Computational Models of Visual Attention within a Probabilistic Inference Framework

by

Justin Theiss

A dissertation submitted in partial satisfaction of the

requirements for the degree of

Doctor of Philosophy

in

Vision Science

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Professor Michael Silver, Chair
Professor Bruno Olshausen
Professor Frederic Theunissen

Spring 2022

Computational Models of Visual Attention within a Probabilistic Inference Framework

Abstract

Computational Models of Visual Attention within a Probabilistic Inference Framework

by

Justin Theiss

Doctor of Philosophy in Vision Science

University of California, Berkeley

Professor Michael Silver, Chair

Attention is a well-studied and complex topic that covers many fields of research. Effects of attention are ubiquitous throughout the brain and can be differentiated among sensory modalities (e.g., visual vs. auditory), volition (e.g., exogenous vs. endogenous), and application (e.g., covert vs. overt). Over the past few decades, researchers have proposed many computational models of visual attention, and with the rise of machine learning tools, more have been proposed to solve computer vision problems as well. In this dissertation, I will focus on a specific subset of models that place visual attention within a probabilistic inference framework in which humans utilize attention to infer the current state of the world from noisy sensory information. Across three experiments, I propose and evaluate computational models of visual attention that address endogenous spatial attention, feature and spatial attention during covert visual search, and bottom-up and top-down attention during free viewing of natural images. Each model builds upon the previous one in an effort to understand the influence of common principles across different tasks and applications.

In the first experiment, I propose a computational model of spatial attention that uses a dynamic pooling mechanism to simulate receptive field changes that have been observed in neurophysiological studies of endogenous spatial attention. The model can be viewed as a spatial prior over a region of the visual field that reduces uncertainty in visual processing by enhancing the local spatial resolution. By reproducing well-characterized perceptual phenomena observed in visual crowding literature, we conclude that reduction in the spatial uncertainty of encoded feature representations relieves crowding. This decrease in uncertainty influences crowding mainly by increasing the redundancy of encoded representations, with effects on fidelity playing a more limited role. In the second experiment, I extend this model by incorporating spatial attention into a hierarchical generative model to simulate a covert visual search task for digits among non-digit distractors. The generative model learns top-down priors over digit features, and these priors disambiguate among low-level target and distractor features during search to highlight regions that are likely to contain the target. By spatially attending predicted target locations that were generated with or without the use of top-down priors, we show a benefit of using top-down priors on downstream target clas-

sification accuracy that is greater than the improvement from spatial attention alone. Finally, in the third experiment, I introduce a model of bottom-up and top-down attention at multiple levels of feature complexity and spatial scale to account for gaze behavior in a free-viewing experiment across many categories of natural images. As an extension of the second experiment, priors in this experiment influence bottom-up as well as top-down attention. In this case, bottom-up attention is measured as the surprise relative to priors (in an information-theoretic sense) when viewing a scene. By learning priors within as well as across categories, the results demonstrate that surprise from category-specific priors over high-level features best accounted for gaze behavior across the majority of scene types.

To Bowie & Beptziey.

Contents

List of Figures

List of Tables

Acknowledgments

There are many people I would like to thank for their support over the past several years. First, I'd like to thank my advisor, Michael Silver, for his mentorship and for providing me with this opportunity. I would similarly like to thank Bruno Olshausen, Frederic Theunissen, as well as the rest of my lab for their valuable feedback during presentations and discussions. Furthermore, I'd like to thank my co-author Joel Bowen for the long hours spent brainstorming, writing, and revising our paper (Chapter 2). I'd also like to thank my parents, Dora and Richard, for always encouraging me to pursue my interests. Finally, I would like to thank Erin Brosey for her continued support throughout this process: moving across the country, building a life in SF, and keeping me healthy and sane. I could not have done any of this without her.

Previously published work included in this dissertation:

Justin D Theiss, Joel D Bowen, and Michael A Silver (2022). "Spatial Attention Enhances Crowded Stimulus Encoding Across Modeled Receptive Fields by Increasing Redundancy of Feature Representations". In: *Neural Computation* 34.1, pp. 190–218

Chapter 1

General Introduction

1.1 Perception as Bayesian inference

The Bayesian framework for perception dates back to Helmholtz's (Von Helmholtz, 1867) perspective of vision as "unconscious inference", which is based on using prior information to infer causes from observations. This is exemplified by the different ways a three-dimensional object can project onto the retina as the same two-dimensional image. Several psychophysical studies further explored this view, demonstrating that humans encode and utilize uncertainty during visual perception (Carpenter and M. Williams, 1995; Deneve, Latham, and Pouget, 1999; Ernst and M. S. Banks, 2002; Nakayama and Shimojo, 1992; Pouget, Dayan, and R. Zemel, 2000; Y. Weiss, Simoncelli, and Adelson, 2002). Indeed, the Helmholtz machine (Dayan, Hinton, et al., 1995) was proposed as a model of the visual system whereby a generative model of visual inputs would train a recognition model in order to infer the probability distribution over underlying causes of those inputs. More formal theoretical proposals followed, formulating putative mechanisms for how the cortex may encode probabilities and compute Bayesian inference (Kersten, 1999; Kersten, Mamassian, and Yuille, 2004; Knill and Richards, 1996; Rao, B. A. Olshausen, and Lewicki, 2002). Taking inspiration from neurophysiological and computer vision research, T. S. Lee and Mumford (2003) posited that the neural activity in visual cortex could represent hypotheses for underlying properties of the visual input, with each visual area propagating information based on sensory evidence (likelihood) and top-down feedback from higher visual areas (prior). This theory of hierarchical Bayesian inference in visual cortex was naturally extended to include attention as top-down priors over spatial locations and features (Chikkerur et al., 2010; Rao, 2005; Yu and Dayan, 2004).

1.2 Attention in the visual cortex

Effects of spatial attention on neural responses in visual cortex have been well-studied, leading to improved models of attentional mechanisms. The biased competition model (Desimone and Duncan, 1995) explained how spatial attention to one of multiple stimuli in a neuronal receptive field biases processing in favor of the attended stimulus. For a preferred stimulus, this leads to enhanced neural responses, whereas attention to a non-preferred stimulus suppresses responses to an unattended preferred stimulus within the same receptive field (Reynolds, Chelazzi, and Desimone, 1999). Similarly, spatial attention within V4 neuronal receptive fields causes a multiplicative scaling of the orientation tuning curve (McAdams and Maunsell, 1999). Yu and Dayan (2004) simulated a spatial cueing task (Posner, 1980) in order to model spatial attention within a Bayesian inference framework as a Gaussian prior over spatial locations, which can in turn affect the posterior probability over simulated orientation-tuned responses. Rao (2005) proposed a similar probabilistic attention model, which additionally reproduced the attentional effects on single-cell responses observed in Reynolds, Chelazzi, and Desimone (1999) and Connor et al. (1997).

In order to account for varying effects of attentional modulation observed across studies manipulating spatial and feature attention (Martinez-Trujillo and Treue, 2004; McAdams and Maunsell, 1999; Treue and Trujillo, 1999), Reynolds and Heeger (2009) proposed the normalization model of attention that models population-level neural responses as a multiplicative effect of an attention field on stimulus-driven responses normalized across spatial and feature dimensions. In addition to explaining several disparate effects of attentional modulation on single-cell responses, the model has since been shown to account for attention-related changes in properties of population-level receptive fields measured in humans using fMRI (Klein, Harvey, and Dumoulin, 2014). Although Reynolds and Heeger (2009) did not define their model in probabilistic terms, Chikkerur et al. (2010) noted the similarities between the normalization model and their probabilistic formulation of perception as Bayesian inference with top-down priors modulating bottom-up stimulus-driven responses.

1.3 Computational models of visual attention

Over the past few decades, many computational models of visual attention have been proposed to explain various neurophysiological and behavioral phenomena. Early models incorporated the known properties of early visual cortical processing, including orientation and color contrast as well as lateral inhibition (Itti, Koch, and Niebur, 1998; Koch and Ullman, 1985; Z. Li, 1999). These approaches have since become known as "bottom-up image saliency" models, which produce saliency maps for predicting human gaze behavior in natural images (Borji, D. N. Sihite, and Itti, 2012; Harel, Koch, and Perona, 2006; Parkhurst, Law, and Niebur, 2002). Other computational models attempted to explain effects of spatial attention on neuronal processing within a probabilistic inference framework (Rao, 2005; Yu and Dayan, 2004). As computer vision and machine learning techniques progressed, more complex probabilistic models were introduced that could predict gaze behavior in natural images (Chikkerur et al., 2010; J. Li, Tian, and T. Huang, 2014; Oliva et al., 2003; Torralba, Oliva, et al., 2006; Zhang et al., 2008). Such models learned priors over image features or scene-level gist descriptors in order to generate priority maps corresponding to regions in natural images containing particular objects (Chikkerur et al., 2010; Oliva et al., 2003; Torralba, Oliva, et al., 2006). More recently, deep learning models have provided state-of-the-art performance in predicting gaze behavior by learning associations between features of pre-trained deep neural networks and human gaze data (e.g., Kümmerer, Wallis, and Bethge, 2016).

In addition to modeling top-down attention, Itti and Baldi (2005) proposed a theoretical formulation of saliency within the probabilistic setting by computing the surprise between prior and posterior probability distributions. This form of saliency, later termed "Bayesian surprise" (Itti and Baldi, 2009), is also closely related to the free-energy principle proposed by Friston (2009) as a unifying theory of neural structure and function. When using bottom-up saliency models to extract features from images, Bayesian surprise is computed over a family of models that represent predictions regarding orientation, color, etc. at different spatial scales (Itti and Baldi, 2005). More recent research has subsequently found evidence of Bayesian surprise represented in cortex as well (Gijsen et al., 2021; Kolossa, Kopp, and Fingscheidt, 2015; Ostwald et al., 2012), lending further support to this probabilistic theory of saliency.

1.4 Outline of the dissertation

In this dissertation, I propose and evaluate three computational models of visual attention in order to study various components of attention within a probabilistic framework. In Chapter 2, I designed a computational model simulating the effects of spatial attention on receptive field (RF) organization based on the normalization model of attention (Reynolds and Heeger, 2009) and neurophysiological findings (Klein, Harvey, and Dumoulin, 2014; Womelsdorf et al., 2006). As in Chikkerur et al. (2010), I interpreted the attention field in the normalization model as a top-down spatial prior, which in the RF pooling model influences stimulus-driven processing through local spatial resolution enhancement. In Chapter 3, I modeled top-down attention during simulated visual search by incorporating the RF pooling model from Chapter 2 within a hierarchical generative model that learned a probability distribution over target features. This probabilistic visual search model used top-down feature priors to disambiguate among target and distractor features in order to instantiate a spatial prior over the predicted target location, which enhanced local spatial resolution for downstream processing. In Chapter 4, I used normalizing flow models to learn feature and spatial priors over extracted features of natural images in order to predict gaze behavior. Using a public dataset containing 20 categories of scenes, I evaluated the effect of category-specific (learned within category) vs. category-agnostic (learned across categories) priors on performance of gaze predictions. I then characterized the relative contributions of bottom-up Bayesian surprise and top-down feature and spatial priors at varying levels of feature complexity and spatial scale.

Chapter 2

Spatial attention enhances crowded stimulus encoding across modeled receptive fields by increasing redundancy of feature representations

2.1 Abstract

Any visual system — biological or artificial — must make a trade-off between the number of units used to represent the visual environment and the spatial resolution of the sampling array. Humans and some other animals are able to allocate attention to spatial locations to reconfigure the sampling array of receptive fields (RFs), thereby enhancing the spatial resolution of representations without changing the overall number of sampling units. Here, we examine how representations of visual features in a fully-convolutional neural network interact and interfere with each other in an eccentricity-dependent RF pooling array and how these interactions are influenced by dynamic changes in spatial resolution across the array. We study these feature interactions within the framework of visual crowding, a well-characterized perceptual phenomenon in which target objects in the visual periphery that are easily identified in isolation are much more difficult to identify when flanked by similar nearby objects. By separately simulating effects of spatial attention on RF size and on the density of the pooling array, we demonstrate that the increase in RF density due to attention is more beneficial than changes in RF size for enhancing target classification for crowded stimuli. Furthermore, by varying target/flanker spacing as well as the spatial extent of attention, we find that feature redundancy across RFs has more influence on target classification than the fidelity of the feature representations themselves. Based on these findings, we propose a candidate mechanism by which spatial attention relieves visual crowding through enhanced feature redundancy that is mostly due to increased RF density.

2.2 Introduction

The cerebral cortex is composed of a hierarchy of processing areas, each containing overlapping neuronal receptive fields (RFs) that tile the visual field at different spatial scales. The visual systems of humans and other animals use spatial attention to dynamically reconfigure the size and density of RFs (Klein, Harvey, and Dumoulin, 2014; Womelsdorf et al., 2006) to enhance sampling of stimuli (Anton-Erxleben and Carrasco, 2013) and perception (Carrasco, 2011) at attended locations.

Physiologically, directing spatial attention to one of multiple objects within a single RF can bias responses in favor of the attended object (Desimone and Duncan, 1995). Specifically, attending to a preferred object reduces the suppressive effect of simultaneous presentation of a nonpreferred object in the RF, whereas attending to a nonpreferred object enhances its suppressive effect (Reynolds, Chelazzi, and Desimone, 1999). Such attentional effects have been observed at the single-cell level as a scaling of neuronal responses to an attended stimulus by a gain factor (McAdams and Maunsell, 1999) as well as a shrinking of the neuronal RF around an attended stimulus (Anton-Erxleben, Stephan, and Treue, 2009).

In an fMRI study in humans, Vo, Sprague, and Serences (2017) found that attention-related shifts in RF position were more important than changes in RF size for population-level encoding of fine spatial information. Reynolds and Heeger (2009) provided a unifying model of attention in which the neuronal responses to a stimulus are normalized by a suppressive population response and multiplied by a spatial attention field. In addition to predicting neuronal responses, the model also accounts for the observed changes in RF properties with spatial attention in both humans and monkeys (Klein, Harvey, and Dumoulin, 2014; Womelsdorf et al., 2006) by modeling attention as a Gaussian multiplication of an attention field with individual RFs. The normalization model of attention therefore provides a computational framework for studying the effects of spatial attention on RF properties, stimulus encoding, and task performance.

Reconfiguration of RFs by spatial attention is perhaps more relevant to stimulus encoding in the visual periphery, where RFs are larger and less densely arranged compared to foveal RFs (Gattass, Gross, and Sandell, 1981; Gattass, Sousa, and Gross, 1988). As such, limits on the size and density of RFs have been theorized to contribute to the perceptual phenomenon known as visual crowding (Levi, 2008; Rosenholtz, 2016; Whitney and Levi, 2011), in which target objects in the visual periphery that are easily identified in isolation are more difficult to identify when flanked by similar nearby objects. Interestingly, flanking stimuli that are presented more peripherally, relative to a target stimulus location, crowd more than those that are presented more foveally (W. P. Banks, Bachrach, and Larson, 1977; Petrov and Meleshkevich, 2011), which suggests that target and flanker features encoded in larger RFs may be spatially over-integrated. Indeed, visual crowding has been modeled as a pooling mechanism in which relative spatial information of features is discarded (Balas, Nakano, and Ruth Rosenholtz, 2009; Freeman and Simoncelli, 2011; Keshvari and Ruth Rosenholtz, 2016; Van den Berg, Roerdink, and Cornelissen, 2010). However, there are additional aspects of crowding that cannot be explained by a simple pooling model, such as substitution errors in which subjects report one of the flankers instead of the target (Coates, Bernard, and

Chung, 2019; Ester, Klee, and Awh, 2014; Hanus and Vul, 2013), categorical target/flanker effects (Reuther and Chakravarthi, 2014), global/contextual effects (Herzog et al., 2015; Manassi, Sayim, and Herzog, 2012), and holistic effects (Farzin, Rivera, and Whitney, 2009).

It has further been shown that pre-cueing spatial attention to the target location can relieve crowding in humans (Albonico et al., 2018; Scolari et al., 2007; Yeshurun and Rashal, 2010) and improve performance on other peripheral visual tasks (Barbot and Carrasco, 2017; Yeshurun and Carrasco, 1998; Yeshurun, Montagna, and Carrasco, 2008). Conceptually, these effects of attention can be viewed as changing the spatial extent of a "perceptual window" (Sun, Chung, and Tjan, 2010) or as an attraction of RFs (Baruch and Yeshurun, 2014) to enhance stimulus encoding, similar to the Gaussian attention field that has been used to account for modulation of RF properties by attention (Klein, Harvey, and Dumoulin, 2014; Womelsdorf et al., 2006). Moreover, similar studies have shown that the size of an attention cue significantly impacts performance on peripheral tasks (Albonico et al., 2018; Yeshurun and Carrasco, 2008). In addition, He, Y. Wang, and Fang (2019) recently demonstrated that following perceptual learning, decreases in RF size of individual fMRI voxels in cortical area V2 correlated with improved performance on a crowding task. However, a mechanistic account of how spatial attention alleviates visual crowding has not yet been established.

When flanker and target features are within the same set of RFs, this should result in greater competition for processing compared to cases in which the flanker and target are not in the same set of RFs. We define two metrics, fidelity and redundancy, to characterize this competition and its contributions to performance on a crowding task. *Feature fidelity* is the similarity of the encoded features of an isolated target compared to those of a target crowded by flankers. *Feature redundancy* is the average number of RFs that sample a target feature in a crowded stimulus, regardless of its fidelity.

There are multiple ways that structural properties of an array of RFs might enhance encoding or performance on a visual crowding task. At one extreme, signals from individual small and minimally overlapping RFs could have strong feature fidelity within individual RFs due to low levels of competition between target and flanker features, which would be expected to result in good performance. At the other extreme, signals from large and highly overlapping RFs could have poor fidelity at the level of individual RFs, but when combined, might maintain a high-quality encoding based on redundant representation of features across pools of RFs, which would also lead to good performance. Although multiple studies have described the effects of spatial attention on RF properties, it is currently unclear how changes in feature fidelity and redundancy due to spatial attention may affect downstream processing and perception.

In the current study, we extend the conceptual framework of the normalization model of attention (Reynolds and Heeger, 2009) to investigate how attention-dependent changes in RF size and position relate to the fidelity and redundancy of feature representations and to downstream processing of crowded stimuli. Specifically, we simulated a visual crowding task in which a target stimulus in the peripheral visual field was surrounded by various flanking stimuli. We measured target classification accuracy, feature fidelity, and feature redundancy over a range of target/flanker spacings and spatial extents of a 2-D Gaussian attention field. Using a novel technique for simulating cortical RFs within a convolutional neural network (CNN), we characterized the independent contributions of feature fidelity and redundancy to perception of crowded stimuli. Following the conventions proposed by Kording et al. (2020), we aimed to create a computational model that inspires experiments and provides macroscopic realism. We discuss and interpret our findings within the context of previous neurophysiological, psychophysical, and computational modeling studies.

2.3 Materials and Methods

Model Description

Theoretical Framework. Following the normalization model of attention (Reynolds and Heeger, 2009), we assume that changes in position and size of RFs reflect changes in the responses of populations of neurons. As such, we used a dynamic RF pooling mechanism in order to model attention-dependent effects on visual processing and representations. Furthermore, in order to assess performance on a target identification task, we defined a selection mechanism that simulated a population of neurons that process RF outputs via Gaussian "cortical" weights. Finally, the pooling mechanism in our model is based on the assumption that competition for processing within and across RFs is the driving force of crowding. However, we acknowledge there are other aspects of crowding, such as global/context effects (Manassi and Whitney, 2018) that are not addressed in our model.

Convolutional neural network model. We trained a three-layer fully-convolutional feedforward neural network to classify grayscale handwritten digits (28 × 28 pixels; MNIST) (LeCun et al., 1998). Each convolutional layer in the neural network takes an image (or stack of images) as input and decomposes it into a set of feature maps, with each pixel in the feature map indicating the relative presence or absence of that feature. These feature maps are then passed through a non-linear activation function (rectified linear unit [ReLU] or softmax; see Table 2.1). Finally, a pooling operation is applied to reduce the image size of the feature maps. Typically, this involves taking the maximum value within subsets of pixels (e.g., 2 × 2 subsets of pixels to reduce image height and width by 50%).

Although max-pooling is commonly used in the machine learning literature, it is worth noting that it is unlikely to be an optimal pooling mechanism used by populations of visual cortical neurons (Y. Chen, Geisler, and Seidemann, 2006; Simoncelli and B. A. Olshausen, 2001). Instead, Y. Chen, Geisler, and Seidemann (2006) determined that an optimal pooling mechanism would have spatial antagonism (e.g., center-surround) in RFs in order to decorrelate neural responses. However, given substantial differences between the number of neurons in a given visual cortical area and the number of pixels in a given layer of a CNN representing a portion of the visual field, it is unclear how to implement a center-surround pooling mechanism within 2 × 2 subsets of pixels.

We trained our model for ten epochs (ten full passes through training set of 60,000 images), with a mini-batch size of ten, using supervised learning for digit classification with backpropagation (stochastic gradient descent with a learning rate of 0.001 and momentum of 0.9). The trained model achieved a test set error rate of 0.96% (100 - classification accuracy) on a held-out test

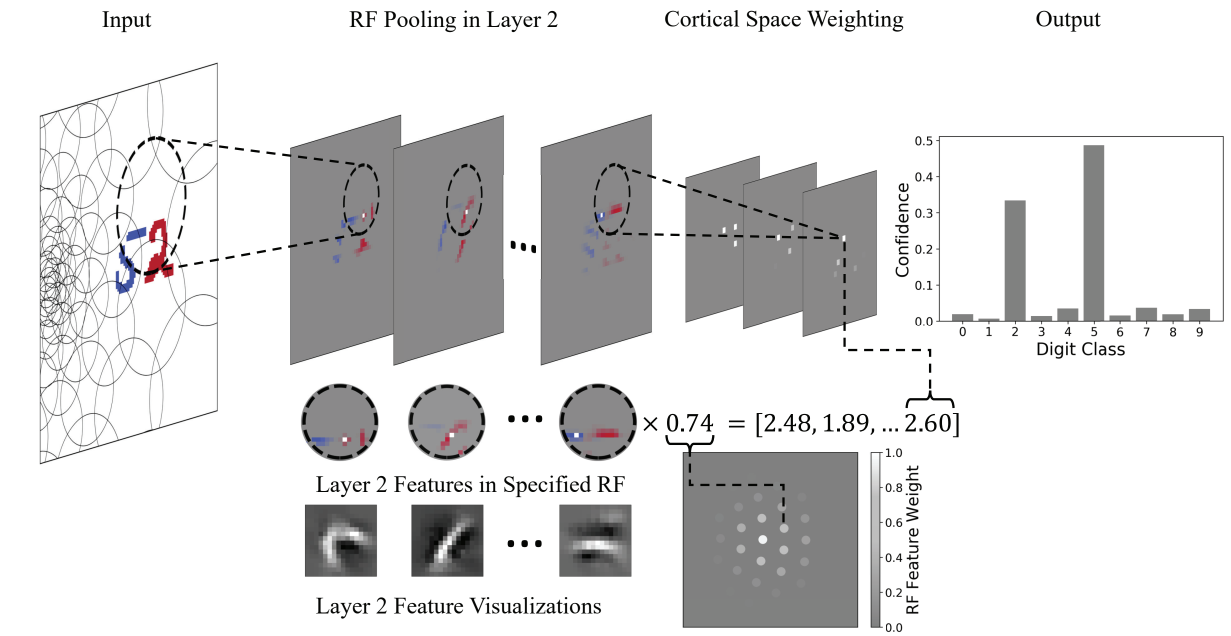

Figure 2.1: The three-layer fully-convolutional neural network was trained to classify handwritten digits, with the softmaxed values in the output vector signifying the confidence of the classification for each digit. Flanking digits (red) were presented at various locations around the central target digit (blue). To model peripheral visual processing, we implemented a pooling operation on an RF array that simulates the eccentricity dependence of RFs in human visual cortex (here, eccentricity increases from left to right, with the fovea centered on the left edge of the input image). Feature maps in the second layer were spatially pooled within each RF separately (as shown for the highlighted example RF). Target (blue) and flanker (red) features compete within each RF, and the maximum value (shown here in white) in each masked feature map is retained while all other pixels were set to zero. In order to classify the target, the second-layer features in each RF were proportionately weighted based on the RF's cortical distance from the target (Equation 2.3), simulating 2-D Gaussian connections to a third-layer population of neurons that is centered on the target digit. Values within the brackets for the given RF indicate estimates of the relative presence or absence of the respective feature representations shown in the figure.

set of 10,000 images. Table 2.1 shows the number of channels, activation functions, and pooling operations for each layer used during training.

The trained model was then used to extract features to be studied in crowding experiments in which multiple digits are simultaneously presented. In order to simulate peripheral vision for these crowding experiments and therefore provide the model with macroscopic realism (Kording et al., 2020), we replaced the max-pooling function in the second layer with an RF pooling array (Figure

2.1, described below). We chose the second layer for this because the weights in this layer are more likely to represent unique fragments of the target digit that are shared across different digits, compared to the first-layer weights (which convolve over an area much smaller than a digit) and the third-layer weights (which convolve over an entire digit). Therefore, the second-layer weights better reflect competition between features within RFs. By training the model on individual 28×28 pixel digits without the RF pooling array, we ensured that only the size and density of RFs would affect stimulus encoding during the crowding experiments.

Table 2.1: Model architecture used for training

Input	Output	Conv	Activation	Pool
1	32	5×5	ReLU	Max 2×2
32	64	5×5	ReLU	Max 2×2
64	10	4×4	Softmax	None

Receptive field pooling. Unlike a typical max-pooling layer, RF pooling occurs within RFs of variable size. As shown in Figure 2.1 for an example RF, responses in the second-layer feature maps are pooled separately per RF to obtain the maximum response per channel within the RF. In order to pool across each individual RF, we define an array with shape $(receptive\ fields \times height \times width)$ that contains a mask that represents the center location (μ) and size (σ) for each RF (i.e., a value of 1 for pixels corresponding to the RF and 0 elsewhere). An input of shape $(batch \times channels \times height \times width)$ can then be masked by the pooling array to obtain the responses for each RF separately, with a resulting shape of $(batch \times channels \times receptive fields \times height \times width)$. We then retain only the pixel with the maximum value within each channel of the RF, maintaining its spatial location, while setting all other pixels within each channel to zero (Figure 2.1). The output of this RF pooling operation is therefore a sparse array of feature maps, with each feature map containing a single value per RF. As a result, features within the same RF compete for processing within, but not across, channels. The RF pooling step is followed by a typical max-pooling operation to obtain a subsampled output that matches the output size of the original layer used during training (i.e., 2×2 max-pooling, Table 2.1).

Using this approach, we maintain the spatial organization of the feature maps while pooling information with variable spatial resolution across the image. This allows us to separately examine the outputs across RFs (to assess redundancy of stimulus encoding) as well as the interactions within individual RFs (to assess fidelity of stimulus encoding). Finally, since each RF is defined by its μ and σ values, the RF array can be dynamically updated by allocation of attention to change the center positions and/or sizes of each RF using Equation 2.4 (described in Section 2.3).

Spatial organization of the RF pooling array. The RF pooling array is organized into concentric rings that expand from a central point (fovea; left edge of Input image in Figure 2.1), with the circular RFs in each ring increasing exponentially in size as a function of eccentricity. Each RF center μ and size σ is determined by the following equations:

$$\mu = \left(\frac{1+s}{1-s}\right) e_{n-1} \cdot \begin{bmatrix}\cos(\theta)\\ \sin(\theta)\end{bmatrix}, \quad \sigma = \left(\frac{1+s}{1-s}\right) e_{n-1} * s \tag{2.1}$$

where θ is the polar angle of the RF with respect to a reference axis expanding from the fovea, e_{n-1} is the eccentricity of the radially-adjacent and more foveal RF, and s is the eccentricity-based scaling factor. For our model, the scaling factor is 0.2, based on fMRI population-level RF measurements from human visual cortical area V2 (Wandell and Winawer, 2015). However, we do not assume or require a one-to-one relationship between feature maps in our model and neural responses in visual cortex. In our model, increasing the scaling factor would simply lead to an increase in size and a decrease in density of RFs as a function of eccentricity.

We presented stimuli at different locations in the visual periphery by applying a horizontal or vertical offset of the RF pooling array (Figure 2.1). Specifically, we shifted the RF pooling array by 60 pixels in the image space, resulting in a target eccentricity of 3 degrees of visual angle (DVA), with 1 DVA defined as 20 pixels, or the approximate width of an MNIST digit. In order to reduce bias related to the initial organization of RFs in the pooling array, we randomly rotated the RFs about the fovea (maximum rotation was half the angle between two eccentrically-adjacent RFs), and we randomly jittered the input image (maximum jitter was 5 pixels, or 0.25 DVA) for each stimulus image.

Weighting of RF features for digit classification. In order to simulate peripheral vision while maintaining spatial relationships among features in image space, we weighted RF features based on their respective locations in "cortical space" (Figure 2.1). This weighting procedure simulates a selection mechanism in which pooled features in one location are enhanced relative to pooled features from other locations in the image, allowing the model to selectively classify a target object among flanking objects. To convert from eccentricity e_{image} and polar angle θ_{image} values in image space to eccentricity e_{cortical} and polar angle θ_{cortical} values in cortical space, we used the following relationships, which are derived from the inversion of the exponential expansion in equation 2.1:

$$\begin{aligned} e_{\text{cortical}} &= \frac{1}{\ln\left(\frac{1+s}{1-s}\right)} * \ln\left(e_{\text{image}}\right), \\ \theta_{\text{cortical}} &= \frac{1}{e_{\text{cortical}}} * \frac{\theta_{\text{image}}}{\theta_{\text{ring}}} \end{aligned} \tag{2.2}$$

where θ_{ring} is the polar angle between the centers of adjacent RFs in the same ring. Note that in image space, the arc length between adjacent RFs in a more peripheral ring is larger than the arc length between RFs in a more foveal ring, but in cortical space, these arc lengths are independent of eccentricity.

With this approach, we assume that the third convolutional layer represents a population of neurons that is centered on the target digit location and has 2-D Gaussian connections (in cortical space) to the second-layer RF outputs. Specifically, we computed digit classification by first passing a stimulus image through the first two convolutional layers of the model and the RF pooling array. We then weighted the outputs of each RF by a 2-D Gaussian in cortical space (Figure 2.1):

$$w_{RF}(x, y) = \frac{1}{2\pi\sigma_w^2} \exp\left(-\frac{(x - x_w)^2 + (y - y_w)^2}{2\sigma_w^2}\right) \tag{2.3}$$

where x and y are the cortical space coordinates for a given RF center, x_w and y_w represent the cortical space coordinates of the 2-D Gaussian weighting function, and σ_w represents the size of the weighting function in cortical space. We set the values of x_w and y_w to be the target location (in cortical space) and σ_w to 1. The weighted features were then passed through the third layer of the network, and we computed target classification by selecting the feature class with the greatest value across the image space in the output layer (also known as global max pooling). This approach provides three benefits: 1) it is a mechanism of selection of the target digit that could be employed in visual cortex, 2) it does not require additional training or manipulation of the data set, and 3) it facilitates comparison of equivalently-weighted RFs across experiments that vary target/flanker spacing and the spatial extent of attention (see Section 2.4).

Model Summary. In summary, a fully-convolutional neural network (Table 2.1) was first trained to classify 28×28 individual MNIST digits, and the learned weights were then fixed for all experiments. To simulate peripheral vision, we created an eccentricity-dependent RF pooling array (Figure 2.1) with an eccentricity scaling factor of 0.2 and a horizontal or vertical offset. For all experiments, the RF pooling array replaced the max-pooling operation after the second convolutional layer (Table 2.1). Although the exponential nature of the RF pooling array is important for accurately simulating peripheral vision, the specific value of the eccentricity scaling factor and the pooling operation are not important factors for studying the effects of crowding on task performance in our model.

In order to simulate a target-identification task in the periphery (e.g., Input in Figure 2.1), we used a 2-D Gaussian weighting function in "cortical space" as a selection mechanism to classify the target among flanking digits. The use of cortical weighting in our model is similar to asking a human participant to report the identity of the central digit as opposed to the flankers. Therefore, we weighted values pooled by RFs inversely proportional to the cortical distance from the target (i.e., RFs closer to the target had greater weights than RFs further from the target). Importantly, we used the same weights for all experiments (i.e., we did not recalculate the weights following attentional modulation of RF properties) to ensure that any changes in the model's ability to classify a target digit were driven primarily by the structural properties of the RF pooling array. The weighted features were then passed through the final convolutional layer, and we computed target classification by selecting the feature class with the greatest value across the image space in the output layer.

Experimental Design and Statistical Analyses

Visual crowding experiment. Inspired by stimuli used in perceptual experiments on visual crowding, we employed a classification task in which the target object is closely surrounded by flanking objects. We constructed crowded stimuli from a balanced test set of 10,000 MNIST digits that were not used during training. We randomly chose target digits and placed them at the center of the stimulus image, and we randomly chose flankers from non-target classes. Target/flanker spacing was measured center-to-center. Figure 2.2 illustrates the four configurations we used in this study (outlined by colored boxes). In this example, the RF pooling array is offset horizontally. The *inner* (yellow), *outer* (blue), and *radial* (green) configurations have flankers at different eccentricities than the target, and the *tangential* (red) configuration has flankers at approximately the same eccentricity as the target. Note that the RFs that sample the peripheral flanking digit are larger and the RFs that sample the foveal flanking digit are smaller.

We offset the RF pooling array relative to the stimulus image so that the location of the target at the center of the image corresponds to 3 degrees eccentricity. Additionally, we averaged all results over the simulated right and left horizontal meridians (an array offset horizontally to the left or right respectively) and the lower and upper vertical meridians (an array offset vertically up or down respectively) to account for asymmetries in the handwritten MNIST digits.

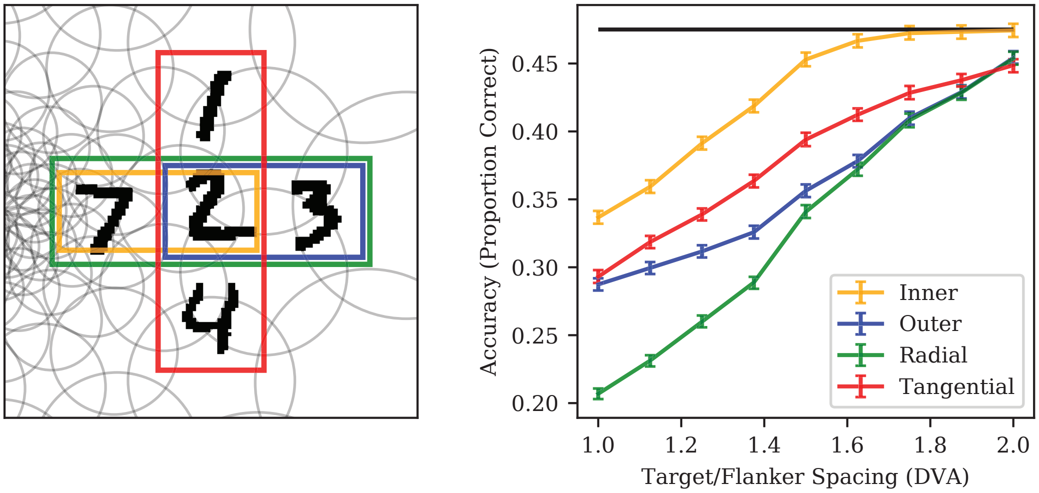

Figure 2.2: (*left*) Examples of crowded stimuli (target/flanker spacing = 1.5 DVA). The fixation point in this example is three DVA to the left of the central target digit (i.e., the left edge of the image). Gray circles show locations and sizes of individual RFs, and colored boxes outline the four unique configurations. (*right*) Target classification accuracy as a function of target/flanker spacing for each configuration. Line and symbol colors correspond to the box colors on the left. The black line indicates accuracy for targets presented without flankers. Chance performance is 0.1 (one out of ten possible digits). Error bars are bootstrapped 95% confidence intervals.

Attentional modulation of RF properties. We simulated spatial attention in our model by modifying the center locations (μ) and sizes (σ) of the RFs in the array. Following the normalization model of attention (Reynolds and Heeger, 2009), Klein, Harvey, and Dumoulin (2014) demonstrated that multiplying a 2-D Gaussian attention field by a 2-D Gaussian population-level (single fMRI voxel) RF provides a good model of the effects of spatial attention on voxel RF locations and sizes in human visual cortex. Specifically, they modeled the effects of spatial attention as changes in the σs and μs for the set of voxel RFs within a given cortical region:

$$\mu = \frac{\mu_{RF}\sigma^2_{AF} + \mu_{AF}\sigma^2_{RF}}{\sigma^2_{AF} + \sigma^2_{RF}}, \quad \sigma^2 = \frac{\sigma^2_{RF}\sigma^2_{AF}}{\sigma^2_{RF} + \sigma^2_{AF}} \tag{2.4}$$

where AF and RF index the attention field and RF kernels, respectively. Decreases in the spatial extent of the attention field (i.e., smaller values of σ_{AF}) cause RFs to become smaller and more densely packed around the center of the attention field (left and center panels of Figure 2.3). To study the effects of this simulated attentional allocation, we empirically varied the size of σ_{AF} and computed new values of σ_{RF} and μ_{RF} for each RF in the pooling array via the Gaussian multiplication described above. However, we do not assume the range and/or scale of σ_{AF} used in our experiments have a one-to-one relationship with the full range of attentional modulation in humans.

Redundancy and fidelity metrics. For crowded visual displays, RFs containing target representations often also contain flanker representations, leading to competition within individual RFs. RFs with a strong target representation might contribute to target classification because they provide a high-fidelity signal for target features. On the other hand, individual RFs with corrupted target representations might still contribute to target classification by sampling the target features in a manner that is redundant with other RFs. For simplicity, we call these two types of target feature interactions fidelity and redundancy, respectively.

We used the outputs of the RF pooling array to obtain vectorized sets of the pooled features for each target-containing RF when the target was presented alone u_t, when the flankers were presented alone u_f, and when the target was crowded by flankers $u_{(t+f)}$. In order to make comparisons across changes in attentional allocation using the same RFs, the indices of the target-containing RFs for these metrics were calculated from the baseline condition with no attention (equivalent to infinite attention field extent). For the fidelity metric, we measured how similar the target signal was in the absence of flankers compared to when it was corrupted by the flanker features for each RF. Specifically, we defined feature fidelity (F) as the cosine similarity between the uncorrupted (no flankers) target features u_t and the corrupted target features $u_{(t+f)} - u_f$, concatenated across target-containing RFs:

$$F = \frac{\langle u_t, (u_{(t+f)} - u_f)\rangle}{\|u_t\|_2 \|(u_{(t+f)} - u_f)\|_2}, \tag{2.5}$$

where $\|\cdot\|_2$ is the Euclidean norm and $\langle\cdot,\cdot\rangle$ represents the dot product of two vectors. Fidelity values closer to 1 indicate that the pooled target features were less corrupted by flanker features across target-containing RFs.

For the redundancy metric, we computed the average number of RFs that represented the corrupted target features $u_{(t+f)} - u_f$. We first selected the activated (i.e., non-zero) target features for each RF using an indicator function that sets the value of each element in the vector to 1 if it is greater than 0 and to 0 otherwise:

$$a_{RF} = \mathbb{1}_{X>0}(u_{(t+f)} - u_f), \quad \text{where } \mathbb{1}_{X>0}(X) = \begin{cases} 1, & \text{if } x > 0 \\ 0, & \text{otherwise} \end{cases} \forall x \in X. \tag{2.6}$$

Then we computed the average number of RFs per activated target feature from the corrupted target signal:

$$\begin{aligned} a &= \sum^{N_{RF}} a_{RF} \\ R &= \frac{1}{\|a\|_0} \sum^{N_a} a, \end{aligned} \tag{2.7}$$

where $\|\cdot\|_0$ returns the number of non-zero values in a vector, and N_a represents the number of features in the pooling layer (i.e., 64). Larger redundancy scores indicate that, on average, more RFs represent an activated target feature within the corrupted target signal.

Statistical procedures. In order to obtain 95% confidence intervals for our estimates, we used 1000 iterations of bootstrap resampling of the data with replacement. For statistical comparisons between two distributions, we first centered each distribution's mean at the combined mean of the two distributions and then bootstrap resampled (again with 1000 iterations) from the centered distributions. We report p-values as the proportion of observed mean differences that were greater than the original mean difference (Efron and Tibshirani, 1994). Additionally, to measure the unique variance in target classification accuracy that was explained by feature fidelity or redundancy, we performed multiple linear regression for fidelity and redundancy combined as well as for each factor alone. The difference in variance explained between the linear model that included both factors and the single-factor model is the unique variance explained by the excluded factor.

Code/software. We implemented all training and computation in PyTorch (Paszke et al., 2017) as well as custom Python and C++ code. The code used to produce the results described in this paper is available upon request.

2.4 Results

Replication of visual crowding effects. Unlike visual acuity, which is typically limited by representations of single features, visual crowding can occur as a result of mixing of high-contrast features within the crowded stimulus, making it difficult to match objects with the individual features that comprise them (Whitney and Levi, 2011).

We examined how representations of features of crowded stimuli interact within the RFs of the pooling array. Both targets and flankers were grayscale handwritten digits (MNIST) (LeCun et al., 1998). We compared target classification accuracy (see Section 2.3) for crowded stimuli over a range of nine target/flanker spacings (equally spaced between 1 and 2 DVA). If portions of multiple objects that are represented within individual RFs lead to feature interference, then increasing target/flanker spacing should relieve crowding (i.e., increase target classification accuracy).

We manipulated spacing for four unique target/flanker configurations (*inner*, *outer*, *radial*, and *tangential*; Figure 2.2). In humans, crowding is influenced by target/flanker configuration: a single *inner* flanker presented foveally to the target causes less crowding than the same *outer* flanker presented peripherally to the target (W. P. Banks, Bachrach, and Larson, 1977). Additionally, crowding is anisotropic: flankers presented on either side of the target along a *radial* axis emanating from the fovea cause more crowding than flankers presented along a *tangential* axis that is perpendicular to the radial axis (J. Chen et al., 2014; Toet and Levi, 1992). In the current experiment, we measured target classification accuracy in each of these target/flanker configurations to determine if our simple (relative to previous models, e.g., Chaney, Fischer, and Whitney, 2014; Nandy and Tjan, 2012) model of RF pooling could reproduce the effects observed in the literature that are described above.

Figure 2.2 shows that for all four configurations, target classification accuracy increased as a function of target/flanker spacing and that, at large target/flanker spacings, all four configurations approached accuracy levels observed in the target-alone condition (black line). Generally, crowding is greatest for target/flanker spacings that are less than one-half of the target eccentricity (in our model, 3 DVA) (Bouma, 1970). We also found that accuracy was lower for the *radial* configuration than for the *tangential* configuration for spacings at or below 1.5 DVA (green and red lines; bootstrapped p-value [1000 samples] = 0.), consistent with previously reported anisotropies of crowding in human subjects (Toet and Levi, 1992). Moreover, accuracy was lower for the *outer* configuration compared to the *inner* configuration for spacings at or below 1.5 DVA (blue and yellow lines; bootstrapped p-value [1000 samples] = 0.), again consistent with asymmetries that have been reported in human subjects (W. P. Banks, Bachrach, and Larson, 1977).

Smaller attention field extent relieves visual crowding. In the previous section, we showed that our model reproduced known effects of target/flanker spacing and configuration on human visual crowding. In this experiment, we fixed the target/flanker spacing at 1 DVA and applied a spatial attention field (2-D Gaussian centered on the target) that modified the sizes (σ) and center locations (μ) of RFs in the pooling array. Specifically, we calculated the product of this spatial attention field with each of the RFs in the pooling array (Equation 2.4). Although modulating RFs in this way effectively describes how spatial attention influences visual representations in the brain (Klein, Harvey, and Dumoulin, 2014; Womelsdorf et al., 2006), it is not known how these effects of attention influence feature interference in visual crowding. If decreasing the size of the attention field at the target location increases the spatial resolution of the target representation at that location, this should relieve crowding.

We varied the spatial extent of the attention field from 1 to 3 DVA, resulting in the RF pooling arrays depicted in Figure 2.3. Specifically, the Gaussian attention field acts to pull RF locations towards its center and to reduce their size. We chose a minimum attention field extent that was large enough to ensure that all flanking stimuli were still completely covered by the RF pooling array after Gaussian multiplication. The maximum attention field extent that we used roughly corresponds to the point at which target classification accuracy no longer decreased significantly with increases in the spatial extent of attention. We picked this range of attention field extents to examine the relative performance across the full range of attentional modulation in our model; however, it likely does not have a one-to-one relationship with the full range of attentional modulation in

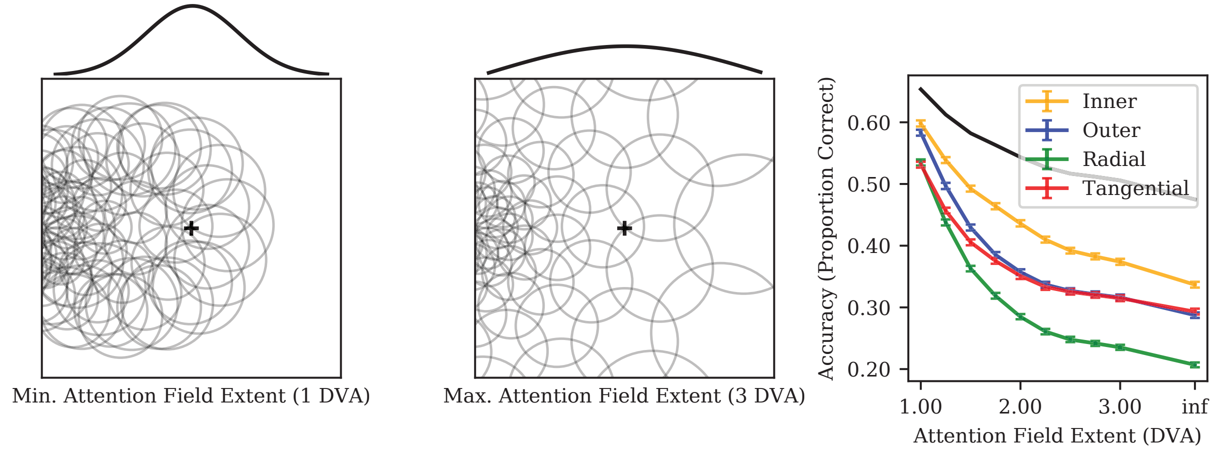

Figure 2.3: (*left*) RF pooling array at the minimum attention field extent. (*middle*) RF pooling array at the maximum attention field extent. Black cross indicates the attended target location. A 1D slice of the Gaussian attention field is displayed above the RF array for both 1 and 3 DVA examples. (*right*) Target classification accuracy as a function of the spatial extent of the attention field for the four target/flanker configurations. The "infinity" point corresponds to no attention field applied to the RF array. Target/flanker spacing was fixed at 1 DVA. The black line indicates accuracy for targets presented without flankers. Error bars are bootstrapped 95% confidence intervals.

human psychophysics (i.e. from pre-cueing the target location).

As expected, more precise attention (smaller spatial extent) centered at the target location resulted in greater target classification accuracy for every target/flanker configuration (Figure 2.3). Moreover, the effect of increasing attention field extent on target classification decayed exponentially. Furthermore, the relationships among the four configurations remained the same as that observed in Figure 2.2, with *outer/radial* having lower target classification accuracy than *inner/tangential* configurations.

Substitution errors occur at above chance levels when crowding is strong. Both increasing target/flanker spacing (Figure 2.2) and decreasing the attention field extent (Figure 2.3) had positive effects on our model's ability to correctly classify the target digit. These increases in performance are consistent with what has been shown in previous human studies. However, target classification is not the only metric that has been used to study visual crowding in human subjects. Substitution errors – or the phenomenon of incorrectly reporting the flanker's identity instead of the target's at an above-chance rate – is an additional metric used to characterize target/flanker interactions in crowding (Coates, Bernard, and Chung, 2019; Ester, Klee, and Awh, 2014; Hanus and Vul, 2013). In this experiment, we analyzed the results from the same target/flanker spacings and attention field extents as before. However, instead of reporting target classification accuracy, we present the number of flanker responses for each configuration as a proportion of incorrect trials (i.e., trials in which the target was not reported). Under strong crowding conditions, RFs that contain both

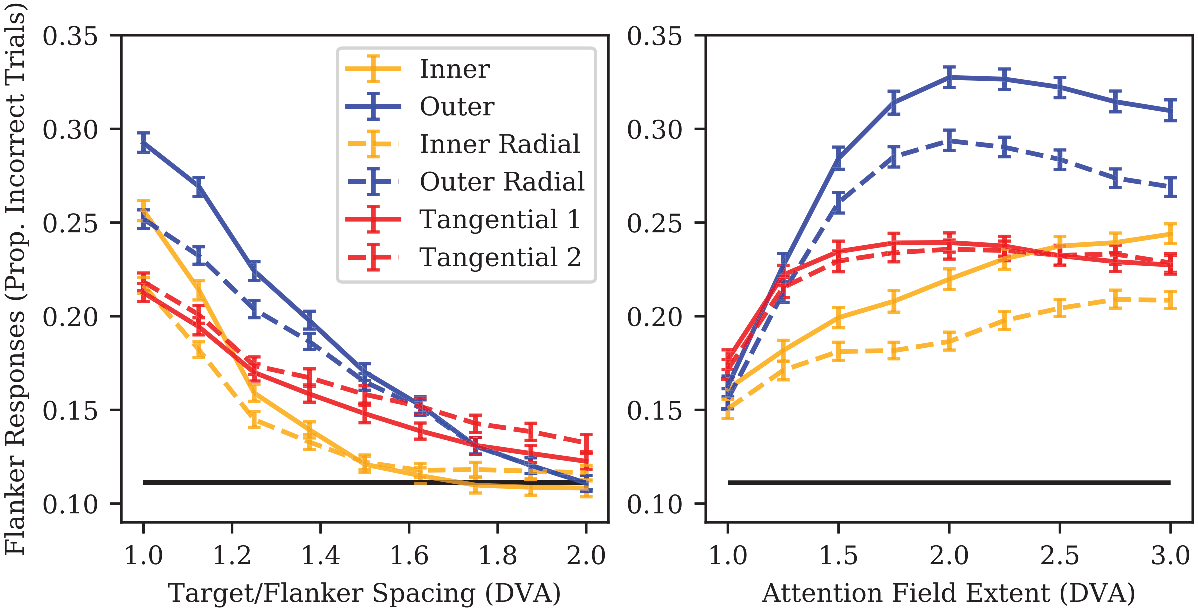

Figure 2.4: Proportion of incorrect trials for which the flanker digit was erroneously reported as a function of (*left*) target/flanker spacing and (*right*) attention field extent. The same four configurations were used as before. The *radial/tangential* configurations each had two possible flanker choices (*Inner/Outer*, *Radial*, and *Tangential 1/2*, respectively). *Tangential 1/2* correspond to the perpendicular flankers placed below and above the radial axis in Figure 2.2, respectively. Black lines indicate chance probability for incorrectly reporting a non-target digit (one out of nine possible digits). Error bars are bootstrapped 95% confidence intervals.

target and flanker features will exhibit competition and therefore have feature interference. This interference should lead to the identities of the flankers being reported at above chance levels on incorrect trials, compared to all other non-target digits.

Figure 2.4 shows that under the strongest crowding conditions (*left*: 1 DVA spacing; *right*: 3 DVA extent), the proportion of trials in which the flankers were identified for each configuration was significantly above chance (black lines). Furthermore, the rate of incorrectly reporting the flanker decreased as the target/flanker spacing increased and as the attention field extent decreased. Interestingly, the *outer* flanker was reported more often than the *inner* flanker across the majority of target/flanker spacings and attention field extents, and this asymmetry was observed both when the *inner*/*outer* flankers were presented as a single flanker with the target (solid yellow and blue lines, respectively) as well as when they were presented as pairs of flankers in the *inner radial*/*outer radial* conditions (dashed yellow and blue lines, respectively). These results suggest that when there is substantial crowding, representations of the identities of the specific flankers are stronger than those of the identities of all other non-target classes. Furthermore, these findings indicate

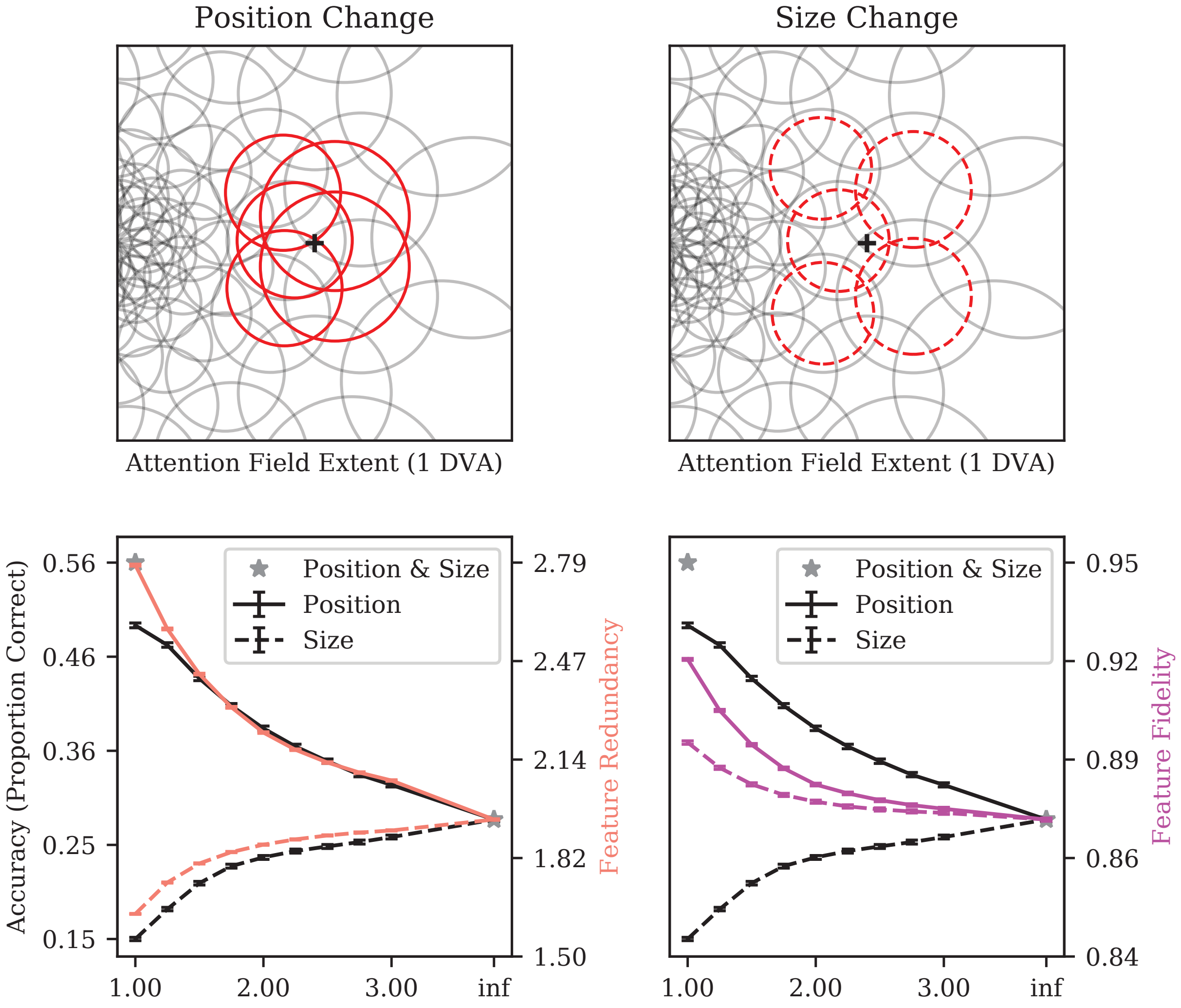

Figure 2.5: Changes in RF position with more focused spatial attention increase the density of RFs at the attended target location (indicated by the black cross, *top left*), whereas changes in RF size alone decrease RF density (*top right*). Target classification accuracy and feature redundancy (Equation 2.7) both increase with attention-related position changes but decrease with size changes (*bottom left*). In contrast, the fidelity of feature representations (as measured by cosine similarity; Equation 2.5) increases with more focused attention for both RF location and size changes (*bottom right*). Note that each of the y-axes has been scaled so that the corresponding metric is plotted relative to the value obtained for that metric following changes in both RF position and size with an attention field extent of 1 DVA (left gray star in each bottom panel) and infinity (i.e., "no attention"; right gray star in each bottom panel). All metrics depicted were averaged across the four target/flanker configurations (Figure 2.2). Error bars are bootstrapped 95% confidence intervals.

that in our model, crowding is due to competition between representations of target and flanker features.

Increases in target classification accuracy depend largely on RF position shifts. We have shown that reconfiguration of the RF pooling array by attention modifies both RF locations and sizes in our model (Figure 2.3). In this experiment, we limited the effects of attention to changes in either the positions or the sizes of the RFs in our pooling array by separately applying updates to either μ or σ from Equation 2.4, respectively. Previous fMRI research in humans indicates that shifts in RF position by attention are more important than changes in RF size for population-level encoding of fine spatial information (Vo, Sprague, and Serences, 2017). This suggests that shifting RFs in our pooling array towards the attended target location, without changing their size, should increase target classification accuracy more than decreasing the sizes of RFs without changing their positions.

We employed the same target/flanker configurations and range of attention field extents as before (Figure 2.3), but here we applied attention effects separately for RF position and size. Target/flanker spacing was fixed at 1 DVA for this experiment. Figure 2.5 shows an example RF pooling array for updated position (top left) and size (top right). As expected, shifts in fixed-size RFs toward the target location with attention increased target classification accuracy (Figure 2.5; black solid line indicates average across configurations). Interestingly, decreasing the size of stationary RFs with attention decreased target classification accuracy (black dashed line). Note that allowing attention to affect both RF position and size together resulted in greater target classification accuracy (i.e., the value of 0.56 indicated by the gray stars at 1 DVA in the bottom panels of Figure 2.5) than either position or size changes alone.

In the second part of this experiment, we characterized the effects of shifts in RF position and size by attention on the redundancy and fidelity of feature representations. As described in Section 2.3, we define redundancy as the average number of RFs that represent an activated target feature when corrupted by the flankers and fidelity as a measure (cosine similarity) of how corrupted the target features are by the flanker features. In order to visualize the relationships among these variables with each other and with target classification accuracy, we plotted each metric in Figure 2.5 relative to the same metric obtained for changes in *both* RF size and position for an attention field extent at infinity ("no attention") and at 1 DVA target/flanker spacing. Each metric is therefore relative to these matched points, which are shown as gray stars in the bottom panels of Figure 2.5.

Shifts in the positions of RFs toward the target location increased the density of target-containing RFs (and therefore the redundancy of feature representations; [Figure 2.5, top left panel]), whereas reductions in the size of RFs decreased redundancy (Figure 2.5, top right panel). We found that feature redundancy (salmon lines) was tightly coupled with target classification accuracy (black lines) for both RF position and size changes (Figure 2.5, bottom left panel) across a range of attention field extents, suggesting that RF density at the target location (i.e., feature redundancy) is strongly related to downstream effects on target classification accuracy.

On the other hand, fidelity of feature representations (magenta lines) increased both when the positions of fixed-size RFs were shifted toward the target location and when stationary RFs shrunk with attention (Figure 2.5, bottom right panel). Decreasing RF size results in less competition for processing between the target and flankers within a single RF, and this is reflected by increased feature fidelity values for smaller spatial extents of attention (magenta dashed line). However, attention field size has a very different relationship with feature fidelity than it has with target classification accuracy, which is worse for smaller attention field size (and therefore for smaller RFs; black dashed line). Together, these results suggest that target classification is more closely related to feature redundancy than it is to the fidelity of feature representations.

Feature redundancy has greater influence than feature fidelity on target classification. As demonstrated by the results of the previous experiment, attentional modulation of RF properties has divergent effects on feature redundancy and fidelity. Intuitively, redundancy of feature representations correlates strongly with RF density (the amount of overlap of RFs), with shifts in RF location towards the attended location increasing redundancy and reductions in RF size decreasing it. In contrast, feature fidelity increases with more focused attention, and this occurs for both effects of attention: RFs moving towards the attended location and shrinking in size.

Although these results indicate a stronger relationship between target classification accuracy and feature redundancy compared to the relationship with feature fidelity (Figure 2.5, bottom panels), interactions between features are dependent on both RF sampling and the relative distance between the target and flanker. In the previous experiment, all results were obtained with a target/flanker spacing of 1 DVA. We therefore conducted an additional experiment to more fully characterize the effects of feature redundancy and fidelity on target classification accuracy.

We selected a subsample of 1,000 test images (from the original 10,000) for each combination of attention field extent and target/flanker spacing values used in the previous experiments. This enabled us to characterize the effects of both of these factors on the full range of observed variation in target classification accuracy that we studied. As shown in Figure 2.6, both redundancy and fidelity were highly correlated with target classification accuracy ($R^2 = 0.96, R^2 = 0.69$, respectively). We then computed the unique variance explained by each factor with multiple linear regression and found that the unique variance explained by redundancy was substantially greater than the unique variance explained by fidelity ($R^2 = 0.28$ vs. $R^2 = 4.96 \times 10^{-3}$, bootstrapped p-value of the difference in explained variance [1000 samples] = 0). These results indicate that redundancy of target feature representations is likely to be more important than fidelity for target classification in visual crowding.

Taken together, the results of all of our experiments provide a more complete understanding of the mechanistic relationships among feature redundancy, fidelity, and target classification for crowded stimuli. Specifically, spatial attention enhances target classification accuracy by increasing the redundancy of sampling of the corrupted target signal, and this greater redundancy is mostly due to increased RF density resulting from position shifts.

2.5 Discussion

Inspired by the normalization model of attention (Reynolds and Heeger, 2009), we constructed a model with a dynamic pooling array of RFs that were modulated by spatial attention in order to characterize how feature redundancy and fidelity relate to downstream target classification during a visual crowding task. Our model reproduced patterns of target classification for different tar-

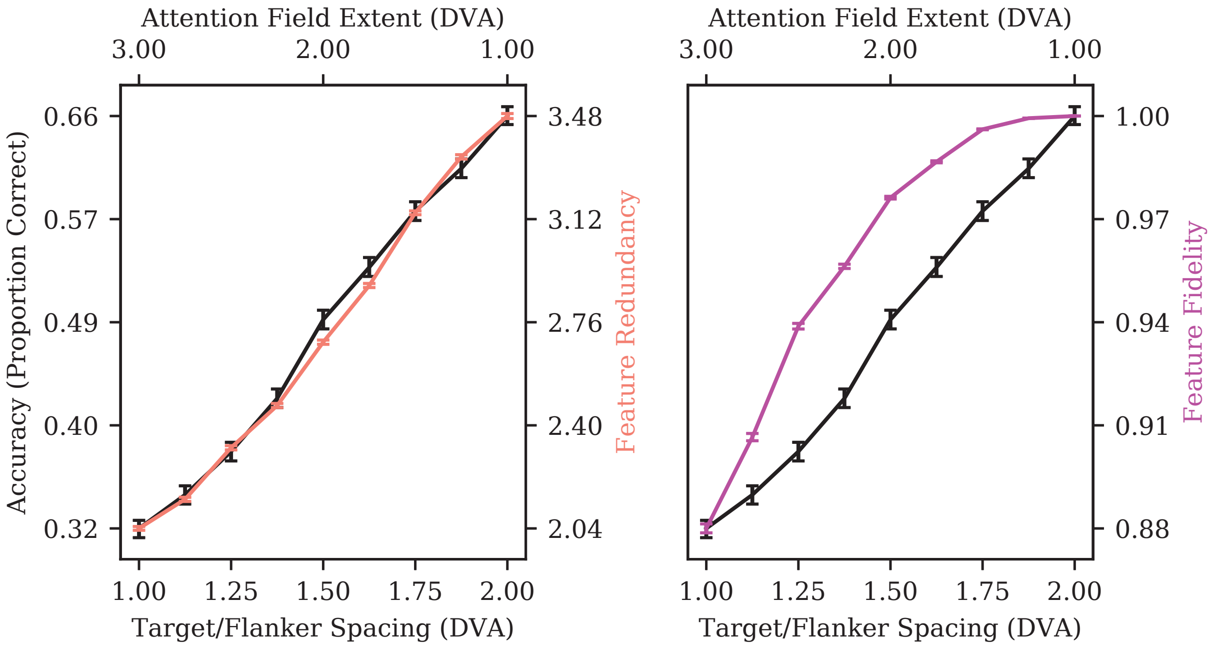

Figure 2.6: Target classification accuracy plotted with feature redundancy (*left*; Equation 2.7) or feature fidelity (*right*; Equation 2.5) for a diagonal cross-section of the combined factors of attention field extent and target/flanker spacing. Target classification accuracy is much more closely related to feature redundancy than it is to feature fidelity. All metrics plotted here were averaged across the four target/flanker configurations. Error bars are bootstrapped 95% confidence intervals.

get/flanker spacings and configurations that have been reported in psychophysical visual crowding experiments (Whitney and Levi, 2011). Next, by separately manipulating the effects of spatial attention on RF size and location, we demonstrated a plausible mechanism by which visual crowding is relieved by position shifts in RFs that increase their density at the attended target location. Finally, by varying target/flanker spacing and the spatial extent of attention, we revealed that feature redundancy explained far more unique variance in target classification accuracy than was explained by feature fidelity (Figure 2.6).

A model of spatial attention effects on downstream processing and perception. In our model, spatial attention increases RF density at the attended target location, resulting in an increase in feature redundancy across populations of RFs that improves target classification in crowded stimuli (Figure 2.5, bottom left panel). Our model does not explicitly contain a metric of response amplitude per se but instead quantifies feature representations in individual RFs. Therefore, we did not explore the effects of attention on response gain in the current study. However, our model is conceptually compatible with literature demonstrating gain modulation by spatial attention (e.g., Moran and Desimone, 1985). The RF pooling operation in our model encodes information in a lossy manner relative to the total information available in the second-layer feature maps. However, more information is preserved with the smaller and more densely organized RFs that are produced by attention, demonstrating an increase in information gain with more precise attention. This is similar to the effect of attention on feature fidelity observed in the bottom right panel of Figure 2.5, in which spatial attention directed toward the target digit increased the fidelity of the encoded target signal.

Similar to Vo, Sprague, and Serences (2017) and Baruch and Yeshurun (2014), we found that shifts in RF position with attention are more important than changes in RF size for improving feature representations. Interestingly, we found that at the minimum attention field extent, target classification accuracy based only on changes in RF position was considerably lower than what would be predicted by its relationship with feature redundancy (Figure 2.5, bottom left panel). This discrepancy may be explained by differences in the effects of feature redundancy measured across partially versus completely overlapping RFs. In our model, as RFs approach complete spatial overlap, they are more likely to represent the exact same pixel locations for a given feature, which does not provide any benefits for target classification. Indeed, Nigam, Pojoga, and Dragoi (2019) demonstrated that synergistic connections within a cortical column in V1 (i.e., connections between nearby neurons sharing very similar RFs) allow for greater decoding of stimulus information than do redundant connections. This physiological result is consistent with our interpretation of our modeling results that feature redundancy across partially overlapping RFs is more beneficial for perception than redundancy within highly overlapping RFs.

RF models of visual crowding. Other models have also utilized biologically-plausible RF pooling arrays to model peripheral vision (Deza and M. Eckstein, 2016; Deza, Jonnalagadda, and M. P. Eckstein, 2019; Volokitin, Roig, and Poggio, 2017), and these types of models have also been shown to reproduce known effects of both target/flanker spacing (Freeman and Simoncelli, 2011) and configuration (Chaney, Fischer, and Whitney, 2014; F. X. Chen et al., 2017; Nandy and Tjan, 2012). Nandy and Tjan (2012) theorized that the *radial/tangential* anisotropy in crowding is caused by a radial bias in image statistics that is attributable to patterns of eye movements that occur during natural vision throughout development. Chaney, Fischer, and Whitney (2014), inspired by the finding that primate V4 RFs have elliptical shapes that reflect V1 cortical magnification (Motter, 2009), observed a *radial/tangential* anisotropy in crowding in their model that is based on a bias in the orientation and length of elliptical RFs that have a major axis in the radial direction. In contrast to this previous work, the *radial/tangential* anisotropy in our model arises from an RF array with eccentricity-dependent and concentric organization that is based on the known properties of human visual cortical area V2 (Wandell and Winawer, 2015). These simple RF organizing principles can also be applied to the study of other visual cortical areas and to encoding of any feature dimension.

One noteworthy challenge for visual crowding models is to incorporate a biologically-plausible method for prioritizing selection of the target over the flankers. F. X. Chen et al. (2017) implemented eccentricity-dependent pooling within a CNN by creating a "multi-scale input" from crops that had different size but identical resolution. However, the authors specifically note that their model did not include a procedure for explicitly selecting target over flanker features. Instead, they computed classification accuracy for crowded digits by using odd MNIST digits as targets and even digits as flankers. In an alternative approach, Chaney, Fischer, and Whitney (2014) trained a different classifier for each target/flanker configuration and spacing based on the outputs of the final layer of a neural network model. Unlike these previous approaches, our model contains a direct target selection mechanism that is based on weighting the pooled features from the RF array as a function of their distance from the target location in cortical space. Because we trained a single classifier only once for all of our experiments, as opposed to multiple classifiers for each experimental condition, our model takes less time to implement, is easily scalable for the study of more complex tasks and stimuli, and avoids possible biases that can occur when employing multiple classifiers (e.g., variability in initial parameter values, local minima in the loss surface, etc.).

More recently, Lonnqvist, A. D. Clarke, and Chakravarthi (2020) reported a study of visual crowding in deep neural networks. Although the authors observed striking differences between the pattern of visual crowding observed in CNNs and what has typically been observed in human studies, there are important differences between their study and ours. Lonnqvist, A. D. Clarke, and Chakravarthi (2020) logarithmically downsampled images in order to simulate peripheral vision, whereas our model used eccentricity-dependent RF pooling of feature maps. However, downsampling the image simulates peripheral visual input rather than peripheral visual processing, and it is inconsistent with the interpretation of visual crowding as a high-contrast mixing of stimulus features. Additionally, Lonnqvist, A. D. Clarke, and Chakravarthi (2020) did not incorporate a selection mechanism for classifying target objects separate from flankers but instead trained their model to classify a single object at a target location, followed by testing with both target and flanking objects. It is possible that their inability to observe increased performance as a function of target/flanker spacing (e.g., our Figure 2.2, right panel) was due to overfitting during the target-alone training procedure in their model. These differences highlight the importance of eccentricity-dependent pooling and selection mechanisms for successfully modeling visual crowding.

Computational models of attention. Many existing models have studied spatial attention in the context of bottom-up saliency (e.g., Itti, Koch, and Niebur, 1998). While such models have been useful for characterizing which aspects of visual features attract attention, our model instead focuses on how attention affects feature representations. Jia, C. Huang, and Darrell (2012) and Cheung, E. Weiss, and B. Olshausen (2016) both used an approach that is similar to our RF pooling mechanism by sampling images with a mutable array of RFs. However, in both of these studies, spatial information was disregarded following the pooling operation. In contrast, we believe that our model will more effectively generalize to other tasks by maintaining spatial information after RF pooling, since this allows the pooling operation to occur at any level of a CNN.

In Jia, C. Huang, and Darrell (2012), the spatial organization of RFs was learned in order to optimize image classification, which in the context of our study can be viewed as optimizing covert spatial attention (directing attention to a peripheral visual field location without eye movements). On the other hand, Cheung, E. Weiss, and B. Olshausen (2016) employed overt attention (shifts of attention that are accompanied by eye movements to the attended location) during a visual search task to learn an optimal sampling lattice. Interestingly, they found that the optimal lattice for target classification contains a foveated region that is similar to that observed in the human retina. A strength of our RF pooling method is that the attention field or RF parameters can be learned through gradient descent, which future researchers can use to explore similar hypotheses regarding optimal biological structures and mechanisms. Moreover, the specific pooling operation (e.g., max-pooling) in our model can be changed to better reflect biological mechanisms, such as a stochastic pooling operation to study how noise might interact with the effects of spatial attention.

Our model's RF reconfiguration by attention is probably most similar to the Attentional Attraction Field (AAF) model described by Baruch and Yeshurun (2014). They showed that attraction of RFs towards an attended location accounts for a number of known spatial and temporal aspects of attention, such as enhanced resolution, gain modulation, and biased competition. We build from the results of the AAF model by quantitatively characterizing the differential contributions of changes in RF size and position to performance on a perceptual task and to the redundancy and fidelity of feature representations.

There are also several models in which spatial attention has been implemented through enhanced responses (e.g., B. A. Olshausen, Anderson, and Van Essen, 1993; Mozer and Sitton, 1998; Hamker, 2004). For instance, Deco and T. S. Lee (2002) used a set of Gaussian weights similar to our cortical weighting mechanism (see Section 2.3) to enhance responses within an attended region. However, our model uses cortical weighting as a method for selecting target features for classification, not for gain modulation.

Increasingly, attention has been implemented in deep neural networks (e.g., Sabour, Frosst, and Hinton, 2017; Vaswani et al., 2017) to selectively sample and enhance information in a task-agnostic manner. This is an important challenge in machine learning, since it is notoriously difficult to train neural networks to generalize to multiple tasks without a significant decrease in performance on the original task for which the network was trained (French, 1999). However, humans can dynamically change the relative weights of feature representations for a given task via spatial and/or feature-based attention. In our model, Gaussian multiplication is an effective implementation of a circular "spotlight" of spatial attention. However, it currently does not allow updating of RF properties for more complex attention fields (e.g., curved contours, shapes, or objects; Somers et al., 1999). Perhaps the effects of more complex attention fields on RF properties would be similar to object detection techniques that are commonly used in machine learning (Ren et al., 2015), in which the appropriate resolution is dictated by the current task and/or local features. Therefore, future research could treat the size, position, and other parameters of the attention field used in the current study as parameters that could be adapted for specific tasks. Our modeling approach is very compatible with this direction, as the parameters of the attention field could be directly optimized during the neural network training process. Such an approach could be used to make predictions of RF changes measured via fMRI for perceptual tasks in which greater spatial resolution of attention can paradoxically lessen performance (e.g., Yeshurun and Carrasco, 1998; Barbot and Carrasco, 2017). In this way, combining predictions made by our model with experimental data could provide further insights into the adaptability of spatial attention and its consequences for perception.

Chapter 3

Top-down priors disambiguate among target and distractor features in simulated covert visual search

3.1 Abstract

Several models of attention during visual search have been proposed to study gaze behavior by considering visual attention as part of a perceptual inference process, where top-down priors disambiguate bottom-up sensory information. Relatively fewer models have been proposed to study covert attention, during which attention is directed to a region of the image without shifting one's gaze. Here, we propose a biologically plausible model of covert attention during visual search using top-down priors learned over target features and spatial resampling of modeled cortical receptive fields to enhance local spatial resolution for downstream target classification. We evaluate this model during simulated visual search for handwritten digits among non-digit distractors, finding that top-down priors improve target location and classification accuracy relative to bottom-up signals alone. Our results support previous literature demonstrating the effect of top-down priors on visual search performance, while extending the literature to incorporate known mechanisms of spatial attention to study covert attention.

3.2 Introduction

Due to dynamic environmental factors (e.g., lighting, motion, occlusion, etc.), humans frequently encounter noisy and/or ambiguous visual stimuli in everyday life. For example, the same object viewed from different angles can project wildly different geometries onto the retina. There is inherent uncertainty in visual perception of many natural stimuli, but humans are rarely disoriented when navigating complex environments. For example, humans encode and utilize uncertainty in making predictions of object speed (Y. Weiss, Simoncelli, and Adelson, 2002) and size (Ernst and M. S. Banks, 2002). The Bayesian coding hypothesis (Knill and Pouget, 2004) suggests that encoding uncertainty allows humans to infer complex properties from noisy sensory information. Within this framework, the cortex is hypothesized to encode the conditional probability of features, given a sensory input.

In line with this hypothesis, Rao (2005) proposed a probabilistic generative model of attention, which assumed that the visual system converges to probable explanations of observations through a combination of bottom-up likelihoods of sensory information and top-down priors over spatial locations and features by using Bayes' rule. In artificial experiments, the generative model was trained to represent probability distributions of stimuli over location and orientation dimensions. In the feedforward direction, the posterior probabilities of location and orientation were inferred from an image. In the feedback direction, the prior probabilities over features or locations were used to influence an intermediate representation of the stimulus and to update the posterior probabilities. Importantly, the model of Rao (2005) demonstrated that feedback of prior probabilities over spatial locations can mimic effects of top-down attention that have been well-characterized in neurophysiological research.

Previous Bayesian models of attention have also successfully modeled human eye movements during visual search and free-viewing conditions with natural images. The contextual guidance model (Torralba, Oliva, et al., 2006) combined bottom-up saliency computed by a local pathway with scene priors computed by a global pathway. Impressively, the model was able to predict eye movements during visual search for people, paintings, and mugs in natural scenes. Chikkerur et al. (2010) used a similar approach but instead modeled top-down attention during visual search as a combination of both spatial and feature priors. This model was used to demonstrate how a Bayesian framework of attention can account for various known effects: feature pop-out (Bravo and Nakayama, 1992), multiplicative modulation of response amplitude (McAdams and Maunsell, 1999), as well as shift and gain effects of the contrast response function (Martınez-Trujillo and Treue, 2002; Treue and Trujillo, 1999). Furthermore, the model of Chikkerur et al. (2010) accounted for eye movements during both visual search and free viewing with natural images.

Since eye movements are a common outcome measure of visual search tasks, most studies have focused on modeling overt as opposed to covert visual attention (i.e., directing spatial attention to a particular location without altering gaze position). Although the premotor theory of attention posits that the same process underlies both covert and overt attention (Rizzolatti et al., 1987), sustained covert attention directly enhances the representations of encoded features, whereas overt attention uses the structural advantage of central vision to improve spatial sampling. During covert spatial attention, receptive fields (RFs) in early visual cortex, which are smaller in central vision and larger in the periphery, shift toward the attended location and decrease in size (Klein, Harvey, and Dumoulin, 2014; Womelsdorf et al., 2006).

Theiss, Bowen, and Silver (2022) introduced a computational model of cortical RFs as a dynamic pooling array within a convolutional neural network. This RF pooling array was updated by Gaussian multiplication with an attention field, modeling known effects of spatial attention on properties of neuronal and population-level RFs in visual cortex (Klein, Harvey, and Dumoulin, 2014; Womelsdorf et al., 2006). The validity of this model was demonstrated across multiple experiments that replicated results observed in psychophysical studies of visual crowding in humans. For visual search, the RF pooling array can be used to simulate allocation of covert spatial attention to a predicted target in order to enhance local spatial processing and improve downstream target classification.

In the current study, we propose a Bayesian model of attention that learns priors over target features and employs endogenous spatial attention in order to simulate covert visual search. The model uses these feature priors to dismabiguate bottom-up signals related to target and distractor features, simulating feature-based attention to highlight the location of the target. This induces a spatial prior at the predicted target location, which is then used to enhance the encoded representation of the target features for classification. We test this model using a search task for handwritten digits among non-digit distractors to evaluate target location and classification accuracy. Although we focus on visual search with artificial images, we discuss how the model could be extended to more complex tasks with natural images.

3.3 Methods

Model Description

Hierarchical generative model. The current model builds on previous hierarchical Bayesian models (Chikkerur et al., 2010; Rao, 2005; Torralba, Oliva, et al., 2006) but learns priors over features using a 2-layer convolutional deep belief network (CDBN; H. Lee et al., 2009). The CDBN (Figure 3.1) is a hierarchical generative model composed of multiple restricted Boltzmann machine (RBM) layers (Smolensky, 1986). Each RBM layer models its input with a set of hidden units, which are active with the following probability:

$$P(h_j = 1|\mathbf{v}) = \sigma(b_j + \sum_i v_i w_{ij}) \tag{3.1}$$

$$P(v_i = 1|\mathbf{h}) = \sigma(c_i + \sum_j h_j w_{ij}), \tag{3.2}$$

where h_j represents a single hidden unit, v_i represents a visible unit of the input $\mathbf{v}$, w_{ij} represents the weight between v_i and h_j, b_j represents the bias for hidden unit h_j, c_i represents the bias for visible unit v_i, and σ is the sigmoid function. The above equations are used to obtain conditional probabilities for hidden units given an input and vice versa for visible units. The model is trained to represent a data distribution by increasing the probability assigned to data examples while decreasing the probability assigned to model-generated examples using an algorithm known as contrastive divergence (CD; Hinton, 2002):

$$\Delta w_{ij} \propto \langle v_i h_j \rangle_{\text{data}} - \langle v_i h_j \rangle_{\text{model}} \tag{3.3}$$

Although the expected values for the first term can be obtained directly using Equation 3.1, it is intractable to obtain expected values over the model distribution (2^{M+N} combinations of M visible and N hidden units). However, samples can be estimated efficiently using block Gibbs sampling, whereby visible and hidden units are sampled alternately (Bernoulli sampling using Equations 3.1

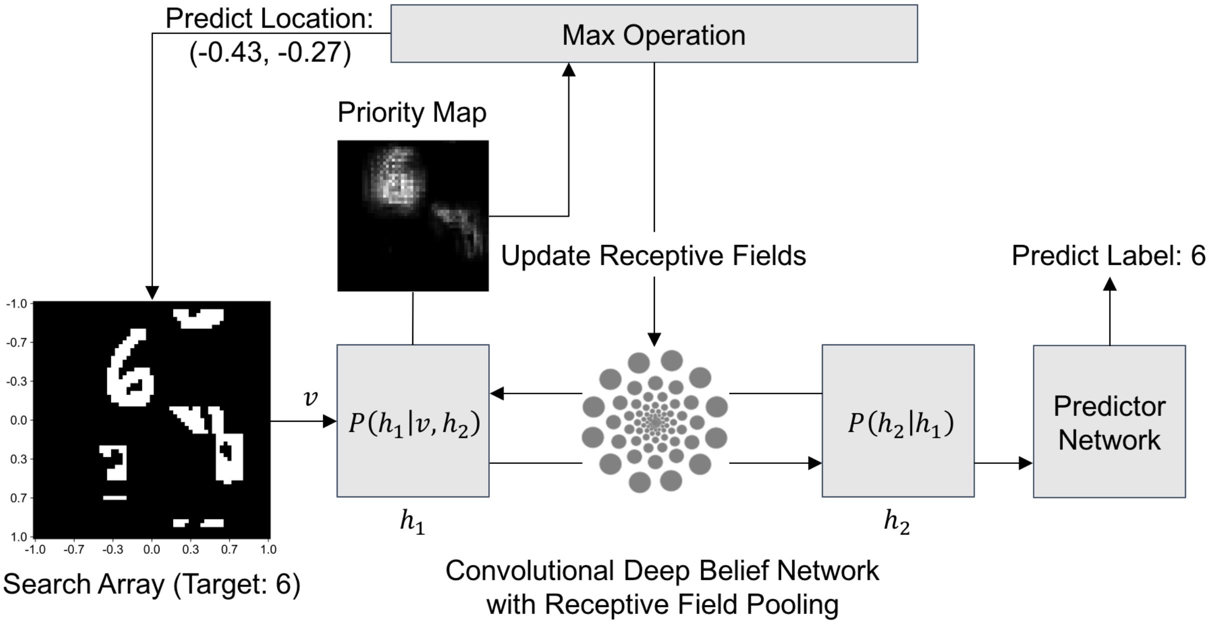

Figure 3.1: Model of covert attention during visual search using top-down priors and spatial resampling. Visual search for handwritten digits among non-digit distractors (Search Array) was evaluated for location (Predict Location) and classification (Predict Label) accuracy. During training, priors were learned over digit features in the first layer (h_1), which were combined with bottom-up signals to generate the priority map ($P(h_1|v, h_2)$ during evaluation. The location with the maximum value in the priority map was selected as the target prediction and used to spatially resample (Equation 3.5) the receptive pooling array to enhance the spatial resolution of feature representations at that location for downstream classification.

and 3.2). With more Gibbs sampling steps, the distribution of the model can be more closely approximated (known as CD$_n$ where n denotes the number of Gibbs sampling steps). Once a single RBM layer is trained, its weights are fixed, and additional layers can then be trained on the outputs of the previous layer in a layer-wise manner, thereby forming a Convolutional Deep Belief Network (CDBN; Hinton and Salakhutdinov, 2006). Since the first layer learns to represent the probability of the data distribution, training a subsequent layer forms a prior over the first layer, $P(h_1|h_2)$.

Within the context of the CDBN, the convolution operation results in a set of feature maps, within which each hidden unit has a receptive field covering a specific portion of the input. When trained on handwritten digits (MNIST; LeCun et al., 1998), the first layer of the CDBN learns to represent oriented and curved lines that are characteristic of parts of MNIST digits (Figure 3.2). Each layer's hidden units then constitute a set of feature maps, with the second layer representing complex features such as entire digits. In order to obtain a vector representation of a digit in the second-layer hidden units, a probabilistic max-pooling operation is performed at the first layer to

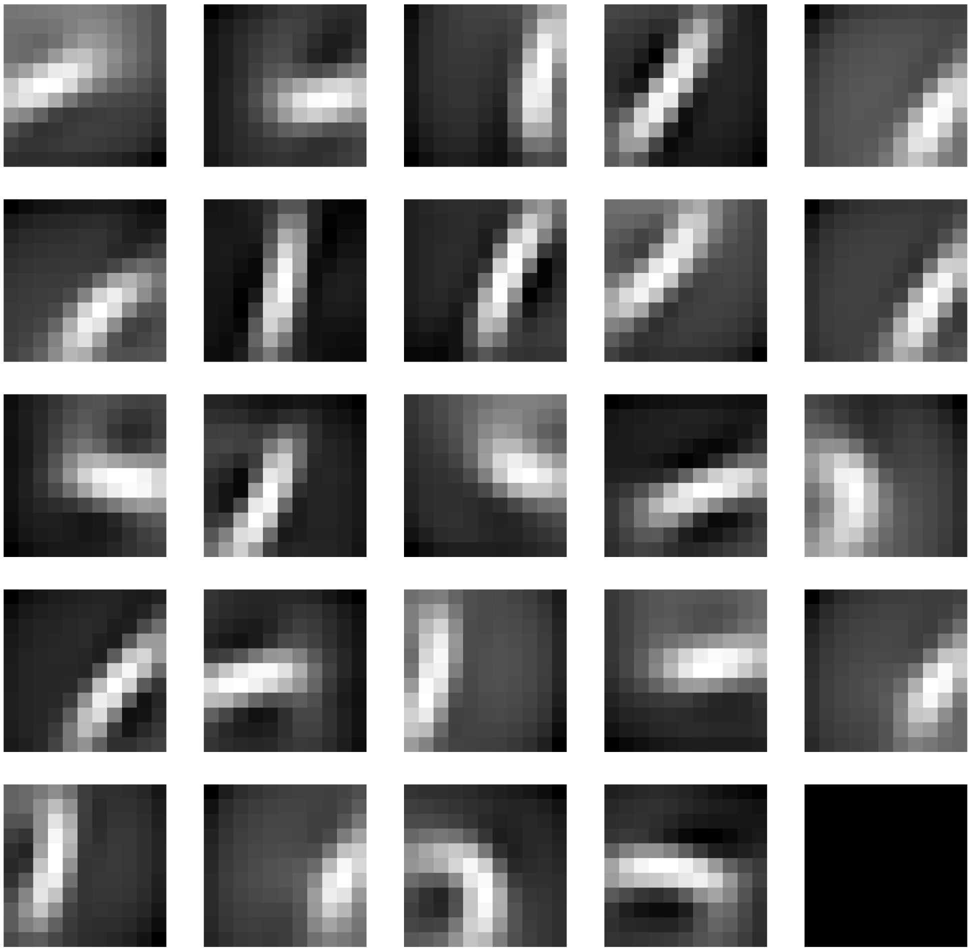

Figure 3.2: Weights in the first layer of the CDBN (h_1 in Figure 3.1) following unsupervised learning with handwritten digits. Whereas first-layer weights comprise oriented and curved lines characteristic of digit parts, second-layer weights (not shown) are combinations of the first-layer weights that can represent entire digits.

reduce the image size of its hidden units (H. Lee et al., 2009). Using probabilistic max-pooling, blocks of hidden units (e.g., 2×2) are modeled as multinomial units in which a single unit is "on" or all units within the block are "off". This provides a straightforward way of obtaining the posterior probability of hidden units given the input and top-down feedback from the layer above:

$$P(h_j^k = 1|\mathbf{v}, \mathbf{h}') = \frac{\exp(I(h_j^k) + I(p_\alpha^k))}{1 + \sum_{B_\alpha} \exp(I(h_j^k) + I(p_\alpha^k))} \tag{3.4}$$

where $\mathbf{h}'$ represents the second-layer hidden units, B_α represents the block (indexed by α) containing hidden unit h_j^k (with feature map index k and pixel index j), $I(h_j^k)$ represents the bottom-up contribution (convolution of first-layer weights with the input plus bias), and $I(p_\alpha^k)$ represents the top-down contribution (transposed convolution of the second-layer weights with the second-layer hidden units).

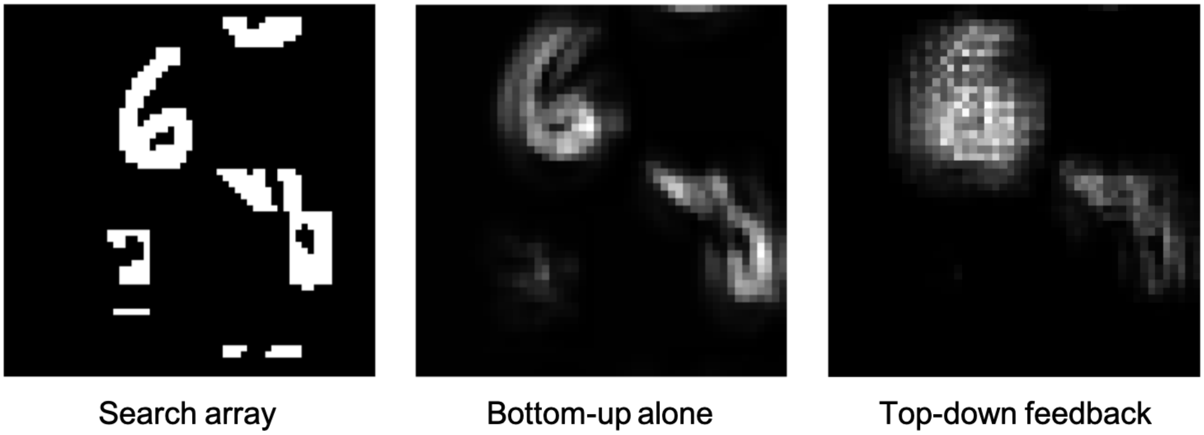

Figure 3.3: Example search array and priority maps for simulated visual search. The search array (*left*) contained a single target (here, 6) among non-digit distractors. Comparing the "bottom-up alone" (*middle*) and "top-down feedback" (*right*) priority maps demonstrates the effect of top-down priors to disambiguate among first-layer target and distractor features by assigning greater priority to the target relative to distractors.

Attention priority maps. Using Equations 3.1 and 3.4, we computed priority maps representing the bottom-up conditional probabilities over first-layer features (i.e., $P(\mathbf{h_1}|\mathbf{v})$; "bottom-up alone") as well as the posterior probability of those features given the top-down prior (i.e., $P(\mathbf{h_1}|\mathbf{v}, \mathbf{h_2})$; "top-down feedback"), respectively. Since the probabilities across first-layer features correspond to a vector at each pixel location, we normalized the sum across probabilities to the maximum value to generate the priority map for both conditions. In order to spatially attend to the predicted target during visual search, the location with the maximum value in the priority map was selected, as demonstrated in Figure 3.1.

Visual search for digits should be facilitated by priority maps that represent the relative probability of digit features present in an image. However, since the first-layer features comprise digit fragments (Figure 3.2), non-digit distractors could be confused with digits when using a priority map of bottom-up conditional probabilities alone. Therefore, the priority map incorporating the top-down prior across first-layer features should help to disambiguate which regions of the image contain a digit versus non-digit distractor. This effect is demonstrated in Figure 3.3, where the relative priority of the target digit and rightmost distractor are similar in the bottom-up priority map (middle panel) but are more divergent in favor of the digit in the priority map incorporating top-down priors (right panel).

Receptive field pooling array. When attention is covertly directed to a particular region of the visual field (without changing gaze position), RFs in visual cortex tend to shrink and shift toward the attended location (Klein, Harvey, and Dumoulin, 2014; Theiss, Bowen, and Silver, 2022; Womelsdorf et al., 2006). This has the effect of locally enhancing the spatial resolution of feature representations at the attended location. In order to model cortical RFs during visual search (but not during training), we replaced the pooling blocks described in Equation 3.4 with receptive fields of variable size that can be dynamically updated. This dynamic receptive field pooling array maintains a location and size for each receptive field individually (see Theiss, Bowen, and Silver (2022) for more details). The locations and distributions of RFs in the pooling array were selected to mimic a centrally fixated gaze, with greater density and smaller RFs at the center of the search array (Figure 3.1). In order to simulate the eccentricity dependence observed in visual cortical RFs, a scaling rate was used to determine the change in size and spacing of RFs as a function of eccentricity (Theiss, Bowen, and Silver, 2022). For the current study, the scaling rate was set to 0.1 in order to mimic V1 cortical RFs (Wandell and Winawer, 2015).

Prior to implementing probabilistic max-pooling, we first obtained the hidden unit outputs ($(I(h_j^k)$ in Equation 3.4) for a given RF using a masked array (receptive fields × height × width) with values of 1 corresponding to pixels contained within the RF and 0 for pixels outside the RF. A Gaussian attention field centered at the predicted target location was then multiplied with the RF array such that the location and size of each RF was updated:

$$\mu = \frac{\mu_{RF}\sigma_{AF}^2 + \mu_{AF}\sigma_{RF}^2}{\sigma_{AF}^2 + \sigma_{RF}^2}, \quad \sigma^2 = \frac{\sigma_{RF}^2\sigma_{AF}^2}{\sigma_{RF}^2 + \sigma_{AF}^2} \tag{3.5}$$

where μ_{RF} and σ_{RF} represent the location and size of each Gaussian RF, and μ_{AF} and σ_{AF} represent the same parameters for the attention field. For the current study, μ_{AF} was set to the predicted location on each trial and σ_{AF} was set to 8 pixels in the feature map space, with $2\sigma_{AF}$ approximating the size of an MNIST digit in image space.

Predictor network. In order to evaluate the effect of spatial attention on intermediate feature representations, we trained a predictor network using the extracted features from the second layer of the CDBN. As shown in Figure 3.1, the predictor network classified the digit after the RF pooling array in the first layer was updated via spatial attention at the predicted target location (Equation 3.5). Unlike the CDBN, the predictor network is a strictly feedforward neural network with two convolutional layers with ReLU and softmax activation functions, respectively (see Table 3.1). The convolutional filter sizes were chosen such that the output for the 60×60 search array would be a 10-dimensional vector, corresponding to the softmax values for each digit class (i.e., 0–9).

Table 3.1: Model architecture used for training

Network	Input	Output	Conv	Activation	Pool
CDBN	1	24	11×11	None	ProbMax 2×2
	24	40	9×9	Sigmoid	None
Predictor Network	40	64	8×8	ReLU	Max 2×2
	64	10	5×5	Softmax	None

Predictions were made based on the index of the 10-dimensional vector containing the maximum value.

Model training. The complete model as shown in Figure 3.1 and Table 3.1 was trained in two steps. First, the CDBN was trained layer-wise with unsupervised learning to model the data distribution using the contrastive divergence algorithm described above (CD$_1$; Hinton, 2002). For this portion of the training paradigm, the input to the model was a 28×28 pixel image of a handwritten digit (LeCun et al., 1998). Each layer was trained with a mini-batch size of one for 40 epochs (i.e., 40 passes through the training set of 60,000 images) using an initial learning rate of 0.02 and initial momentum of 0.5 (set to 0.9 after four epochs). The learning rate was decayed after each epoch with a time-based schedule and decay rate set to 0.01 as done in H. Lee et al. (2009). In order to reduce overfitting and encourage sparsely active hidden units, $L2$ weight-decay and sparsity constraints were used during training (Hinton, 2012).

The predictor network was subsequently trained for ten epochs using supervised learning for digit classification with backpropagation (stochastic gradient descent with a learning rate of 0.001 and momentum of 0.9). In order to train the predictor network to classify digits presented anywhere in the search array, the 28×28 pixel MNIST digit was first padded on each side with zeros to match the size of the search array (60×60 pixels). Then the digit was randomly translated horizontally and vertically up to a maximum of 15 pixels (0.25 of the search array size). The image was then passed through the CDBN, and the predictor network was trained on the second-layer extracted features. The trained predictor network achieved a classification accuracy of 81.26% on the held-out test set of 10,000 images padded to 60×60 pixels (chance-level accuracy is 10%). For all further experiments, the 2×2 pooling operation in the first layer of the CDBN (Table 3.1) was replaced with the RF pooling array, which performed probabilistic max-pooling across each RF instead of 2×2 blocks of pixels.

Model overview. In summary, the model shown in Figure 3.1 contained three main components: a Bayesian model of attention (CDBN), an RF pooling layer, and a predictor network. During visual search, features were extracted by the CDBN, the target location was predicted from a priority map in order to update the RF pooling array, and the enhanced features were used to classify the target digit. Following Theiss, Bowen, and Silver (2022), we multiplied a Gaussian attention field with the RF pooling array to model the effects of spatial attention on feature representations. Similar to Chikkerur et al. (2010), we considered the Gaussian attention field to be a spatial prior over the predicted target location from the priority map.

Two priority maps were evaluated for target location and classification accuracy. The bottom-up priority map (middle panel of Figure 3.3) represented the conditional probability of first-layer features, given the visual search array. The priority map with top-down feedback (right panel Figure 3.3) incorporated second-layer priors over first-layer features to help disambiguate features representing both targets and distractors. These priors represented the probability distribution over combinations of curved and oriented features (Figure 3.2) that constitute handwritten digits. By comparing the differences in location and classification accuracy between these two priority maps, we quantified the effect of top-down feature priors on visual search performance.

Experimental Design and Statistical Analyses

Visual search experiment. The visual search experiment for digits among non-digit distractors contained 10,000 search arrays using the held-out MNIST test set (LeCun et al., 1998). Each 60×60 pixel search array contained a single 28×28 target MNIST digit placed in a random location among various distractors (left panel of Figure 3.3). In order to encourage ambiguity in first-layer feature representations, the distractors were generated from fragments of digits (described in detail below). As shown in the middle panel of Figure 3.3, this had the effect of increasing the uncertainty of the target location in the “bottom-up alone” priority map. On each trial, the location of the maximum value in the “bottom-up alone” and “top-down feedback” priority maps was selected as the predicted target location, which was then used to update the RF pooling array using Equation 3.5, with μ_{AF} set to the predicted location. Following this update, the predictor network classified the target digit using the second-layer features, separately for the two priority map conditions.

Non-digit distractors. In order to generate distractors that contained similar first-level features as target digits, we manipulated portions of four randomly-selected MNIST digits (per search array) from the test set of 10,000 digits. For each randomly-selected distractor digit, we cropped the 28×28 MNIST image to the central 14×14 pixels, randomly rotated the cropped image by one of $[0, 90, 180, 270]$ degrees, and randomly zeroed half of the resulting image along either the horizontal or vertical axis. As shown in the left panel of Figure 3.3, the resulting distractors contain digit fragments but are not identifiable as any particular digit. In order to avoid spatial overlap with the target digit, each distractor was randomly placed in the search array such that the center-to-center distance to the target digit was greater than 7.5 pixels (0.125 of search array size) along either axis. Note that this still allows distractors to overlap parts of the target digit as well as other distractors.

Statistical procedures. Target location accuracy was evaluated by computing precision and recall for each search array by varying the threshold of the priority map between 0 and 1 with a step size of 0.01. Although location accuracy could also be measured as the Euclidean distance between the target center and predicted location (among other metrics), we chose precision and recall in order to obtain a more complete account of location accuracy performance. The method described below has been used to evaluate saliency model performance for fixation predictions (e.g., J. Wang et al., 2016) as well as visual saliency detection (Xie and Lu, 2011), which is perhaps more relevant to the current evaluation.

For each threshold between 0 and 1, pixels in the priority map with values above the threshold were considered positive (i.e., target) predictions, whereas those below the threshold were considered negative (i.e., background) predictions. Above-threshold pixels were considered true positives if they overlapped a 16×16 block of pixels at the center of the target location in the priority map (approximately the size of the MNIST digit in image space). Precision is the proportion of above-threshold pixels overlapping the target relative to all above-threshold pixels, which in this case indicates the relative priority of target vs. distractor locations. Recall is the proportion of above-threshold pixels overlapping the target relative to the target area, which indicates the sensitivity for detecting the target within the priority map. Average precision (AP) was then computed for each trial using the following equation:

$$AP = \sum_{n} (R_n - R_{n-1}) P_n \tag{3.6}$$

where $(R_n - R_{n-1})$ represents the change in recall rates between thresholds n and $n - 1$, and P_n represents the precision at threshold n (Zhu, 2004). We then averaged across trials to calculate the AP for a given condition. Chance level for precision was 0.1024 (i.e., the proportion of ground-truth target pixels).

Target classification accuracy was evaluated as the proportion of trials correctly classifying the target digit. In addition to evaluating target classification accuracy for the two priority maps, we also evaluated classification accuracy for each trial without updating the RF pooling array as a control condition.

In order to obtain 95$\%$ confidence intervals for our estimates, we used 1000 iterations of bootstrap resampling of the data with replacement. For statistical comparisons between two distributions, we first centered each distribution's mean at the combined mean of the two distributions and then bootstrap resampled (again with 1000 iterations) from the centered distributions. We report p-values as the proportion of observed mean differences that were greater than the original mean difference (Efron and Tibshirani, 1994).

Code/software. We implemented all training and computation in PyTorch (Paszke et al., 2017) as well as custom Python code. The code used to produce the results described in this paper is available upon request.

3.4 Results

Top-down feature priors improve target location accuracy. To evaluate the effects of top-down priors on visual search performance, we tested target location and classification accuracy with either the "bottom-up alone" (i.e., $p(h_1|v)$) or "top-down feedback" (i.e., $p(h_1|v, h_2)$) priority maps. For target location accuracy, we plotted the precision-recall curve (Figure 3.4), which displays the predictive performance for both priority maps relative to a random baseline, where greater area under the curve (AUC) indicates a better model for predicting the target location.

As described in Section 3.3 above, the precision-recall curve is computed by thresholding the priority map between 0 and 1, where a high threshold preserves only the greatest values in the

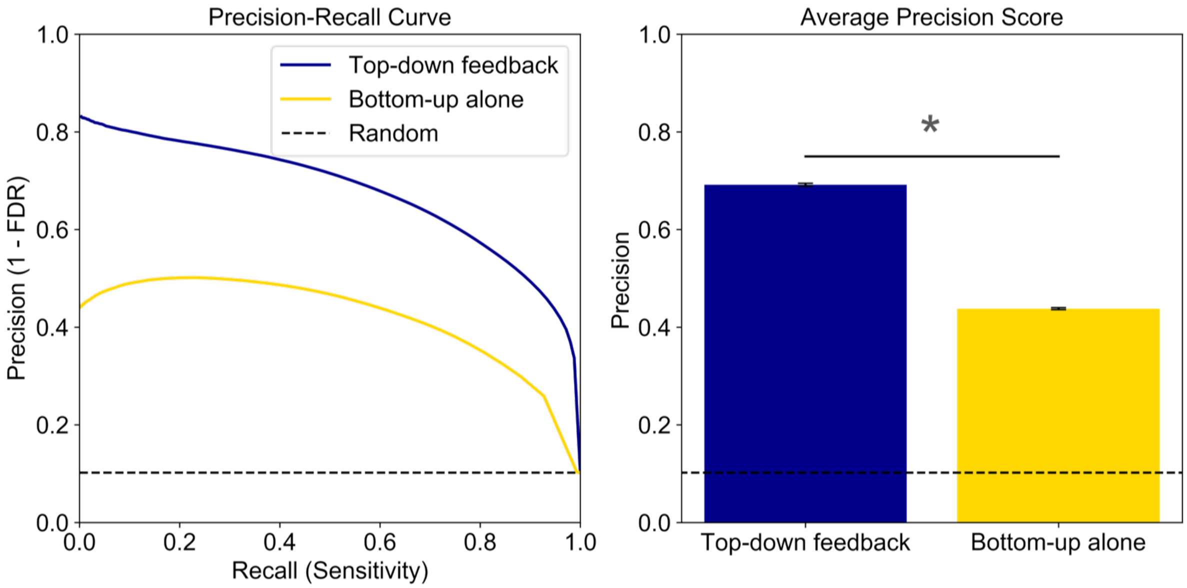

Figure 3.4: Precision-recall curve (*left*) and average precision (*right*) evaluating location accuracy for "bottom-up alone" and "top-down feedback" priority maps. By binarizing the priority maps at various thresholds, the precision-recall curve indicates the relative proportion of pixels assigned to target vs. distractors (precision) as a function of target overlap (recall) in each thresholded priority map. Average precision (Equation 3.6) is a summary metric of the precision-recall curve that is equivalent to area under the curve. Together, these results demonstrate that top-down priors helped disambiguate bottom-up signals by highlighting regions associated with the target relative to distractors. Chance performance is indicated by the dashed line (i.e., proportion of ground-truth target pixels). Error bars are bootstrapped 95% confidence intervals. *Bootstrapped p-value [1000 samples] = 0.

priority map. Therefore, high precision at low recall (as seen for the "top-down feedback" priority map in Figure 3.4) indicates that the greatest values in the priority map were more likely to overlap the target digit than distractors. Alternatively, the precision-recall curve for the "bottom-up alone" priority map indicates that as fewer above-threshold pixels overlapped the target (low recall), the proportion of pixels overlapping distractors increased (low precision). The "random" baseline (dotted line) can be viewed as the performance of a model that predicts target locations with a probability equal to the proportion of ground-truth target pixels (i.e., 0.1024). Both the "bottom-up alone" and "top-down feedback" priority maps clearly surpassed baseline performance.

Average precision (Equation 3.6) is a summary metric of the precision-recall curve that is equivalent to AUC. High average precision therefore indicates greater priority for target vs. distractor locations across all recall rates. As shown in the right panel of Figure 3.4, incorporating top-down priors in the priority map improved target location accuracy relative to using the bottom-

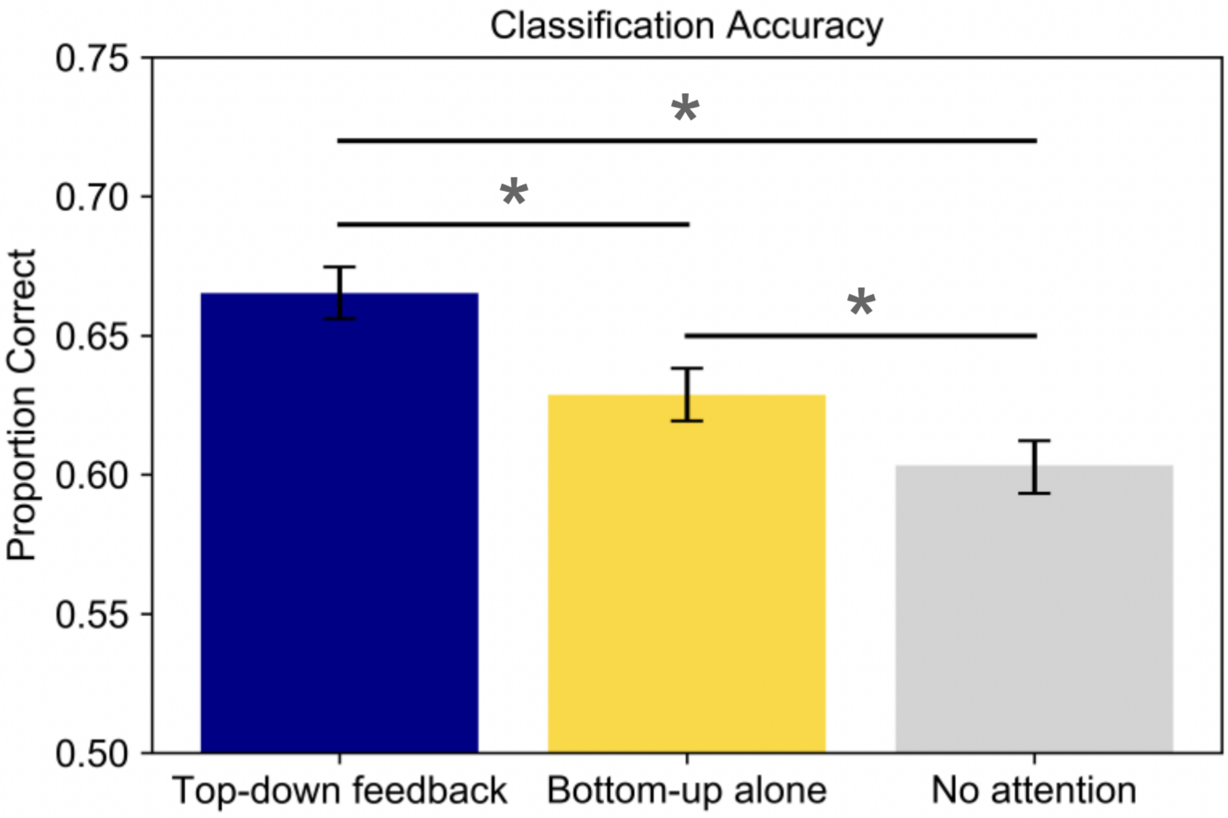

Figure 3.5: Classification accuracy during simulated visual search using spatial resampling of RFs at locations predicted by “bottom-up alone” or “top-down feedback” priority maps. “No attention” indicates classification accuracy without updating the RF array. The results demonstrate that spatial attention to predicted locations improved classification accuracy, with “top-down feedback” priority maps resulting in greater performance relative to “bottom-up alone”. Chance performance is 0.1 (one out of ten possible digits). Error bars are bootstrapped 95% confidence intervals. *Bootstrapped p-value [1000 samples] = 0.

up conditional probabilities alone (0.69 versus 0.44; bootstrapped p-value [1000 samples] = 0.).

Top-down priors disambiguate among target and distractor features. In order to evaluate the effect of spatial attention on encoded features at the predicted location, we used a predictor network to classify the target digit based on the updated features from the second layer of the CDBN (Figure 3.1). As shown in the middle panel of Figure 3.3, it’s possible that the “bottom-up alone” priority map could confuse distractor and target features, generating strong predictions for both locations. Under the assumption of a single “spotlight of attention”, the bottom-up conditional probabilities would then often lead to spatial attention being directed towards the distractor instead of the target. Meanwhile, if top-down priors disambiguate among target and distractor features, the resulting priority map would more clearly favor directing spatial attention to the target location (right panel Figure 3.3).

Similar to our evaluation of location accuracy, we compared target classification accuracy to chance-level performance of a random classifier. For MNIST digits, a random classifier would be expected to have an accuracy level of 0.1 (i.e., the probability of selecting one out of ten digit classes). However, it is more useful to compare the effects of spatial attention on classification accuracy relative to a "no-attention" condition. For this condition, we obtained target classification predictions for each trial without updating the RF pooling array. As shown in Figure 3.5, performance was well above chance for all three conditions. Furthermore, although both attention conditions performed better than the "no-attention" condition (0.67 and 0.63 versus 0.60; bootstrapped p-value [1000 samples] = 0.), the priority map with top-down priors better localized the target in order to enhance the encoded feature representations (0.67 versus 0.63; bootstrapped p-value [1000 samples] = 0.). It is worth noting that since performance of the predictor network on the test set was relatively low (0.81) at full spatial resolution (i.e., no RF pooling array) and without distractors, spatial attention using the "top-down feedback" priority map achieved 82% of the maximal classification accuracy.

3.5 Discussion

In the current study, we proposed a simple Bayesian model of covert visual attention and evaluated the model using a visual search task with handwritten digits among non-digit distractors. In contrast to previous models, our model learned priors over target features using an update rule that is similar to Hebbian learning and enhanced spatially-attended features using a neurobiologically plausible mechanism. Comparing the average precision for predicting target locations between priority maps with or without top-down priors, we observed that Bayesian priors over target features significantly improved target location accuracy. Furthermore, by modeling spatial attention as an interaction of an RF pooling array with an attention field at the predicted location (Theiss, Bowen, and Silver, 2022), we demonstrated that top-down priors help to disambiguate among distractor and target features such that the target is more likely to be attended. The study provides further support for the Bayesian brain hypothesis in the specific case of covert visual search.

Several Bayesian models of attention have been proposed over the past two decades to explain overt attention during free viewing and visual search with natural images (e.g., Chikkerur et al., 2010; Itti and Baldi, 2009; Torralba, Oliva, et al., 2006), while others have addressed covert attention with artificial stimuli (see Vincent (2015) for review). The common assumption in studying overt attention is that eye movements reflect covert attention through common networks, as posited by the premotor theory of attention (Rizzolatti et al., 1987). However, covert and overt attention can be untethered during visual search, which appears to be a task-dependent strategy (MacInnes et al., 2020). It is therefore important to develop models of visual search that can account for both overt and covert attention. By using a foveated RF pooling array, our model has the potential to account for covert attention by spatial resampling of RFs (Theiss, Bowen, and Silver, 2022) and overt attention by translating the pooling array to direct the "fovea" to a different part of the image (Cheung, E. Weiss, and B. Olshausen, 2016; Larochelle and Hinton, 2010). For example, rather than maintaining central fixation as done in the current study, our model can be used to systematically evaluate the contributions of covert and overt attention to task-dependent performance differences in a psychophysical experiment with humans.

One component of many visual search models that was not addressed in the current study is bottom-up saliency. For most computational models of visual attention, saliency is defined as a contrast of local features (Itti, Koch, and Niebur, 1998) or in the Bayesian framework as a measure of the difference between prior and posterior distributions (Itti and Baldi, 2009). Torralba, Oliva, et al. (2006) specifically fit hyperparameters to appropriately combine bottom-up and top-down attention to optimize eye movement predictions. However, for the current study it is unclear how bottom-up attention should be weighted relative to top-down priors. Indeed, it is generally an open question of how bottom-up and top-down attention are weighted across tasks. For example, Chikkerur et al. (2010) assumed uniform priors to model eye movements during free viewing. However, since it is unlikely that all priors—such as a light-from-above prior (Stone, Kerrigan, and Porrill, 2009)—would be uniform during free viewing, the combination of bottom-up and top-down attention is likely dynamic and task-dependent.

Recent developments in both machine learning and fMRI research have provided insight into the relationship between features learned in convolutional neural networks and patterns of activity in visual cortex (e.g., Devereux, A. Clarke, and Tyler, 2018; O'Connell and Chun, 2018; St-Yves and Naselaris, 2018). Combined with the known population-level effects of attention (Klein, Harvey, and Dumoulin, 2014; Womelsdorf et al., 2006) as well as the distributed nature of attention across the visual cortex (Melloni et al., 2012; Serences and Yantis, 2007; Sprague and Serences, 2013), we contend that the model we have proposed is well-suited to further evaluate the Bayesian brain hypothesis with neuroimaging data and natural images. Extended to multiple levels of feature complexity, our model of Bayesian priors is dynamic and local. This allows us to study the dynamics of top-down attention by updating priors across trials and at multiple levels of the visual hierarchy. Under free-viewing conditions with natural images (i.e., without an explicit task), we predict that the influence of spatial and feature priors on gaze behavior should be reflected at the respective level of feature encoding that best accounts for the statistical regularities across similar scenes (S. C.-H. Yang, Lengyel, and Wolpert, 2016). However, during visual search we predict that task-based attention acts as a hyperprior, giving stronger weight to task-relevant priors across the visual hierarchy.

Although the model described in the current study was relatively simple and evaluated with artificial stimuli, it can easily be extended to more complex features and visual tasks. The main challenge when using a CDBN is the amount of training time required to model natural images with many RBM layers, since each layer is trained in sequence. However, if we assume that features learned in deep neural networks can approximate those represented in visual cortex (Yamins et al., 2014), we can instead train a single RBM at multiple layers of a pre-trained neural network in parallel, considering each RBM to learn priors over local features and spatial locations. Not only does this reduce the training time, but it also allows for studying relative effects of priors at different levels of feature complexity and spatial scale.

Chapter 4

Bayesian surprise from category-specific priors accounts for gaze behavior during free viewing

4.1 Abstract

The statistical regularities that exist across natural images within the same category (e.g., city streets) have previously been modeled as Bayesian priors over feature and spatial configurations to predict eye movements during visual search. However, whereas priors may be useful for guiding attention to relevant features or locations in visual search, during free viewing (i.e., without an explicit task), humans tend to fixate salient features that are surprising (in an information-theoretic sense) relative to priors. In the current study, we characterize the relative contributions of bottom-up surprise and top-down priors at various levels of feature complexity and spatial scale to predictions of free-viewing gaze behavior across different scene categories. By comparing category-specific and category-agnostic priors, we found that Bayesian surprise explained more variance in gaze behavior than within-category priors over high-level features for most categories. Our findings suggest that humans utilize category-specific priors when viewing a scene, even in the absence of a category-specific task.

4.2 Introduction

When viewing a scene, we typically have expectations about the distribution of features and objects based on previous experience with similar scenes (Torralba and Oliva, 2003). For example, we expect to find cars outside on roads and paintings inside on walls. These associations influence our perception at various scales and complexities, providing contextual information about where objects are typically found (Torralba, Oliva, et al., 2006). By formulating these associations within a Bayesian inference framework, we can consider selective visual attention as a top-down prior that reduces uncertainty in perception (Dayan and R. S. Zemel, 1999; Pelli, 1985). Within this framework, perception is viewed as a process of probabilistic inference, where the brain constructs a generative model of its visual inputs in order to infer the current state of the world from noisy sensory information (Dayan, Hinton, et al., 1995; Friston, 2009; T. S. Lee and Mumford, 2003).

Over the past two decades, several Bayesian models have been introduced to account for attentional effects observed in neurophysiological and behavioral experiments (Borji, D. N. Sihite, and Itti, 2013; Borji, D. Sihite, and Itti, 2012; Chikkerur et al., 2010; Dayan and R. S. Zemel, 1999; Feldman and Friston, 2010; Itti and Baldi, 2009; T. S. Lee and Mumford, 2003; Rao, 2005; Torralba, Oliva, et al., 2006; Vossel et al., 2015; Whiteley and Sahani, 2012; Yu and Dayan, 2004). Although there are some variations across implementations of these models, most represent visual processing with a hierarchical generative model, whereby top-down priors help to disambiguate noisy sensory information. For modeling gaze behavior during visual search, a single priority map is typically generated that represents the posterior probability, given priors over target-specific features and locations (e.g., Chikkerur et al., 2010; Torralba, Oliva, et al., 2006). For free-viewing experiments where a task is not specified, it is typically assumed that bottom-up attention and saliency of image features drive gaze behavior. For example, Chikkerur et al. (2010) used uniform priors in order to account for fixations in natural scenes. However, Itti and Baldi (2009) provided an alternative account of bottom-up attention within the perceptual inference framework by introducing a metric termed Bayesian surprise, which measures the Kullback-Leibler (KL) divergence between posterior and prior probability distributions. Bayesian surprise reflects the relative change in beliefs incurred by observing new data, with greater surprise indicating that new information differs from that expected by the prior.

Much of the previous work related to gaze behavior within the Bayesian attention framework has considered spatial prior and feature priors over target-relevant features using scene-level gist descriptors (e.g., Torralba and Oliva, 2003). However, it is less clear from the literature how priors over different levels of feature complexity or spatial scales influence gaze behavior. Although early models of gaze behavior used hand-tuned filters to simulate low-level features represented in visual cortex (e.g., Itti, Koch, and Niebur, 1998), more recent state-of-the-art approaches have used deep neural networks (DNNs) to incorporate more complex features and to learn associations between features and human gaze data (Kümmerer, Wallis, and Bethge, 2016). Hayes and Henderson (2021) found that predictions of these "deep saliency models" have strong associations with both low-level saliency (Itti, Koch, and Niebur, 1998) and high-level meaning (Henderson and Hayes, 2017), suggesting that these features are important in gaze behavior, relative to mid-level features. However, these authors noted that the diversity of scenes and mid-level features explored in their work was limited and that the relative weight of low-level saliency and high-level meaning in predicting gaze behavior likely depends on the task being performed by the observer.

In the current study, we characterize how priors at different levels of feature complexity and spatial scale account for human gaze behavior across different scene categories. Specifically, we study free-viewing behavior, in which specific visual targets are not defined. The expectations that humans have about likely features and locations for different categories of scenes (Torralba and Oliva, 2003) likely influence gaze behavior even in task-free settings. Therefore, we evaluated the influence of priors using a publicly available dataset containing fixation data across 20 different scene categories (CAT2000; Borji and Itti, 2015). First, we extracted features from images at multiple levels of complexity and spatial scale using a pre-trained DNN (VGG16 (Simonyan and Zisserman, 2014)). We then learned spatial and feature priors with normalizing flows (Dinh, Krueger, and Y. Bengio, 2014; Dinh, Sohl-Dickstein, and S. Bengio, 2016; Durk P Kingma and Dhariwal, 2018) that were trained separately for each feature level and image category. We tested the effect of using category-specific priors on predicting gaze behavior, relative to a baseline model with category-agnostic features. Finally, we evaluated the unique contributions of bottom-up Bayesian surprise and top-down priors at varying levels of feature complexity and spatial scale to gaze behavior predictions.

4.3 Methods

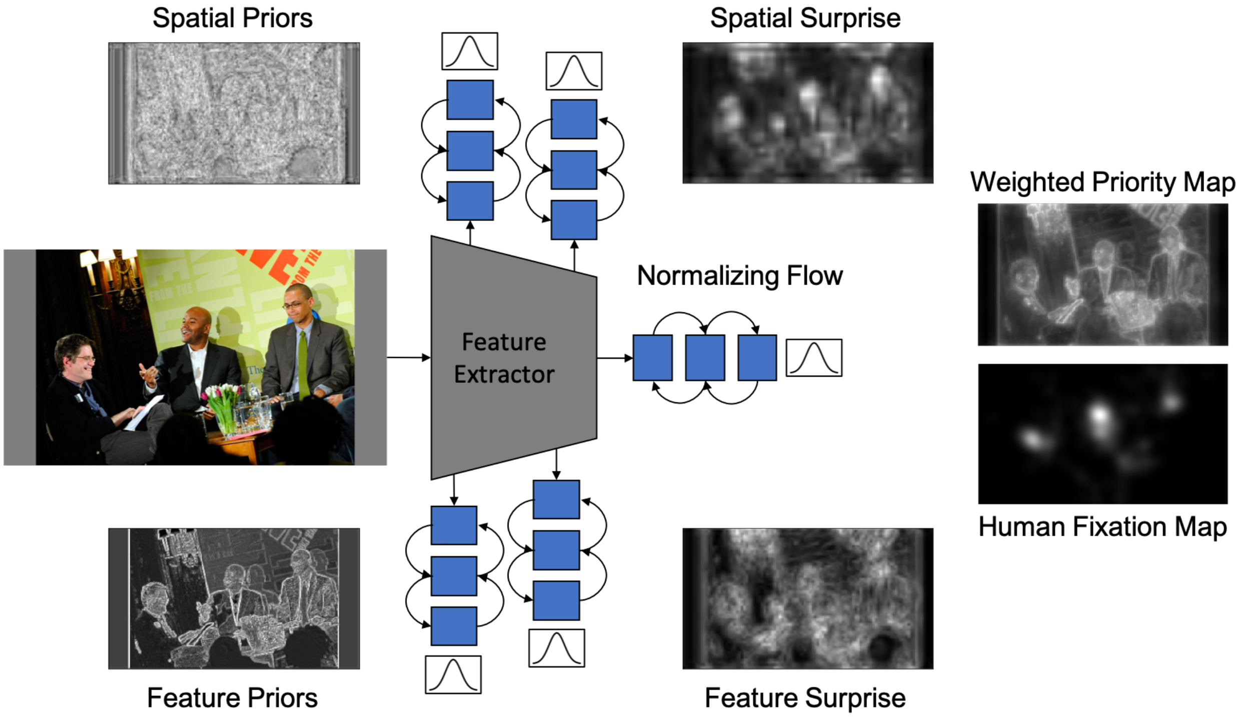

Figure 4.1: Method for generating priority maps from a given image using category-specific priors. We extracted features from each image at multiple layers of a pre-trained deep neural network (Feature Extractor). Prior to evaluation, we trained normalizing flow models at each of five layers to represent category-specific feature and spatial priors. For each layer, we then generated four different priority maps, corresponding to top-down feature and spatial priors as well as bottom-up feature and spatial surprise (here depicted at different levels of feature extraction). Finally, we linearly combined all priority maps to obtain a single prediction for fixations for the given image (Weighted Priority Map).

In order to evaluate the effects of feature and spatial priors on gaze behavior, we used normalizing flow models to learn category-specific priors over features extracted from natural images. Our goal was to characterize the relative contributions of bottom-up surprise and top-down priors as well as feature complexity and spatial scale to predicting gaze behavior during free viewing. Figure 4.1 shows an example of our method, which we describe in more detail below.

We evaluated performance on the publicly-available CAT2000 dataset (Borji and Itti, 2015), which contains 20 categories of images with 200 images per category (equally split between training and test sets). In this dataset, each image was viewed by 24 different observers for 5 seconds without specific instruction (i.e., free viewing). Since images were presented in a random order across categories, we assume that feature and spatial priors for specific categories are from previous experience (i.e., not accumulated during the experiment). For each category, we derived feature and spatial priors from a training set of 100 images without fixation data and evaluated performance on a held-out test set of 100 images with fixations.

Feature and Spatial Priors

For each category of the CAT2000 dataset, we obtained several feature and spatial priors over extracted features of VGG16 (Simonyan and Zisserman, 2014), pre-trained on the ImageNet dataset (Deng et al., 2009) (Figure 4.1). Specifically, we trained normalizing flow models (Dinh, Krueger, and Y. Bengio, 2014; Dinh, Sohl-Dickstein, and S. Bengio, 2016; Durk P Kingma and Dhariwal, 2018) to learn separate priors for features extracted from five different layers of VGG. Since we are interested in learning a given prior over features that are relevant to a given layer in VGG, we sampled feature vectors proportional to their magnitude. This ensures that vectors with a small magnitude (e.g., low-contrast background pixels), which are less representative of the features encoded at the given layer, will not be strongly represented in the prior distribution. Furthermore, we used locality sensitive hashing (LSH; Charikar, 2002) to randomly project the features to the vector space $\mathcal{V} \in [-1, 1]^n$ (where n is equal to the number of feature maps), which reduces the complexity of the normalizing flow transformation by making the distribution of feature vectors more similar to its output distribution (i.e., a multivariate Gaussian).

For spatial priors, we followed the same procedure as above except that we sampled random 11×11 overlapping patches of feature vectors. These patch sizes account for progressively larger receptive fields from early to late VGG layers (ranging from approximately 0.1% image size in the first layer to 36% in the final layer). We then used LSH to project the spatial configurations within each patch to lower dimensionality (n, as above), resulting in a single vector that represents the spatial relationships within each patch. By learning feature and spatial priors at five layers across VGG (each ReLU layer prior to Max-Pooling, with indices $\{3, 8, 15, 22, 29\}$), we were able to evaluate the effects of feature complexity and spatial scale, respectively, on gaze behavior.

Normalizing Flows

As shown in Figure 4.1, we used normalizing flow models at each of our five extracted-feature layers to represent feature and spatial priors for a given category. A normalizing flow model

(Dinh, Krueger, and Y. Bengio, 2014; Dinh, Sohl-Dickstein, and S. Bengio, 2016; Durk P Kingma and Dhariwal, 2018) is a set of invertible transformations that can be applied to a relatively simple distribution (e.g., $\mathcal{N}(0; \mathbf{I})$) to obtain a more complex distribution (i.e., the data distribution). Simple transformations are chosen to ensure that the log-likelihood is tractable and efficient to compute. Specifically, normalizing flows use the change-of-variable theorem to define a probability density function over the observed data distribution $\mathcal{X}$ using an invertible function f_ϕ and base distribution $p_\mathcal{Z}(z)$:

$$p_\mathcal{X}(x) = p_\mathcal{Z}\left(f_\phi^{-1}(x)\right)\left|\det\frac{\partial f_\phi^{-1}(x)}{\partial x}\right|, \tag{4.1}$$

where $p_\mathcal{Z}(z) = \mathcal{N}(0; \mathbf{I})$ and $f_\phi : \mathcal{Z} \to \mathcal{X}$ is the set of transformations used in the Glow architecture (Durk P Kingma and Dhariwal, 2018). Each transformation in Glow contains an "activation normalization" (actnorm) layer, 1×1 invertible convolution layer, and an affine coupling layer. The actnorm layer normalizes inputs by subtracting a location parameter and multiplying a scaling parameter. The 1×1 convolution permutes the order of the data to ensure that each channel of the input is transformed across the normalizing flow. Finally, the affine coupling layer splits the input into two parts across the channel dimension and applies an affine transformation to one part conditioned on the other:

$$\begin{aligned} s, t &= \mathit{NN}(x_{1:d}) \\ y_{d:D} &= x_{d:D} \odot \exp(s) + t \\ y_{1:d} &= x_{1:d}, \end{aligned} \tag{4.2}$$

where D (resp. d) is the dimensionality (resp. half the dimensionality) of the input vector x and $\mathit{NN}(\cdot)$ is a neural network that generates the log scale and translation parameters s and t, respectively. As seen in the equation above, half of the input is unchanged during a given transformation, which highlights the importance of the permutation step to ensure that all channels are transformed across the normalizing flow. The normalizing flow composed of K transformations is then trained to maximize the log-likelihood of the data:

$$\log p_\mathcal{X}(x) = \log p_\mathcal{Z}(z_0) - \sum_{i=1}^{K} \log\left|\det\frac{\partial f_i(z_i)}{\partial z_{i-1}}\right|, \tag{4.3}$$

where $z_K = x$ in the forward direction $\mathcal{Z} \to \mathcal{X}$. For all experiments, we use $K = 8$ transformations with the affine coupling layer's $\mathit{NN}(\cdot)$ parameterized as a three-layer convolutional network with 512 intermediate channels (input and output dimensionality is constant throughout the normalizing flow). We trained each normalizing flow model for 150 epochs using the Adam optimizer (Diederik P Kingma and Ba, 2014) with a batch size of 16 and learning rates of 5×10^{-5} for feature priors and 5×10^{-6} for spatial priors. Learning rates were chosen to prevent overfitting by monitoring the negative log-likelihood computed for the held-out test set. During training, images were resized from 1920×1080 to a height of 224 pixels while maintaining the original aspect ratio. As described below, priority maps were upsampled to the original image size during evaluation.

Bayesian Surprise

Using a normalizing flow with the multivariate Gaussian base distribution $p_{\mathcal{Z}}(z) = \mathcal{N}(0; \mathbf{I})$, we can further consider this method as transforming the extracted features into a set of Gaussian variables with zero mean and unit variance. This simplifies the calculation of Bayesian surprise compared to other normalizing flow methods that allow for greater complexity in the base distribution. Following Itti and Baldi (2009), we computed Bayesian surprise as the KL divergence between the prior and posterior Gaussian distributions:

$$S(D, \mathcal{M}) = \frac{1}{N} \sum_{i=1}^{N} \frac{1}{2\sigma_i^2} \left[\sigma_i^2 + (\mu_i - \overline{m}_i)^2 \right], \tag{4.4}$$

where $S(D, \mathcal{M})$ is the surprise between the prior distribution for a family of models $\mathcal{M}$ and posterior distribution given data D, each of the N models in $\mathcal{M}$ is a Gaussian variable parameterized by (μ_i, σ_i^2), and $\overline{m}_i$ is the observed mean given the data. As described above, each Gaussian variable in the base distribution of the normalizing flow used in the current study was designed to have $\mu_i = 0$ and $\sigma_i^2 = 1$. We computed Bayesian surprise per location for the features extracted from each VGG layer, thereby generating a set of bottom-up priority maps for feature and spatial priors separately.

Priority Maps

Separate bottom-up and top-down priority maps were generated for the learned feature and spatial priors per layer (20 priority maps in total). For each test image, we generated the priority map for a given VGG layer by extracting features at that layer, transforming the extracted feature vectors into the $\mathcal{Z}$ space of the normalizing flow, and computing either the average probability or Bayesian surprise across the vector, considering each Gaussian variable to be a separate "model", as in Itti and Baldi (2009).

Top-down priority maps highlight regions of the image that are more probable under the learned priors for the given scene category. Bottom-up priority maps, on the other hand, highlight regions that are surprising relative to the priors learned for that category. Since we are only interested in features that are representative of a given VGG layer, we scaled the bottom-up priority maps proportional to the magnitude of each feature vector. This scaling reduced surprise due to non-feature vectors that were not used while training the normalizing flows (e.g., pixels corresponding to a uniform background). Priority maps were then resized to match the original image size and normalized by the sum across the priority map to obtain a valid probability density. Finally, we linearly combined each of the 20 priority maps [2 (bottom-up vs. top-down) $\times$ 2 (feature vs. spatial priors) $\times$ 5 (layers)] by summing across layers for bottom-up ($M_{BottomUp}$) and top-down ($M_{TopDown}$) priority maps separately and taking a weighted combination:

$$M_{Total} = \alpha M_{BottomUp} + (1 - \alpha) M_{TopDown}, \tag{4.5}$$

where α is a weight in $[0, 1]$, which we varied with a step size of 0.1 as done in Mahdi, Qin, and Crosby (2019) to evaluate the influence of bottom-up vs. top-down attention.

Metrics and Baselines

Normalized Scanpath Saliency (NSS) (Peters et al., 2005) is a measure of the correspondence between priority maps and human fixations and is defined as the average of normalized values in the priority map at each fixated location. Since the priority map is normalized to have zero mean and unit variance, NSS is measured in units of standard deviation, where zero indicates chance-level prediction of fixations. Values above zero indicate above-chance correspondence with gaze behavior, whereas values below zero indicate anti-correspondence.

Information Gain (IG) (Kümmerer, Wallis, and Bethge, 2015) is a measure of the predictive performance of a model relative to a baseline model. It is the difference in log-likelihoods averaged across fixations with units of $\frac{\text{bits}}{\text{fixation}}$. Additionally, Information Gain Explained (IGE) is the proportion of potential Information Gain that can be explained relative to a "gold standard" model:

$$\begin{aligned} IG(\hat{p}\|p_{baseline}) &= \frac{1}{N}\sum_i \log_2 \hat{p}(x_i, y_i|I_i) - \log_2 p_{baseline}(x_i, y_i) \\ IGE &= \frac{IG(\hat{p}\|p_{baseline})}{IG(p_{gold}\|p_{baseline})}, \end{aligned} \tag{4.6}$$

where the log-likelihood of each model is evaluated at (x, y) locations for N observer fixations indexed by i. Typically, the gold standard model is defined as the prediction using other observers' fixations, which therefore accounts for the relative correspondence of fixations among observers (Kümmerer, Wallis, and Bethge, 2015). Here, we use the normalized 2D histogram of fixations as the gold standard model.

We evaluated IG and IGE performance relative to two baseline models. First, we compared model priority maps to a random baseline model that assigns a probability sampled from a uniform distribution to each location across the image. We used this baseline model to compute IGE for each category, where a value of 1 indicates that the model's priority map explains all of the potential IG of the fixation distribution, and a value of 0 indicates equivalence to the random baseline model. Next, we computed IG for our model using category-specific priors relative to a model that used a single prior across all categories. This baseline used the same method as described above, but with feature and spatial priors learned over all categories. To do this, we trained each normalizing flow model using a dataset of 100 images with 5 images randomly sampled from each category (matching the size of each category's training set). Importantly, this baseline allows us to quantify the degree to which category-specific priors account for gaze behavior, with positive IG values indicating that the model with category-specific priors assigned greater probability to fixated locations compared to the baseline model (and vice versa for negative values).

In order to measure the unique contribution of a particular priority map to overall model performance, we computed semi-partial linear correlations as done in Henderson and Hayes (2017).

This approach evaluates the unique variance explained by a particular priority map while controlling for the shared variance explained by all other priority maps in the full model prediction (M_{Total}). Specifically, we used this approach to determine the unique contributions of bottom-up and top-down attention as well as layer-specific contributions to gaze behavior predictions.

Statistical Procedures

For hypothesis testing with a single distribution, we used 1000 iterations of bootstrap resampling with replacement from the null distribution by centering the sample distribution at the mean for the null hypothesis (e.g., 0). For statistical comparisons between two distributions, we first centered each distribution's mean at the combined mean of the two distributions and then bootstrap resampled (again with 1000 iterations) from the centered distributions. We report p-values as the proportion of test statistics observed through bootstrap resampling from the null distribution that were greater than the sample statistic (Efron and Tibshirani, 1994).

Code/Software

We implemented all training and computation in PyTorch (Paszke et al., 2017) as well as custom Python code. The code used to produce the results described in this paper is available upon request.

4.4 Results

Category-specific priors improved gaze predictions relative to category-agnostic priors. We first evaluated performance across categories for the full model by combining bottom-up Bayesian surprise and top-down priors for each layer while varying the relative weights of the bottom-up and top-down contributions (α value in Equation 4.5). As shown in the left panel of Figure 4.2, we observed an average NSS value of 0.93 at $\alpha = 0.5$, which indicates that the average priority across fixation locations was about a standard deviation above chance level. For most categories, we observed similar NSS values for either the equally-weighted model (hatched bars, $\alpha = 0.5$) or bottom-up surprise alone (red bars, $\alpha = 1$), but we observed lower values when using only top-down priors (blue bars, $\alpha = 0$). We then evaluated the Information Gain Explained using the random baseline model described above ($p_{baseline}$ in Equation 4.6). As shown in the middle panel of Figure 4.2, across all categories we found that the equally-weighted model (i.e., $\alpha = 0.5$) accounted for roughly 29% of the total explainable Information Gain. Similar to NSS, we observed lower performance on average when using only top-down priors (blue bars, $\alpha = 0$) and higher performance when using only bottom-up surprise (red bars, $\alpha = 1$).

In addition to quantifying model performance across categories, we next evaluated the degree to which category-specific priors improved predictions relative to a model with a single prior across all categories (henceforth referred to as category-agnostic). If category-specific priors better account for gaze behavior, we would observe positive Information Gain relative to the category-agnostic model. We examined this using bootstrap hypothesis testing against a null hypothesis

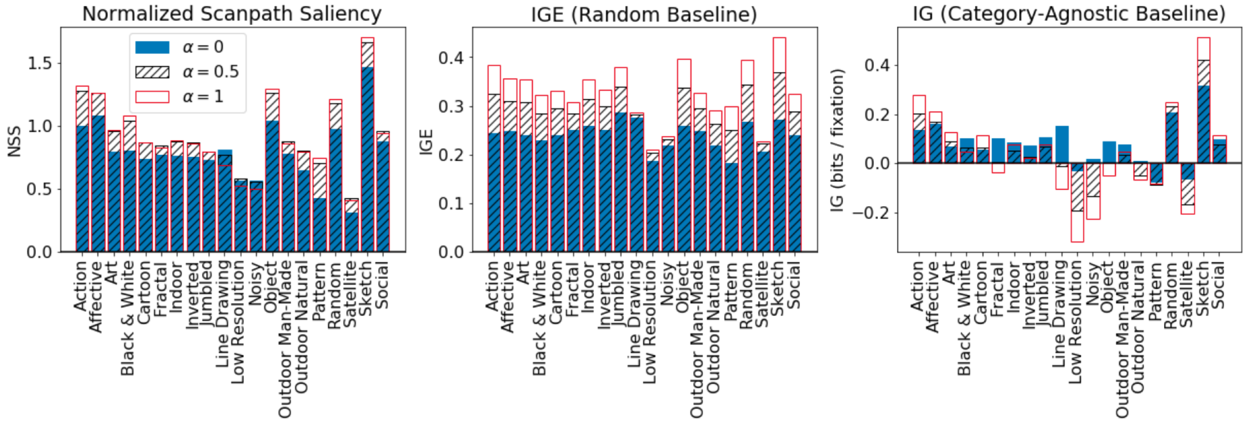

Figure 4.2: Evaluation across categories for full model performance as measured by (*left*) Normalized Scanpath Saliency, (*middle*) Information Gain Explained relative to the random baseline model, and (*right*) Information Gain relative to the category-agnostic baseline model. We report performance in each panel for three linear combinations of bottom-up and top-down priority maps by varying the α values in Equation 4.5. Blue bars correspond to top-down priority maps alone ($\alpha = 0$), hatched bars correspond to equally-weighted priority maps ($\alpha = 0.5$), and red bars correspond to bottom-up priority maps alone ($\alpha = 1$).

where the IG relative to the category-agnostic model across categories is 0, which would indicate similar performance for each model. Indeed, the right panel of Figure 4.2 shows that predictions for the majority of categories benefited from using category-specific priors (at $\alpha = 0.5$, bootstrapped p-value [1000 samples] = 0.). As might be expected, the Low Resolution and Noisy categories had among the worst performance relative to the category-agnostic model. The other categories with negative values included Outdoor Natural, Pattern, and Satellite images. Each of these categories may have less consistent feature representations compared to the other categories, given that the feature extractor is trained for object classification. In these categories, the prior distributions over features across all categories may have provided some benefit for the category-agnostic model when well-represented features were present in the images.

Next, we wanted to understand the relative importance of bottom-up and top-down components of our model to NSS performance. To do this, we averaged the optimal α that maximized NSS values across categories, obtaining a value of 0.68, which indicates that on average bottom-up surprise contributed to greater NSS values (average value of 0.94 when using optimal α) compared to top-down priors. It is worth noting that Equation 4.5 uses equal weighting for feature and spatial priors as well as across layers, and it does not consider more complex weighting schemes that may better account for overall gaze behavior.

Bottom-up surprise accounted for greater unique variance in gaze behavior compared to top-down priors. In order to further characterize the influence of bottom-up surprise and top-down feature and spatial priors during free viewing, we computed the unique variance explained by mea-

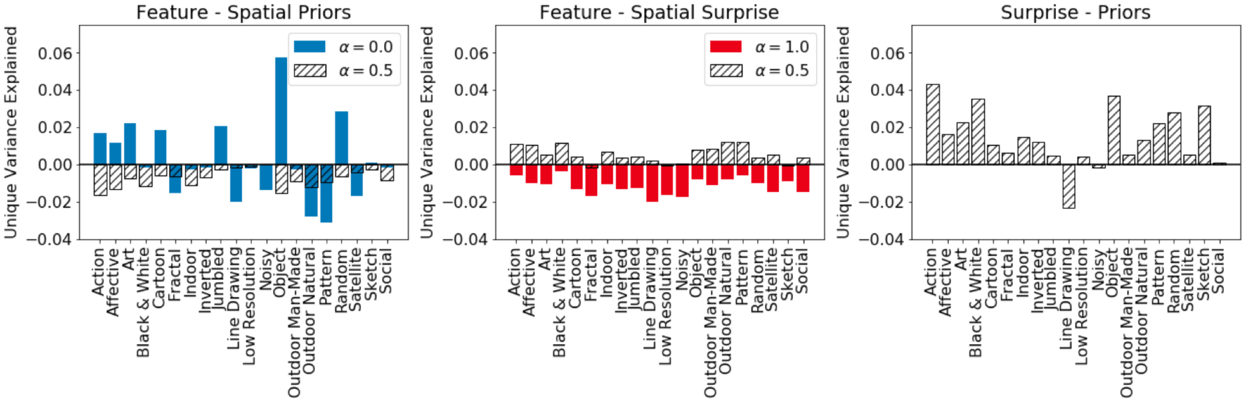

Figure 4.3: Unique variance explained by feature priors relative to spatial priors (*left*), feature surprise relative to spatial surprise (*middle*), and bottom-up surprise relative to top-down priors (*right*) across categories. Unique variance explained is reported relative to the equally-weighted model (hatched bars, $\alpha = 0.5$) as well as relative to bottom-up or top-down priority maps separately (color bars, $\alpha \in \{0, 1\}$).

suring the change incurred by removing each model component in turn from the equally-weighted priority map (i.e., $\alpha = 0.5$). If bottom-up and top-down priority maps contribute similar predictive performance, the unique variance explained by each should be close to zero as the removal of one component would be compensated by the others.

Figure 4.3 shows the unique variance explained for each of top-down feature and spatial priors and bottom-up feature and spatial surprise. As before, we used bootstrap resampling methods to test for differences between unique variance explained by bottom-up surprise and top-down priors against a null hypothesis of no difference. When comparing unique variance explained relative to the equally-weighted priority map, bottom-up feature and spatial surprise had greater unique variance explained compared to top-down priors across most categories (right panel in Figure 4.3; bootstrapped p-value [1000 samples] = 0.). As will be seen in later results, this does not mean that top-down priors provide poor predictions, but instead that bottom-up surprise predicted aspects of gaze behavior that were not otherwise accounted for in the equally-weighted priority map. Furthermore, when comparing feature and spatial surprise (hatched bars in middle panel of Figure 4.3), we observed that feature surprise accounted for more unique variance in gaze behavior for most categories relative to spatial surprise. Overall, these results provide further context to our understanding of the contributions of bottom-up surprise and top-down priors to predicting gaze behavior: bottom-up feature surprise, more so than spatial surprise or top-down priors, accounted for the most unique variance explained across many categories.

In order to evaluate the relative contributions of feature and spatial priors to either bottom-up or top-down priority maps, we additionally computed unique variance explained within bottom-up and top-down priority maps separately. This was done by setting α to either 0 or 1 (Equation 4.5), reflecting priority maps comprising only top-down priors or only Bayesian surprise, respectively.

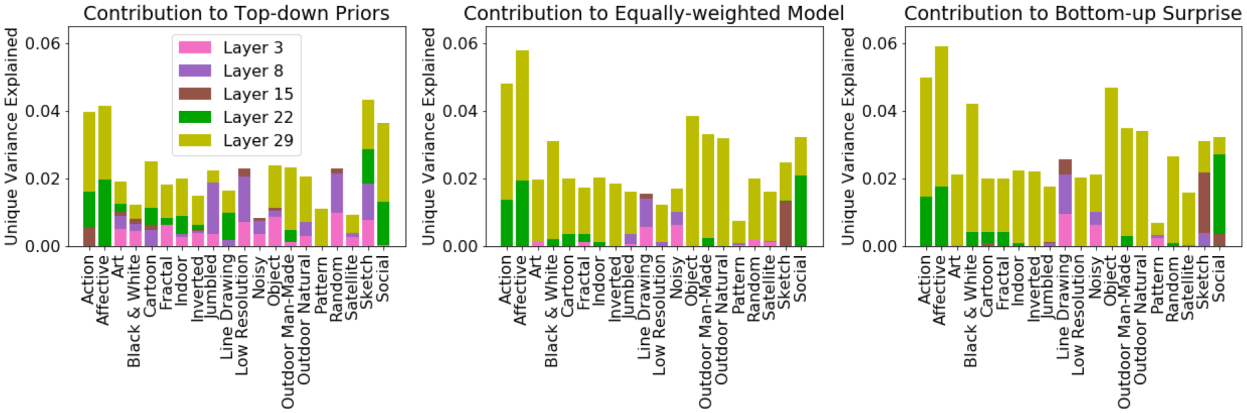

Figure 4.4: Unique variance explained by layer across categories for top-down priors ($\alpha = 0$, *left*), equally-weighted priority maps ($\alpha = 0.5$, *middle*), and bottom-up surprise ($\alpha = 1$, *right*).

As shown in Figure 4.3 (color bars in left and middle panels, $\alpha \in \{0, 1\}$), we observed different relationships between feature and spatial surprise and feature and spatial priors when computing unique variance explained within bottom-up or top-down priority maps separately. When greater weight was given to top-down priors (blue bars in left panel of Figure 4.3), we observed category-specific differences between feature priors and spatial priors, with feature priors explaining more unique variance for some categories (Object, Random, Art, Jumbled, Cartoon, Action, and Affective), and spatial priors explaining more unique variance for other categories (Pattern, Outdoor Natural, Line Drawing, Fractal).

Meanwhile, bottom-up spatial surprise explained greater unique variance in gaze behavior compared to feature surprise when only considering bottom-up priority maps (red bars in middle panel of Figure 4.3). These comparisons are markedly different from the unique variance explained for each component in the equally-weighted model (i.e., $\alpha = 0.5$), where top-down spatial priors and bottom-up feature surprise explained more unique variance in gaze behavior relative to feature priors and spatial surprise, respectively (hatched bars in left and middle panels of Figure 4.3). This suggests that there are category-specific differences in the effects of feature and spatial priors on gaze behavior that are dependent on the relative weighting of bottom-up and top-down attention.

High-level features explained greater unique variance in gaze behavior relative to lower-level features. In order to understand the roles of low-, mid-, and high-level features in our predictions of gaze behavior, we computed the unique variance explained for each model layer separately. Specifically, we computed unique contributions relative to the equally-weighted priority map (i.e., $\alpha = 0.5$) as well as for bottom-up and top-down attention separately. Again, if a priority map at a specific level of feature complexity and spatial scale is redundant with other priority maps in the overall model prediction, we expect that the unique variance explained by that priority map would be close to zero. As shown in the middle panel of Figure 4.4, we found that across most categories the final layer (layer 29) contributed the greatest amount of unique variance explained relative to the other four layers (bootstrapped p-value [1000 samples] = 0.). Meanwhile, the middle layers

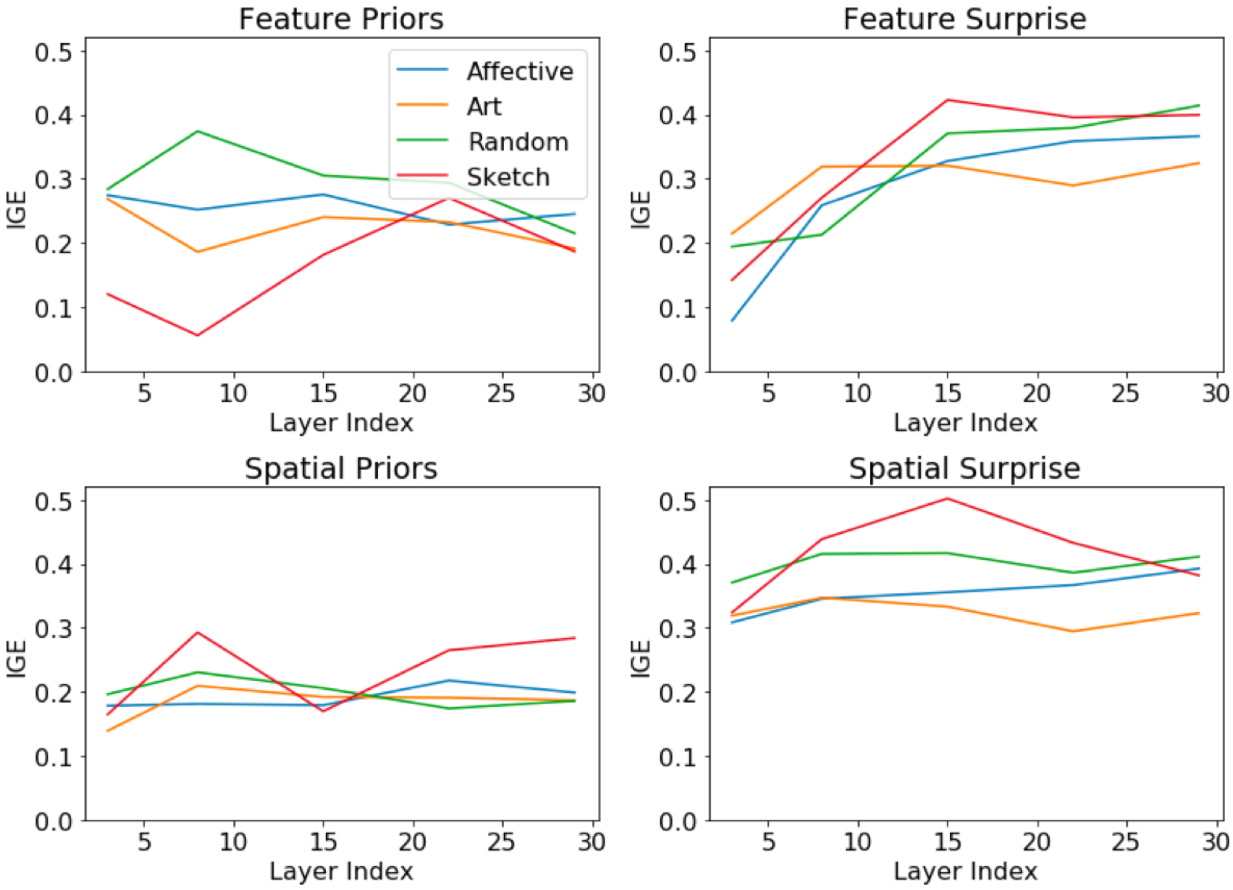

Figure 4.5: Information Gain Explained per layer for feature priors (*top left*), feature surprise (*top right*), spatial priors (*bottom left*), and spatial surprise (*bottom right*). IGE is reported for the selected categories relative to the random baseline model.

(layers 8 through 22) contributed most for only a few other categories (e.g., Line Drawing and Social). This suggests that high-level feature complexity and spatial scale best explained gaze behavior when equally-weighing bottom-up and top-down attention for most of the image categories in CAT2000. Interestingly, we observed greater differences between layer-specific contributions when considering top-down priors separately. Specifically, the contributions among different layers in top-down priority maps were more equally distributed (left panel of Figure 4.4). Note the overall similarity between the middle and right panels of Figure 4.4, which supports the previous finding that bottom-up surprise contributed more unique variance to the equally-weighted model prediction compared to top-down priors. These results suggest that surprise among larger spatial scales and more complex features contributed more than top-down priors to the overall model prediction across categories.

Characterizing differences in bottom-up surprise and top-down priors across layers. In

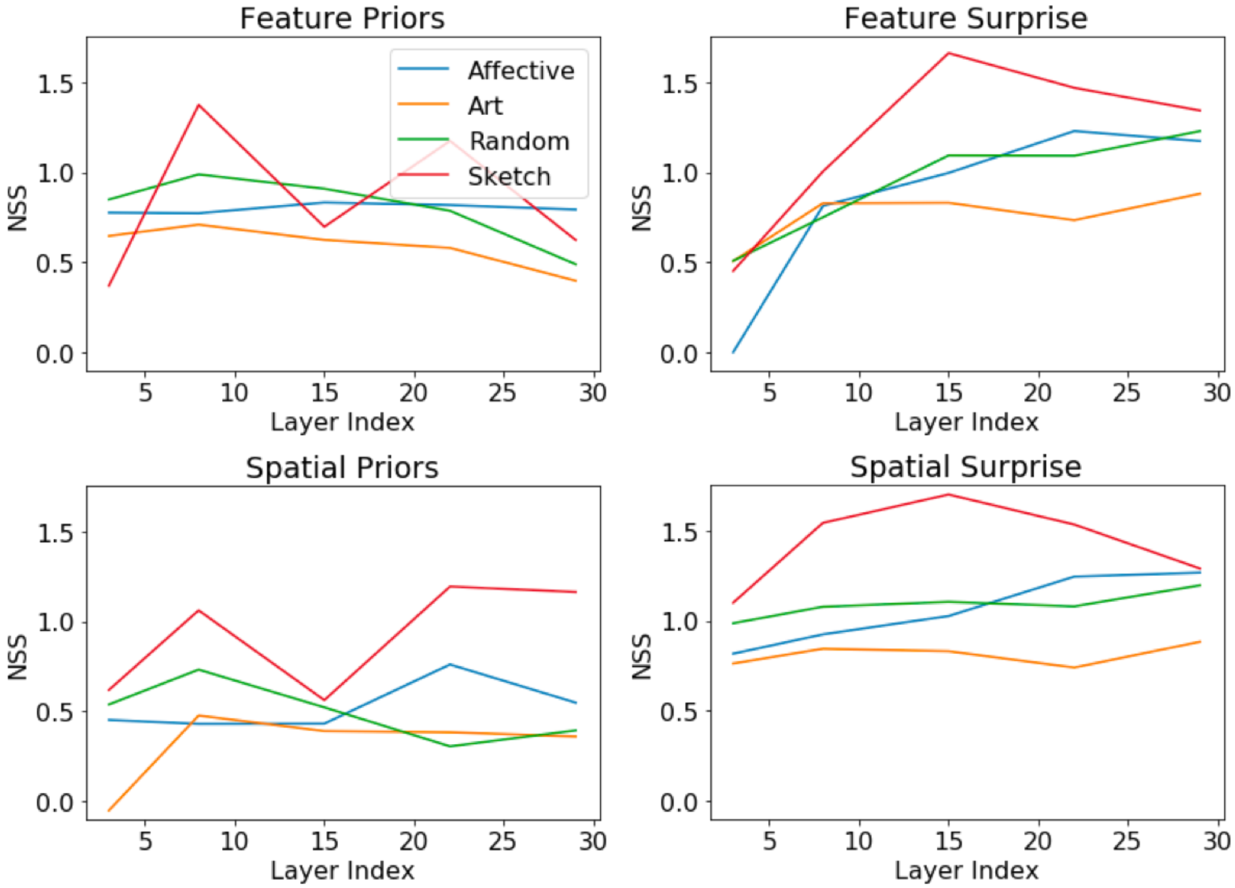

Figure 4.6: Normalized Scanpath Saliency per layer for feature priors (*top left*), feature surprise (*top right*), spatial priors (*bottom left*), and spatial surprise (*bottom right*).

order to further understand the predictive value of bottom-up surprise and top-down priors across layers, we computed IGE for each priority map separately as shown for select categories in Figure 4.5. Although we cannot entirely disentangle feature complexity from spatial scale in each layer, this evaluation allows us to better understand the associations among feature complexity, spatial scale, bottom-up surprise, and top-down priors. When considering top-down feature priors (top-left panel of Figure 4.5), we observed large differences across categories in early layers, suggesting that low-level features were more informative for some categories (e.g., Random) compared to others (e.g., Sketch). In contrast, feature surprise was much less informative in early layers compared to later layers across these categories (top-right panel in Figure 4.5). Compared to feature complexity, we observed relatively fewer differences in IGE across different spatial scales (i.e., bottom panels of Figure 4.5). However, spatial surprise within the middle layer (layer 15) accounted for roughly half the total explainable Information Gain for images in the Sketch category. This indicates that differences from the expected spatial configuration of features at this scale best predicted gaze behavior for the category. Note that these results demonstrate the predictive value of each priority map individually, which could provide insight regarding putative attentional mechanisms at a particular level of feature complexity and spatial scale. For instance, one could use this approach to test hypotheses about the influences of bottom-up or top-down attention at specific levels of feature processing, given a particular scene category.

As noted previously, whereas NSS indicates the priority assigned to each fixation relative to other locations in the image, IGE indicates the proportion of Information Gain that can be explained for a given distribution of fixations. This difference is demonstrated by the comparison between feature priors in the second layer of extracted features (layer 8) for the Sketch and Art categories. Although category-specific feature priors in this layer better match the distribution of fixations in Art compared to Sketch images, the average priority for fixations in the Sketch category is nearly double that in the Art category (compare red and yellow lines in the top-left panels of Figure 4.5 and Figure 4.6). In most cases, however, NSS and IGE were well-correlated. Overall, these results suggest that the degree to which bottom-up surprise and top-down priors varied with feature complexity and spatial scale was dependent on the scene category.

4.5 Discussion

In the current study, we used normalizing flows to learn category-specific feature and spatial priors at multiple layers of a pre-trained DNN in order to characterize contributions of bottom-up vs. top-down attention to gaze behavior during free viewing. We compared our model with category-specific priors to one using a single prior across all categories, finding that for most scene types, the category-specific model improved performance. Furthermore, by evaluating the unique variance explained by each component of our model, we found that bottom-up Bayesian surprise—particularly, feature surprise—accounted for more unique variance compared to top-down priors. We also quantified the unique variance explained by feature complexity and spatial scale, revealing that most categories had unique contributions from the final layer of extracted features, which contributed more to bottom-up surprise priority maps compared to top-down priority maps. Finally, we characterized the individual predictive value across feature complexity and spatial scale, bottom-up surprise, and top-down priors for different scene types. Taken together, these results suggest that Bayesian surprise from category-specific priors influences gaze behavior with the greatest contributions from features with larger spatial scales and greater complexity.

Many models of visual attention contain a combination of bottom-up saliency and top-down task-relevant control (e.g., Wolfe, Cave, and Franzel, 1989). In the current study, we found that bottom-up surprise accounted for more unique variance in gaze behavior compared to top-down priors. However, the conceptual difference between bottom-up and top-down attention is somewhat ambiguous within our model, since feature and spatial priors influence both top-down and bottom-up priority maps, with Bayesian surprise computed relative to these priors. Indeed, accumulating evidence indicates that top-down attention influences gaze behavior even in the absence of a specific task (A. Li, Wolfe, and Z. Chen, 2020; Wolfe and Horowitz, 2017). In the past decade, several studies have described methods for incorporating top-down attention for modeling gaze behavior during free viewing (Betz et al., 2010; Borji, 2012; Chakraborty, Samaras, and Zelinsky, 2022; Großekathöfer, Suchotzki, and Gamer, 2020; Hua et al., 2013; Murabito et al., 2018; J. Yang and M.-H. Yang, 2016). Our approach is perhaps most similar to the SUN model (Zhang et al., 2008), which used natural image statistics to define bottom-up saliency as the self-information of low-level filter responses. Roy, S. Ghosh, and A. Ghosh (2018) used a similar approach by training a restricted Boltzmann machine (RBM) to learn a probability distribution over random image patches, which highlighted salient regions based on the Bayesian surprise between observed patches and the prior distribution learned by the RBM. In a different probabilistic approach, J. Li, Tian, and T. Huang (2014) learned priors over the correlations of salient patches from millions of natural images in order to enhance salient regions. However, whereas these other approaches used natural image statistics from randomly-sampled natural images, we derived category-specific priors in order to estimate the Bayesian surprise with respect to priors that were relevant for a given scene.

Our finding that category-specific priors improved performance for most scene types relative to category-agnostic priors contributes to a complicated literature regarding the effects of priors (Chikkerur et al., 2010; Torralba, Oliva, et al., 2006; Yu and Dayan, 2004), priming (Maljkovic and Nakayama, 1994, 1996, 2000), and selection history (Awh, Belopolsky, and Theeuwes, 2012; Failing and Theeuwes, 2018; Theeuwes, 2019) on perception and gaze behavior. Each of these areas of research overlap in their study of attentional effects based on statistical regularities, but they tend to differ in the temporal window within which these effects influence perception: priming for trials (Maljkovic and Nakayama, 1994); selection history for experimental sessions (Della Libera and Chelazzi, 2009); and priors over years (Adams, 2007). Since images from all categories were presented in a random order to each observer in the CAT2000 dataset, it is unlikely that our category-specific prior model reflects priming or selection history effects. Our model was trained and tested on different image sets for each category, but effects due to priming and selection history would depend on images that each observer had seen. Although a model utilizing previous trial or fixation data from individual observers would likely account for additional variance in gaze behavior, a probability distribution over the features of preceding images averaged across observers would be similar to the category-agnostic prior used in the baseline model of the current study. Therefore, a promising future direction from this work is to investigate the unique variance in gaze behavior explained by priming, selection history, and priors.

Our finding that high-level features, compared to low- and mid-level features, contributed most to predictions of gaze behavior during free viewing supports previous literature that has demonstrated the importance of high-level semantic information relative to low-level saliency in natural scenes (Henderson and Hayes, 2017, 2018; C. C. Williams and Castelhano, 2019; Wu, Wick, and Pomplun, 2014). Chakraborty, Samaras, and Zelinsky (2022) recently studied the influences of low-level saliency, target features, and object uncertainty on free-viewing and visual search tasks. For free viewing, they found that object uncertainty had more predictive value than low-level saliency, where their definition of object uncertainty is similar to our construct of Bayesian surprise using high-level feature priors. Interestingly, they found that low-level saliency predictions outperformed predictions from target features in target-absent visual search and vice versa for target-present search, concluding that target-absent search may be more exploratory than target-guided (i.e., guided by features similar to those of the target representation). Although the current study only predicted gaze behavior during free viewing, our finding that low-level features had greater unique contribution to top-down compared to bottom-up priority maps suggests an alternative hypothesis that top-down attention may have more complementary influence across the visual hierarchy than is assumed by models that only consider object-level features in top-down target prediction.

Although our goal was not to outperform current state-of-the-art methods, there were a few limitations in the current study that could be improved upon in future research. Given the complexity of our approach, we did not include methods used in other state-of-the-art models, such as center biasing (Tatler, 2007) or learning associations between extracted features and human gaze data (Kümmerer, Wallis, and Bethge, 2016), both of which would likely improve performance. Furthermore, we evaluated performance assuming linear combination with equal weights for the feature and spatial components of our model. Future work should evaluate biologically-informed or data-driven combinations that may provide further insights into the contributions of spatial and feature priors to attention. Finally, by using training sets of 100 images per category for learning priors, it is likely that these priors were suboptimal in their representations of features and spatial associations compared to priors learned from larger training sets. With large publicly available datasets used for training DNNs, it should be possible to obtain many more images per category in CAT2000 in order to improve feature and spatial priors in future work.

In summary, we characterized the contributions of bottom-up and top-down attention using category-specific priors for predicting gaze behavior during free viewing. For most categories, Bayesian surprise from priors over features that were relevant to a scene's category best explained variance in gaze behavior predicted by our model. Although high-level features were most important in our model, we also observed differences across categories with respect to the predictive value of individual priority maps. Overall, our results provide new insights regarding the category-dependent relationships among bottom-up Bayesian surprise and top-down priors at varying levels of feature complexity and spatial scale.

Chapter 5

General Conclusions

In this dissertation I proposed and evaluated three computational models of visual attention, defining attention's role as influencing perceptual inference via top-down spatial and feature priors as well as bottom-up Bayesian surprise. Each experiment modeled different effects of attentional modulation (top-down spatial attention, feature-based attention, and bottom-up surprise) on visual perception across three behavioral tasks (visual crowding, visual search, and free viewing). Furthermore, the models incorporated principles (e.g., Bayesian surprise; Gijsen et al., 2021), mechanisms (e.g., Gaussian multiplication of cortical receptive fields; Klein, Harvey, and Dumoulin, 2014) and structures (e.g., distributed attentional priority maps; Serences and Yantis, 2007) for which supporting evidence has been found in neurophysiological studies of humans.

In Chapter 2, I designed a receptive field pooling model that mimics covert spatial attention in peripheral vision during a visual crowding task. Inspired by theoretical work positing spatial attention as a prior over regions of the visual field that reduces uncertainty in visual processing (Rao, 2005; Yu and Dayan, 2004), the pooling model also simulated known attention-related changes in cortical receptive field size and position measured in macaques and humans (Klein, Harvey, and Dumoulin, 2014; Womelsdorf et al., 2006). The model reproduced patterns of target classification performance in human subjects that have been reported previously in the visual crowding literature and further provided predictions regarding how spatial attention influences downstream perception during crowding.

In Chapter 3, I used a hierarchical generative model to simulate feature-based attention with probabilistic priors learned over digit features, and I evaluated performance on a visual search task for a single digit in an array of non-digit distractors. The feature priors in this study were used to disambiguate among target and distractor features by increasing the relative priority associated with digit features across the image space when predicting target location. Using the receptive field pooling model from Chapter 2, spatial attention was then simulated at the predicted target location, which enhanced local spatial resolution, thereby improving digit classification at the predicted location. By comparing performance using priority maps with and without top-down priors, the results demonstrate that implementing priors over digit features improved target location and subsequently classification accuracy, relative to the bottom-up priority map.

In Chapter 4, I modeled category-specific feature and spatial priors over features extracted from natural images by using normalizing flow models at different levels of a pre-trained deep neural network. I characterized the relative contributions of bottom-up Bayesian surprise and top-down priors at varying levels of feature complexity and spatial scale to gaze predictions for a free-viewing experiment across 20 categories of images. I compared the model using category-specific priors (i.e., trained within each category separately) to one with category-agnostic priors (i.e., trained across all categories), demonstrating performance improvement across the majority of categories when using category-specific priors. I further evaluated the unique variance in gaze behavior explained by bottom-up Bayesian surprise and top-down priors, finding that Bayesian surprise accounted for greater unique variance relative to top-down priors. Finally, I additionally evaluated the unique variance explained by different layers of extracted features, which revealed that higher-level features accounted for greater unique variance in gaze behavior compared to lower-level features.

Bibliography

Adams, Wendy J (2007). “A common light-prior for visual search, shape, and reflectance judgments”. In: *Journal of Vision* 7.11, pp. 11–11.

Albonico, Andrea et al. (2018). “Focusing and orienting spatial attention differently modulate crowding in central and peripheral vision”. In: *Journal of vision* 18.3, pp. 4–4.

Anton-Erxleben, Katharina and Marisa Carrasco (2013). “Attentional enhancement of spatial resolution: Linking behavioural and neurophysiological evidence”. In: *Nature Reviews Neuroscience* 14.3, p. 188.

Anton-Erxleben, Katharina, Valeska M Stephan, and Stefan Treue (2009). “Attention reshapes center-surround receptive field structure in macaque cortical area MT”. In: *Cerebral Cortex* 19.10, pp. 2466–2478.

Awh, Edward, Artem V Belopolsky, and Jan Theeuwes (2012). “Top-down versus bottom-up attentional control: A failed theoretical dichotomy”. In: *Trends in cognitive sciences* 16.8, pp. 437–443.

Balas, Benjamin, Lisa Nakano, and Ruth Rosenholtz (2009). “A summary-statistic representation in peripheral vision explains visual crowding”. In: *Journal of vision* 9.12, pp. 13.1–13.18.

Banks, William P, Kenneth M Bachrach, and Douglas W Larson (1977). “The asymmetry of lateral interference in visual letter identification”. In: *Perception & Psychophysics* 22.3, pp. 232–240.

Barbot, Antoine and Marisa Carrasco (2017). “Attention modifies spatial resolution according to task demands”. In: *Psychological science* 28.3, pp. 285–296.

Baruch, Orit and Yaffa Yeshurun (2014). “Attentional attraction of receptive fields can explain spatial and temporal effects of attention”. In: *Visual Cognition* 22.5, pp. 704–736.

Betz, Torsten et al. (2010). “Investigating task-dependent top-down effects on overt visual attention”. In: *Journal of vision* 10.3, pp. 15–15.

Borji, Ali (2012). “Boosting bottom-up and top-down visual features for saliency estimation”. In: *2012 ieee conference on computer vision and pattern recognition*. IEEE, pp. 438–445.

Borji, Ali and Laurent Itti (2015). “Cat2000: A large scale fixation dataset for boosting saliency research”. In: *arXiv preprint arXiv:1505.03581*.

Borji, Ali, Dicky N Sihite, and Laurent Itti (2012). “Quantitative analysis of human-model agreement in visual saliency modeling: A comparative study”. In: *IEEE Transactions on Image Processing* 22.1, pp. 55–69.

Borji, Ali, Dicky N Sihite, and Laurent Itti (2013). “What/where to look next? Modeling top-down visual attention in complex interactive environments”. In: *IEEE Transactions on Systems, Man, and Cybernetics: Systems* 44.5, pp. 523–538.

Borji, Ali, Dicky Sihite, and Laurent Itti (2012). “An object-based bayesian framework for top-down visual attention”. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 26. 1, pp. 1529–1535.

Bouma, Herman (1970). “Interaction effects in parafoveal letter recognition”. In: *Nature* 226.5241, pp. 177–178.

Bravo, Mary J and Ken Nakayama (1992). “The role of attention in different visual-search tasks”. In: *Perception & psychophysics* 51.5, pp. 465–472.

Carpenter, Roger HS and MLL Williams (1995). “Neural computation of log likelihood in control of saccadic eye movements”. In: *Nature* 377.6544, pp. 59–62.

Carrasco, Marisa (2011). “Visual attention: The past 25 years”. In: *Vision research* 51.13, pp. 1484–1525.

Chakraborty, Souradeep, Dimitris Samaras, and Gregory J Zelinsky (2022). “Weighting the factors affecting attention guidance during free viewing and visual search: The unexpected role of object recognition uncertainty”. In: *Journal of Vision* 22.4, pp. 13–13.

Chaney, Wesley, Jason Fischer, and David Whitney (2014). “The hierarchical sparse selection model of visual crowding”. In: *Frontiers in integrative neuroscience* 8, pp. 73.1–73.11.

Charikar, Moses S (2002). “Similarity estimation techniques from rounding algorithms”. In: *Proceedings of the thirty-fourth annual ACM symposium on Theory of computing*, pp. 380–388.

Chen, Francis X et al. (2017). “Eccentricity dependent deep neural networks: Modeling invariance in human vision.” In: *AAAI spring symposium series*, pp. 541–546.

Chen, Juan et al. (2014). “Attention-dependent early cortical suppression contributes to crowding”. In: *Journal of Neuroscience* 34.32, pp. 10465–10474.

Chen, Yuzhi, Wilson S Geisler, and Eyal Seidemann (2006). “Optimal decoding of correlated neural population responses in the primate visual cortex”. In: *Nature neuroscience* 9.11, pp. 1412–1420.

Cheung, Brian, Eric Weiss, and Bruno Olshausen (2016). “Emergence of foveal image sampling from learning to attend in visual scenes”. In: *arXiv preprint arXiv:1611.09430*.

Chikkerur, Sharat et al. (2010). “What and where: A Bayesian inference theory of attention”. In: *Vision research* 50.22, pp. 2233–2247.

Coates, Daniel R, Jean-Baptiste Bernard, and Susana TL Chung (2019). “Feature contingencies when reading letter strings”. In: *Vision research* 156, pp. 84–95.

Connor, Charles E et al. (1997). “Spatial attention effects in macaque area V4”. In: *Journal of Neuroscience* 17.9, pp. 3201–3214.

Dayan, Peter, Geoffrey E Hinton, et al. (1995). “The helmholtz machine”. In: *Neural computation* 7.5, pp. 889–904.

Dayan, Peter and Richard S Zemel (1999). “Statistical models and sensory attention”. In: *1999 Ninth International Conference on Artificial Neural Networks ICANN 99.(Conf. Publ. No. 470)*. Vol. 2. IET, pp. 1017–1022.

Deco, Gustavo and Tai Sing Lee (2002). “A unified model of spatial and object attention based on inter-cortical biased competition”. In: *Neurocomputing* 44, pp. 775–781.

Della Libera, Chiara and Leonardo Chelazzi (2009). “Learning to attend and to ignore is a matter of gains and losses”. In: *Psychological science* 20.6, pp. 778–784.

Deneve, Sophie, Peter E Latham, and Alexandre Pouget (1999). “Reading population codes: a neural implementation of ideal observers”. In: *Nature neuroscience* 2.8, pp. 740–745.

Deng, Jia et al. (2009). “Imagenet: A large-scale hierarchical image database”. In: *2009 IEEE conference on computer vision and pattern recognition*. Ieee, pp. 248–255.

Desimone, Robert and John Duncan (1995). “Neural mechanisms of selective visual attention”. In: *Annual review of neuroscience* 18.1, pp. 193–222.

Devereux, Barry J, Alex Clarke, and Lorraine K Tyler (2018). “Integrated deep visual and semantic attractor neural networks predict fMRI pattern-information along the ventral object processing pathway”. In: *Scientific reports* 8.1, pp. 1–12.

Deza, Arturo and Miguel Eckstein (2016). “Can peripheral representations improve clutter metrics on complex scenes?” In: *Advances in neural information processing systems*, pp. 2847–2855.

Deza, Arturo, Aditya Jonnalagadda, and Miguel P. Eckstein (2019). “Towards Metamerism via Foveated Style Transfer.” In: *International Conference on Learning Representations*.

Dinh, Laurent, David Krueger, and Yoshua Bengio (2014). “Nice: Non-linear independent components estimation”. In: *arXiv preprint arXiv:1410.8516*.

Dinh, Laurent, Jascha Sohl-Dickstein, and Samy Bengio (2016). “Density estimation using real nvp”. In: *arXiv preprint arXiv:1605.08803*.

Efron, Bradley and Robert J Tibshirani (1994). *An introduction to the bootstrap*. CRC press.

Ernst, Marc O and Martin S Banks (2002). “Humans integrate visual and haptic information in a statistically optimal fashion”. In: *Nature* 415.6870, pp. 429–433.

Ester, Edward F, Daniel Klee, and Edward Awh (2014). “Visual crowding cannot be wholly explained by feature pooling.” In: *Journal of Experimental Psychology: Human Perception and Performance* 40.3, p. 1022.

Failing, Michel and Jan Theeuwes (2018). “Selection history: How reward modulates selectivity of visual attention”. In: *Psychonomic bulletin & review* 25.2, pp. 514–538.

Farzin, Faraz, Susan M Rivera, and David Whitney (2009). “Holistic crowding of Mooney faces”. In: *Journal of vision* 9.6, pp. 18–18.

Feldman, Harriet and Karl Friston (2010). “Attention, uncertainty, and free-energy”. In: *Frontiers in human neuroscience* 4, p. 215.

Freeman, Jeremy and Eero P Simoncelli (2011). “Metamers of the ventral stream”. In: *Nature neuroscience* 14.9, pp. 1195–1201.

French, Robert M (1999). “Catastrophic forgetting in connectionist networks”. In: *Trends in cognitive sciences* 3.4, pp. 128–135.

Friston, Karl (2009). “The free-energy principle: a rough guide to the brain?” In: *Trends in cognitive sciences* 13.7, pp. 293–301.

Gattass, Ricardo, CG Gross, and JH Sandell (1981). “Visual topography of V2 in the macaque”. In: *Journal of Comparative Neurology* 201.4, pp. 519–539.

Gattass, Ricardo, AP Sousa, and CG Gross (1988). "Visuotopic organization and extent of V3 and V4 of the macaque". In: *Journal of neuroscience* 8.6, pp. 1831–1845.

Gijsen, Sam et al. (2021). "Neural surprise in somatosensory Bayesian learning". In: *PLoS computational biology* 17.2, e1008068.

Großekathöfer, Jonas D, Kristina Suchotzki, and Matthias Gamer (2020). "Gaze cueing in naturalistic scenes under top-down modulation–Effects on gaze behaviour and memory performance". In: *Visual Cognition* 28.2, pp. 135–147.

Hamker, Fred H (2004). "A dynamic model of how feature cues guide spatial attention". In: *Vision research* 44.5, pp. 501–521.

Hanus, Deborah and Edward Vul (2013). "Quantifying error distributions in crowding". In: *Journal of Vision* 13.4, pp. 17–17.

Harel, Jonathan, Christof Koch, and Pietro Perona (2006). "Graph-based visual saliency". In: *Advances in neural information processing systems* 19.

Hayes, Taylor and John Henderson (2021). "Deep saliency models learn low-, mid-, and high-level features to predict scene attention". In: *Scientific Reports* 11.1, pp. 1–13.

He, Dongjun, Yingying Wang, and Fang Fang (2019). "The critical role of V2 population receptive fields in visual orientation crowding". In: *Current Biology* 29.13, pp. 2229–2236.

Henderson, John and Taylor Hayes (2017). "Meaning-based guidance of attention in scenes as revealed by meaning maps". In: *Nature human behaviour* 1.10, pp. 743–747.

Henderson, John and Taylor Hayes (2018). "Meaning guides attention in real-world scene images: Evidence from eye". In:

Herzog, Michael H et al. (2015). "Crowding, grouping, and object recognition: A matter of appearance". In: *Journal of vision* 15.6, pp. 5–5.

Hinton, Geoffrey E (2002). "Training products of experts by minimizing contrastive divergence". In: *Neural computation* 14.8, pp. 1771–1800.

Hinton, Geoffrey E (2012). "A practical guide to training restricted Boltzmann machines". In: *Neural networks: Tricks of the trade*. Springer, pp. 599–619.

Hinton, Geoffrey E and Ruslan R Salakhutdinov (2006). "Reducing the dimensionality of data with neural networks". In: *science* 313.5786, pp. 504–507.

Hua, Yan et al. (2013). "A probabilistic saliency model with memory-guided top-down cues for free-viewing". In: *2013 IEEE International Conference on Multimedia and Expo (ICME)*. IEEE, pp. 1–6.

Itti, Laurent and Pierre Baldi (2005). "A principled approach to detecting surprising events in video". In: *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*. Vol. 1. IEEE, pp. 631–637.

Itti, Laurent and Pierre Baldi (2009). "Bayesian surprise attracts human attention". In: *Vision research* 49.10, pp. 1295–1306.

Itti, Laurent, Christof Koch, and Ernst Niebur (1998). "A model of saliency-based visual attention for rapid scene analysis". In: *IEEE Transactions on pattern analysis and machine intelligence* 20.11, pp. 1254–1259.

Jia, Yangqing, Chang Huang, and Trevor Darrell (2012). "Beyond spatial pyramids: Receptive field learning for pooled image features". In: *2012 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, pp. 3370–3377.

Kersten, Daniel (1999). "High-level vision as statistical inference". In: *The new cognitive neurosciences* 2, pp. 353–363.

Kersten, Daniel, Pascal Mamassian, and Alan Yuille (2004). "Object perception as Bayesian inference". In: *Annu. Rev. Psychol.* 55, pp. 271–304.

Keshvari, Shaiyan and Ruth Rosenholtz (2016). "Pooling of continuous features provides a unifying account of crowding". In: *Journal of Vision* 16.3, pp. 39–39.

Kingma, Diederik P and Jimmy Ba (2014). "Adam: A method for stochastic optimization". In: *arXiv preprint arXiv:1412.6980*.

Kingma, Durk P and Prafulla Dhariwal (2018). "Glow: Generative flow with invertible 1x1 convolutions". In: *Advances in neural information processing systems* 31.

Klein, Barrie P, Ben M Harvey, and Serge O Dumoulin (2014). "Attraction of position preference by spatial attention throughout human visual cortex". In: *Neuron* 84.1, pp. 227–237.

Knill, David C and Alexandre Pouget (2004). "The Bayesian brain: the role of uncertainty in neural coding and computation". In: *TRENDS in Neurosciences* 27.12, pp. 712–719.

Knill, David C and Whitman Richards (1996). *Perception as Bayesian inference*. Cambridge University Press.

Koch, Christof and Shimon Ullman (1985). "Shifts in selective visual attention: towards the underlying neural circuitry". In: *Human neurobiology* 4.4, pp. 219–227.

Kolossa, Antonio, Bruno Kopp, and Tim Fingscheidt (2015). "A computational analysis of the neural bases of Bayesian inference". In: *Neuroimage* 106, pp. 222–237.

Kording, Konrad P et al. (2020). "Appreciating the variety of goals in computational neuroscience". In: *arXiv preprint arXiv:2002.03211*.

Kümmerer, Matthias, Thomas SA Wallis, and Matthias Bethge (2015). "Information-theoretic model comparison unifies saliency metrics". In: *Proceedings of the National Academy of Sciences* 112.52, pp. 16054–16059.

Kümmerer, Matthias, Thomas SA Wallis, and Matthias Bethge (2016). "DeepGaze II: Reading fixations from deep features trained on object recognition". In: *arXiv preprint arXiv:1610.01563*.

Larochelle, Hugo and Geoffrey E Hinton (2010). "Learning to combine foveal glimpses with a third-order Boltzmann machine". In: *Advances in neural information processing systems* 23.

LeCun, Yann et al. (1998). "Gradient-based learning applied to document recognition". In: *Proceedings of the IEEE* 86.11, pp. 2278–2324.

Lee, Honglak et al. (2009). "Convolutional deep belief networks for scalable unsupervised learning of hierarchical representations". In: *Proceedings of the 26th annual international conference on machine learning*, pp. 609–616.

Lee, Tai Sing and David Mumford (2003). "Hierarchical Bayesian inference in the visual cortex". In: *JOSA A* 20.7, pp. 1434–1448.

Levi, Dennis M (2008). "Crowding—An essential bottleneck for object recognition: A mini-review". In: *Vision research* 48.5, pp. 635–654.

Li, Aoqi, Jeremy M Wolfe, and Zhenzhong Chen (2020). “Implicitly and explicitly encoded features can guide attention in free viewing”. In: *Journal of Vision* 20.6, pp. 8–8.

Li, Jia, Yonghong Tian, and Tiejun Huang (2014). “Visual saliency with statistical priors”. In: *International journal of computer vision* 107.3, pp. 239–253.

Li, Zhaoping (1999). “Contextual influences in V1 as a basis for pop out and asymmetry in visual search”. In: *Proceedings of the National Academy of Sciences* 96.18, pp. 10530–10535.

Lonnqvist, Ben, Alasdair DF Clarke, and Ramakrishna Chakravarthi (2020). “Crowding in humans is unlike that in convolutional neural networks”. In: *Neural Networks*.

MacInnes, W Joseph et al. (2020). “No advantage for separating overt and covert attention in visual search”. In: *Vision* 4.2, p. 28.

Mahdi, Ali, Jun Qin, and Garth Crosby (2019). “DeepFeat: A bottom-up and top-down saliency model based on deep features of convolutional neural networks”. In: *IEEE Transactions on Cognitive and Developmental Systems* 12.1, pp. 54–63.

Maljkovic, Vera and Ken Nakayama (1994). “Priming of pop-out: I. Role of features”. In: *Memory & cognition* 22.6, pp. 657–672.

Maljkovic, Vera and Ken Nakayama (1996). “Priming of pop-out: II. The role of position”. In: *Perception & psychophysics* 58.7, pp. 977–991.

Maljkovic, Vera and Ken Nakayama (2000). “Priming of popout: III. A short-term implicit memory system beneficial for rapid target selection”. In: *Visual cognition* 7.5, pp. 571–595.

Manassi, Mauro, Bilge Sayim, and Michael H Herzog (2012). “Grouping, pooling, and when bigger is better in visual crowding”. In: *Journal of Vision* 12.10, pp. 13–13.

Manassi, Mauro and David Whitney (2018). “Multi-level crowding and the paradox of object recognition in clutter”. In: *Current Biology* 28.3, R127–R133.

Martinez-Trujillo, Julio C and Stefan Treue (2004). “Feature-based attention increases the selectivity of population responses in primate visual cortex”. In: *Current biology* 14.9, pp. 744–751.

Martınez-Trujillo, Julio C and Stefan Treue (2002). “Attentional modulation strength in cortical area MT depends on stimulus contrast”. In: *Neuron* 35.2, pp. 365–370.

McAdams, Carrie J and John HR Maunsell (1999). “Effects of attention on orientation-tuning functions of single neurons in macaque cortical area V4”. In: *Journal of Neuroscience* 19.1, pp. 431–441.

Melloni, Lucia et al. (2012). “Interaction between bottom-up saliency and top-down control: how saliency maps are created in the human brain”. In: *Cerebral cortex* 22.12, pp. 2943–2952.

Moran, Jeffrey and Robert Desimone (1985). “Selective attention gates visual processing in the extrastriate cortex”. In: *Science* 229.4715, pp. 782–784.

Motter, Brad C (2009). “Central V4 receptive fields are scaled by the V1 cortical magnification and correspond to a constant-sized sampling of the V1 surface”. In: *Journal of Neuroscience* 29.18, pp. 5749–5757.

Mozer, Michael C and Mark Sitton (1998). “Computational modeling of spatial attention”. In: *Attention* 9, pp. 341–393.

Murabito, Francesca et al. (2018). “Top-down saliency detection driven by visual classification”. In: *Computer Vision and Image Understanding* 172, pp. 67–76.

Nakayama, Ken and Shinsuke Shimojo (1992). “Experiencing and perceiving visual surfaces”. In: *Science* 257.5075, pp. 1357–1363.

Nandy, Anirvan S and Bosco S Tjan (2012). “Saccade-confounded image statistics explain visual crowding”. In: *Nature neuroscience* 15.3, pp. 463–469.

Nigam, Sunny, Sorin Pojoga, and Valentin Dragoi (2019). “Synergistic Coding of Visual Information in Columnar Networks”. In: *Neuron* 104.2, pp. 402–411.

O’Connell, Thomas P and Marvin M Chun (2018). “Predicting eye movement patterns from fMRI responses to natural scenes”. In: *Nature communications* 9.1, pp. 1–15.

Oliva, Aude et al. (2003). “Top-down control of visual attention in object detection”. In: *Proceedings 2003 International Conference on Image Processing (Cat. No. 03CH37429)*. Vol. 1. IEEE, pp. I–253.

Olshausen, Bruno A, Charles H Anderson, and David C Van Essen (1993). “A neurobiological model of visual attention and invariant pattern recognition based on dynamic routing of information”. In: *Journal of Neuroscience* 13.11, pp. 4700–4719.

Ostwald, Dirk et al. (2012). “Evidence for neural encoding of Bayesian surprise in human somatosensation”. In: *NeuroImage* 62.1, pp. 177–188.

Parkhurst, Derrick, Klinton Law, and Ernst Niebur (2002). “Modeling the role of salience in the allocation of overt visual attention”. In: *Vision research* 42.1, pp. 107–123.

Paszke, Adam et al. (2017). “Automatic Differentiation in PyTorch”. In: *NIPS Autodiff Workshop*.

Pelli, Denis G (1985). “Uncertainty explains many aspects of visual contrast detection and discrimination”. In: *JOSA A* 2.9, pp. 1508–1532.

Peters, Robert J et al. (2005). “Components of bottom-up gaze allocation in natural images”. In: *Vision research* 45.18, pp. 2397–2416.

Petrov, Yury and Olga Meleshkevich (2011). “Asymmetries and idiosyncratic hot spots in crowding”. In: *Vision research* 51.10, pp. 1117–1123.

Posner, Michael I (1980). “Orienting of attention”. In: *Quarterly journal of experimental psychology* 32.1, pp. 3–25.

Pouget, Alexandre, Peter Dayan, and Richard Zemel (2000). “Information processing with population codes”. In: *Nature Reviews Neuroscience* 1.2, pp. 125–132.

Rao, Rajesh PN (2005). “Bayesian inference and attentional modulation in the visual cortex”. In: *Neuroreport* 16.16, pp. 1843–1848.

Rao, Rajesh PN, Bruno A Olshausen, and Michael S Lewicki (2002). *Probabilistic models of the brain: Perception and neural function*. MIT press.

Ren, Shaoqing et al. (2015). “Faster r-cnn: Towards real-time object detection with region proposal networks”. In: *Advances in neural information processing systems*, pp. 91–99.

Reuther, Josephine and Ramakrishna Chakravarthi (2014). “Categorical membership modulates crowding: Evidence from characters”. In: *Journal of Vision* 14.6, pp. 5–5.

Reynolds, John H, Leonardo Chelazzi, and Robert Desimone (1999). “Competitive mechanisms subserve attention in macaque areas V2 and V4”. In: *Journal of Neuroscience* 19.5, pp. 1736–1753.

Reynolds, John H and David J Heeger (2009). “The normalization model of attention”. In: *Neuron* 61.2, pp. 168–185.

Rizzolatti, Giacomo et al. (1987). "Reorienting attention across the horizontal and vertical meridians: evidence in favor of a premotor theory of attention". In: *Neuropsychologia* 25.1, pp. 31–40.

Rosenholtz, R (2016). "Capabilities and Limitations of Peripheral Vision." In: *Annual review of vision science* 2, pp. 437–457.

Roy, Rahul, Susmita Ghosh, and Ashish Ghosh (2018). "Salient Object Detection based on Bayesian Surprise of Restricted Boltzmann Machine". In: *Proceedings of the 11th Indian Conference on Computer Vision, Graphics and Image Processing*, pp. 1–8.

Sabour, Sara, Nicholas Frosst, and Geoffrey E Hinton (2017). "Dynamic routing between capsules". In: *Advances in neural information processing systems*, pp. 3856–3866.

Scolari, Miranda et al. (2007). "Spatial attention, preview, and popout: Which factors influence critical spacing in crowded displays?" In: *Journal of Vision* 7.2, pp. 7.1–7.23.

Serences, John T and Steven Yantis (2007). "Spatially selective representations of voluntary and stimulus-driven attentional priority in human occipital, parietal, and frontal cortex". In: *Cerebral cortex* 17.2, pp. 284–293.

Simoncelli, Eero P and Bruno A Olshausen (2001). "Natural image statistics and neural representation". In: *Annual review of neuroscience* 24.1, pp. 1193–1216.

Simonyan, Karen and Andrew Zisserman (2014). "Very deep convolutional networks for large-scale image recognition". In: *arXiv preprint arXiv:1409.1556*.

Smolensky, Paul (1986). *Information processing in dynamical systems: Foundations of harmony theory*. Tech. rep. Colorado Univ at Boulder Dept of Computer Science.

Somers, David C et al. (1999). "Functional MRI reveals spatially specific attentional modulation in human primary visual cortex". In: *Proceedings of the National Academy of Sciences* 96.4, pp. 1663–1668.

Sprague, Thomas C and John T Serences (2013). "Attention modulates spatial priority maps in the human occipital, parietal and frontal cortices". In: *Nature neuroscience* 16.12, pp. 1879–1887.

Stone, JV, IS Kerrigan, and J Porrill (2009). "Where is the light? Bayesian perceptual priors for lighting direction". In: *Proceedings of the Royal Society B: Biological Sciences* 276.1663, pp. 1797–1804.

Sun, Gerald J, Susana TL Chung, and Bosco S Tjan (2010). "Ideal observer analysis of crowding and the reduction of crowding through learning". In: *Journal of vision* 10.5, pp. 16.1–16.14.

Tatler, Benjamin W (2007). "The central fixation bias in scene viewing: Selecting an optimal viewing position independently of motor biases and image feature distributions". In: *Journal of vision* 7.14, pp. 4–4.

Theeuwes, Jan (2019). "Goal-driven, stimulus-driven, and history-driven selection". In: *Current opinion in psychology* 29, pp. 97–101.

Theiss, Justin D, Joel D Bowen, and Michael A Silver (2022). "Spatial Attention Enhances Crowded Stimulus Encoding Across Modeled Receptive Fields by Increasing Redundancy of Feature Representations". In: *Neural Computation* 34.1, pp. 190–218.

Toet, Alexander and Dennis M Levi (1992). "The two-dimensional shape of spatial interaction zones in the parafovea". In: *Vision research* 32.7, pp. 1349–1357.

Torralba, Antonio and Aude Oliva (2003). “Statistics of natural image categories”. In: *Network: computation in neural systems* 14.3, p. 391.

Torralba, Antonio, Aude Oliva, et al. (2006). “Contextual guidance of eye movements and attention in real-world scenes: the role of global features in object search.” In: *Psychological review* 113.4, p. 766.

Treue, Stefan and Julio C Martinez Trujillo (1999). “Feature-based attention influences motion processing gain in macaque visual cortex”. In: *Nature* 399.6736, pp. 575–579.

Van den Berg, Ronald, Jos BTM Roerdink, and Frans W Cornelissen (2010). “A neurophysiologically plausible population code model for feature integration explains visual crowding”. In: *PLoS Comput Biol* 6.1, e1000646.

Vaswani, Ashish et al. (2017). “Attention is all you need”. In: *Advances in neural information processing systems*, pp. 5998–6008.

Vincent, Benjamin T (2015). “Bayesian accounts of covert selective attention: a tutorial review”. In: *Attention, Perception, & Psychophysics* 77.4, pp. 1013–1032.

Vo, Vy A, Thomas C Sprague, and John T Serences (2017). “Spatial tuning shifts increase the discriminability and fidelity of population codes in visual cortex”. In: *Journal of Neuroscience* 37.12, pp. 3386–3401.

Volokitin, Anna, Gemma Roig, and Tomaso A Poggio (2017). “Do deep neural networks suffer from crowding?” In: *Advances in Neural Information Processing Systems*, pp. 5628–5638.

Von Helmholtz, Hermann (1867). *Handbuch der physiologischen Optik: mit 213 in den Text eingedruckten Holzschnitten und 11 Tafeln*. Vol. 9. Voss.

Vossel, Simone et al. (2015). “Cortical coupling reflects Bayesian belief updating in the deployment of spatial attention”. In: *Journal of Neuroscience* 35.33, pp. 11532–11542.

Wandell, Brian A and Jonathan Winawer (2015). “Computational neuroimaging and population receptive fields”. In: *Trends in cognitive sciences* 19.6, pp. 349–357.

Wang, Jingwei et al. (2016). “Learning a combined model of visual saliency for fixation prediction”. In: *IEEE Transactions on Image Processing* 25.4, pp. 1566–1579.

Weiss, Yair, Eero P Simoncelli, and Edward H Adelson (2002). “Motion illusions as optimal percepts”. In: *Nature neuroscience* 5.6, pp. 598–604.

Whiteley, Louise and Maneesh Sahani (2012). “Attention in a Bayesian framework”. In: *Frontiers in human neuroscience* 6, p. 100.

Whitney, David and Dennis M Levi (2011). “Visual crowding: A fundamental limit on conscious perception and object recognition”. In: *Trends in cognitive sciences* 15.4, pp. 160–168.

Williams, Carrick C and Monica S Castelhano (2019). “The changing landscape: High-level influences on eye movement guidance in scenes”. In: *vision* 3.3, p. 33.

Wolfe, Jeremy M, Kyle R Cave, and Susan L Franzel (1989). “Guided search: an alternative to the feature integration model for visual search.” In: *Journal of Experimental Psychology: Human perception and performance* 15.3, p. 419.

Wolfe, Jeremy M and Todd S Horowitz (2017). “Five factors that guide attention in visual search”. In: *Nature Human Behaviour* 1.3, pp. 1–8.

Womelsdorf, Thilo et al. (2006). “Dynamic shifts of visual receptive fields in cortical area MT by spatial attention”. In: *Nature neuroscience* 9.9, pp. 1156–1160.

Wu, Chia-Chien, Farahnaz Ahmed Wick, and Marc Pomplun (2014). “Guidance of visual attention by semantic information in real-world scenes”. In: *Frontiers in psychology* 5, p. 54.

Xie, Yulin and Huchuan Lu (2011). “Visual saliency detection based on Bayesian model”. In: *2011 18th IEEE International Conference on Image Processing*. IEEE, pp. 645–648.

Yamins, Daniel LK et al. (2014). “Performance-optimized hierarchical models predict neural responses in higher visual cortex”. In: *Proceedings of the national academy of sciences* 111.23, pp. 8619–8624.

Yang, Jimei and Ming-Hsuan Yang (2016). “Top-down visual saliency via joint CRF and dictionary learning”. In: *IEEE transactions on pattern analysis and machine intelligence* 39.3, pp. 576–588.

Yang, Scott Cheng-Hsin, Mate Lengyel, and Daniel M Wolpert (2016). “Active sensing in the categorization of visual patterns”. In: *Elife* 5, e12215.

Yeshurun, Yaffa and Marisa Carrasco (1998). “Attention improves or impairs visual performance by enhancing spatial resolution”. In: *Nature* 396.6706, pp. 72–75.

Yeshurun, Yaffa and Marisa Carrasco (2008). “The effects of transient attention on spatial resolution and the size of the attentional cue”. In: *Perception & Psychophysics* 70.1, pp. 104–113.

Yeshurun, Yaffa, Barbara Montagna, and Marisa Carrasco (2008). “On the flexibility of sustained attention and its effects on a texture segmentation task”. In: *Vision research* 48.1, pp. 80–95.

Yeshurun, Yaffa and Einat Rashal (2010). “Precueing attention to the target location diminishes crowding and reduces the critical distance”. In: *Journal of Vision* 10.10, pp. 16.1–16.12.

Yu, Angela J and Peter Dayan (2004). “Inference, attention, and decision in a Bayesian neural architecture”. In: *Advances in neural information processing systems* 17.

St-Yves, Ghislain and Thomas Naselaris (2018). “The feature-weighted receptive field: an interpretable encoding model for complex feature spaces”. In: *NeuroImage* 180, pp. 188–202.

Zhang, Lingyun et al. (2008). “SUN: A Bayesian framework for saliency using natural statistics”. In: *Journal of vision* 8.7, pp. 32–32.

Zhu, M (2004). *Recall, Precision, and Average Precision*. Tech. rep. University of Waterloo, Waterloo.