Title

Top-Down Priors Disambiguate Target and Distractor Features in Simulated Covert Visual Search

Permalink

https://escholarship.org/uc/item/0fh3c28d

Journal

Neural Computation, 36(10)

ISSN

0899-7667

Authors

Theiss, Justin D
Silver, Michael A

Publication Date

2024-09-17

DOI

10.1162/neco_a_01700

Peer reviewed

Top-Down Priors Disambiguate Target and Distractor Features in Simulated Covert Visual Search

Justin D. Theiss
theissjd@berkeley.edu
Michael A. Silver
masilver@berkeley.edu
University of California, Berkeley, CA 94720, U.S.A.

Several models of visual search consider visual attention as part of a perceptual inference process, in which top-down priors disambiguate bottom-up sensory information. Many of these models have focused on gaze behavior, but there are relatively fewer models of covert spatial attention, in which attention is directed to a peripheral location in visual space without a shift in gaze direction. Here, we propose a biologically plausible model of covert attention during visual search that helps to bridge the gap between Bayesian modeling and neurophysiological modeling by using (1) top-down priors over target features that are acquired through Hebbian learning, and (2) spatial resampling of modeled cortical receptive fields to enhance local spatial resolution of image representations for downstream target classification. By training a simple generative model using a Hebbian update rule, top-down priors for target features naturally emerge without the need for hand-tuned or predetermined priors. Furthermore, the implementation of covert spatial attention in our model is based on a known neurobiological mechanism, providing a plausible process through which Bayesian priors could locally enhance the spatial resolution of image representations. We validate this model during simulated visual search for handwritten digits among nondigit distractors, demonstrating that top-down priors improve accuracy for estimation of target location and classification, relative to bottom-up signals alone. Our results support previous reports in the literature that demonstrated beneficial effects of top-down priors on visual search performance, while extending this literature to incorporate known neural mechanisms of covert spatial attention.

1 Introduction

Due to various dynamic environmental factors (e.g., lighting, motion, occlusion), humans frequently encounter noisy and/or ambiguous visual stimuli in everyday life. For example, the same object viewed from different angles can project widely varying geometries onto the retina. There is inherent

Neural Computation **36, 2201–2224** (2024)
https://doi.org/10.1162/neco_a_01700

uncertainty in visual perception of many natural stimuli, but humans often easily overcome this uncertainty when viewing complex environments. For example, humans encode and account for uncertainty in making predictions of object speed (Weiss et al., 2002) and size (Ernst & Banks, 2002). The Bayesian coding hypothesis (Knill & Pouget, 2004) suggests that humans represent sensory information probabilistically. Within this framework, the cortex is hypothesized to encode the conditional probability of features, given a set of sensory inputs.

Early computational models of perception introduced the notion of reducing uncertainty in observations (Pelli, 1985; Dayan et al., 1995; Dayan & Zemel, 1999; Lee & Mumford, 2003). In addition, work studying covert attention with the Posner cueing paradigm (Posner, 1980) employed Bayesian models to explain attentional effects on psychophysical performance in humans (Eckstein et al., 2002; Shimozaki et al., 2003) and proposed neuromodulators that could implement uncertainty computations (Yu & Dayan, 2005). Ma et al. (2011) later proposed a neural network model using biologically plausible operations via divisive normalization to implement visual search with probabilistic population coding, and this model could account for human performance on the same task. These and similar models (see Eckstein, 2017, for review) apply the Bayesian probabilistic framework (Knill & Pouget, 2004; Geisler, 2011) to explain psychophysical and neurobiological phenomena related to visual attention.

In line with these works, Rao (2005) proposed a probabilistic generative model of attention in which the visual system uses Bayes's rule to converge to probable explanations of the visual environment by combining bottom-up likelihoods of sensory information and top-down priors over spatial locations and features. In artificial experiments, the generative model was trained to represent probability distributions of stimuli over location and orientation dimensions. In the bottom-up direction, the posterior probabilities of location and orientation were inferred from an image. In the top-down direction, the prior probabilities over features or locations were used to influence an intermediate level of 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 reproduce effects of top-down attention that have been well characterized in neurophysiological studies.

Previous Bayesian models of attention have also successfully modeled patterns of human eye movements during both visual search and free-viewing of natural images. The contextual guidance model (Torralba et al., 2006) combines bottom-up saliency that is computed by a local pathway with scene priors that are computed by a global pathway. Impressively, the model was able to predict humans' 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 multiple known effects: feature pop-out (Bravo & Nakayama, 1992), multiplicative modulation of response amplitude (McAdams & Maunsell, 1999), and shifts and changes in gain of the contrast response function (Treue & Martínez Trujillo, 1999; Martínez-Trujillo & Treue, 2002). Furthermore, the model of Chikkerur et al. (2010) accounted for eye movements during both visual search and free viewing of natural images.

Since eye movements are commonly used to study visual search, most studies have focused on modeling overt as opposed to covert visual attention (i.e., directing spatial attention to a particular peripheral location without altering gaze position). Although the premotor theory of attention posits that covert and overt attention share many processes (Rizzolatti et al., 1987), overt attention uses the structural advantages of central over peripheral vision to improve spatial sampling at attended locations, while sustained covert spatial attention directly enhances the representations of encoded features. Specifically, 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 (Womelsdorf et al., 2006; Klein et al., 2014). Previous visual search models have incorporated differences between central and peripheral vision in spatial resolution: Zelinsky (2008) used a retina transform function to progressively blur the peripheral regions of the image, and Akbas and Eckstein (2017) used a foveated visual field based on the feature pooling method described in Freeman and Simoncelli (2011). However, these models studied overt attention using static foveated image processing applied at different locations in order to mimic eye movements and did not include covert spatial attention.

Taking inspiration from the normalization model of attention (Reynolds & Heeger, 2009), Theiss et al. (2022) described 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 that was centered at the attended location, modeling known effects of spatial attention on properties of neuronal and population-level RFs in visual cortex (Womelsdorf et al., 2006; Klein et al., 2014). The validity of this model was demonstrated across multiple experiments that replicated results from psychophysical studies of visual crowding in humans (Bouma, 1970; Banks et al., 1977; Toet & Levi, 1992). For visual search, the RF pooling array of Theiss et al. (2022) can be used to simulate allocation of covert spatial attention to a predicted target location in order to enhance local spatial processing at the attended location and improve downstream target classification.

In this study, we describe a biologically plausible Bayesian model of attention that learns priors over target features through Hebbian mechanisms and employs top-down spatial attention in order to simulate covert visual search. The model uses these feature priors to disambiguate bottom-up

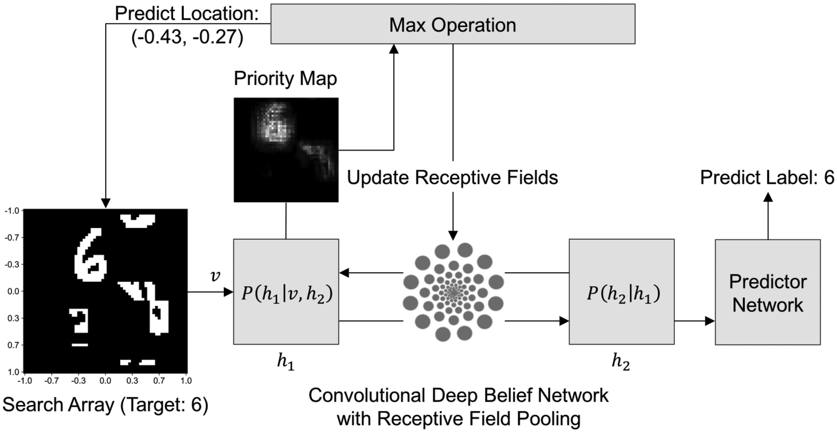

Figure 1: Model of covert attention during visual search using top-down priors and spatial resampling. Visual search for handwritten digits (approximately 20 × 20 pixels) among nondigit distractors within a 60 × 60 search array is evaluated for target location (Predict Location) and classification (Predict Label) accuracy. During training, priors (comprising 40 feature maps) are learned over digit features in the first layer (h_1, 24 feature maps) and combined with bottom-up signals (v) to generate the priority map ($P(h_1|v, h_2)$) during evaluation. The location with the maximum value in the priority map is defined as the predicted target location and is then used to spatially resample (see equation 2.5) the receptive field pooling array to enhance the spatial resolution of feature representations at the predicted target location to facilitate downstream classification via a two-layer convolutional network.

signals that contain target and distractor features, simulating feature-based attention that highlights 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 nondigit distractors and evaluate both target location and classification accuracy. Although we focus on visual search with artificial images, we also discuss how the model could be extended to more complex tasks with natural images.

2 Method

2.1 Overview. In the following sections, we describe our model of covert attention during visual search (shown in Figure 1). This model comprises both top-down feature-based and spatial attention using biologically plausible mechanisms. Feature-based attention is implemented with a hierarchical generative model that acquires top-down priors over target features during training via a Hebbian learning rule. Through feedback of top-down priors, attention priority maps are updated to represent likely target locations. Spatial attention within the model is then implemented as an enhancement of the spatial resolution of sampled features at the predicted target location, reflecting known effects of attention on cortical receptive fields of populations of neurons.

We evaluate this model using a visual search task for handwritten digits among nondigit distractors. Although the model is general and not specific to a particular visual cortical area, we employ a task in which ambiguity is expected between target and distractor features in the first-layer representations and therefore requires top-down feedback to accurately predict the target. By using scrambled digits as distractors, we ensure that low-level features within small receptive fields will be similar, while higher-level features in larger receptive fields will be distinct (see Figure 2).

2.2 Model Description

2.2.1 Hierarchical Generative Model. We constructed a model that is conceptually related to previous hierarchical Bayesian models (Rao, 2005; Torralba et al., 2006; Chikkerur et al., 2010), but unlike these models, it learns priors over target features using a biologically plausible updating rule that is based on Hebb's theory of synaptic plasticity (Hebb, 1949). To implement this model, we use a simple two-layer convolutional deep belief network (CDBN) (Figure 1) (Lee et al., 2009) that is a hierarchical generative model composed of multiple restricted Boltzmann machine (RBM) layers (Smolensky, 1986). Each RBM layer models its input using a set of hidden units that are active with the following probabilities:

$$P(h_j = 1|\mathbf{v}) = \sigma\left(b_j + \sum_i v_i w_{ij}\right), \tag{2.1}$$

$$P(v_i = 1|\mathbf{h}) = \sigma\left(c_i + \sum_j h_j w_{ij}\right), \tag{2.2}$$

where h_j represents a single hidden unit from the set $\mathbf{h}$, 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. During training, the model weights are updated with new values that are proportional to the simultaneous activation of the visible and hidden units, using an algorithm known as contrastive divergence (CD; Hinton, 2002),

$$\Delta w_{ij} \propto \langle v_i h_j \rangle_{\text{data}} - \langle v_i h_j \rangle_{\text{model}}, \tag{2.3}$$

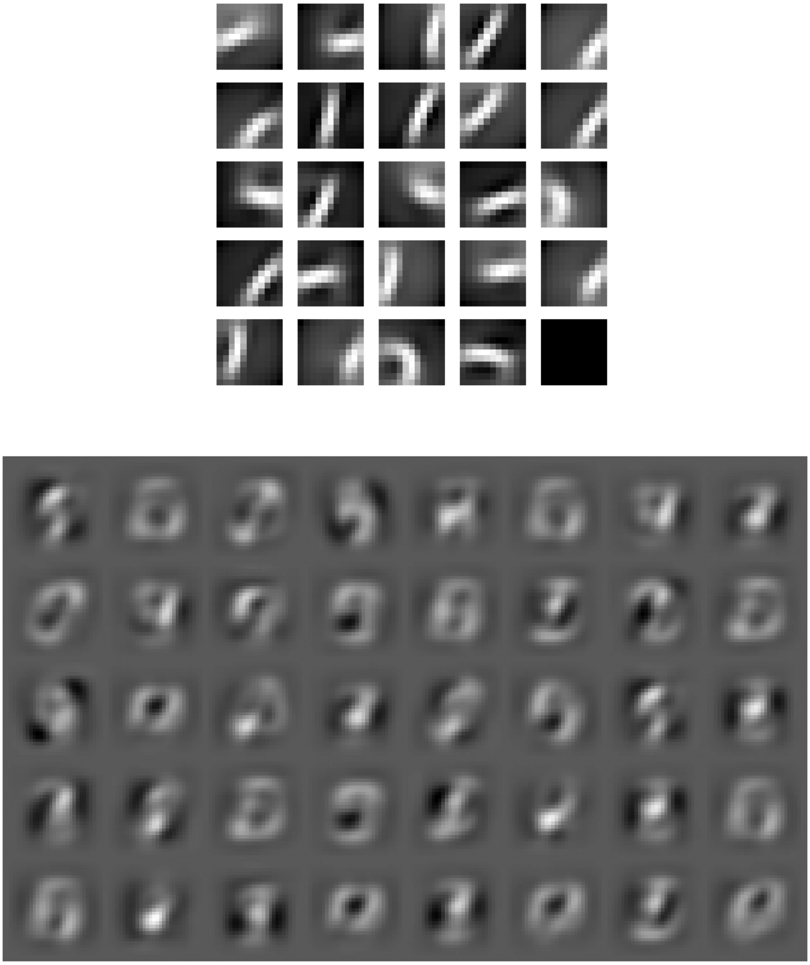

Figure 2: Top: Weights in the first layer of the CDBN (h_1 in Figure 1), following unsupervised learning with handwritten digits. Bottom: Whereas first-layer weights comprise oriented and curved lines that are characteristic of parts of digits, second-layer weights (shown projected into image space) are combinations of the first-layer weights that can resemble entire digits. Therefore, first-layer responses do not reliably disambiguate targets and distractors.

where the expected values from the data distribution are compared to those generated by the model. Choosing an update rule based on contrastive divergence means that visible and hidden units that are simultaneously active are given stronger weights, much like Hebbian learning for synaptic connections.

Once trained, the model can generate samples from the data distribution using Gibbs sampling (Geman & Geman, 1984), a sampling method in which visible and hidden units are alternately sampled (Bernoulli sampling using equations 2.1 and 2.2). After 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 constructing a deep belief network (DBN; Hinton & Salakhutdinov, 2006). Since the first layer learns to represent the probability of the data distribution, subsequent training of a higher layer induces a prior over the first layer, $P(h_1|h_2)$ (i.e., the likely relationships among first-layer features).

In the context of digits, these priors represent the likely combinations of low-level features (e.g., oriented line segments) that form the common components of digits. The second-layer hidden units therefore represent different parts of handwritten digits such that in the top-down direction, full digits could be generated by combining first-layer features.

Within the CDBN, the convolution operation results in a set of feature maps. Within each of these feature maps, each hidden unit has a receptive field covering a specific portion of the input space. We designed the model to represent digits using two layers, comprising smaller RFs (11 × 11) corresponding to digit fragments (layer 1) and larger RFs (28 × 28) encompassing entire digits (layer 2).

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 2, top), while the second layer represents more complex features that resemble entire digits (Figure 2, bottom). Since our goal is to induce ambiguity among low-level target and distractor features, we used a relatively small number of weights for each layer (24 and 40, respectively) and a sparsity constraint during training to prevent overfitting, thereby reducing the likelihood of first-layer feature responses being biased toward the target versus the distractors.

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 reduce the image size of its hidden units (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 deriving the posterior probability of hidden units, given the input and the top-down feedback from the layer above:

$$P(h_j^k = 1|\mathbf{v}, \mathbf{h}') = \frac{\exp(I(h_j^k) + I(p_\alpha^k))}{1 + \sum_{B_\alpha} \exp(I(h_j^k) + I(p_\alpha^k))} \tag{2.4}$$

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

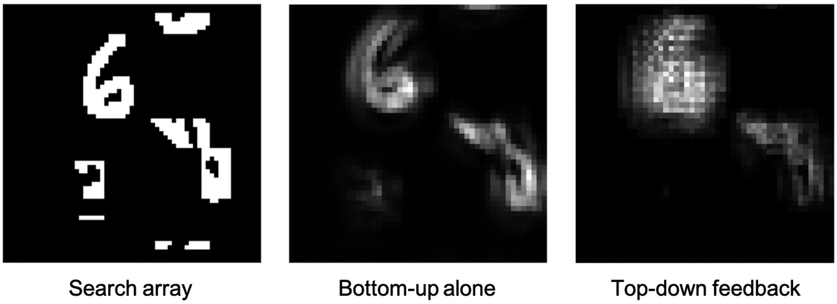

Figure 3: Example search array and priority maps for simulated visual search. The search array (left) contains a single target (here, "6") among nondigit distractors. Comparing the bottom-up alone (middle) and top-down feedback (right) priority maps demonstrates how top-down priors disambiguate first-layer target and distractor features by assigning greater priority to the target relative to distractors.

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).

2.2.2 Attention Priority Maps. Using equations 2.1 and 2.4, 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 probabilities of those features, given top-down priors (i.e., $P(\mathbf{h_1}|\mathbf{v}, \mathbf{h_2})$; "top-down feedback"). Since the probabilities across first-layer features correspond to a vector for each pixel location, we normalized the sum across probabilities to the maximum value to generate priority maps for both bottom-up alone and top-down feedback conditions. Allocation of spatial attention to the predicted target location during visual search was modeled as the location with the maximum value in the priority map (see Figure 1).

Visual search for digits should be facilitated by priority maps that represent relative probabilities of digit features that are present in an image. However, since the first-layer features comprise digit fragments (see Figure 2), nondigit distractors will be confused with digits when using only a priority map of bottom-up conditional probabilities. Therefore, use of the priority map that incorporates top-down priors across first-layer features should help disambiguate which regions of the image contain a digit versus nondigit distractor. This is demonstrated in Figure 3, where the relative priorities of the target digit and right-most distractor are similar in the bottom-up–alone priority map (middle panel) but are more distinct in the priority map incorporating top-down priors (right panel), highlighting the digit over the distractor.

2.2.3 Receptive Field Pooling Array. When attention is covertly directed to a particular peripheral region of the visual field (without changing gaze position), RFs in visual cortex shrink and shift toward the attended location (Klein et al., 2014; Theiss et al., 2022; Womelsdorf et al., 2006), resulting in local enhancement of the spatial resolution of feature representations at the attended location. In order to model cortical RFs during visual search (but not during training), we replace the pooling blocks described in equation 2.4 with receptive fields of variable size that can be dynamically updated. This dynamic receptive field pooling array maintains a location and size for each individual receptive field (see Theiss et al., 2022, for more details). The locations and distributions of RFs in the pooling array are selected to mimic gaze that is fixed at the center of the image, with greater density and smaller RFs at the center of the search array (see Figure 1). In order to simulate the eccentricity dependence of visual cortical RF properties, a scaling rate is used to define the size and spacing of RFs as a function of eccentricity (Theiss et al., 2022). For the current study, this scaling rate was set to 0.1, based on the properties of voxel RFs estimated from fMRI measurements of human visual cortical area V1 (Kay et al., 2013).

Before implementing probabilistic max-pooling, we first obtain the hidden unit outputs ($I(h_j^k)$ in equation 2.4) for each RF using a masked array (receptive fields × height × width), with values of 1 for pixels within the RF and 0 for pixels outside the RF. A two-dimensional gaussian attention field centered at the predicted target location is then multiplied with the RF array to update the location and size of each RF:

$$\mu = \frac{\mu_{RF}\sigma_{AF}^2 + \mu_{AF}\sigma_{RF}^2}{\sigma_{AF}^2 + \sigma_{RF}^2}, \quad \sigma^2 = \frac{\sigma_{RF}^2\sigma_{AF}^2}{\sigma_{RF}^2 + \sigma_{AF}^2}, \tag{2.5}$$

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} is set to the predicted target location for each trial, and σ_{AF} is set to eight pixels in the feature map space, with $2\sigma_{AF}$ approximating the size of a MNIST digit in image space.

2.2.4 Predictor Network. In order to evaluate the effects of spatial attention on the first-layer feature representations, we train a predictor network using the extracted features from the second layer of the CDBN. As shown in Figure 1, the predictor network classifies the digit after the RF pooling array in the first layer is updated via spatial attention that is centered at the predicted target location (see equation 2.5). Unlike the CDBN, the predictor network is a strictly feedforward neural network with two convolutional layers (ReLU and softmax activation functions, respectively, as shown in Table 1). The convolutional filter sizes are chosen such that the output for the 60 × 60 search array is a 10-dimensional vector, corresponding to the

Table 1: Model Architecture Used for Training.

Network	Input Shape	Output Shape	Convolution	Activation	Pool
CDBN	1 × 60 × 60	24 × 50 × 50	11 × 11	None	ProbMax 2 × 2
	24 × 25 × 25	40 × 17 × 17	9 × 9	Sigmoid	None
Predictor	40 × 17 × 17	64 × 10 × 10	8 × 8	ReLU	Max 2 × 2
network	64 × 5 × 5	10 × 1 × 1	5 × 5	Softmax	None

softmax values for each digit class (i.e., 0–9). Predictions of the target class label are based on the index of the 10-dimensional vector containing the maximum value.

2.2.5 Model Training. The complete model shown in Figure 1 and Table 1 is trained in two steps. First, the CDBN is trained layer-wise with unsupervised learning to model the data distribution using the contrastive divergence algorithm described above (Hinton, 2002). For this portion of training, the inputs to the model are 28 × 28 pixel images of a handwritten digit (LeCun et al., 1998). Each layer is 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 (increasing to 0.9 after four epochs). The learning rate is decayed after each epoch using a time-based schedule and a decay rate of 0.01, as described in Lee et al. (2009). In order to reduce overfitting and encourage sparsely active hidden units, $L2$ weight-decay and sparsity constraints are used during training (Hinton, 2012).

Next, the predictor network is trained for 10 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 is first padded on each side with zeros to match the size of the search array (60 × 60 pixels). The digit is then randomly translated horizontally and vertically up to a maximum of 15 pixels in each direction (25% of the search array size).

The image is then passed through the CDBN, and the predictor network is 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 additional experiments, the 2 × 2 pooling operation in the first layer of the CDBN (see Table 1) was replaced with the RF pooling array, which performs probabilistic max-pooling across each RF instead of on 2 × 2 blocks of pixels.

2.2.6 Model Overview. In summary, the model shown in Figure 1 contains three main components: a Bayesian attention model (CDBN), an RF pooling layer, and a predictor network. During visual search, features are first extracted by the CDBN; then the target location is predicted from a priority map in order to update the RF pooling array, and finally, the enhanced features are used to classify the target digit. Following Theiss et al. (2022), we multiply a two-dimensional 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 consider the gaussian attention field to be a spatial prior over the predicted target location that is derived from the priority map.

Two priority maps are evaluated for target location and classification accuracy. The bottom-up priority map (see the middle panel of Figure 3) represents the conditional probability of first-layer features, given the visual search array. The priority map with top-down feedback (see the right panel of Figure 3) incorporates second-layer top-down priors over first-layer features to help disambiguate features that could represent both targets and distractors. These priors reflect the probability distribution over combinations of curved and oriented features (see Figure 2) that constitute handwritten digits. By comparing the differences in performance between these two priority maps for both location and classification accuracy, we quantify the effects of top-down feature priors on visual search performance.

2.3 Experimental Design and Statistical Analyses.

2.3.1 Visual Search Experiment. The visual search experiment for a digit among non-digit distractors contains 10,000 search arrays using the held-out MNIST test set (LeCun et al., 1998). Each 60×60 pixel search array contains a single 28×28 target MNIST digit placed in a random location among various distractors (see the left panel of Figure 3). In order to increase ambiguity in first-layer feature representations, the distractors are generated from fragments of digits (described in detail below). As shown in the middle panel of Figure 3, this increases the uncertainty of target location in the bottom-up alone priority map. For each trial and condition (bottom-up alone versus top-down feedback), the location of the maximum value in the respective priority map is selected as the predicted target location, and this is then used to update the RF pooling array using equation 2.5, with μ_{AF} set to the predicted target location. Following this update, the predictor network classifies the target digit using the second-layer features, separately for the two priority map conditions.

2.3.2 Nondigit Distractors. In order to generate distractors that contain similar first-level features as target digits, we manipulate 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 crop the 28×28 MNIST image to the central 14×14 pixels, randomly rotate the cropped image by one of $[0, 90, 180, 270]$ degrees, and randomly zero half of the resulting image along either the horizontal or vertical axis. As shown in the left panel of Figure 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 in the search array, each distractor is randomly placed such that the center-to-center distance to the target digit was greater than 7.5 pixels (12.5% of the search array size) along both horizontal and vertical axes. Note that this allows distractors to overlap with parts of the target digit as well as with other distractors. However, as a result of the relative size of the targets and distractors with respect to the size of the search array, there is an inherent limitation regarding the number of distractors that can be added to the array. This unfortunately prevents set size analysis in our study, which is often included in psychophysical visual search studies in humans.

2.3.3 Statistical Procedures. Target location accuracy is evaluated by computing precision and recall (defined below) 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 we describe has previously been used to evaluate saliency model performance for locations of fixations (Wang et al., 2016) as well as for visual saliency detection (Xie & Lu, 2011), which is relevant to our study.

For each threshold value between 0 and 1, pixels in the priority map with values above the threshold are considered positive (i.e., target) predictions, whereas those below the threshold are considered negative (i.e., background) predictions. Above-threshold pixels are considered true positives if they overlap a 16 × 16 block of pixels centered at the target location in the priority map (approximately the size of the MNIST digit in image space).

Precision is defined as the proportion of above-threshold pixels overlapping the target relative to all above-threshold pixels (quantifying the relative priority of target versus distractor locations). Recall is defined as the proportion of above-threshold pixels overlapping the target relative to the 16 × 16 pixel target area (quantifying the sensitivity for detecting the target within the priority map). Average precision (AP) is then computed for each trial using the following equation,

$$AP = \sum_{n} (R_n - R_{n-1}) P_n, \tag{2.6}$$

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 average across trials to calculate the AP for a given condition (bottom-up alone versus top-down feedback). Chance level for precision is 0.1024 (i.e., the proportion of ground-truth target pixels).

Target classification accuracy is evaluated as the proportion of trials correctly classifying the target digit. We also evaluate a control condition of classification accuracy for each trial without updating the RF pooling array (no changes in RF location or size due to covert visual spatial attention).

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

2.3.4 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 article is available on request.

3 Results

3.1 Top Down Feature Priors Improve Target Location Accuracy. To evaluate the effects of top-down priors on visual search performance, we tested both target location and classification accuracy for the bottom-up alone (i.e., $P(\mathbf{h_1}|\mathbf{v})$) and the top-down feedback (i.e., $P(\mathbf{h_1}|\mathbf{v}, \mathbf{h_2})$) priority maps. To quantify target location accuracy, we plotted the precision-recall curve (see Figure 4), which displays performance for each of the two priority maps, relative to a random baseline (defined below). Greater area under the curve (AUC) indicates better performance for predicting target location.

As described in section 2.3.3 above, the precision-recall curve is computed by thresholding the priority map between 0 and 1, where a high threshold preserves only the largest values in the priority map. Therefore, high precision at low recall (as seen for the top-down feedback priority map in Figure 4) indicates that the locations with the largest values in the priority map are more likely to overlap with the target digit than with distractors.

In contrast, the precision-recall curve for the bottom-up alone priority map indicates that as fewer above-threshold pixels overlapped with the target (low recall), the proportion of pixels that overlapped with 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 surpass baseline performance.

Average precision (see equation 2.6) is a summary metric of the precision-recall curve that is equivalent to the area under the curve. High

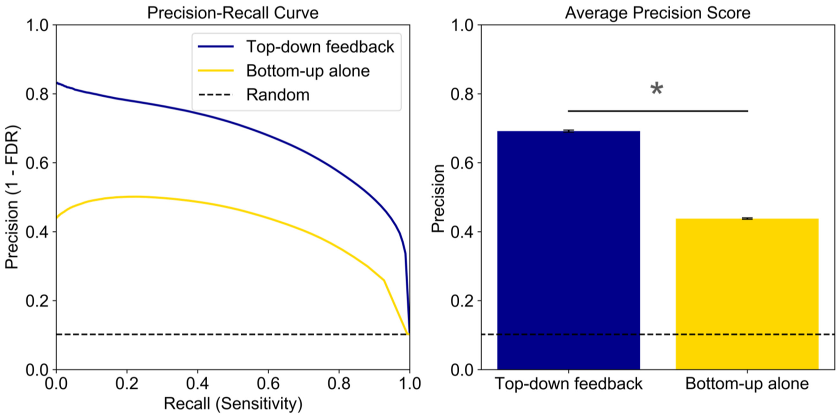

Figure 4: Precision-recall curve (left) and average precision (right) quantify target location accuracy for bottom-up alone and top-down feedback priority maps. The precision-recall curve is based on binarizing the priority maps at various thresholds and reflects the relative proportion of pixels assigned to the target versus the distractors (precision) as a function of overlap with the target (recall) in each thresholded priority map. Average precision (see equation 2.6) is a summary metric of the precision-recall curve that is equivalent to the area under the curve. Together, these results demonstrate that top-down priors disambiguate bottom-up signals by highlighting regions associated with the target relative to the 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.

average precision therefore indicates greater priority for target versus distractor locations across all recall levels. As shown in the right panel of Figure 4, incorporating top-down priors in the priority map substantially improves target location accuracy relative to using the bottom-up conditional probabilities alone (0.69 versus 0.44).

3.2 Top Down Priors Disambiguate among Target and Distractor Features. In order to evaluate the effect of spatial attention on encoded features at the predicted target location, we use a predictor network to classify the target digit based on the updated features from the second layer of the CDBN (see Figure 1). As shown in the middle panel of Figure 3, the bottom-up alone priority map can 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 toward the distractor instead of the target. However, if top-down priors disambiguate target and distractor

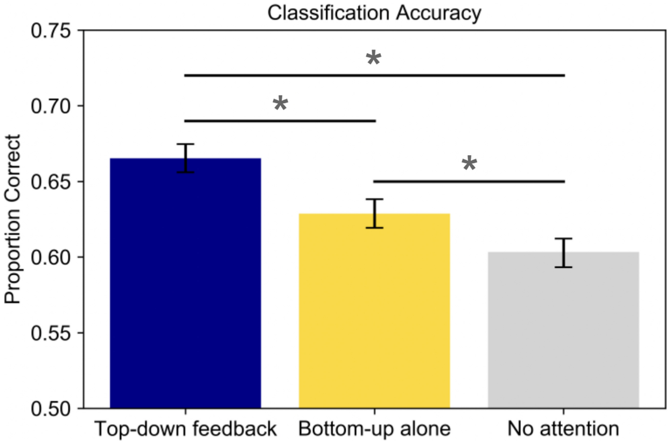

Figure 5: Classification accuracy during simulated visual search using spatial resampling of RFs at locations predicted by bottom-up alone or top-down feedback priority maps. “No attention” indicates classification accuracy without updating the RF array based on attention. Spatial attention to predicted locations improves classification accuracy, with top-down feedback priority maps resulting in greater performance relative to bottom-up alone. Chance performance is 0.1 (1 out of 10 possible digit classes). Error bars are bootstrapped 95% confidence intervals. *Bootstrapped p-value (1000 samples) = 0.

features, the resulting priority map would be more likely to direct spatial attention to the target location (see the right panel of Figure 3).

As for our analysis of target location accuracy, we compare 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 1 out of 10 digit classes). However, since the predictor network is trained to classify digits anywhere in the image and distractors should have poor representation in the output of the CDBN, the RF pooling array is the primary bottleneck for digit information passing to the predictor network. Indeed, as shown in Figure 5, performance is well above chance (0.1 proportion correct) for all three conditions. It is therefore more informative to compare the effects of spatial attention on classification accuracy relative to a “no-attention” condition. For this condition, we obtain target classification predictions for each trial without updating the RF pooling array based on spatial attention.

Although performance is better in both attention conditions compared to the no-attention condition (0.67 and 0.63 versus 0.60), classification accuracy is greater for the priority map with top-down priors that enhance the encoded feature representations (0.67 versus 0.63). To provide further intuition for this result, we note that digits placed in the center of the search array should have equal performance in all three conditions, whereas spatial attention will influence performance for digits placed at the periphery. Attention to target digits or even nearby distractors in the bottom-up alone condition could lead to greater spatial resolution and therefore improved classification compared to the no-attention condition. More accurate selection of the target digit by attention results in better classification in the top-down feedback condition compared to the bottom-up alone condition (see Figure 4).

Finally, it is worth noting that performance of the predictor network on the test set is relatively low (0.81) at full spatial resolution (i.e., no RF pooling array) in the absence of any distractors. Spatial attention using the top-down feedback priority map in the presence of distractors achieves 82% of this maximal classification accuracy.

4 Discussion

In this study, we present a simple Bayesian model of covert visual attention and evaluate the model using a visual search task with handwritten digits among nondigit distractors. In contrast to previous models, our model learns priors over target features using an update rule that is similar to Hebbian learning, and it enhances representations of features at attended spatial locations 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 improve target location accuracy. Furthermore, by modeling spatial attention as an interaction of an RF pooling array with an attention field centered at the predicted target location (Theiss et al., 2022), we demonstrate that top-down priors disambiguate distractor and target features, resulting in the target being more likely to be attended. The study provides further support for a Bayesian conceptual framework to explain covert visual search.

The enhancing effects of covert spatial attention on local spatial resolution of visual stimuli has been well known for over two decades (Yeshurun & Carrasco, 1998; Carrasco et al., 2000). Although other models have implemented eccentricity-dependent processing to emulate the foveated sampling observed in the retina and visual cortex (Zelinsky, 2008; Akbas & Eckstein, 2017), the dynamic RF pooling array in our model implements known changes in populations of neuronal receptive fields due to spatial attention (Womelsdorf et al., 2006; Klein et al., 2014). These changes in RF position and size have been observed across the visual cortical hierarchy (Klein et al., 2014), which suggests that there is not a single biological substrate that underlies the attentional field that is centered at the attended location. It further suggests that our proposed model is not dependent on a particular visual cortical area and can be extended to represent various areas across the visual processing hierarchy.

Whereas other previous Bayesian models have altered top-down priors to demonstrate attentional effects (Rao, 2005; Torralba et al., 2006; Chikkerur et al., 2010) or accumulated priors through Bayesian updates trial-by-trial (Droll et al., 2009; Itti & Baldi, 2009), our model acquires priors through pretask experience (i.e., training) and then uses top-down priors to disambiguate feature representations and deploy spatial attention to a selected location. This approach shares some similarities with the model described in Chalk et al. (2013), which also takes inspiration from the normalization model of attention (Reynolds & Heeger, 2009). However, in Chalk et al. (2013), the task is solely detection of simple binary stimuli, while we have characterized the effects of attention on classification in our model. More important, whereas their model is trained to optimize task rewards, ours acquires top-down priors through a biologically plausible update rule that is similar to Hebbian learning.

Several Bayesian models of overt attention during free viewing and visual search of natural images have been described over the past two decades (Torralba et al., 2006; Itti & Baldi, 2009; Chikkerur et al., 2010), while others similar to Chalk et al. (2013) have studied covert attention with simple artificial stimuli (see Vincent, 2015; Eckstein, 2017, for reviews). A common assumption in studies of overt attention is that highly overlapping networks underlie both eye movements and covert attention, as posited by the premotor theory of attention (Rizzolatti et al., 1987). However, recent studies have investigated the degree to which humans dissociate overt and covert attention, finding that doing so improved performance for change detection (Chetverikov et al., 2018) but not visual search (MacInnes et al., 2020). Interestingly, participants actively tried to uncouple overt and covert attention during the visual search task even though it hindered performance.

It is important for the field to develop models of visual search that can account for both overt and covert attention, including the ability to model dissociated overt and covert attention. By using a foveated RF pooling array, our model treats covert attention as a spatial resampling of RFs (Theiss et al., 2022) and overt attention as a translation of the pooling array that directs the "fovea" to a different part of the image (Cheung et al., 2016; Larochelle & Hinton, 2010). Instead of always maintaining central fixation, as we have done in this study, our model can be extended to systematically evaluate the contributions of covert and overt attention to task-dependent performance in psychophysical experiments with human subjects.

One component of many visual search models that was not addressed in our study is bottom-up saliency. In most computational models of visual attention, saliency is defined as a contrast of local features (Itti et al., 1998) or, in the Bayesian framework, as the difference between prior and posterior distributions (Itti & Baldi, 2009). Torralba et al. (2006) specifically fit hyperparameters to the combination of bottom-up and top-down attention to optimize eye movement predictions. However, for this study, it is unclear how bottom-up attention should be weighted relative to top-down priors.

Indeed, how bottom-up and top-down attention are weighted across tasks is generally an open question. For example, Chikkerur et al. (2010) assumed uniform priors to model eye movements during free viewing. However, since it is unlikely that all priors (e.g., a light-from-above prior; Stone et al., 2009) would be uniform during free viewing, the combination of bottom-up and top-down attention is likely dynamic and task-dependent.

Other recent studies have used more complex neural network models in order to study covert attention in tasks with natural stimuli. These methods take advantage of the fact that features learned in deep neural networks approximate those represented in visual cortex (Yamins et al., 2014). Nicholson and Prinz (2022) used pretrained neural networks to study whether set size effects observed in human psychophysical experiments could be explained by a mismatch between the feature statistics of the simplistic images used in common visual search tasks and those of natural images that the visual system has been optimized to perceive through evolution and experience. The authors indeed observed set size effects typical of human visual search when using a model that had been trained using a different domain from the one used in the task (e.g., natural versus synthetic images), but not when the model was trained on the task images. However, this difference was also observed for two different data sets of natural images, which suggests that other factors besides feature statistics could contribute to these differences. Our model, though tested on simplistic search arrays, is informed by known biological mechanisms, and the generality of our conclusions extends beyond the feature statistics of handwritten digits.

In another study, Lindsay and Miller (2018) used deep neural networks to study the feature similarity gain model of attention by enhancing network activity at different layers within a classification model. The authors demonstrated that applying attention did improve classification performance, although the effects were smaller at earlier model layers compared to later layers. Interestingly, the authors provided an alternative type of attentional modulation by using the gradients of the network's prediction error. In doing so, the authors observed similar effects as with the tuning approach in later layers but greater effects in earlier layers. This is not surprising from a machine learning perspective, but it does provide a novel view of how attention might be implemented across the visual cortex by minimizing some objective function related to the task at hand. In the context of our study, new theoretical directions and predictions could arise from exploring other classes of generative models commonly used in machine learning literature, even if they are not biologically plausible (Goodfellow et al., 2014; Dinh et al., 2014; Sohl-Dickstein et al., 2015; Devlin et al., 2018; Radford et al., 2019).

Recent developments in both machine learning and fMRI research have provided insights into the relationship between features learned in convolutional neural networks and patterns of activity in visual cortex (Devereux et al., 2018; O'Connell & Chun, 2018; St.-Yves & Naselaris, 2018). Combined with the known effects of attention on RFs (Womelsdorf et al., 2006; Klein et al., 2014), as well as the distributed nature of attention across many cortical areas (Serences & Yantis, 2007; Melloni et al., 2012; Sprague & Serences, 2013; Bressler et al., 2020), the model we have proposed is well suited to further investigate the Bayesian brain hypothesis with neuroimaging data and natural images.

When extended to multiple levels of feature complexity, our model of Bayesian priors is dynamic and local. This allows investigations of the dynamics of top-down attention by updating priors across trials and at multiple levels of the visual processing hierarchy. Under free-viewing conditions with natural images (i.e., without an explicit task), we predict that the respective influences of spatial and feature priors on gaze behavior should be reflected at the level of feature encoding that best accounts for the statistical regularities across similar scenes (Yang et al., 2016). However, during visual search, we predict that task-based attention will act as a hyperprior, giving stronger weight to task-relevant priors across the visual processing hierarchy.

Although the model described in this study is relatively simple and was 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 sequentially. Furthermore, the definition of the restricted Boltzmann machine as a bipartite graph precludes the addition of more complex intralayer connections (i.e., connections between hidden units are "restricted"). However, we can instead train a single RBM at multiple layers of a pretrained neural network in parallel, with each RBM learning priors over local features and spatial locations. Not only does this substantially reduce the training time, but it also allows for studying relative effects of priors at different levels of feature complexity and spatial scale.

Acknowledgments

This research was supported by National Eye Institute grant R01 EY025278, National Eye Institute Training Grant T32 EY007043, and National Eye Institute Core Grant P30 EY003176.

References

Akbas, E., & Eckstein, M. P. (2017). Object detection through search with a foveated visual system. *PLOS Computational Biology*, *13*(10), e1005743. 10.1371/journal.pcbi.1005743

Banks, W. P., Bachrach, K. M., & Larson, D. W. (1977). The asymmetry of lateral interference in visual letter identification. *Perception and Psychophysics*, *22*(3), 232–240. 10.3758/BF03199684

Bouma, H. (1970). Interaction effects in parafoveal letter recognition. *Nature*, *226*(5241), 177–178. 10.1038/226177a0

Bravo, M. J., & Nakayama, K. (1992). The role of attention in different visual-search tasks. *Perception and Psychophysics*, *51*(5), 465–472. 10.3758/BF03211642

Bressler, D. W., Rokem, A., & Silver, M. A. (2020). Slow endogenous fluctuations in cortical fMRI signals correlate with reduced performance in a visual detection task and are suppressed by spatial attention. *Journal of Cognitive Neuroscience*, *32*(1), 85–99. 10.1162/jocn_a_01470

Carrasco, M., Penpeci-Talgar, C., & Eckstein, M. (2000). Spatial covert attention increases contrast sensitivity across the CSF: Support for signal enhancement. *Vision Research*, *40*(10–12), 1203–1215. 10.1016/S0042-6989(00)00024-9

Chalk, M., Murray, I., & Seriès, P. (2013). Attention as reward-driven optimization of sensory processing. *Neural Computation*, *25*(11), 2904–2933. 10.1162/NECO_a_00494

Chetverikov, A., Kuvaldina, M., MacInnes, W. J., Jóhannesson, Ó. I., & Kristjánsson, Á. (2018). Implicit processing during change blindness revealed with mouse-contingent and gaze-contingent displays. *Attention, Perception, and Psychophysics*, *80*(4), 844–859. 10.3758/s13414-017-1468-5

Cheung, B., Weiss, E., & Olshausen, B. (2016). *Emergence of foveal image sampling from learning to attend in visual scenes.* arXiv:1611.09430.

Chikkerur, S., Serre, T., Tan, C., & Poggio, T. (2010). What and where: A Bayesian inference theory of attention. *Vision Research*, *50*(22), 2233–2247. 10.1016/j.visres.2010.05.013

Dayan, P., Hinton, G. E., Neal, R. M., & Zemel, R. S. (1995). The Helmholtz machine. *Neural Computation*, *7*(5), 889–904. 10.1162/neco.1995.7.5.889

Dayan, P., & Zemel, R. S. (1999). Statistical models and sensory attention. In *Proceedings of the 1999 Ninth International Conference on Artificial Neural Networks* (vol. 2, pp. 1017–1022).

Devereux, B. J., Clarke, A., & Tyler, L. K. (2018). Integrated deep visual and semantic attractor neural networks predict fMRI pattern-information along the ventral object processing pathway. *Scientific Reports*, *8*(1), 1–12. 10.1038/s41598-018-28865-1

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). *BERT: Pre-training of deep bidirectional transformers for language understanding*. arXiv:1810.04805.

Dinh, L., Krueger, D., & Bengio, Y. (2014). *NICE: Non-linear independent components estimation*. arXiv:1410.8516.

Droll, J. A., Abbey, C. K., & Eckstein, M. P. (2009). Learning cue validity through performance feedback. *Journal of Vision*, *9*(2):18, 1–22. 10.1167/9.2.18

Eckstein, M. P. (2017). Probabilistic computations for attention, eye movements, and search. *Annual Review of Vision Science*, *3*, 319–342. 10.1146/annurev-vision-102016-061220

Eckstein, M. P., Shimozaki, S. S., & Abbey, C. K. (2002). The footprints of visual attention in the Posner cueing paradigm revealed by classification images. *Journal of Vision*, *2*(1), 25–45. 10.1167/2.1.3

Efron, B., & Tibshirani, R. J. (1994). *An introduction to the bootstrap*. CRC Press.

Ernst, M. O., & Banks, M. S. (2002). Humans integrate visual and haptic information in a statistically optimal fashion. *Nature*, *415* (6870), 429–433. 10.1038/415429a

Freeman, J., & Simoncelli, E. P. (2011). Metamers of the ventral stream. *Nature Neuroscience*, *14*(9), 1195–1201. 10.1038/nn.2889

Geisler, W. S. (2011). Contributions of ideal observer theory to vision research. *Vision Research*, *51*(7), 771–781. 10.1016/j.visres.2010.09.027

Geman, S., & Geman, D. (1984). Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *6*, 721–741. 10.1109/TPAMI.1984.4767596

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., . . . Bengio, Y. (2014). Generative adversarial nets. In Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, & K. Q. Weinberger (Eds.), *Advances in neural information processing systems*, *27*. Curran.

Hebb, D. O. (1949). *The organisation of behaviour: A neuropsychological theory*. Science Editions.

Hinton, G. E. (2002). Training products of experts by minimizing contrastive divergence. *Neural Computation*, *14*(8), 1771–1800. 10.1162/089976602760128018

Hinton, G. E. (2012). A practical guide to training restricted Boltzmann machines. In G. Montavon, G. Orr, & K.-R. Müller (Eds.), *Neural networks: Tricks of the trade* (pp. 599–619). Springer. 10.1007/978-3-642-35289-8_32

Hinton, G. E., & Salakhutdinov, R. R. (2006). Reducing the dimensionality of data with neural networks. *Science*, *313*(5786), 504–507. 10.1126/science.1127647

Itti, L., & Baldi, P. (2009). Bayesian surprise attracts human attention. *Vision Research*, *49*(10), 1295–1306. 10.1016/j.visres.2008.09.007

Itti, L., Koch, C., & Niebur, E. (1998). A model of saliency-based visual attention for rapid scene analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *20*(11), 1254–1259. 10.1109/34.730558

Kay, K. N., Winawer, J., Mezer, A., & Wandell, B. A. (2013). Compressive spatial summation in human visual cortex. *Journal of Neurophysiology*, *110*(2), 481–494. 10.1152/jn.00105.2013

Klein, B. P., Harvey, B. M., & Dumoulin, S. O. (2014). Attraction of position preference by spatial attention throughout human visual cortex. *Neuron*, *84*(1), 227–237. 10.1016/j.neuron.2014.08.047

Knill, D. C., & Pouget, A. (2004). The Bayesian brain: The role of uncertainty in neural coding and computation. *Trends in Neurosciences*, *27*(12), 712–719. 10.1016/j.tins.2004.10.007

Larochelle, H., & Hinton, G. E. (2010). Learning to combine foveal glimpses with a third-order Boltzmann machine. In J. D. Lafferty, C. K. I. Williams, J. Shawe-Taylor, R. S. Zemel, & A. Culotta (Eds.), *Advances in neural information processing systems*, *23*. Curran.

LeCun, Y., Bottou, L., Bengio, Y., & Haffner, P. (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, *86*(11), 2278–2324. 10.1109/5.726791

Lee, H., Grosse, R., Ranganath, R., & Ng, A. Y. (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, T. S., & Mumford, D. (2003). Hierarchical Bayesian inference in the visual cortex. *JOSA A*, *20*(7), 1434–1448. 10.1364/JOSAA.20.001434

Lindsay, G. W., & Miller, K. D. (2018). How biological attention mechanisms improve task performance in a large-scale visual system model. *eLife*, *7*, e38105. 10.7554/eLife.38105

Ma, W. J., Navalpakkam, V., Beck, J. M., van den Berg, R., & Pouget, A. (2011). Behavior and neural basis of near-optimal visual search. *Nature Neuroscience*, *14*(6), 783–790. 10.1038/nn.2814

MacInnes, W. J., Jóhannesson, Ó. I., Chetverikov, A., & Kristjánsson, Á. (2020). No advantage for separating overt and covert attention in visual search. *Vision*, *4*(2), 28. 10.3390/vision4020028

Martínez-Trujillo, J. C., & Treue, S. (2002). Attentional modulation strength in cortical area MT depends on stimulus contrast. *Neuron*, *35*(2), 365–370. 10.1016/S0896-6273(02)00778-X

McAdams, C. J., & Maunsell, J. H. (1999). Effects of attention on orientation-tuning functions of single neurons in macaque cortical area V4. *Journal of Neuroscience*, *19*(1), 431–441. 10.1523/JNEUROSCI.19-01-00431.1999

Melloni, L., van Leeuwen, S., Alink, A., & Müller, N. G. (2012). Interaction between bottom-up saliency and top-down control: How saliency maps are created in the human brain. *Cerebral Cortex*, *22*(12), 2943–2952. 10.1093/cercor/bhr384

Nicholson, D. A., & Prinz, A. A. (2022). Could simplified stimuli change how the brain performs visual search tasks? A deep neural network study. *Journal of Vision*, *22*(7):3, 1–22. 10.1167/jov.22.7.3

O'Connell, T. P., & Chun, M. M. (2018). Predicting eye movement patterns from fMRI responses to natural scenes. *Nature Communications*, *9*(1), 1–15.

Paszke, A., Gross, S., Chintala, S., Chanan, G., Yang, E., DeVito, Z., . . . Lerer, A. (2017). Automatic differentiation in PyTorch. In *NIPS Proceedings of the Autodiff Workshop*.

Pelli, D. G. (1985). Uncertainty explains many aspects of visual contrast detection and discrimination. *JOSA A*, *2*(9), 1508–1532. 10.1364/JOSAA.2.001508

Posner, M. I. (1980). Orienting of attention. *Quarterly Journal of Experimental Psychology*, *32*(1), 3–25. 10.1080/00335558008248231

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). Language models are unsupervised multitask learners. *OpenAI blog*, *1*(8), 9.

Rao, R. P. (2005). Bayesian inference and attentional modulation in the visual cortex. *Neuroreport*, *16*(16), 1843–1848. 10.1097/01.wnr.0000183900.92901.fc

Reynolds, J. H., & Heeger, D. J. (2009). The normalization model of attention. *Neuron*, *61*(2), 168–185. 10.1016/j.neuron.2009.01.002

Rizzolatti, G., Riggio, L., Dascola, I., & Umiltá, C. (1987). Reorienting attention across the horizontal and vertical meridians: Evidence in favor of a premotor theory of attention. *Neuropsychologia*, *25*(1A), 31–40. 10.1016/0028-3932(87)90041-8

Serences, J. T., & Yantis, S. (2007). Spatially selective representations of voluntary and stimulus-driven attentional priority in human occipital, parietal, and frontal cortex. *Cerebral Cortex*, *17*(2), 284–293. 10.1093/cercor/bhj146

Shimozaki, S. S., Eckstein, M. P., & Abbey, C. K. (2003). Comparison of two weighted integration models for the cueing task: Linear and likelihood. *Journal of Vision*, *3*(3), 209–229. 10.1167/3.3.3

Smolensky, P. (1986). *Information processing in dynamical systems: Foundations of harmony theory.* Technical report, Colorado University at Boulder.

Sohl-Dickstein, J., Weiss, E., Maheswaranathan, N., & Ganguli, S. (2015). Deep unsupervised learning using nonequilibrium thermodynamics. In *Proceedings of the International Conference on Machine Learning 37*, pp. 2256–2265.

Sprague, T. C., & Serences, J. T. (2013). Attention modulates spatial priority maps in the human occipital, parietal and frontal cortices. *Nature Neuroscience*, *16*(12), 1879–1887. 10.1038/nn.3574

St.-Yves, G., & Naselaris, T. (2018). The feature-weighted receptive field: An interpretable encoding model for complex feature spaces. *NeuroImage*, *180*, 188–202. 10.1016/j.neuroimage.2017.06.035

Stone, J. V., Kerrigan, I. S., & Porrill, J. (2009). Where is the light? Bayesian perceptual priors for lighting direction. *Proceedings of the Royal Society B: Biological Sciences*, *276*(1663), 1797–1804. 10.1098/rspb.2008.1635

Theiss, J. D., Bowen, J. D., & Silver, M. A. (2022). Spatial attention enhances crowded stimulus encoding across modeled receptive fields by increasing redundancy of feature representations. *Neural Computation*, *34*(1), 190–218. 10.1162/neco_a_01447

Toet, A., & Levi, D. M. (1992). The two-dimensional shape of spatial interaction zones in the parafovea. *Vision Research*, *32*(7), 1349–1357. 10.1016/0042-6989(92)90227-A

Torralba, A., Oliva, A., Castelhano, M. S., & Henderson, J. M. (2006). Contextual guidance of eye movements and attention in real-world scenes: The role of global features in object search. *Psychological Review*, *113*(4), 766–786. 10.1037/0033-295X.113.4.766

Treue, S., & Martínez Trujillo, J. C. (1999). Feature-based attention influences motion processing gain in macaque visual cortex. *Nature*, *399*(6736), 575–579. 10.1038/21176

Vincent, B. T. (2015). Bayesian accounts of covert selective attention: A tutorial review. *Attention, Perception, and Psychophysics*, *77*(4), 1013–1032. 10.3758/s13414-014-0830-0

Wang, J., Borji, A., Kuo, C.-C. J., & Itti, L. (2016). Learning a combined model of visual saliency for fixation prediction. *IEEE Transactions on Image Processing*, *25*(4), 1566–1579. 10.1109/TIP.2016.2522380

Weiss, Y., Simoncelli, E. P., & Adelson, E. H. (2002). Motion illusions as optimal percepts. *Nature Neuroscience*, *5*(6), 598–604. 10.1038/nn0602-858

Womelsdorf, T., Anton-Erxleben, K., Pieper, F., & Treue, S. (2006). Dynamic shifts of visual receptive fields in cortical area MT by spatial attention. *Nature Neuroscience*, *9*(9), 1156–1160. 10.1038/nn1748

Xie, Y., & Lu, H. (2011). Visual saliency detection based on Bayesian model. In *Proceedings of the 2011 18th IEEE International Conference on Image Processing* (pp. 645–648). 10.1109/ICIP.2011.6116634

Yamins, D. L., Hong, H., Cadieu, C. F., Solomon, E. A., Seibert, D., & DiCarlo, J. J. (2014). Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proceedings of the National Academy of Sciences*, *111*(23), 8619–8624. 10.1073/pnas.1403112111

Yang, S. C.-H., Lengyel, M., & Wolpert, D. M. (2016). Active sensing in the categorization of visual patterns. *eLife*, *5*, e12215. 10.7554/eLife.12215

Yeshurun, Y., & Carrasco, M. (1998). Attention improves or impairs visual performance by enhancing spatial resolution. *Nature*, *396*(6706), 72–75. 10.1038/23936

Yu, A. J., & Dayan, P. (2005). Uncertainty, neuromodulation, and attention. *Neuron*, *46*(4), 681–692. 10.1016/j.neuron.2005.04.026

Zelinsky, G. J. (2008). A theory of eye movements during target acquisition. *Psychological Review*, *115*(4), 787–835. 10.1037/a0013118

Zhu, M. (2004). *Recall, precision, and average precision*. Technical report, University of Waterloo.

Received December 2, 2023; accepted May 30, 2024.