Neuronal Population Statistics, Computations, and Mechanisms of a Feedforward Convolutional Neural Network that Learns to Covertly Attend

Sudhanshu Srivastava* [1,5], William Yang Wang[3,4,5], Miguel P. Eckstein*[1,2,3,4,5]

[1]Graduate Program in Dynamical Neuroscience

[2]Department of Psychological and Brain Sciences

[3]Department of Computer Science

[4]Department of Electrical and Computer Engineering

[5]Institute for Collaborative Biotechnologies

University of California, Santa Barbara

sudhanshu@ucsb.edu, william@cs.ucsb.edu, miguel.eckstein@psych.ucsb.edu

ABSTRACT

The cueing task is one of the most prominent paradigms to study covert attention. Observers are quicker and more accurate in detecting a target when it appears with a cue (valid trials) than when it appears opposite to the cue (invalid trials). How neuronal populations of cells across the visual hierarchy progressively represent and integrate visual information across the target, cues, and locations to give rise to the behavioral cueing effect is not well understood. To gain a theoretical understanding of the plausible system-wide neuronal population statistics, computations, and mechanisms mediating the cueing effect, we analyze the response properties of 180k neurons per network across layers of ten feedforward Convolutional Neural Networks (CNN). The CNNs are trained from images with luminance noise, without any explicit incorporation of an attention mechanism, and show human-like beneficial influences of cues on detection accuracy. Early visual hierarchy layers show retinotopic neurons separately tuned to target or cue with excitatory or inhibitory responses. Later layers show neurons that are jointly tuned to both target and cue and integrate information across locations. Consistent with physiological findings, we find an increased influence of the cue on target responses in higher layers (areas) in the network and computational stages similar to those of a Bayesian ideal observer, but with more gradual transitions. The cue influences the mean neuronal response to the target and distractor, and changes target sensitivity with two mechanisms: integration of information across locations at the dense layer with neurons more driven by the cued location, and an interaction with the thresholding Rectified Linear Activation Unit (ReLU) in the last convolution layer. We find novel neuronal properties not yet reported in physiological studies: the presence of cue-inhibitory neurons, inhibitory cue influences on target neurons, location-opponent cells, and higher sensitivity to the cue in intermediate layers than later layers. Together, our analyses illustrate a system-wide analysis of the brain computations that might give rise to behavioral cueing effects and provide a theoretical framework to inform neurophysiological studies.

Introduction

Spatial covert visual attention refers to the process of selecting a location in an image or a scene without moving the eyes and can greatly benefit perceptual performance in detection, discrimination, and search. Both behavioral [1,2] and electrophysiological studies [3,4] often operationalize manipulations of covert attention to a location by using cues that co-occur with or point to a likely target location. Subjects tend to be faster and more accurate when the target is at the cued vs. the uncued location. There is vast literature focused on understanding the influences of covert attention (cued locations) using behavioral measures[2,5–7]. Spatial attention can alter appearance[8], contrast sensitivity[9,10], acuity[11], texture segmentation[12], and sensitivity[13] at the cued location compared to the uncued location. Early research focused on descriptive understanding of attention based on analogies such as a zoom or a spotlight[14].

At the level of mechanisms, studies have used psychophysical behavioral results across stimuli conditions (e.g., varying target contrast, image noise) and, by assuming an underlying quantitative model, made inferences about how attention changes the processing of visual information. These studies have proposed that attention can exclude external noise[15] (or improve the ability to integrate signal information, e.g., template sampling efficiency[16]), result in signal enhancement[17], internal noise reduction[17], and optimize the weighting of cued and uncued sensory information[18].

Cell electrophysiological studies with spatial cues have been used to show that attention can increase the amplitude[4], but not the width of the orientation tuning curve at the cued location[4], reduce inter-neuronal correlations[3], and change the contrast response functions[19] of neurons.

These studies have been fundamental in characterizing different signatures of the effects of attention on visual processing, but they remain far from an entire hierarchical pipeline of computational stages required for simple neural circuits to give rise to a behavioral cueing effect. They omit the

computational stages by which brains integrate the response to the cue with the responses to the target, nor do they predict the distribution of cell response properties that one might expect in the brain of biological organisms that manifest behavioral cue effects and covert attention. Even with the progress in multi-neuronal and multi-size recordings, physiological studies probe a small subset of neurons and areas, typically less than 1000 neurons[20–22]. How to map these measurements into computational stages is not straightforward.

The purpose of the current paper is to analyze neurons of a feed-forward neural network trained on images corresponding to a cueing task to characterize the entire multi-stage processing of the visual stimulus and understand how the hierarchy of computations across the neuronal population of each layer can lead to the signature cueing behaviors shown by many organisms from humans and monkeys[23] to crows[24], archerfish[25], and bees[23]. Our analysis seeks to answer specific questions. At which stage are the cue and the target information integrated? When does the cueing effect arise in the network? Does the cueing in the CNN arise in the representation of sensory information at an individual cued and uncued location, or does it appear after neuronal integration across locations? What response properties might neurons show, and do these predictions agree with actual neuronal measurements?

We train a convolutional neural network (CNN) on the Posner cueing task and study the behavioral and neuronal properties that emerge. The network uses images embedded in luminance noise as inputs and is trained to detect whether a target (tilted line) is presented at either location or distractors with less tilt are presented at both locations. The network has no explicit knowledge about the cue nor its predictive probability. We also do not explicitly incorporate any attention mechanism into the network to act differently at the cued location. We compare the properties of 180K cells at the different layers of the CNN to the theoretical computational stages of an image-computable Bayesian ideal observer (BIO) that integrates visual information of the cue, the target, and across locations to maximize perceptual performance detecting the target. The BIO has been used as a normative model to assess how

biological organisms process visual information[16,26–30] and also used as a benchmark model of visual search[31–35] and attention[7,18,33,36–39]. Understanding how a large population of network neurons represents task-relevant information relative to the computational stage of BIO might elucidate how similar probabilistic computations could be carried out in the brain. We interpret the cue, target, and location-tuning properties of the analyzed CNN neurons in the context of existing findings from cell neurophysiology and highlight theoretical predictions that have not yet been reported in the literature.

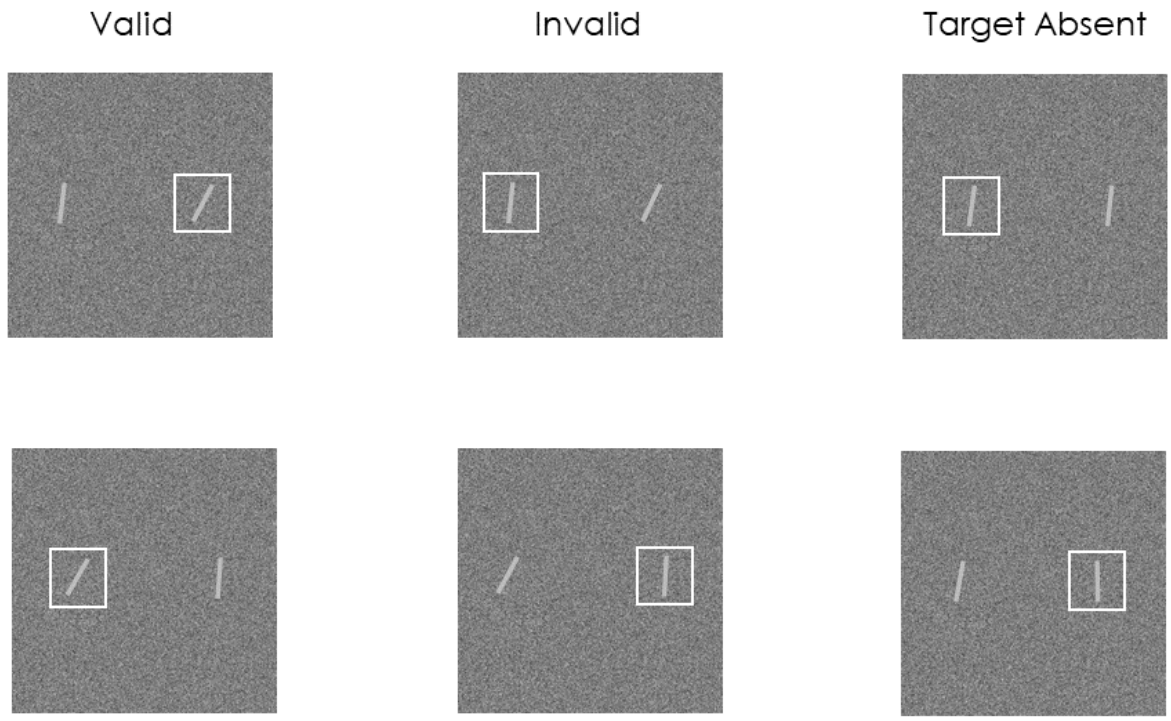

Figure 1: Samples of images used in the cueing paradigm. Valid trials (left column) have the cue (box) appearing with the target (tilted line) and comprise 80% of the target present trials; invalid trials (middle column) have the cue appearing opposite to the target and comprise 20% of target-present trials; while target-absent trials (right column) comprise 50% of all trials. Both the target and distractor angles have angle variability (standard deviation 3.5 deg). The images contain additive independent Gaussian luminance noise. The contrasts of the cue and the target are increased for illustration.

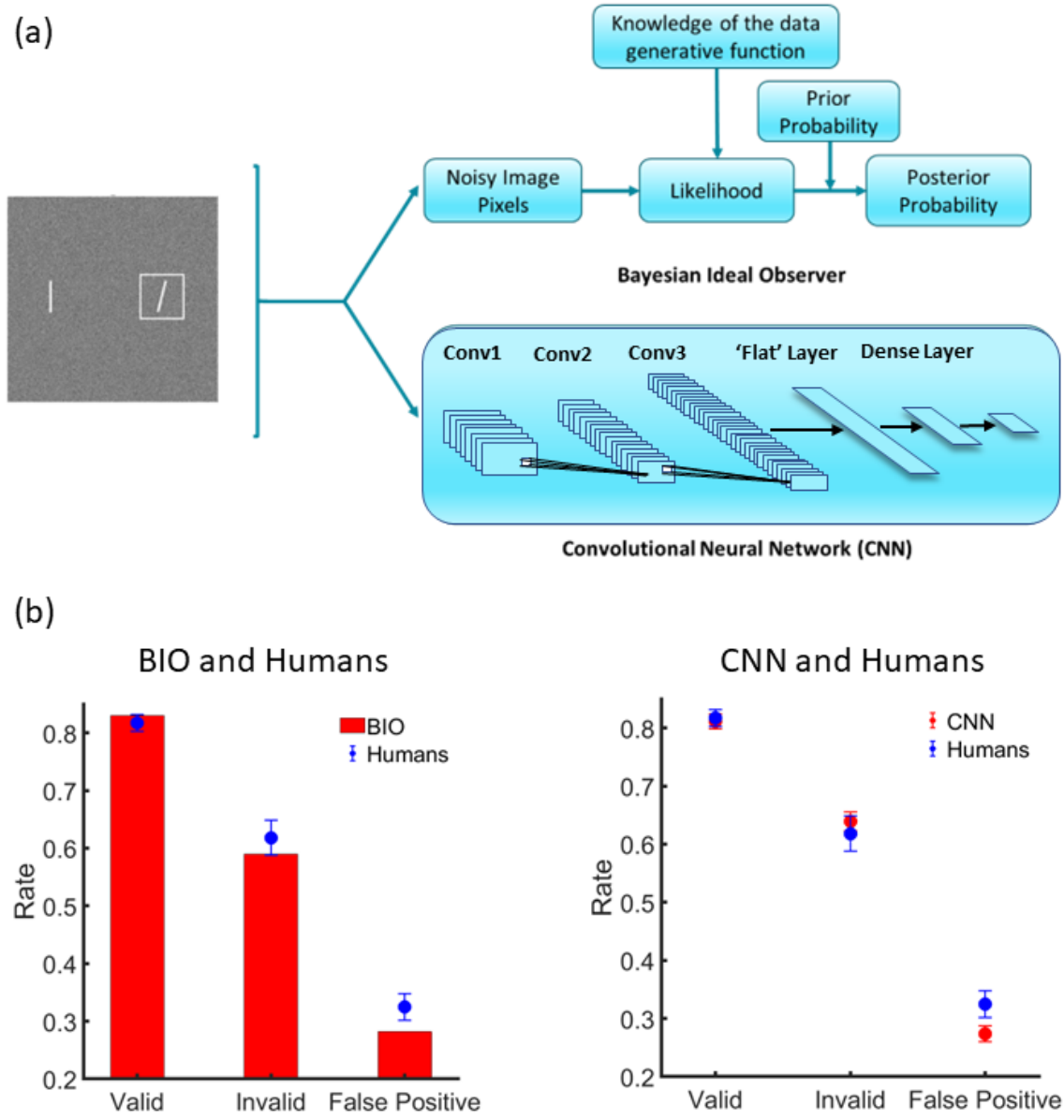

Figure 2 | Model flowchart and detection accuracy in Cueing paradigm (a) Left: Example stimuli used in the Posner cueing experiment with additive external white luminance noise. Right: Stimuli processed by an image-computable Bayesian ideal observer (BIO) and a Convolutional Neural Network (CNN). The flattened third convolution layer is called the Flat layer. (b) Performance of eighteen human observers (blue) compared to the BIO (left) and CNN (right)(red). The RMS of Gaussian luminance noise was increased from 0.14 for human observers to 0.23 for the BIO and 0.21 for the CNN to match human performance (maximize the goodness of fit, AIC). Blue error bars represent the standard error of the mean of n = 18 human observers, and red error bars represent the standard error of the mean of n = 10 CNNs.

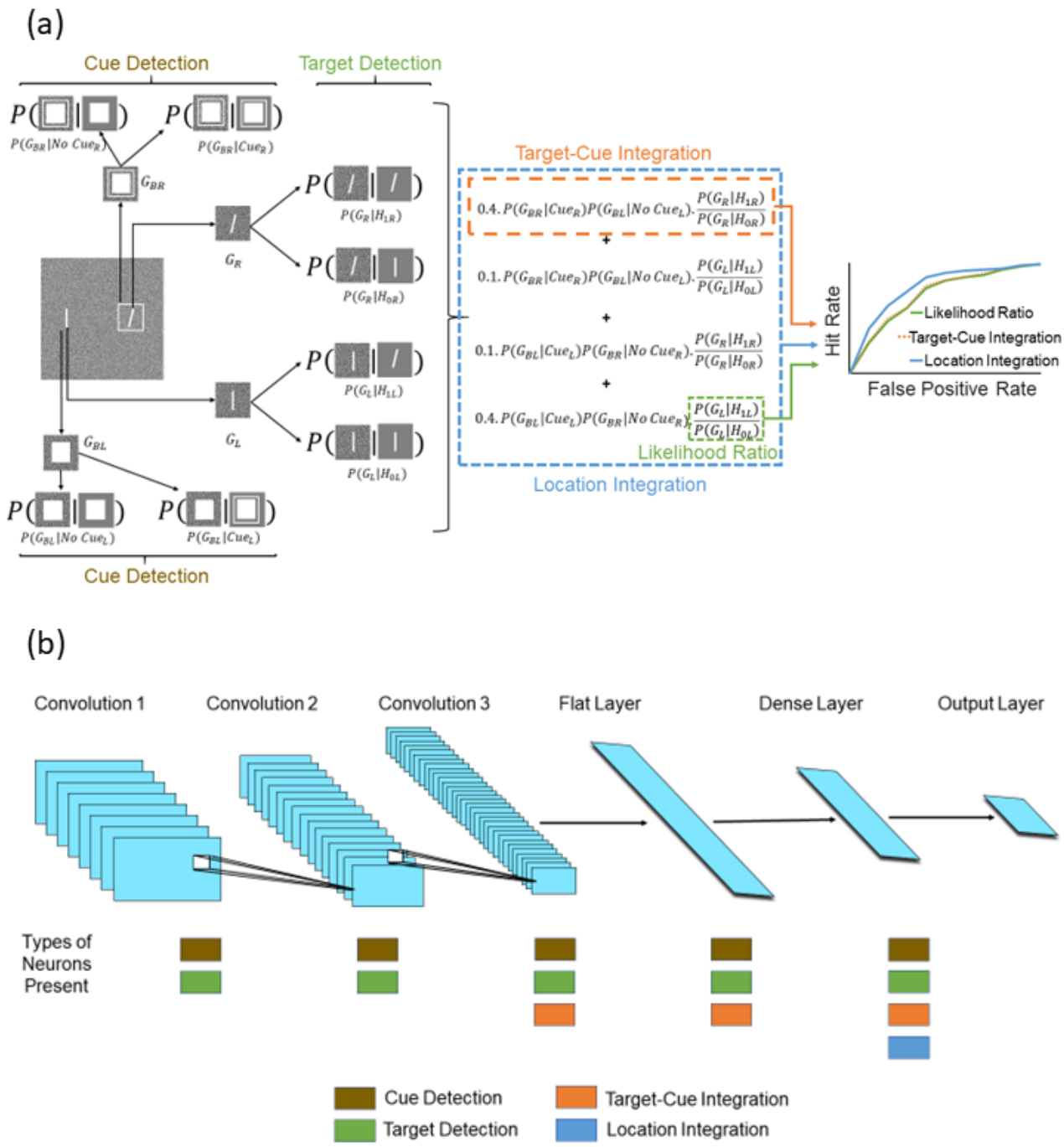

Figure 3| Computational stages in the BIO and equivalent neuronal stages in the CNN. (a) Computational stages of a BIO for a Posner cueing paradigm. The BIO has three stages: the cue/target detection stage (brown/green), the target-cue integration stage (orange), and the location integration stage (blue). The computations involved in each stage are shown in dotted boxes of corresponding colors, and the ROCs detecting the target for each stage are shown on the right panel with the corresponding colors. Note that the orange and green ROC curves overlap while

the blue ROC curve lies above the other two ROCs. (b) Layers and computational stages within the Convolutional Neural Network trained for the cueing paradigm found by our analysis (see Results section). Target-cue integration neurons (orange) are neurons that detect both target and cue at the same location. Location integration neurons can detect the target at both locations.

RESULTS

Behavioral Cueing Task

The task required detecting the presence of a target that appeared with a 50% probability. The target was a tilted line with a tilt angle sampled from a Gaussian distribution with a mean of 14 degrees (deg; Figure 1) and a standard deviation of 3.5 deg. Target present trials contained a target at one location and a distractor at the other location. Distractors were lines with orientation sampled from a Gaussian distribution with a mean of 7 deg and a standard deviation of 3.5 deg. Target absent trials contained independently sampled distractors at both locations. A box cue appeared in all images and appeared with the target with 80% probability when the target was present (i.e., 80 % cue-valid trials vs. 20 % cue-invalid trials). The box cue had an equal probability of being present at either location and was presented simultaneously with the test image (target/distractor). White luminance noise (root mean square contrast, RMS=0.14) was added to the images, including the pixels at the location of the target and the cue. Figure 1 shows examples of stimuli. For the human studies, the stimuli were presented for 250 ms.

Feed Forward Convolutional Neural Network

The CNN was a feedforward convolutional neural network with five layers: three convolutional layers with stride size of 3 with 8, 24, and 32 kernels, respectively, followed by two fully connected layers with 256 and 2 neurons, respectively (Figure 2a). Each convolution kernel was of size 3 by 3. The ReLU

(Rectified Linear Unit) activation function was used in each layer except the output layer, which used the SoftMax activation function. The networks were trained via backpropagation to detect target presence/absence in noisy images until convergence, which usually happened within ten epochs. The images the CNNs were trained on had the target and distractors sampled from the same distributions as human stimuli, but the pixel noise added was increased for the CNNs so that their overall performance was similar to human task accuracy. The CNNs did not use the human trial responses or the BIO output, thus, neither the cueing nor the correspondence with the BIO emerges due to fitting the CNNs to the human or BIO outputs. The CNNs were trained to maximize accuracy across all trial types (valid and invalid cue trials). More details about the training can be found in the Methods section.

Bayesian Ideal Observer

The Bayesian ideal observer for the cueing paradigm[18,33,38] is well-known and consists of weighting sensory evidence about the target (the likelihood) by the prior probability of the target appearing with the cue. That model often omits the process of the detection of the cue and assumes it is detected with no error. Figure 2a shows a more general Bayesian Ideal Observer that uses noise-limited image information to detect the target and the cue in the images. It uses the conditional probability of the presence of the target given the cue to influence the processing of sensory information related to the target presence/absence (see methods for details of the model and relationship to previous instantiations of the BIO that assumed perfect detection of the cue). Figure 3a shows a possible breakdown of the BIO into separate computational stages for the cueing paradigm. The model has full knowledge of the generative function of the data. In its first stage, it detects the presence of the cue and target at each of the two locations by computing the likelihood of the data (noisy luminance values at the pixels along a box at the ith location) given the presence/absence of cue $P(G_{Bi}|Cue_i)$ and the

likelihood of the data (noisy luminance pixels from an ith location) given the presence of the target or a distractor $P(G_i|H_{ji})$, where I denotes the ith location and j denotes the jth hypothesis (j=0, target present and j=1 target absent, distractor present). This stage is denoted as cue detection and target detection in Figure 3a.

In the second stage, the BIO integrates target and cue information by multiplying the cue and target likelihoods and the conditional probability π_i of the target appearing with the cue at that location:

$$\pi_i . P(G_{Bi}|Cue_i) . P(G_{B!i}|No\ Cue_{!i}) .\ P(G_i|H_{1i})/P(G_i|H_{0i}) \qquad (1)$$

Here, !i denotes the location other than the ith location. This calculation is separately computed for each location and is denoted by the target-cue integration stage in Figure 3a.

The third stage integrates the evidence across spatial locations by summing joint probabilities across possible target/cue and location configurations. As we analyze the neuronal population properties of the CNN, we will draw comparisons to similar analyses with the BIO variables at the three computational stages. The details of how to evaluate the values of these variables in the BIO are presented in Supplementary Section S.1, along with the BIO derivation.

Behavioral predictions of the CNN and BIO models

Both the CNN and the BIO attain a higher overall performance than humans when evaluated with the stimuli that human observers viewed. A classic measure to compare the performance of human/model accuracy to the BIO is the statistical efficiency defined as the ratio of noise variance for the BIO to match human/model accuracy[16,27]. For our task, we found statistical efficiencies of 39% for humans and 90% for CNN; see methods). The CNN, trained to optimize overall accuracy across all trial types, shows a cueing effect: higher hit rates for valid vs. invalid cue trials (Figure 2b). The BIO, which attains the upper

bound of accuracy for the task, also shows a cueing effect as demonstrated in previous studies[18,38] (Figure 2b).

To make the overall performances of the CNN and BIO comparable to humans, we increased the external white noise of the images separately for each of the models. The external noise was fit to minimize the Akaike Information Criterion (AIC) with no other fitting parameters (see methods for details). Figure 2(b) shows the performance of the two models compared to the performance accuracy of eighteen human observers in the two-location task with the 250 ms presentation and a central fixation (see methods). Both the CNN and BIO result in cueing effects that are comparable to humans.

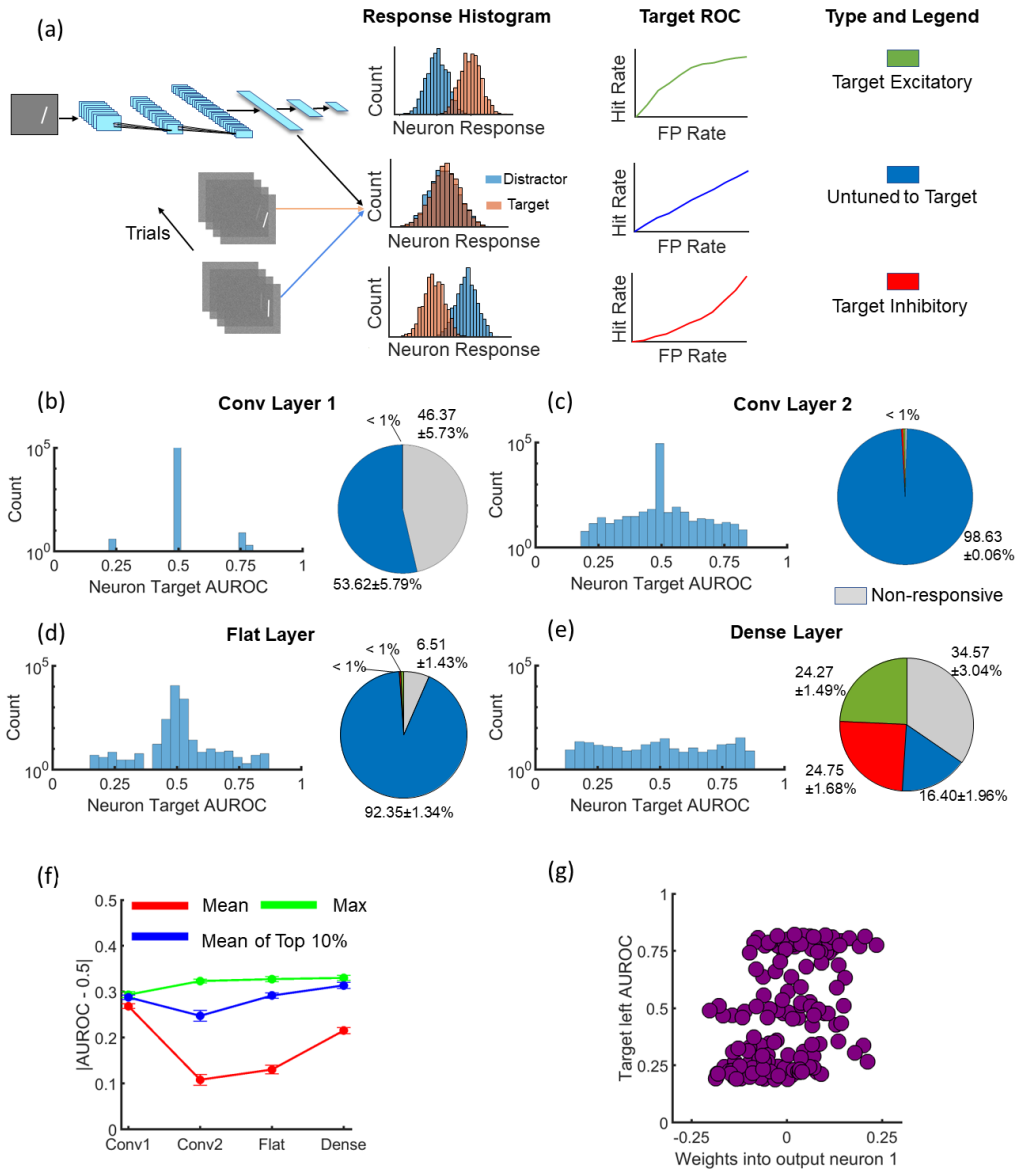

Figure 4 | Receiver operating characteristic (ROC) curves for target detection. (a): Flow diagram for calculating target AUROC using images with noise target present or absent (distractor present) and no cue. Top Right: Example of encountered types of neurons, their response histograms, and ROCs. Second row (b-c): Histograms of target AUROC for the first two convolution layer neurons (the y-axis is in logarithmic scale) and pie charts for various types of neurons based on their target tuning property**.** Error bars are standard errors over ten different trainings. Third row (d-e): Same as the middle row but for the flat and dense layers. Fourth row (f): Layerwise mean, max, and mean of the top 10% of the absolute difference between target AUROC and 0.5 for ten models trained on cue validity of 80%. Error bars depict standard errors of the mean over ten networks. (g) Scatter plot of target AUROC of dense layer target-cue neurons and the weights they project into the output neuron 1 (output neuron 1 denotes the target presence neuron in the output layer, $r = 0.5577$, $p < 10^{-50}$).

Statistical characterization of neuronal properties

To understand the hierarchy of computational stages leading to the cueing effect, we analyzed the response properties of a total of 133.6k neurons for the first convolution layer, 45.6k neurons for the second convolution layer, 7296 neurons for the flat layer (flattened third convolution layer) and 256 neurons for the dense layer. In the next sections, we analyze the neurons' responses to the target, cue, target-cue integration, integration across locations, and the influence of the cue on the target sensitivity. All evaluations included ten separate training repetitions of the networks, each with testing over 10,000 trials. Luminance noise was independent for each training and testing sample. Some analyses (Figures 6, 8) are presented for neurons tuned to the left location (convolution layers and flat layer), but similar results are obtained for neurons tuned to the right spatial location.

Target Sensitivity

To measure the sensitivity of a neuron to the target, we generated a Receiver Operating Characteristic curve (ROC). A neuronal ROC curve was generated by collecting histograms of responses of each neuron in the CNN to a separate set of images that contained the target present or absent (distractor present). We used ten sets of 1,000 images (50 % target present and 50 % target absent with independent samples of white luminance noise). We evaluated each of the two locations separately, and no cue was present in the images (Figure 4a). The analysis was repeated for each layer in the network (conv 1, conv 2, flat layer, and dense layer).

Figure 4a shows histograms of three sample encountered neuron types in the network at the dense layer. Some neurons responded more strongly when the target was present (target excitatory). Other neurons consistently diminished their response when the target was present (target inhibitory), while a

third type was insensitive to the presence of the target (Figure 4a, middle histogram). For each neuron's pair of response distributions (target present and absent), we systematically varied a decision threshold to generate a hit rate (proportion of target present distribution to the right of the decision threshold), a false positive rate (proportion of target-absent distribution to the right of the decision threshold) and traced an ROC curve. The area under the ROC curve (AUROC) can be used as a measure of sensitivity[40]. We use the AUROC analysis to classify the types of neurons: target-excitatory (AUROC > 0.5), target-inhibitory (AUROC < 0.5), and untuned (AUROC not significantly different from 0.5). To assess the statistical significance of these and tests in subsequent sections, we used a one-sample t-test after correcting for multiple comparisons using False Discovery Rate, FDR with $\alpha = 0.01$; two-tailed. A fourth type of cells were unresponsive altogether to either the images with the target present or absent.

Figures 4b-e show histograms of target sensitivities (AUROC) across neurons and pie charts of the percentage of the different neuron types for the four layers. Neurons in the convolution layers and the flat layer are tuned to a single location, while a subset of dense layer neurons are tuned to multiple locations. The results show a greater percentage of neurons with excitatory and inhibitory sensitivities to the target as the processing progresses through the layers of the network (dense layer: 24.3 ± 1.29 % excitatory and 24.3 ± 1.68 % inhibitory neurons). Figure 4f shows the layerwise mean, maximum, and top 10 % (mean) neuronal target sensitivity defined as the absolute difference between a neuron's target AUROC and 0.5. This definition combines AUROC for excitatory and inhibitory neurons. There is a statistically significant increase in the mean target sensitivity for the dense layer compared to the Conv2 and flat layers ($p < 10^{-5}$ for Conv2 vs. dense and for flat vs. dense), and the mean target sensitivity of the top 10% of neurons ($p = 0.0003$ for Conv2 vs dense and $p = 0.0004$ for flat vs dense). For the maximum target sensitivity, the Conv1 layer showed significant differences with all the other layers ($p = 0.0025$ for Conv1 vs. Conv2, $p = 0.0022$ for Conv1 vs. flat, and $p = 0.0016$ for Conv1 vs. dense), but there were no significant differences between Conv2, flat or dense layers.

Finally, in Figure 4g, we present the scatterplot of neuronal target AUROC of dense target-cue neurons and the weights from these neurons to the neuron in the output layer corresponding to the target present class label. We find a moderate positive correlation ($r = 0.5577$, $p < 10^{-50}$) between the target AUROC and the weights to the output layer neuron 1 among dense layer target-cue neurons.

To summarize, the layer's percentage of neurons tuned to the target and their sensitivities increase as visual processing progresses through the network. We find both target-excitatory and target-inhibitory neurons throughout the network.

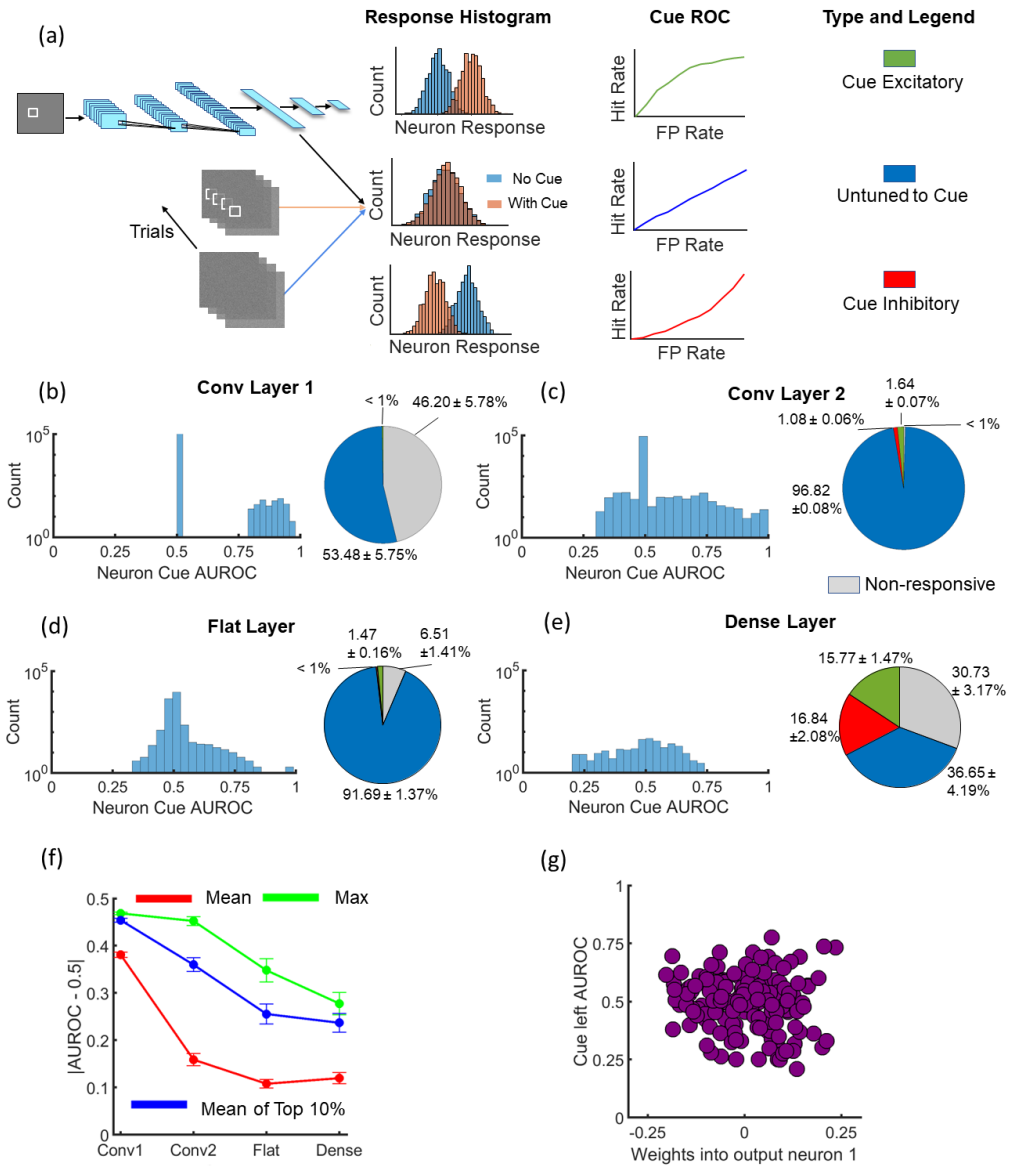

Figure 5| Receiver operating characteristic (ROC) curves for cue detection. (a): Flow diagram for calculating cue AUROC with images with and without the cue, luminance noise but not target or distractors. Top right: Example encountered types of neurons and their ROCs. Second row (b-c): Histograms of cue AUROC for the first two convolution layer neurons (the y-axis is in logarithmic scale) and pie charts for various types of neurons based on their cue tuning**.** Error bars denote the standard error of the mean over ten trained networks. Third row (d-e): Same as the middle row but for the flat and the dense layers. Fourth row (f): Layerwise mean, max, and mean of the top 10% of the absolute difference between cue AUROC and 0.5 for ten models trained on cue validity of 80%. Error bars depict standard errors of the mean over ten networks. (g) Scatter plot of cue AUROC of dense layer target-cue neurons and the weights they project into the output neuron 1 (output neuron 1 denotes the target presence neuron in the output layer, r = -0.1011, p = 0.0001.)

Cue Sensitivity

We conducted a similar analysis to quantify the neurons' sensitivities to the cue by computing the AUROC across an ensemble of noisy images when the cue was present or absent (50 % probability). We separately evaluated the two spatial locations. The evaluation images did not contain targets or distractors. We classified each neuron based on the AUROC as cue-excitatory (increased response with cue presence, AUROC > 0.5), cue-inhibitory (reduced response with cue presence, AUROC < 0.5), and untuned (Figure 5a). If the neuron did not respond at all to the cue, we classified it as unresponsive. Figures 5b-e show histograms of cue sensitivities (AUROC) across neurons and pie charts for the percentage of different neuron types for the four layers. The network showed cue excitatory and cue inhibitory neurons. As with sensitivity to the target, the percentage of neurons with statistically significant sensitivity to the cue increased as processing progressed through the network (dense layer: 15.8 ± 1.5 % cue-excitatory and 16.8 ± 2.1 % cue-inhibitory neurons). However, unlike the neuronal sensitivities to the target, the neurons' mean, max, and top 10 % excitatory and inhibitory sensitivities to the cue (|AUROC- 0.5|, Figure 5f) diminished greatly from the 1st convolution layer to the dense layer ($p < 0.05$ for all comparisons except for mean sensitivity between flat and dense layer and max sensitivity between Conv1 and Conv2). Finally, in Figure 5g, we present the scatterplot of neuronal cue AUROC of

dense target-cue neurons and the weights from these neurons to the neuron in the output layer corresponding to the target present class label. We find a small negative correlation ($r = -0.1011$, $p = 0.0001$) between the cue AUROC and the weights to the output layer neuron 1 among dense layer target-cue neurons.

Target-Cue Integration Neurons

A fundamental step to give rise to a cueing effect must involve units that integrate cue location information with sensory information related to the target. We assessed how different neurons at different layers are jointly tuned to target and cue. Figures 6a-d show all neurons' sensitivity (AUROC) for the target plotted against that of the cue for all the layers of the network. Neurons falling along the two AUROC = 0.5 lines (vertical and horizontal perpendicular lines) correspond to neurons that are separately tuned to only the target or the cue (except for the neurons whose AUROC is 0.5 for both target and cue and are thus tuned to neither). Neurons that fall off the two perpendicular lines correspond to neurons that are jointly tuned and integrate target and cue information. Results show no neurons jointly tuned to the target and cue for the first convolutional layer and a very small percentage of neurons (< 0.01% neurons, Figure 6a-b) and the flat layer (0.16 ± 0.03% jointly tuned vs. 7.50 ± 1.59 % and 1.62±0.21 % responding to the target and cue separately, Figure 6c) but a large increase for the dense layer (28.7 ± 4.5 % jointly tuned vs. 17.83 ± 2.47 % target only and less than 1% for cue only, Figure 6d). Thus, only the neurons in the deeper layers show an integration of the cue and target detection (flat layer onwards). The dense layer also shows an asymmetry with a large percentage of neurons tuned solely to the target but a small percentage tuned only to the cue.

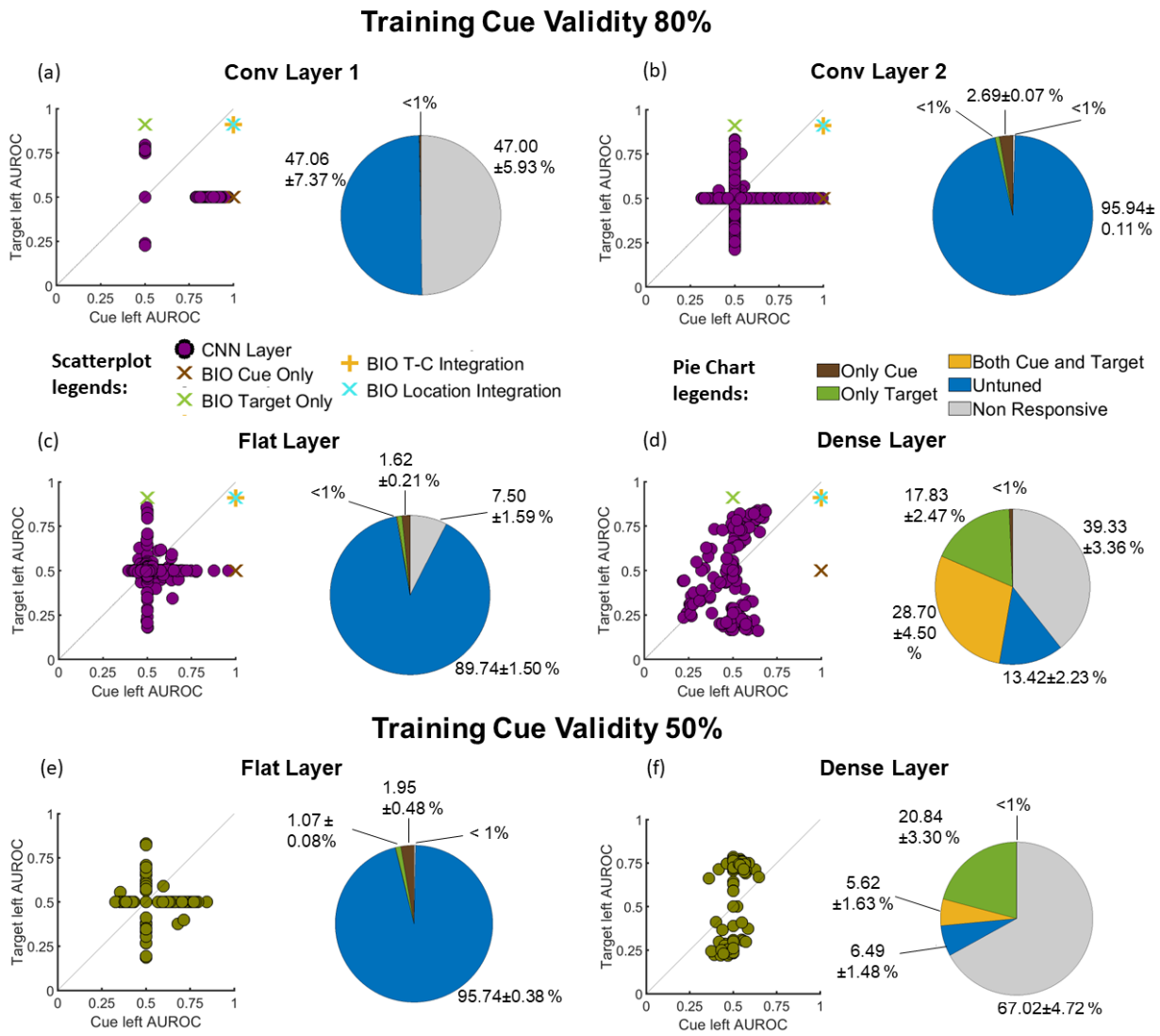

Figure 6 |Target-cue integration at the same location Top row: Scatter plots of neuronal AUROCs for target vs. neuronal AUROCs for cue at the left location for neurons in the first (a) and second convolution layers (b) of CNNs with training cue validity of 80%. The ROCs obtained from the corresponding stage of the BIO are also marked on the scatter plot. Pie charts represent the relative number of the various types of neurons. Errors indicate the standard error of the means over ten different trained neural networks. (c) and (d)same as a) and b)but for the flat and dense layers. Legends for the scatter plots and pie charts are shown in between the graphs. d) and 3) Same as c) and d) but for the flat and dense layer of CNNs with training cue validity of 50%.

Figure 6a-d and Figure 7a-d also plot the AUROC for a BIO's decision variables at different stages of its computational hierarchy (see Figure 3a): cue detection, target detection, target-cue integration, and location integration. There is a correspondence between the various types of neurons that emerge in

the network and the various stages of processing in the BIO. The neurons in the early layers map to the separate cue and target detection stages of the BIO (Figure 6a-b). Neurons in the dense layer (Figure 6d) show properties similar to the BIO decision variable at the target-cue integration stage and the final BIO location integration stage. The neurons at the flat layer (Figure 6c) show an intermediate stage with more neurons that are involved in separate cue and target detection and a small proportion of target-cue integration neurons. There are some important differences. The BIO shows very defined properties with higher AUROCs and abrupt transitions to cue/target integration, while the CNN shows a larger range of AUROCs across its populations and a more nuanced transition towards target-cue integration across layers.

As a control, Figure 6e-f shows analyses for the flat and dense layers of a network trained on a 50% predictive cue (see Supplementary S.2 for the conv1 and 2 layers). Results show similar separate subpopulations of target and cue-tuned neurons in the flat layer (Figure 6e) but a significantly lower number of jointly tuned neurons in the dense layer when compared to the predictive cue (Figure 6f, 5.62 ± 1.63 for 50 % predictive vs. 28.7 ± 4.5 % for the 80% predictive cue trained network, $p < 0.0001$ for a paired t-test across ten networks trained on 80% predictive and ten networks trained on 50% predictive cues). Thus, the 80% predictive cues give rise to more target-cue jointly tuned neurons in the last dense layer of the network.

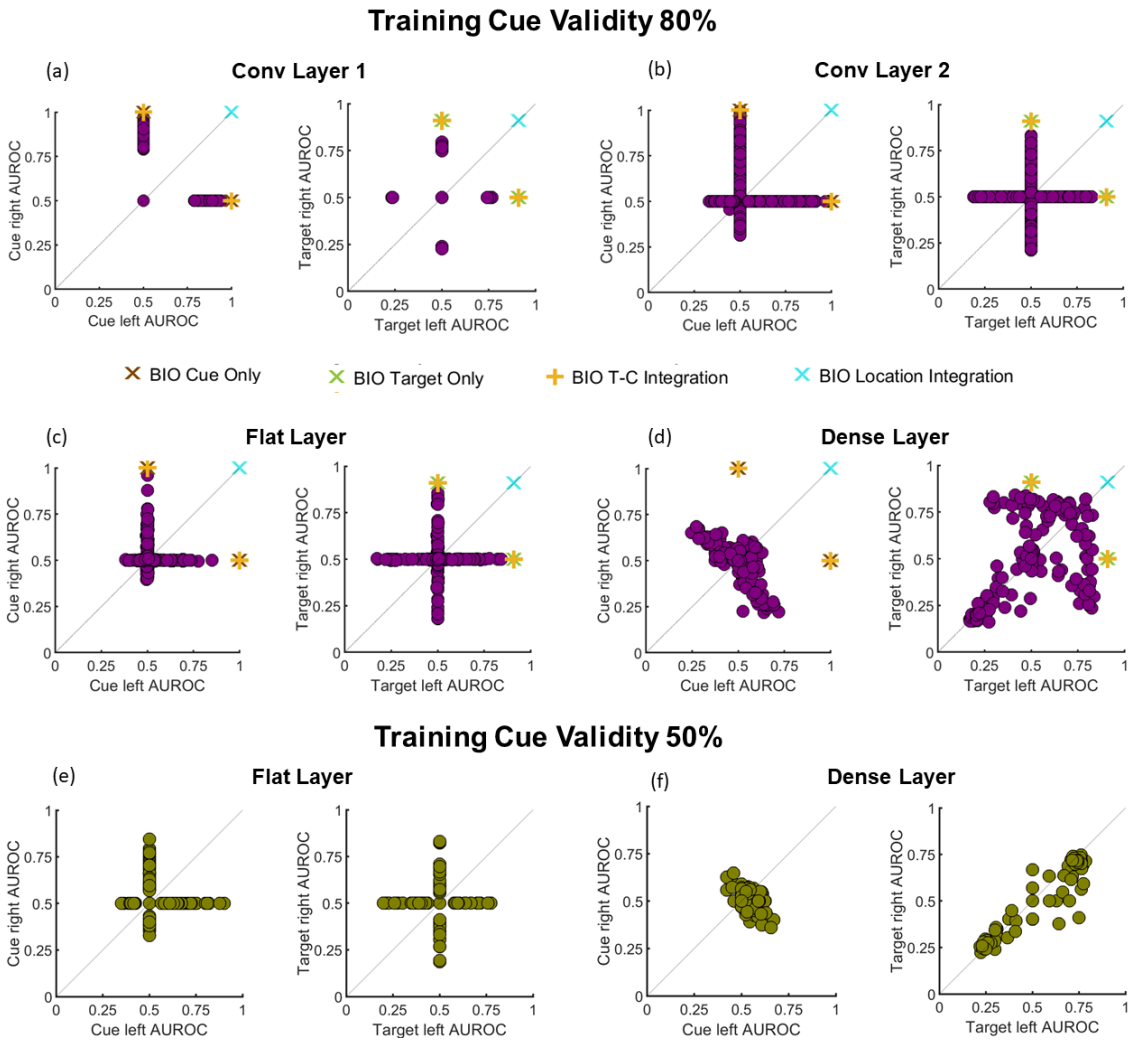

Figure 7|Integration across locations Top row: Scatter plots between neuronal AUROCs for cue detection and target detection at the right and left locations for neurons in the first two convolution layers for CNN with training cue validity of 80%. The AUROCs obtained from the corresponding stage of the BIO are also marked on the scatter plot. Second row: The same as the top row but for the flat layer and the dense layer. The legends are shown in between the first two rows of plots. Third row: The same as the second row but for CNN trained with training cue validity of 50%.

Integration across spatial locations

The integration of target and cue detection locally is not sufficient to successfully perform the Posner cueing task. The organism must also integrate information across the two locations that could contain the target and cue. The subsequent analysis investigated the integration of information across locations

in the CNN. We assessed the relationship between each neuron's AUROC for target or cue detection on the left and the right locations (Figure 7). The early two convolution layers and the flat layer (Figure 7a-c) contained neurons that have target and cue neurons that show no integration across the two locations (data fall along one of the perpendicular axes: AUROC ≠ 0.5 for one location but AUROC= 0.5 for the other location). In contrast, the dense layer (Figure 7d) showed neurons that integrate across both locations (AUROC ≠ 0.5 for both locations). For the target-sensitive neurons, 35.48 ± 1.53 % were target-excitatory at both locations, 43.69 ± 1.79% were target-inhibitory at both locations, and 20.88 ± 1.97% were target-excitatory at one location, and target-inhibitory at the other (location-opponent neurons). The majority of cue neurons (81.25 ±4.41%) at the dense layer were location-opponent.

As a comparison, Figure 7e-f shows analyses for the flat and dense layers of a network trained on a 50 % predictive cue (see Supplementary S.2 for the conv1 and 2 layers). Results show similar separate subpopulations of neurons tuned to each location for the flat layer (Fig 7c vs. 7e) followed by a dense layer integrating target and cue information across locations. However, compared to the network trained with the 80 % predictive cue, the most notable differences are fewer percentage of neurons that are target-excitatory at one location and target inhibitory at the other (2.79 ± 0.71% of all target sensitive neurons vs. 20.88 ± 1.97% for 80% predictive cue network) and lower number of neurons that are cue-excitatory at one location and cue-inhibitory at the other location (54.05 ± 9.95% for the 50% predictive cue vs. 81.25 ±4.41% for the 80%). Thus, the 80 % predictive cue results in a larger subpopulation of location-opponent cells (excitatory at one location and inhibitory at the other).

Influence of cue on target sensitivity

Fundamental to the understanding of cue-mediated covert attention has been elucidating how the presence of the cue influences the processing of the target[3,35–37]. We assessed the location in the

network hierarchy where the cue influences the neural responses to the target. The analyses involve neurons that are tuned only to the target and a single location for the convolution layers and the flat layer. It also involves neurons that integrate target and cue (flat layer and dense layer) as well as across locations (dense layer). For the flat layer, we show separate results for neurons tuned only to the target and those tuned to the target and cue (Flat T-C in Figure 8b, c, d).

We first analyzed the neurons' responses to the target when the cue is valid relative to when the cue is absent. As a control, we compare our neuronal analyses for the network trained with an 80 % predictive cue to the same analyses for a CNN trained with a 50% predictive cue. Figure 8b shows the distribution of the neurons' mean responses to the target with the valid cue relative to the neurons' mean target responses when the cue is absent (no cue). The presence of the cue results in larger increases and decreases in the mean response to the target in a subpopulation of neurons (see tails of distributions) as the processing progresses through the network's layers. Comparison to the distribution of neurons' mean responses to the target for the network trained with 50% predictive cue shows that the influence of the predictive (80 %) cue arises in the flat and dense layer (The Kolmogorov-Smirnov (KS) test compares the statistical significance of differences in distributions across distributions and resulted in: $p=0.9815$ for Conv1, $p=0.8111$ for Conv2, $p= 0.1606$ for flat layer non-target-cue integrating neurons, $p=1.792\times10^{-9}$ for flat layer target-cue integrating neurons and $p= 2.047\times10^{-5}$ for the dense layer).

Second, we compared the neurons' mean responses to the target when the cue was absent relative to when the cue was invalid and appeared opposite to the target (Figure 8c). This analysis assesses any potential inhibitory influence of the appearance of the invalid cue on the response to the target.

Positive differences suggest a decrease in the mean response to the target with the invalid cue (inhibition). The results show that the difference in the distributions of the neurons' mean responses relative to the 50% predictive cue distributions arises at the dense layer (KS test: $p= 0.9578$ for Conv1,

p= 0.7676 for Conv2, p= 0.1238 for flat layer non-target-cue integrating neurons, p= 0.1244 for flat layer target-cue integrating neurons and p= 1.51×10^{-9} for the dense layer). This is consistent with the result that the dense layer contains most of the spatial location integration neurons. If neurons are only responsive to the left location, then the presence of the invalid cue opposite to the target location will not influence the neurons' responses. Figure 8b also shows that there is a larger percentage of target-responsive neurons in the dense layer that are inhibited by the presence of the invalid cue (the tail of the distribution in the positive direction is larger than the negative direction, Figure 8b). The BIO response variables to the target are also inhibited with the presence of the invalid cue only at the stage where information is integrated across spatial locations (see dotted line in Figure 8b).

Assessing the influence of the cue on target sensitivity involves also evaluating the effects of the cue on the neurons' response to the distractor as well as the variability of the response. Our results for the response to the distractor and the standard deviation are included in Supplementary Information Figure S.3. The net effect of the influence of the cue on the neurons' responses to the target, distractor and their variability will impact the neurons' sensitivity to the target. The effect on sensitivity can be quantified using an AUROC analysis comparing the neuronal responses to the target present vs. absent (distractor present) when the valid cue is present vs. when the cue is absent (see Figure 8a, no cue). Figure 8d shows the difference in AUROC Valid vs. AUROC no cue for all analyzed neurons across layers. A comparison of the neuronal AUROC cueing distribution of the network trained with 80 % predictive cue to the neuronal AUROC cueing distribution of the 50% predictive cue networks shows that the influence of the presence of the predictive cue on target sensitivity arises in the flat layer and the dense layer of the network (KS test: p= 0.3343 for Conv1, p= 0.2885 for Conv2, p= 0.1140 for flat layer non-target-cue integrating neurons, p= 0.0132 for flat layer target-cue integrating neurons and p= 1.036×10^{-5} for the dense layer). For comparison, the plot shows the influence of the cue on target sensitivity for the BIO model (dotted lines), which arises only at the stage of integration across locations[41]. Thus, our

analysis shows a qualitative dissimilarity in how the cue influences target sensitivity in the CNN and the BIO.

We observed a similar trend of the increased influence of the cue with the processing stage in the network when measuring the percentage of neurons showing a significant change in AUROC for images with valid cue vs. no cue: 2.49% (80% predictive) and 1.48% (50% predictive) for neurons in Conv1, 15.24% (80% predictive) and 5.78% (50% predictive) for neurons in Conv2, 11.05% (80% predictive) and 11.95% (50% predictive) for neurons in the flat layer non-target-cue integrating neurons, 78.95% (80% predictive) and 11.72% (50% predictive) for neurons in the flat layer target-cue integrating neurons and for 65.28% (80% predictive) and 39.57% (50% predictive) neurons in the dense.

We separately compared the target sensitivity (AUROC) when the cue is invalid and opposite to the target with the sensitivity when the cue is absent (no cue). Figure 8e shows the difference in AUROC no cue and AUROC invalid cue for all analyzed neurons across layers. A comparison of the neuronal AUROC cueing distributions for the network trained with 80 % predictive cue to the neuronal AUROC cueing distributions for the network trained with the 50% predictive cue shows that the invalid cue influences target sensitivity at the dense layer (KS test: $p = 0.2578$ for Conv1, $p= 0.2129$ for Conv2, $p= 0.1088$ for flat layer non-target-cue integrating neurons, $p= 0.0829$ for flat layer target-cue integrating neurons and $p= 0.0021$ for the dense layer). The layerwise percentage of neurons showing significant difference in AUROC for no cue trials relative to invalid cue trials is consistent with these results. We find that 2.15% (80% predictive) and 1.91% (50% predictive) for neurons in Conv1, 15.83% (80% predictive) and 5.82% (50% predictive) for neurons in Conv2, 7.18% (80% predictive) and 12.84% (50% predictive) for non-target-cue integrating neurons in the flat layer, 8.42% (80% predictive) and 12.49% (50% predictive) fortarget-cue integrating neurons in the flat layer and 71.06% (80% predictive) and 38.42% (50% predictive) for neurons in the dense layer. This result is consistent with the concept that it is only neurons that integrate information across the two spatial locations that can show influences of the

invalid cue on the target sensitivity. The AUROC differences are positive and negative, suggesting that there are subpopulations of neurons for which target sensitivity is reduced by the presence of the invalid cue and other subpopulations for which the invalid cue increases the target sensitivity.

Rectified Linear Unit Activation (ReLU) Function and valid cue effect on target sensitivity before integration across locations

An important finding is the CNN's change in target sensitivity with the presence of a valid cue at the cue-integration neurons in the flat layer before integration across spatial locations. This finding differs from the BIO, which only manifests sensitivity changes after the integration across locations stage. What mediates the change in target sensitivity (AUROC valid cue) in the flat layer? Analyses of the neuron responses in the flat layer show that the effect arises from the interaction of the influence of the valid cue on the neurons' mean response with the non-linear ReLU thresholding. Figure 8f illustrates a sample neuron's distributions of responses to the target and distractor before and after the ReLU. The top left graph shows a neuron at the flat layer with pre-ReLU AUROC of 0.6966 without the cue. The ReLU zeroes all negative values, effectively cutting off part of the tails of the target and distractor distributions. This ReLU thresholding reduces the AUROC of the shown neuron when the cue is absent. Figure 8f bottom left shows the same neuron's distribution of responses but with the presence of the valid cue, which has a pre-ReLU AUROC of 0.6844. The valid cue increases the neuron's mean response to the target and distractor, and a larger part of both distributions is greater than zero; thus, less of both distributions are truncated, resulting in a lower decrease in AUROC. While before ReLU, the AUROCs without cue and with valid cue only differed by 0.0122, after the ReLU, this difference increased over seven-fold to 0.0915. The impact of the cue on the flat layer T-C neuron's target sensitivity will depend on how the cue changes the neuron's mean response.

Overall, across all ten trained networks, the ReLU non-linearity of the flat layer T-C integrating neurons increased the influence of the cue on the target AUROC by 217.66%.

Figure 8f right shows a similar analysis for a sample neuron in the dense layer. Neurons in the dense layer have higher means relative to the flat layer T-C neurons. The distributions have a very small fraction of responses below zero, and thus, they are not affected as much by the interaction of the cue's influence on the mean response and the thresholding of the ReLU. While pre-ReLU, the difference in AUROC without and with cue was 0.0418 in the example shown, post-ReLU, this difference is 0.0472. Overall, across all ten trained networks, the ReLU non-linearity of the dense layer T-C integrating neurons increased the influence of the cue on the target AUROC by 40.18%.

Effect of flat layer activation functions on CNNs cueing effects.

To investigate whether the valid cue's influence on target sensitivity at the CNN's target-cue integration stage before the location integration is a crucial mechanism for the cueing effect, we explored various variations of the CNN. We replaced the ReLU activation function with a tanh (hyperbolic tangent function) in the flat layer. The tanh function does not have thresholding (zeroing of responses below a value) and thus should not interact with the changes in the neurons' mean responses to modify the target sensitivity.

Our results confirmed that the flat layer neurons in the tanh network did not show a change in target sensitivity with the presence of the cue. The mean absolute cueing for the T-C integrating neurons in the flat layer with the tanh activation function was 24.20% of the mean absolute cueing for the T-C integrating neurons in the flat layer with the ReLU activation function. At the output level, the CNN still exhibited a comparable cueing effect (valid hit rate – invalid hit rate = 0.16) with the ReLU activation function. Thus, this suggests that the cue-induced target sensitivity change in T-C flat layer neurons

arising from the interaction with the ReLU is not required for the network to result in the behavioral cueing effects as an emergent property of task-optimization

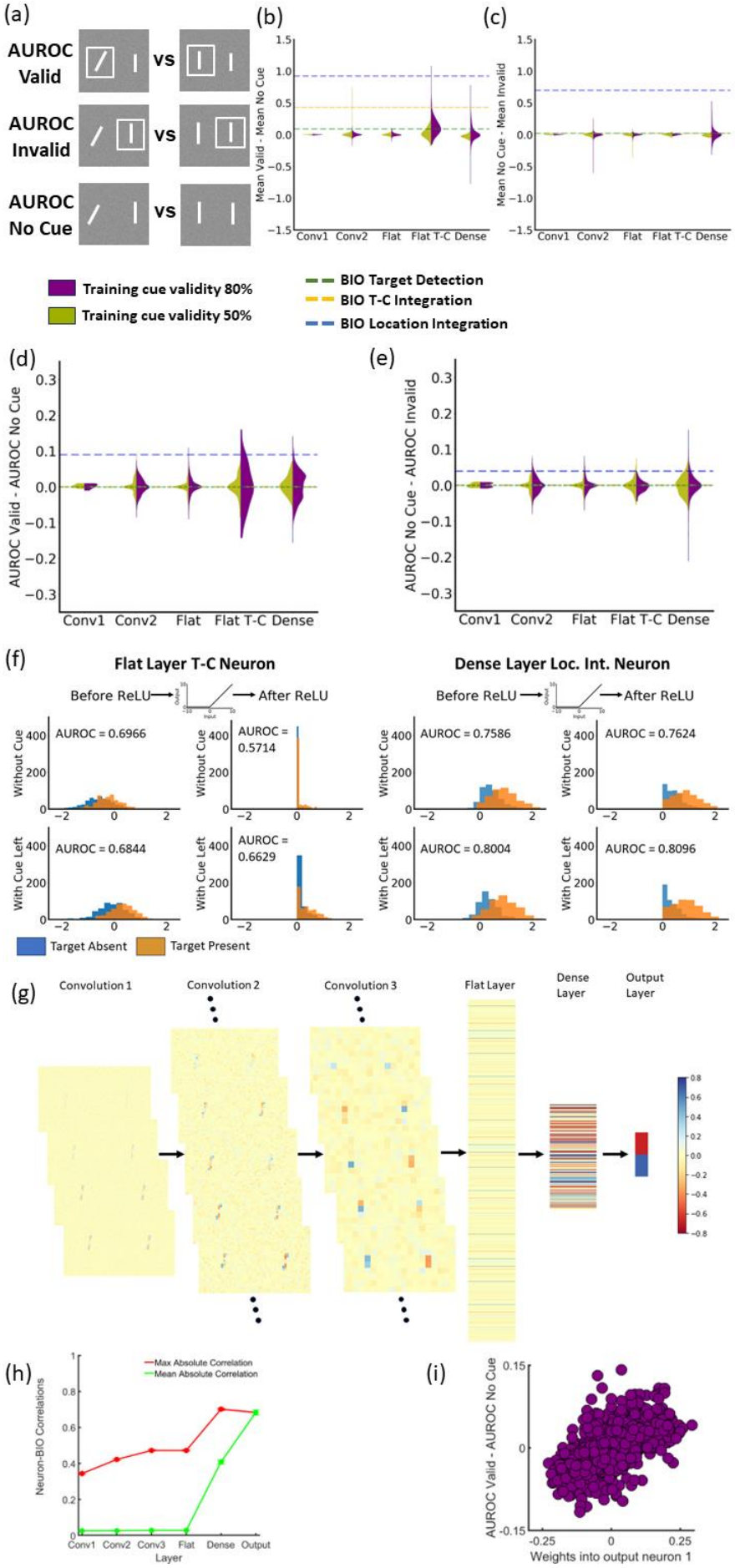
(a)
AUROC Valid
vs
AUROC Invalid
vs
AUROC No Cue
vs
(b)
Mean Valid - Mean No Cue
Conv1 Conv2 Flat Flat T-C Dense
(c)
Mean No Cue - Mean Invalid
Conv1 Conv2 Flat Flat T-C Dense
Training cue validity 80%
Training cue validity 50%
BIO Target Detection
BIO T-C Integration
BIO Location Integration
(d)
AUROC Valid - AUROC No Cue
Conv1 Conv2 Flat Flat T-C Dense
(e)
AUROC No Cue - AUROC Invalid
Conv1 Conv2 Flat Flat T-C Dense
(f)
Flat Layer T-C Neuron
Dense Layer Loc. Int. Neuron
Before ReLU
After ReLU
Without Cue
With Cue Left
AUROC = 0.6966
AUROC = 0.5714
AUROC = 0.6844
AUROC = 0.6629
AUROC = 0.7586
AUROC = 0.7624
AUROC = 0.8004
AUROC = 0.8096
Target Absent
Target Present
(g)
Convolution 1
Convolution 2
Convolution 3
Flat Layer
Dense Layer
Output Layer
(h)
Max Absolute Correlation
Mean Absolute Correlation
Neuron-BIO Correlations
Conv1 Conv2 Conv3 Flat Dense Output
Layer
(i)
AUROC Valid - AUROC No Cue
Weights into output neuron 1

Figure 8 |The effect of hierarchy on neuronal cueing. (a) Example figures depicting stimuli that the ROCs are computed for. (b) Layerwise distributions of differences in mean neuronal responses between images with valid cue at the left location vs. no cue for networks trained with 50% valid cue (green) and 80% valid cue (blue). Dotted lines represent the BIO. (c) Layerwise distributions of differences in mean neuronal responses between images with invalid cue at the right location vs. no cue for networks trained with 50% valid cue (green) and 80% valid cue (blue). (d) Layerwise distributions of differences in neuronal target presence AUROC between images with valid cue at the left location vs. no cue for networks trained with 50% valid cue (green) and 80% valid cue (blue). (e) Layerwise distributions of differences in neuronal target presence AUROC between images with invalid cue at the right location vs. no cue for networks trained with 50% valid cue (green) and 80% valid cue (blue). (f) Example responses before and after ReLU of example flat layer T-C neuron (left half) and dense location integration neuron (right half) when the target is present at the left location(orange) and absent (blue). The top row shows responses when the cue is absent, and the bottom row shows responses when the cue is present at the left location. (g) Neuron-BIO correlations. For each neuron, we calculate the trial-to-trial correlation with the BIO on a test set. A heatmap of these correlations is presented for each layer. For the first convolution layer, all responsive maps are shown. For the next two convolutions, the five activation maps with the highest mean absolute correlations are shown. For the flat layer, a middle strip of 800 neurons with higher correlation is shown, outside which the correlations are small in magnitude. The flat and the dense layer (256 neurons) are approximately equal in scale. The output layer only has two neurons, and their size is exaggerated for visualization. (h) Layer-wise mean and max neuron-BIO correlations**.** Conv 3 and flat layers are the same layers but with different shapes, but both have been included for the sake of consistency. (i) Scatterplot depicting the relation between cueing of neurons in the dense layer and the weights projected by these neurons into neuron 1 (target presence neuron) in the output layer, $r = 0.6349$, $p < 10^{-50}$.

CNN and optimal computations measured by trial correlations

A complementary analysis is to obtain a quantitative measure of the correspondence between neurons' responses and the decision variable of a BIO. We correlated the BIO decision variable (log-likelihood, see methods) and each individual neuron across trials. We extracted the log-likelihood ratio for the BIO across 1000 unseen test images (500 target absent, 400 valid, 100 invalid) and used the same images for the CNN, and obtained the response of each neuron. For the CNN's output layer, we used the SoftMax response. We computed 1.8 million correlations across the ten networks.

The correlation between the output variables of these two models is 0.68 ± 0.015 across the ten models. A map of layer-wise neuron-BIO correlations, as well as the mean absolute and the maximum absolute neuron-BIO correlations, is shown in Figure 8(g) and (h). The convolution layers have nearly the same mean absolute neuron-BIO correlations with a slow increase in the max absolute neuron-BIO correlations from the first convolution layer to the flat layer. The dense layer has both max absolute and

mean absolute correlations significantly greater than previous layers ($p < 10^{-10}$ across ten networks), while the max absolute correlations are not significantly different between the dense and output layers ($p = 0.2451$), but the mean absolute correlation of the dense layer is significantly lower than that in the output layer ($p < 10^{-5}$).

Correlations between neuronal properties of dense layer neurons and the weights they project into the output layer.

To understand the role of the individual neuronal response properties in the final behavioral decision, we computed the correlations between the neuronal target AUROC (Fig 4g), cue AUROC (Fig 5g), and neuronal cueing (AUROC valid minus AUROC No Cue; Fig 8i) of dense layer target-cue neurons and the weights between these neurons and the output neuron used to make target present decisions. We found that across ten networks, the neuronal cue AUROC has a small correlation ($r = -0.1011$, $p = 0.0001$), and the target AUROC and neuronal cueing have higher correlations ($r = 0.5577$ for target AUROC; $r = 0.6349$ for neuronal cueing; $p < 10^{-50}$ for both) with the weights projected on to the target presence neuron in the output layer.

Supplementary analyses

The presented analyses are for a specific network architecture. We further validate the generalization of the results for many variations of the network, including the effect of regularization, the size of the dense layer, the overall number of layers in the network, convolution stride size, kernel size, and varying the noise amplitude. The results are discussed in the Supplementary material (Figure S.4-S.10).

DISCUSSION

For over forty years, researchers have studied visual attention by characterizing the influences of cues on target detection. Although there are numerous theories explaining the cue facilitatory effects in terms of psychological concepts[1,42–44], model-based mechanisms (Bayesian decision weighting by priors[7,18,37,38,41,45,46], probabilistic population codes[47], divisive normalization[48], neural networks[49,50]), or neuronal response properties (correlations[3], contrast response function[19], tuning curve amplitude[4], there are few endeavors that start with the image as input, and characterize a whole system-wide visual processing and computations of the target, cue, integration across locations, and how these lead to behavioral performance and cueing effects. Here, we used a five-layer CNN whose representation can be more easily related to actual neural systems than other modeling approaches[41]. We trained the networks on images with the target, distractor, cues, and luminance noise and optimized target detection. The resulting networks give rise to cueing effects that are similar to those obtained in humans and other animals. We analyzed the 1.8M neurons across ten networks to identify the computational stages of each of the layers, understand how these compare to those of an optimal model (BIO), and relate our findings to existing physiological measurements by previous studies.

Cueing effects can arise as an emerging property of task optimization

A critical aspect of our CNN approach is that we did not incorporate into the network any specialized attention mechanism[49,50] that operates at the cued location nor any explicit resource limitation[36] that had to be shared across locations. In the training of the network parameters to maximize overall task accuracy (proportion of correct decisions) across all trial types, the CNN learned to use the presence of the cue to influence the processing of the target, resulting in cueing effects at the decision stage. This result implies

that many of the simpler biological organisms[20,23–25] showing behavioral cueing effects in recent years might be explained as an emergent property of perceptual task optimization or feedforward networks.

Computational stages along the CNN layers

Our findings showed that the early layers in the CNN are retinotopic, tuned to individual locations, and do not integrate target and cue information. As we move deeper into the network (flat layer and dense layer), the integration of cue and target information at each location emerges first, followed by the integration of information across the two locations (dense layer). Figure 3 shows a summary of the computational stages at each of the layers. In terms of the influence of the cue, our findings show increasing effects on the neural population responses to the target and sensitivity as processing progresses through the layers of the network, consistent with that reported in neurophysiological studies[51,52]. The convolution layers show a small effect, akin to that reported in V1 primary visual cortex single-cell measurements.[4,53] The dense layer shows strong modulation by the cues more similar to the superior colliculus (SC) [53,54] and lateral intraparietal areas (LIP)[53,55]. The flat layer neurons show intermediate effects of the cue, perhaps more similar to those reported in area V4[53]. The increased influence of the cue on the responses to the target occurs even though the sensitivity to the cue itself diminishes with processing across the network layers (Figure 5f).

We also show a correspondence between neuronal properties of the CNN and the stages of the BIO (Figure 3) and a higher correlation with the BIO decision variable as processing progresses through the network, mirroring some experimental work showing correlates of BIO log-likelihood ratio in the lateral intraparietal (LIP) region of the brain[56]. However, there were important differences between the BIO and CNN. The BIO shows defined transitions to target-cue integration and location integration stages,

while the CNN had a distribution of neuronal properties showing different degrees and types of integration, varying sensitivities, and more nuanced transitions across layers.

CNN Neuronal types and subtypes

Through ROC analysis of each neuron for target and cue detection, we tagged the CNN neurons into four types of neurons: cue detection neurons, target detection neurons, target-cue integration (local) neurons, and location-integration (global) neurons. For each type of neuron, we identified different subtypes. These CNN neuron subtypes can be compared to single-cell animal studies and might also serve as theoretical predictors of neurons still not discovered or noted by neuroscientists.

We found neurons that were excited by the presence of the target but other neurons that were inhibited. We also found target-cue integration neurons that were cue and target excitatory but also found other combinations: target excitatory/cue inhibitory, target inhibitory/cue excitatory and target and cue inhibitory. Arguably, the more commonly reported cell type is the target and cue excitatory neuron, whose mean response increases when the cue is present. However, recent studies show a handful of cue-inhibitory neurons in the mouse and primate (SC; Figure 2a in [20]), although they are not discussed. In terms of location integration, the CNN resulted in opponent location neurons that were excited by the cue at one location but inhibited at the other location. We do not know of a report of neurons with such properties.

Mechanisms mediating cue effects in CNNs

An important goal is to understand the mechanism by which the cue influences target sensitivity in the CNN and how it leads to the behavioral cueing effect. Our analyses show that a process by which the

network generates the cueing effects is by influencing the mean responses to the target and distractor with the information in the cued location more strongly than the uncued location. This is the main mechanism in the BIO model using priors equal to the cue validity. The CNN implements the weighting mechanism by increasing the mean neuronal response to visual information (target and distractor) from the cued location relative to the uncued location (Figure 8b and S.3) before integration across locations. This mechanism leads to larger target sensitivity and the cueing effects after location integration, like the BIO. In addition, the CNN shows a second mechanism that changes target sensitivity at the cued location relative to the uncued location before the integration across locations. This mechanism is not present in the BIO. Our investigations showed that this effect arises due to an interaction between the cue-related increase in mean neuronal responses and the thresholding of the ReLU activation function of the network. The increase in a neuronal mean response reduces the detrimental impact of the thresholding on target sensitivity. Is this 2nd mechanism required to generate the network cueing effects? A monotonically increasing transform applied to both distributions doesn't change the AUROC between the two distributions[57]. Replacing the ReLU activation function (which is not monotonically increasing if the input is less than 0) with a monotonically increasing function (tanh activation function) at the flat layer (Supplementary Figure S.11), we found that this 2nd mechanism is not required for the network to generate cue effects. The CNN without the flat layer target sensitivity mechanism (tanh activation function) still generated cueing effects comparable to the CNN with ReLU. However, there might be other important functions for the ReLU cue mechanism that have not been explored in this paper.

Critically, a comparison of our analyses with the 80 % predictive cue to a control condition with a 50 % predictive cue confirmed that the cue-induced influences on target responses are causally related to the predictive probability of the cue.

Limitations of the present approach

There are various shortcomings in our work. Our neuronal analysis pipeline is difficult to scale up to large, deeper CNNs or to higher dimensional tasks where multiple features across multiple locations need to be integrated. The present AUROC analysis was aided by the knowledge of the features necessary for performing the task, which might not be available for more complicated tasks. Another caveat in our analysis is the lack of feedback connections[58,59], which are particularly important[53,60] in selecting task-relevant and inhibiting task-irrelevant locations but were not incorporated into the current network aside from the feedback during training through backpropagation. Future research should evaluate the possible benefits of incorporating feedback mechanisms and recurrence and assess the neuronal properties that arise from such architectures.

Concluding remarks

We presented an image-to-behavior CNN-based framework to understand how a simple biological organism might generate behavioral cueing effects, the resulting neuronal response properties across the hierarchy of visual processing, the mediating mechanisms leading to the cueing effect, and the similarity and differences with optimal computations. The framework can be of potential use to generate theoretical predictions about neuron subtypes that could guide neurophysiologists in their empirical measurements. Since Bayesian models and neural networks are widely used models in neuroscience and psychology, even in domains outside attention, our results and methods are of general interest to these fields.

Methods and Materials

Human Data

We used the Human subjects data from the Posner task in [61]. The demographics and experiment design are described in the following sections.

Subjects

Eighteen undergraduate students from the University of California at Santa Barbara were recruited as subjects. The subjects all have normal or corrected-to-normal vision. Eleven subjects were women, and seven were men, all in the age range of 19-22. The subjects were naïve to the hypotheses of the study. The studies followed the human subject guidelines provided by the Institutional Review Board at the University of California, Santa Barbara. Informed consent was obtained from each participant.

Eye Tracking and Experiment Setup

The left eye of each participant was tracked using an SR Research Eyelink 1000 Tower Mount eye tracker sampling at 1000 Hz. Before each 100-trial session, a 9-point calibration and validation were done, with a mean error of no more than 0.5 degrees of visual angle. In each experiment, participants were asked to fixate at a central cross. If they moved their eyes more than 1 degree from the fixation cross, an error message would appear, and the trial would abort and restart with a new stimulus.

Stimuli and Psychophysics

Stimuli for all experiments were generated using MATLAB. The mean pixel value of the image backgrounds was 128. The minimum luminance was 0.08 cd/m^2 and the maximum 114.7 Cd/m^2. The luminance was linear with greyscale. The display had a resolution of 1920 by 1080 pixels. The viewing

distance was 75 cm. The psychophysics experiment was written in MATLAB using the Psychophysics Toolbox.

Task

We implemented a forced-fixation, yes-no version of the classic Posner Cueing Task (Posner, 1980) with varying levels of cue contrasts. Before each stimulus appeared on the screen, a blank screen with a fixation cross appeared on the screen for 250 ms. Participants were asked to fixate at the fixation cross at the center of the screen. Next, the stimulus was presented on the screen for 250 ms, with the fixation cross at the same location. An error message appeared if participants moved their fixation more than 1 degree away from the fixation cross, either on the pre-stimulus screen or while a stimulus was presented. When broken fixations were detected, the trial was restarted with a different stimulus. Following the stimulus presentation, a response screen was shown, asking participants to press the right arrow key if the target was present in the stimulus and the left arrow key otherwise. The response was recorded, and participants were asked to press the enter key to move on to the next stimulus.

Participants were shown 400 stimuli in total over four blocks of 100 stimuli each. The stimulus contained a target with a 50% probability. When a target was present, it was present with the box with 80% probability and on the other side of the box with 20% probability. These trials are called valid cue trials and invalid cue trials, respectively. The box cue appeared on either side with equal probability. The target angle was sampled from a Gaussian distribution with a mean of 15 degrees and a standard deviation of 3.5 degrees. The distractor angle was sampled from a Gaussian distribution with a mean of 7 degrees and a standard deviation of 3.5 degrees. The noise contrast was 0.14, while the target and the distractors had a contrast of 0.18.

CNN Architecture

We used a feedforward convolutional neural network with three convolutional layers with 8, 24, and 32 kernels, respectively, and kernel strides of size 3 by 3; and two fully connected layers with 256 and 2 (output layer). ReLU activation function was used for all layers except the output layer, where SoftMax activation was used.

The ReLU activation function is defined as:

$$ReLU(x) = \max(0, x) \tag{2}$$

While the SoftMax function for a vector $\boldsymbol{z} = (z_1, z_2, \dots, z_k)$ is defined as:

$$SoftMax(\boldsymbol{z})_i = \frac{e^{z_i}}{\sum_{j=1}^{K} e^{z_j}} \; for \; i = 1,2, \dots, K. \tag{3}$$

Training details

The training was done via gradient descent to minimize cross-entropy loss, with batch sizes of 32, until convergence, which was typically around ten epochs.

The cross-entropy loss for CNN output $\hat{\boldsymbol{y}}$ and ground truth labels $\boldsymbol{y}$ for a batch of size 32 is defined as:

$$\sum_{n=1}^{32} y_{n1} \log \hat{y}_{n1} + (1 - y_{n2}) \log (1 - \hat{y}_{n2}) \tag{4}$$

Where $\hat{y}_{ni}$ is the output of the ith neuron in the CNN to the nth image in the batch, and y_{ni} is the ith component of the nth ground truth one-hot vector in the batch.

We used 10000 training images, with half of the images containing no target, 4000 images with valid target, and 1000 images with invalid target for the main result with 80% cue validity. We tested on ten sets of 1000 images with the same relative proportion of the various stimuli types. For the models trained with 50% predictive cues, the number of target absent images was kept the same (50%), but the numbers of images with valid target and invalid target were both increased to 2500 each.

Ten different models were trained from scratch with random weight initializations for 80% predictive cues, and another ten different models were trained from scratch with random weight initializations for 50% predictive cues. No training history or weights were shared between any two models.

Statistical Efficiency

The external Gaussian noise added to the stimuli shown to the BIO and the CNN was varied to obtain the best fit to human observers. The goodness of fit was determined using the Akaike Information Criterion (AIC). The statistical efficiency of an observer (humans or CNN) is then calculated as the ratio of the variance of the noise added to the observer and the BIO.

Goodness of fit measure for the models

The negative log-likelihood of the model predictions given human data was minimized to determine the best-fit model, while the external noise was varied using grid search. The Akaike Information Criterion was calculated using the log-likelihoods of the best-fit models using:

$$AIC = 2 * k - 2 * \log(\hat{L}) \tag{5}$$

Where $\hat{L}$ denotes the likelihood of getting values equal to the model predictions if they were sampled from Gaussian distributions with means equal to the human performances and standard deviation equal to the human standard errors, and k denotes the number of parameters varied ($k = 1$).

ROC Analysis

Target vs. distractor at each location

The CNN was shown ten sets of 500 noisy images, each with no box and containing the following: only the target at the right location, only the distractor at the right location, only the target at the left location, and only the distractor at the left location. The response of each neuron in the network to these stimuli was recorded. Using these responses, we constructed ROC curves for detecting whether the target or the distractor was present at each location.

Unresponsive neurons have only one point on their ROC curve since their response range is just one point. Therefore, the area under the ROC of such a neuron could technically be classified as zero, but since a zero AUROC also indicates perfect inhibition, we choose the convention to assign an AUROC of 0.5 to such neurons since they can't discriminate between any two categories.

Cue presence vs. absence at each location

The CNN was shown ten sets of 500 noisy images, each with no target or distractor and containing the following: only the box at the right location, image with just the background plus noise, only the box at the left location, image with just the background plus noise. The response of each neuron in the network to these stimuli was recorded. Using these responses, we constructed ROC curves for detecting whether the box cue was present or absent at each location.

Obtaining p-values of the neuronal AUROC

Each AUROC for each neuron was computed on ten sets of 1000 images each (500 for present, 500 for absent). We used the mean and standard error obtained over these ten iterations to perform one-sided t-tests of being different from 0.5. Multiple comparison correction was performed, correcting for the False Discovery rate at $\alpha = 0.01$.

Computing the AUROC for stages of the BIO

The BIO computation has three stages: target/cue detection, target-cue integration, and location integration. These stages and the computations involved in each stage are depicted in Figure 3. Given any image, each stage outputs a value. For the same images used for obtaining the neuronal AUROCs, we compute the values from each computational stage and use these values to construct ROC curves for detecting the cue or target presence at each location.

Comparing the cueing distributions of models trained on 80% and 50% predictive cues

We obtained the distributions of neuronal cueing across ten models, each for 80% and 50% predictive cues. These distributions were then compared using the two-sample Kolmogorov-Smirnov test (KS test). The KS test was implemented using the MATLAB function *kstest2.*

Data and Materials Availability

All data gathered from this study, as well as images used in the experiments are available upon request from the corresponding author.

Code Availability

All scripts, including training, testing, and analyzing the models and running and analyzing the psychophysics experiments, are available upon request from the corresponding author.

Acknowledgments

The Institute for Collaborative Biotechnologies through US Army Research Office Contract W911NF-19-D-001. (MPE)

US Army Research Office Cooperative Agreement W911NF-19-2-0026. (MPE)

Guggenheim Foundation Fellowship. (MPE)

Regents in Dynamical Neuroscience Fellowship. (SS)

The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the US Government. The US Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation herein.

Author Contributions

S.S. and M.P.E. conceived the study, designed the experiments, and interpreted the model and human results. S.S. implemented all the models and collected data from human observers. W.Y.W. provided expertise in CNN architecture selection. All the authors contributed to the writing and editing of the manuscript.

Competing Interests Statement

The authors declare no competing interests.

References

1. Posner, M. I. Orienting of attention. *Q J Exp Psychol* **32**, 3–25 (1980).
2. Carrasco, M. Visual attention: The past 25 years. *Vision Res* **51**, 1484–1525 (2011).
3. Cohen, M. R. & Maunsell, J. H. R. Attention improves performance primarily by reducing interneuronal correlations. *Nature Neuroscience* **12**, 1594–1600 (2009).
4. McAdams, C. J. & Maunsell, J. H. R. Effects of Attention on Orientation-Tuning Functions of Single Neurons in Macaque Cortical Area V4. *J. Neurosci.* **19**, 431–441 (1999).
5. Luck, S. J. *et al.* Effects of spatial cuing on luminance detectability: Psychophysical and electrophysiological evidence for early selection. *Journal of Experimental Psychology: Human Perception and Performance* **20**, 887–904 (1994).
6. Dosher, B. A. & Lu, Z. L. Mechanisms of perceptual attention in precuing of location. *Vision research* **40**, 1269–1292 (2000).
7. Eckstein, M. Probabilistic Computations for Attention, Eye Movements, and Search. *Annu. Rev. Vis. Sci.* **3:18.1–18.24**, (2017).
8. Carrasco, M., Ling, S. & Read, S. Attention alters appearance. *Nat Neurosci* **7**, 308–313 (2004).
9. Pestilli, F. & Carrasco, M. Attention enhances contrast sensitivity at cued and impairs it at uncued locations. *Vision research* **45**, 1867–1875 (2005).
10. Carrasco, M., Penpeci-Talgar, C. & Eckstein, M. Spatial covert attention increases contrast sensitivity across the CSF: support for signal enhancement. *Vision Res* **40**, 1203–1215 (2000).

11. Golla, H., Ignashchenkova, A., Haarmeier, T. & Thier, P. Improvement of visual acuity by spatial cueing: a comparative study in human and non-human primates. *Vision Research* **44**, 1589–1600 (2004).
12. Yeshurun, Y., Montagna, B. & Carrasco, M. On the flexibility of sustained attention and its effects on a texture segmentation task. *Vision Res* **48**, 80–95 (2008).
13. Herrmann, K., Montaser-Kouhsari, L., Carrasco, M. & Heeger, D. J. When size matters: attention affects performance by contrast or response gain. *Nat Neurosci* **13**, 1554–1559 (2010).
14. Eriksen, C. W. & St. James, J. D. Visual attention within and around the field of focal attention: A zoom lens model. *Attention, Perception, & Psychophysics* **40**, 225–240 (1986).
15. Dosher, B. A. & Lu, Z.-L. Noise Exclusion in Spatial Attention. *Psychol Sci* **11**, 139–146 (2000).
16. Burgess, A., Wagner, R., Jennings, R. & Barlow, H. B. Efficiency of human visual signal discrimination. *Science* **214**, 93–94 (1981).
17. Lu, Z.-L. & Dosher, B. A. External noise distinguishes attention mechanisms. *Vision Research* **38**, 1183–1198 (1998).
18. Eckstein, M. P., Shimozaki, S. S. & Abbey, C. K. The footprints of visual attention in the Posner cueing paradigm revealed by classification images. *J Vis* **2**, 25–45 (2002).
19. Reynolds, J. H. & Chelazzi, L. Attentional modulation of visual processing. *Annu. Rev. Neurosci* **27**, 611–647 (2004).
20. Wang, L., Herman, J. P. & Krauzlis, R. J. Neuronal modulation in the mouse superior colliculus during covert visual selective attention. *Sci Rep* **12**, 2482 (2022).

21. Srinath, R., Ruff, D. A. & Cohen, M. R. Attention improves information flow between neuronal populations without changing the communication subspace. *Current Biology* **31**, 5299-5313.e4 (2021).
22. Ruff, D. A. & Cohen, M. R. Simultaneous multi-area recordings suggest that attention improves performance by reshaping stimulus representations. *Nat Neurosci* **22**, 1669–1676 (2019).
23. Eckstein, M. P. *et al.* Rethinking human visual attention: Spatial cueing effects and optimality of decisions by honeybees, monkeys and humans. *Vision Res.* (2013) doi:10.1016/j.visres.2012.12.011.
24. Quest, M., Rinnert, P., Hahner, L. & Nieder, A. Exogenous and endogenous spatial attention in crows. *Proc. Natl. Acad. Sci. U.S.A.* **119**, e2205515119 (2022).
25. Saban, W., Sekely, L., Klein, R. M. & Gabay, S. Endogenous orienting in the archer fish. *Proceedings of the National Academy of Sciences* **114**, 7577–7581 (2017).
26. Geisler, W. S. *Ideal Observer analysis*. (MIT Press, 2003).
27. Barlow, H. B. The absolute efficiency of perceptual decisions. *Philos. Trans. R. Soc. Lond., B, Biol. Sci* **290**, 71–82 (1980).
28. Abbey, C. K. & Eckstein, M. P. Classification images for detection, contrast discrimination, and identification tasks with a common ideal observer. *J Vis* **6**, 335–355 (2006).
29. Burge, J. Image-Computable Ideal Observers for Tasks with Natural Stimuli. *Annual Review of Vision Science* **6**, 491–517 (2020).
30. Wei, X.-X. & Stocker, A. A. A Bayesian observer model constrained by efficient coding can explain ‘anti-Bayesian’ percepts. *Nat Neurosci* **18**, 1509–1517 (2015).

31. Ma, W. J., Beck, J. M., Latham, P. E. & Pouget, A. Bayesian inference with probabilistic population codes. *Nat Neurosci* **9**, 1432–1438 (2006).
32. Palmer, J., Verghese, P. & Pavel, M. The psychophysics of visual search. *Vision Res* **40**, 1227–1268 (2000).
33. Shimozaki, S. S., Schoonveld, W. A. & Eckstein, M. P. A unified Bayesian observer analysis for set size and cueing effects on perceptual decisions and saccades. *J Vis* **12**, (2012).
34. Eckstein, M. P., Thomas, J. P., Palmer, J. & Shimozaki, S. S. A signal detection model predicts the effects of set size on visual search accuracy for feature, conjunction, triple conjunction, and disjunction displays. *Percept Psychophys* **62**, 425–451 (2000).
35. Verghese, P. Visual search and attention: a signal detection theory approach. *Neuron* **31**, 523–535 (2001).
36. Palmer, J. Set-size effects in visual search: the effect of attention is independent of the stimulus for simple tasks. *Vision Res* **34**, 1703–1721 (1994).
37. Vincent, B. T. Bayesian accounts of covert selective attention: A tutorial review. *Atten Percept Psychophys* **77**, 1013–1032 (2015).
38. Yu, A. J. & Dayan, P. Uncertainty, neuromodulation, and attention. *Neuron* **46**, 681–692 (2005).
39. Yu, A. J., Dayan, P. & Cohen, J. D. Dynamics of attentional selection under conflict: toward a rational Bayesian account. *J Exp Psychol Hum Percept Perform* **35**, 700–717 (2009).
40. Green, D. M. & Swets, J. A. *Signal Detection Theory and Psychophysics*. (Peninsula Pub, 1989).

41. Eckstein, M. P., Peterson, M. F., Pham, B. T. & Droll, J. A. Statistical decision theory to relate neurons to behavior in the study of covert visual attention. *Vision Res* **49**, 1097–1128 (2009).

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

43. Treisman, A. & Gelade, G. A feature-integration theory of attention. *Cogn Psychol* **12**, 97–136 (1980).

44. Chun, M. M. & Jiang, Y. Contextual Cueing: Implicit Learning and Memory of Visual Context Guides Spatial Attention,. *Cognitive Psychology* **36**, 28–71 (1998).

45. Eckstein, M. P., Pham, B. T. & Shimozaki, S. S. The footprints of visual attention during search with 100% valid and 100% invalid cues. *Vision Res* **44**, 1193–1207 (2004).

46. Yu, A. J. & Dayan, P. Inference, Attention, and Decision in a Bayesian Neural Architecture. in *Advances in Neural Information Processing Systems* vol. 17 (MIT Press, 2004).

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

48. Reynolds, J. H. & Heeger, D. J. The Normalization Model of Attention. *Neuron* **61**, 168–185 (2009).

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

50. Sörensen, L. K. A., Zambrano, D., Slagter, H. A., Bohté, S. M. & Scholte, H. S. *Leveraging spiking deep neural networks to understand the neural mechanisms underlying selective attention*. http://biorxiv.org/lookup/doi/10.1101/2020.12.15.422863 (2020) doi:10.1101/2020.12.15.422863.

51. Cook, E. P. & Maunsell, J. H. R. Attentional Modulation of Behavioral Performance and Neuronal Responses in Middle Temporal and Ventral Intraparietal Areas of Macaque Monkey. *J. Neurosci.* **22**, 1994–2004 (2002).

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

53. Noudoost, B., Chang, M. H., Steinmetz, N. A. & Moore, T. Top-down control of visual attention. *Curr. Opin. Neurobiol* **20**, 183–190 (2010).

54. Ignashchenkova, A., Dicke, P. W., Haarmeier, T. & Thier, P. Neuron-specific contribution of the superior colliculus to overt and covert shifts of attention. *Nature Neuroscience* **7**, 56–64 (2004).

55. Bisley, J. W. & Goldberg, M. E. Neuronal activity in the lateral intraparietal area and spatial attention. *Science* **299**, 81–86 (2003).

56. Yang, T. & Shadlen, M. N. Probabilistic reasoning by neurons. *Nature* **447**, 1075–1080 (2007).

57. Hand, W. J. K., David J. *ROC Curves for Continuous Data*. (Chapman and Hall/CRC, 2009). doi:10.1201/9781439800225.

58. Kreiman, G. & Serre, T. Beyond the feedforward sweep: feedback computations in the visual cortex. *Ann. N. Y. Acad. Sci.* **1464**, 222–241 (2020).

59. Lamme, V. A. & Roelfsema, P. R. The distinct modes of vision offered by feedforward and recurrent processing. *Trends Neurosci* **23**, 571–579 (2000).

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

61. Srivastava, S., Wang, W. & Eckstein, M. P. A Feedforward Convolutional Neural Network with a Few Million Neurons Learns from Images to Covertly Attend to Cues and Context like Humans and an Optimal Bayesian Observer. Preprint at https://doi.org/10.31234/osf.io/tymfn (2021).