathway. In D. S. Touretzky, M. C. Mozer, & M. E. Hasselmo (Vol. Eds.), *Advances in neural information processing systems: Vol. 8*, (pp. 802–808). Cambridge, MA: MIT Press.
Niebur, E., Itti, L., & Koch, C. (2002). Controlling the focus of visual selective attention. In J. L. Van Hemmen, J. D. Cowan, & E. Domany (Eds.). *Models of neural networks IV: Early vision and attention* (pp. 247–276). New York: Springer Verlag.
Nothdurft, H.-C. (2006). Salience-controlled visual search: Are the brightest and the least bright targets found by different processes? *Visual Cognition, 13*(6), 700–732. https://doi.org/10.1080/13506280544000237 URLhttp://www.tandfonline.com/doi/abs/10.1080/13506280544000237.
Oliva A. (2005). Gist of the scene. In L. Itti, G. Rees, J. K. Tsotsos (Ed.), Neurobiology of attention (pp 251–257). chapter 41,http://cvcl.mit.edu/papers/oliva04.pdf.
Parkhurst, D., Law, K., & Niebur, E. (2002). Modelling the role of salience in the allocation of visual selective attention. *Vision Research, 42*(1), 107–123.
Pashler, H., & Harris, C. R. (2001). Spontaneous allocation of visual attention: Dominant role of uniqueness. *Psychonomic Bulletin & Review, 8*(4), 747–752.
Pentland, A. P. (1986). Perceptual organization and the representation of natural form. *Artificial Intelligence, 28*, 293–331.
Perazzi, F., Krahenbuhl, P., Pritch, Y., & Hornung, A. (2012). Saliency filters: Contrast based filtering for salient region detection. *2012 IEEE Conference on Computer Vision and Pattern Recognition* (pp. 733–740). IEEE. https://doi.org/10.1109/CVPR.2012.6247743 ISBN 978-1-4673-1228-8.http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=6247743 http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=6247743.
Qiu, F. T., & von der Heydt, R. (2005). Figure and ground in the visual cortex: V2 combines stereoscopic cues with Gestalt rules. *Neuron, 47*, 155–166.
Qiu, F. T., & von der Heydt, R. (2007). Neural representation of transparent overlay. *Nature Neuroscience, 10*(3), 283–284.
Qiu, F. T., Sugihara, T., & von der Heydt, R. (2007). Figure-ground mechanisms provide structure for selective attention. *Nature Neuroscience, 10*(11), 1492–1499.
Ramenahalli, S., Mihalas, S., & Niebur, E. (2014). Local spectral anisotropy is a valid cue for figure-ground organization in natural scenes. *Vision Research, 103*, 116–126. https://doi.org/10.1016/j.visres.2014.08.012.
Rensink, R. A. (2000). The dynamic representation of scenes. *Visual Cognition, 7*(1/2/3), 17–42.
Reynolds, J. H., & Heeger, D. J. (2009). The normalization model of attention. *Neuron, 61*(2), 168–185. https://doi.org/10.1016/j.neuron.2009.01.002.
Russell, A. F., Mihalas, S., von der Heydt, R., Niebur, E., & Etienne-Cummings, R. (2014). A model of proto-object based saliency. *Vision Research, 94*, 1–15.
Theeuwes, J. (1992). Perceptual selectivity for color and form. *Perception & Psychophysics, 51*(6), 599–606.
Theeuwes, J. (1994). Stimulus-driven capture and attentional set: Selective search for color and visual abrupt onsets. *Journal of Experimental Psychology: Human Perception and Performance, 20*, 799–806.
Theeuwes, J. (2010). Top-down and bottom-up control of visual selection. *Acta Psychologica, 135*(2), 77–99.
Theeuwes, J., & Van der Burg, E. (2011). On the limits of top-down control of visual selection. *Attention, Perception, & Psychophysics, 73*(7), 2092.
Tomasello, M., Carpenter, M., & Liszkowski, U. (2007). A new look at infant pointing. *Child Development, 78*(3), 705–722.
Treisman, A., & Gormican, S. (1988). Feature analysis in early vision: Evidence from search asymetries. *Psychological Review, 95*, 15–48.
Tseng, P.-H., Carmi, R., Cameron, I. G. M., Munoz, D. P., & Itti, L. (2009). Quantifying center bias of observers in free viewing of dynamic natural scenes. *Journal of Vision, 9*(7), 1–16 7, ISSN 1534-7362.http://journalofvision.org/9/7/4/.
Tsotsos, J. K. (1990). Analyzing vision at the complexity level. *Behavioral and Brain Sciences, 13*(3), 423–445.
Wagatsuma, N., von der Heydt, R., & Niebur, E. (2016). Spike synchrony generated by modulatory common input through NMDA-type synapses. *Journal of Neurophysiology, 116*(3), 1418–1433.
Walther, D., & Koch, C. (2006). Modeling attention to salient proto-objects. *Neural Networks, 19*, 1395–1407.
Williford, J. R., & von der Heydt, R. (2016). Figure-ground organization in visual cortex for natural scenes. *eNeuro, 3*(6) ENEURO-0127.
Yarbus, A. L. (1967). *Eye movements and vision.* New York: Plenum Press.
Zhou, H., Friedman, H. S., & von der Heydt, R. (2000). Coding of border ownership in monkey visual cortex. *Journal of Neuroscience, 20*(17), 6594–6611.