Bridging Neurons and Behavior in a Convolutional Neural Network with Emergent Human-like Covert Attention

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

Covert visual attention allows the brain to select different regions of the visual world without eye movements. Cues predictive of a target location orient covert attention and improve perceptual performance. How entire neuronal populations represent and integrate target, cues, and location information to result in behavioral signatures of covert attention is not understood. We analyze 1.8M neurons of feedforward Convolutional Neural Networks (CNNs) that show human-like attentional cueing effects. Consistent with neurophysiology, we show early layers with retinotopic neurons separately tuned to target or cue, and later layers with neurons with joint tuning and increased influence of the cue on target responses. We show cue-inhibitory and location-opponent neurons that are unreported by neurophysiologists and identify the neuronal circuits that result in neuronal cueing effects. The cue influences the mean neuronal responses and changes target sensitivity through three mechanisms: cue-weighted summation and opponency across locations, and interaction with the thresholding Rectified Linear Unit (ReLU). Some CNN computational stages mirror a Bayesian ideal observer (BIO), but with more gradual transitions, while the opponency and ReLU interaction are distinct from the BIO. Together, the findings establish a likely system-wide characterization of the brain computations that mediate the behavioral signatures of covert attention.

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 location (valid cue trials) vs. when appearing at the uncued location (invalid cue trials).

There is a vast literature focused on understanding the influences of covert attention (cueing) using behavioral measures[2,5–7]. Early research focused on a descriptive understanding of attention in cueing tasks based on analogies such as a zoom or a spotlight[8,9]. Subsequent studies in the last two decades have used psychophysical results across stimuli conditions (e.g., varying target contrast[10–12], image noise[6,13,14]) and, by assuming an underlying quantitative model, made inferences about the mechanisms by which attention changes the processing of visual information. These studies have proposed that attention can exclude external noise[13] (or improve the ability to integrate signal information, e.g., template sampling efficiency[15]), result in signal enhancement[14], internal noise reduction[14], arise from interactions between a multiplicative attention field and a divisive suppression[10,12,16], and optimize the weighting of cued and uncued sensory information[17]. 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[18] of neurons. Recently, studies have incorporated neurally-inspired attention mechanisms into convolutional neural networks[19,20] to assess the impact on model perceptual accuracy.

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 that can be mapped to neural circuits and give rise to behavioral cueing effects. The models 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[21–23]. How to map these measurements into computational stages is not straightforward[24–26].

The purpose of this paper is to analyze neurons of a feedforward neural network that learns to covertly attend in a cueing paradigm. We aim 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[1,17,27] and monkeys[27] to crows[28], archerfish[29], and bees[27]. Our analysis seeks to answer the following 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 of the existing vast literature[21,30–33]? How might a neuron showing a cueing effect be built from the combination of input neurons from earlier layers? Are there any novel predictions about neuronal properties that might be inferred from the analyses of CNN units?

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. A cue (a box around one of the two lines) indicates the likely target location when present. The network has no explicit knowledge about the cue nor its predictive probability. We also do not explicitly incorporate any attention mechanism (e.g., gain[20], prior[17,34], feature similarity[19]) into the network to act differently at the cued location. We compare the properties of 180K neuronal units at the different layers of ten trained CNNs (1.8 M neurons) 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[15,35–39] and as a benchmark model of visual search[40–44] and attention[7,17,42,45–48]. Understanding how a large population of network neurons represents task-relevant information relative to the computational stages 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 not yet reported in the literature.

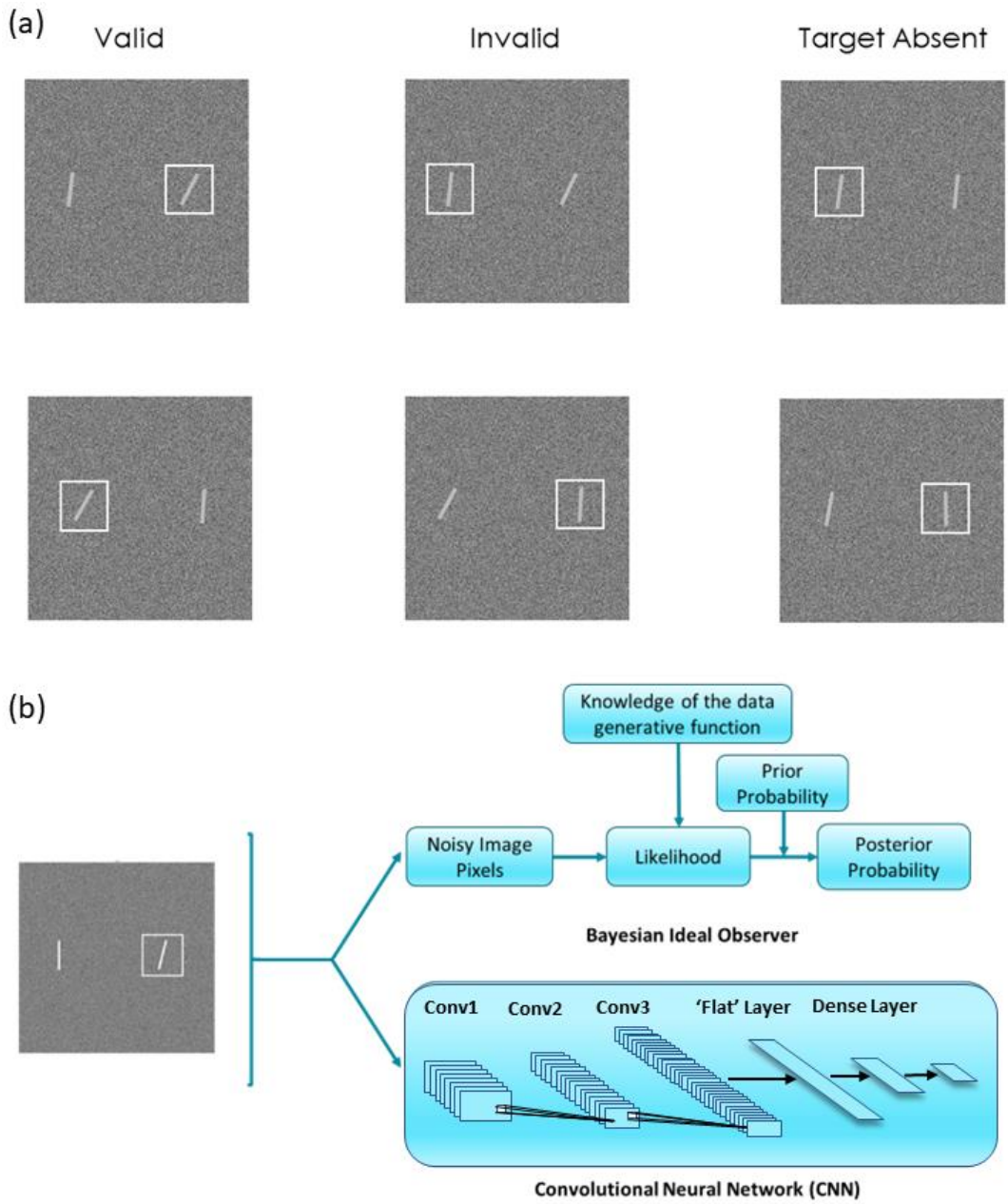


Figure 1 | Example stimuli, model flowchart(a) Valid trials (left column) have the cue (box) appearing with the target (tilted line) and comprise 80% of the target present trials; invalid trials (middle column) have the cue appearing opposite to the target and comprise 20% of target-present trials, while 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. (b) 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 referred to as the Flat layer.

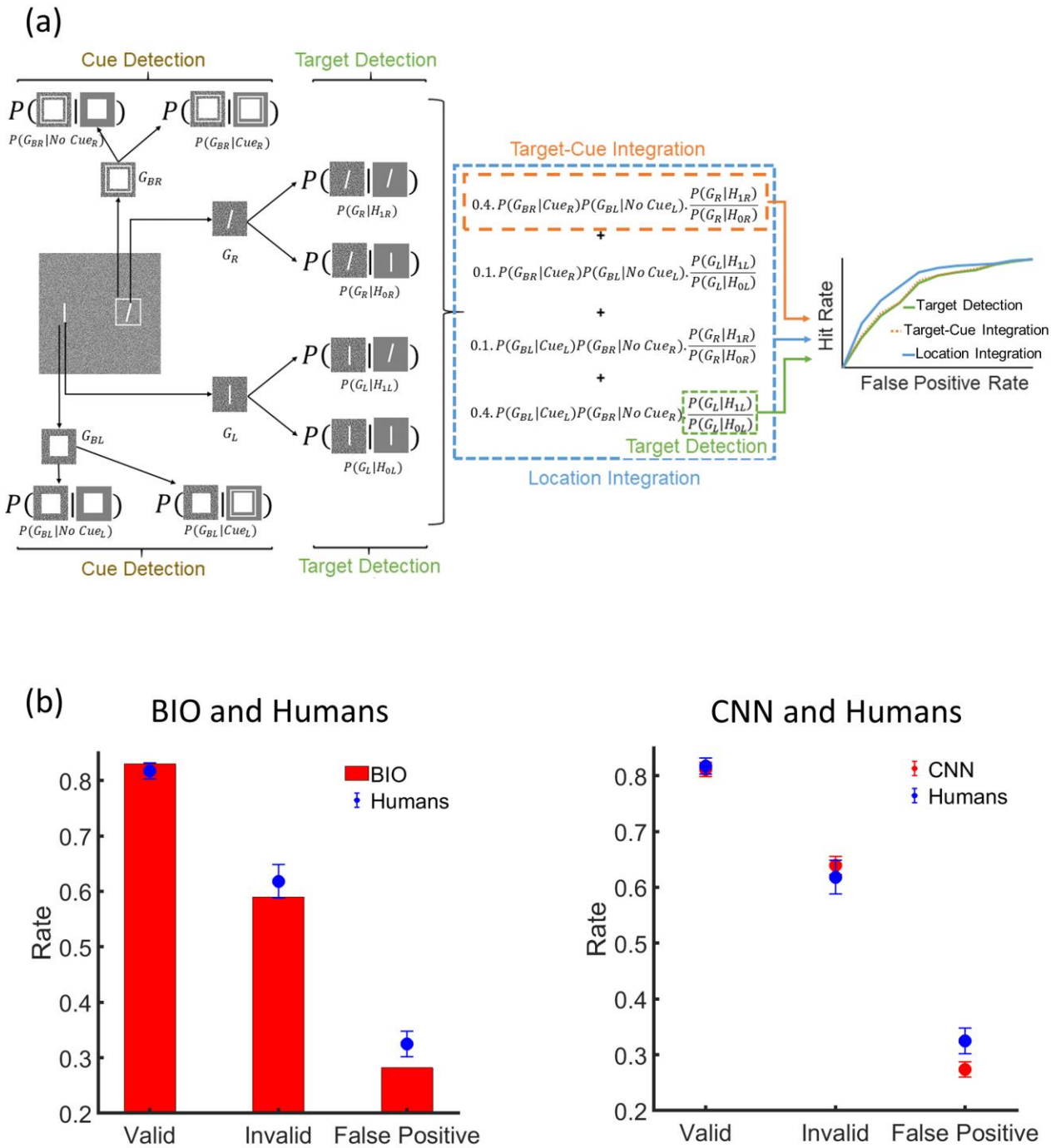


Figure 2| Computational stages in the BIO and decision accuracy of humans, BIO, and CNN. (a) Computational stages of a BIO for a Posner cueing paradigm. The BIO has three stages: the cue/target detection stage (brown/green) where noisy stimulus pixels G_{Bi} (where subscript B refers to surrounding background) and G_i are compared to the template images of the cue and the target to compute the likelihood of cue ($P(G_{Bi}|Cue_i)$ or $P(G_{Bi}|No\ Cue_i)$)or target presence and absence $P(G_i|H_{ji})$, given the noisy stimulus pixels at each location, after which we compute the ratio of the target and distractor likelihoods, LR= $P(G_i|H_{1i})/P(G_i|H_{0i})$; the target-cue integration stage (orange) where the likelihoods for the cue and the target are multiplied with the prior

probabilities π_i of target presence and absence given cue presence and absence: $\pi_i . P(G_{Bi}|Cue_i). P(G_{B!i}|No\ Cue_{!i}). P(G_i|H_{1i})/P(G_i|H_{0i})$, where, !i denotes the location other than the ith location; and the location integration stage (blue) where the local target-cue integration stages are summed across both locations. 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. See Methods for more details. (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 eighteen human observers, and red error bars represent the standard error of the mean across ten trained CNNs.

RESULTS

Behavioral Cueing Task

The task required that humans and models detect 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 1a) 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 cue had an equal probability of being present at either location and was presented simultaneously with the test image (target/distractor). White luminance noise was added to the images (see Methods for details).

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 1b). Each convolution kernel was of size 3 by 3. The flattened third convolution layer is referred to as the flat layer and the first fully-connected layer is referred to as the dense layer throughout the paper. 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. The CNNs did not use 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, see methods for details).

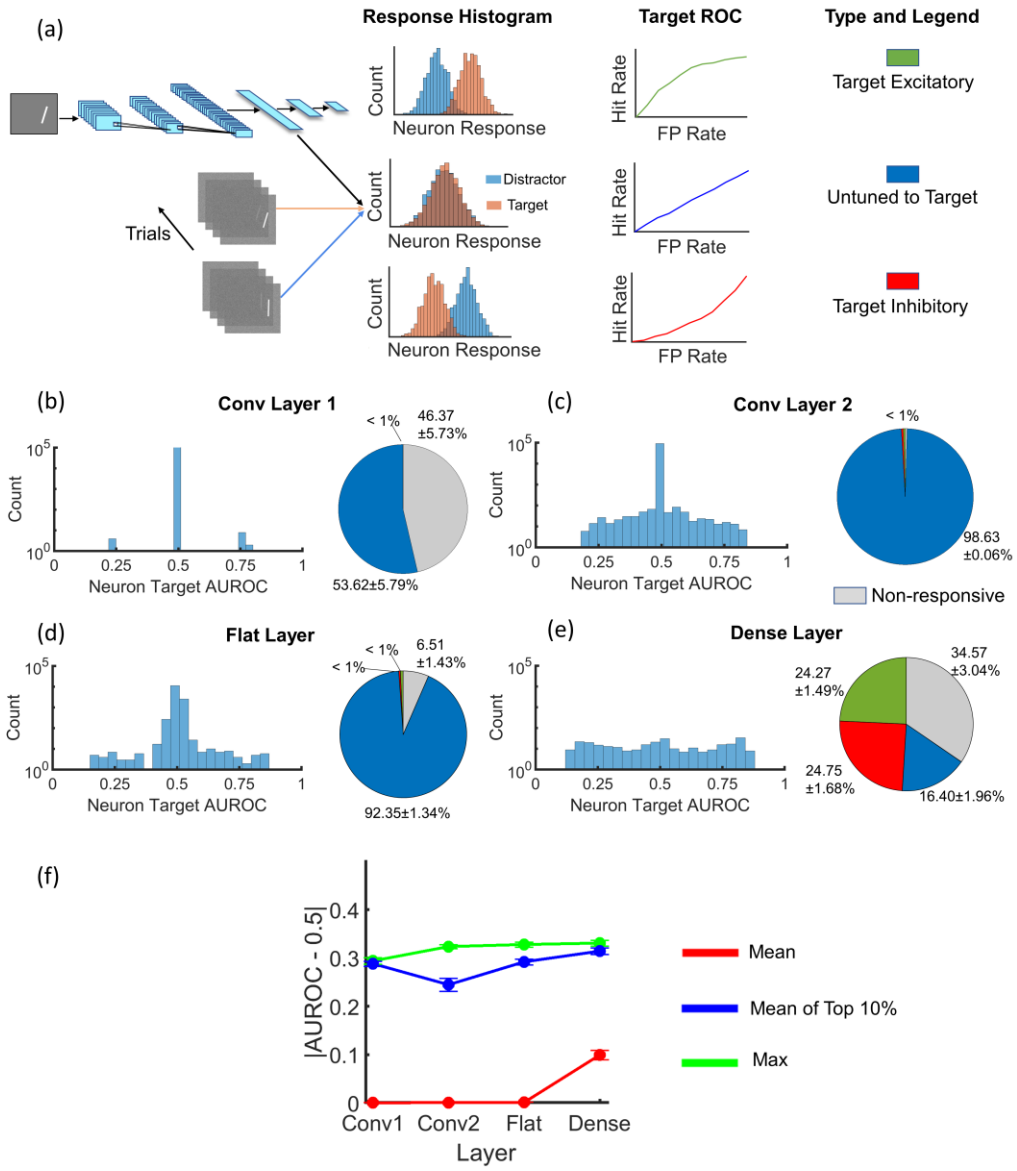
Bayesian Ideal Observer

The image-computable Bayesian ideal observer for the cueing paradigm[17,42,47] (Figure 2a) 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. The model often omits the process of cue detection and assumes it is detected with no error but has been recently generalized to incorporate the detection of the cue which is also presented in additive luminance noise[49]. The model has full knowledge of the generative function of the data (see Methods and Supplementary Section S.1 for details).

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 humans viewed (see Supplementary section S.2 for statistical efficiency). To make the overall performances of the CNN and BIO fit that of humans (minimize the Akaike Information Criterion, AIC;

see methods for details), we increased the external white noise of the images separately for each model. The CNNs, trained to optimize overall accuracy across all trial types, and the BIO, show a cueing effect[17,47] (higher hit rates for valid vs. invalid cue trials; Figure 2b) comparable to that of eighteen human observers in the two-location task with the 250 ms presentation and a central fixation (see methods).



.

Figure 3 | 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.

Statistical characterization of neuronal properties

In the next sections, we analyze the neurons' responses to the target, cue, the target-cue integration, integration across locations, the influence of the cue on target sensitivity, and neural circuits resulting in neurons that manifest cueing effects. 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.

Target Sensitivity

We measured the sensitivity of a neuron to the target by generating a Receiver Operating Characteristic curve (ROC) using response distributions of each neuron in the CNN to separate sets of images that contained the target present (50% probability) or absent (distractor present and no cue in the images (Figure 3a)). We evaluated each of the two locations separately.

Figure 3a shows sample histograms of three neuron types encountered in the network at the dense layer. Some neurons responded more strongly when the target was present (target-excitatory, T+). Other neurons consistently diminished their response when the target was present (target-inhibitory, T-), while a third type was insensitive to the presence of the target (Figure 3a, middle histogram). We use the area under the ROC curve (AUROC)[50] analysis to classify the types of neurons: target-excitatory (AUROC > 0.5, T+), target-inhibitory (AUROC < 0.5, T-), 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.

Figures 3b-e show histograms of target AUROC across neurons and pie charts of the percentage of the different neuron types for the four layers. The results show a greater percentage of T+ and T- neurons as the processing progresses through the layers of the network (dense layer: 24.3 ± 1.29% T+ and 24.3 ± 1.68% T- neurons). Figure 3f shows higher target sensitivity with processing across layers: increasing layer-wise mean AUROC ($p < 10^{-5}$ for all layers vs. dense), maximum ($p < 0.01$ for Conv1 vs. all layers) and top 10% neuronal AUROC ($p = 0.0022$ for Conv2 vs. flat, $< 10^{-5}$ for Conv2 vs. dense and 0.0005 for flat vs. dense).

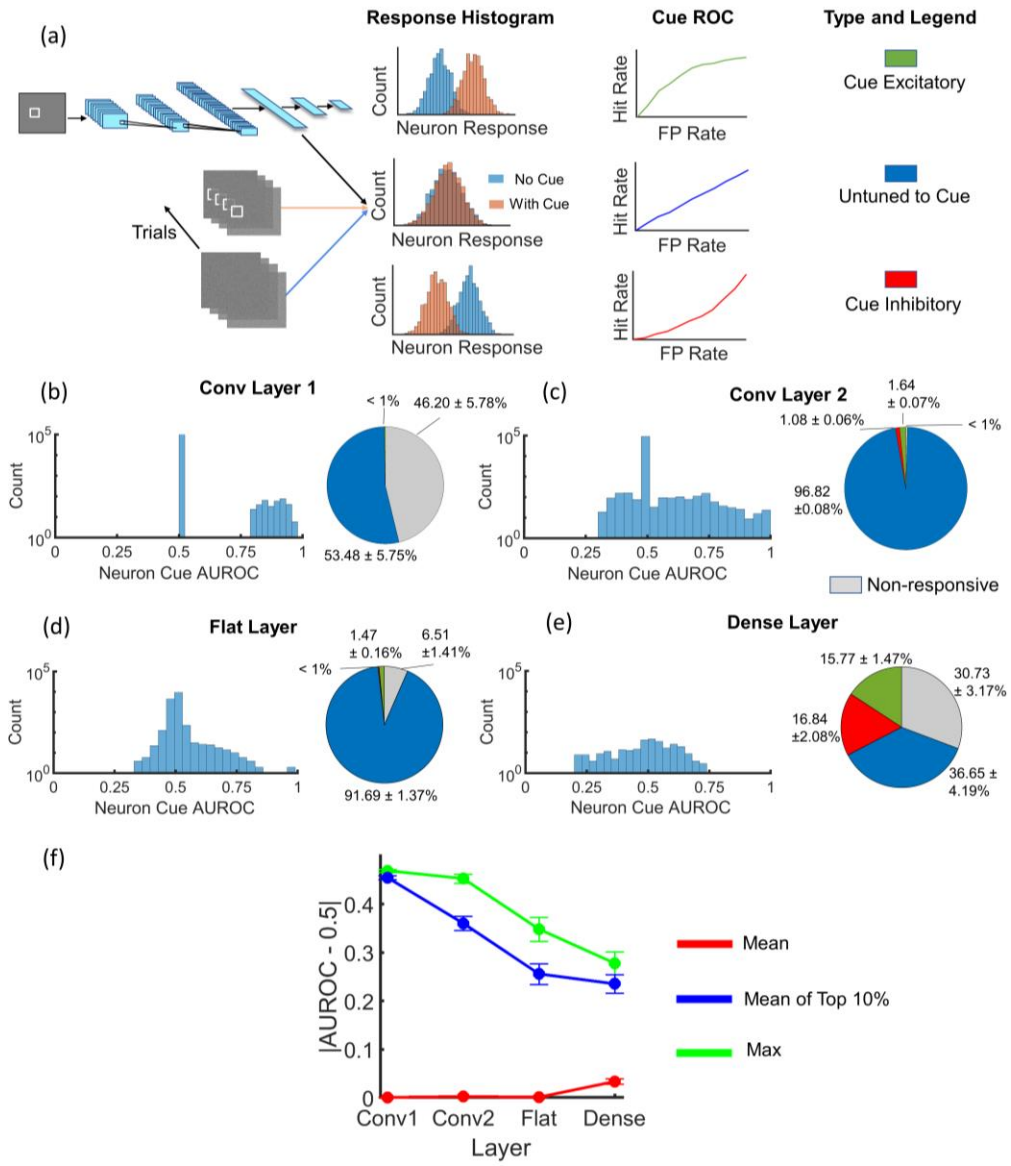


Figure 4| Receiver operating characteristic (ROC) curves for cue detection. (a): Flow diagram for calculating cue AUROC with images with and without the cue, luminance noise but not target or distractors. Top right: Example encountered types of neurons and their ROCs. Second row (b-c): Histograms of 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.

Cue Sensitivity

We conducted a similar analysis to quantify the neurons' cue sensitivities 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 (C+, AUROC > 0.5), cue-inhibitory (C-, AUROC < 0.5), cue untuned (AUROC = 0.5) and unresponsive (Figure 4a). Figures 4b-e show histograms of cue AUROC across neurons and pie charts with the percentage of different neuron types for the four layers. The percentage of neurons tuned to the cue (C+ and C-) increased as processing progressed through the network. However, unlike the neuronal target sensitivities, the neurons' max and top 10% excitatory and inhibitory cue sensitivities (|AUROC- 0.5|, Figure 4f) diminished greatly from the Conv1 to the dense layer ($p < 0.05$ for all comparisons except for max sensitivity between Conv1 and Conv2).

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 neurons at different layers are jointly tuned to target and cue. Figures 5a-d show all neurons' AUROC for the target plotted against that for 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 separately tuned to only the target or the cue. Neurons that fall off the two perpendicular lines correspond to neurons that are jointly tuned and integrate target and cue information (TC neurons). Results show no TC neurons for Conv1, a very small percentage of neurons for Conv2 (< 0.01%, Figure 5a-b) and the flat layer (0.16 ± 0.03% TC vs. 7.50 ± 1.59% and 1.62± 0.21% responding to the target and cue separately, Figure 5c), but

a large increase for the dense layer (28.7 ± 4.5% TC vs. 17.83 ± 2.47% target only and < 1% for cue only, Figure 5d). 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 5d).

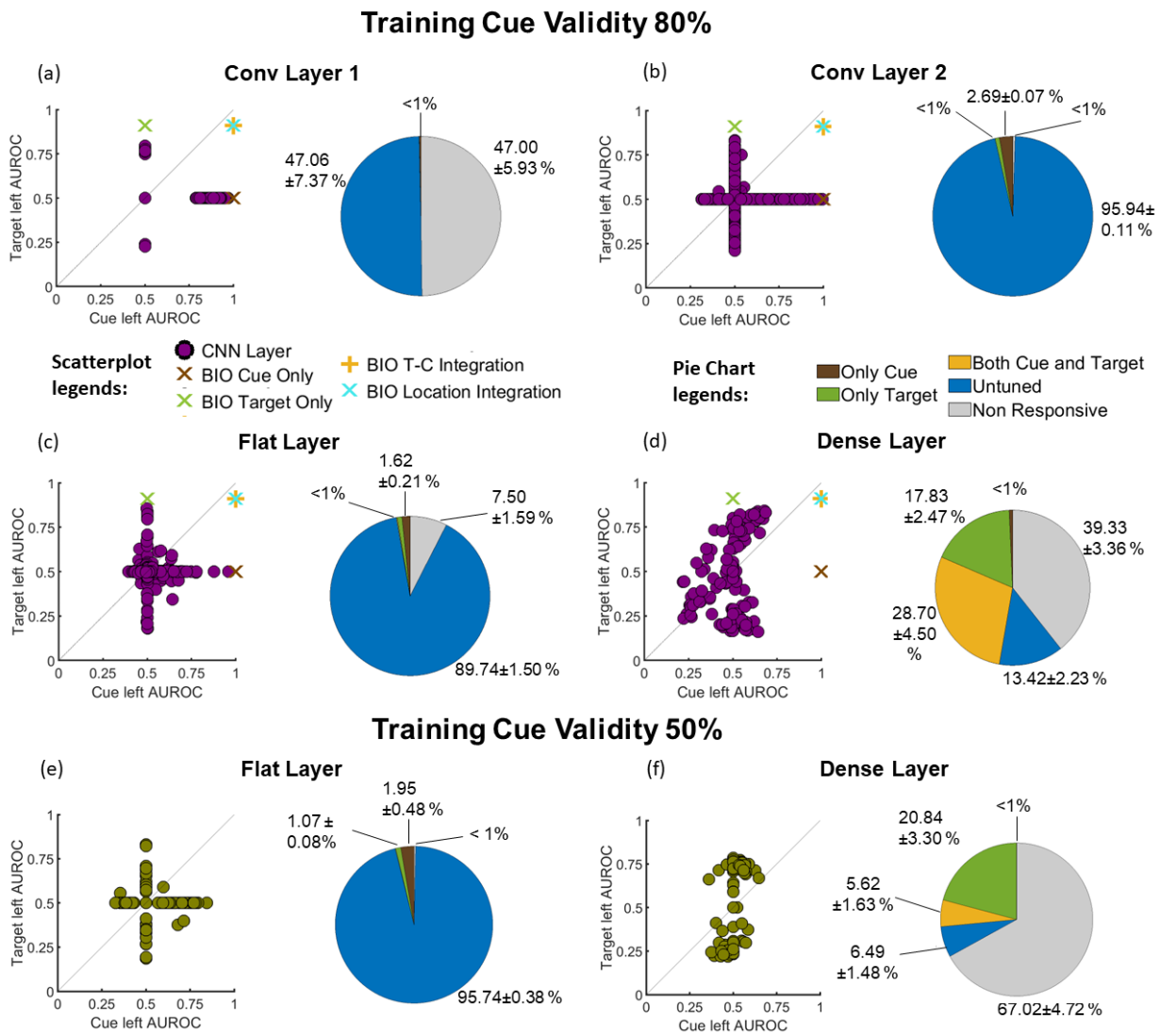


Figure 5 |Target-cue integration at the same location Top row: Scatter plots of neuronal AUROCs for target vs. neuronal AUROCs for cue at the left location for neurons in the first (a) and second convolution layers (b) of CNNs with training cue validity of 80%. The ROCs obtained from the corresponding 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) are the 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. (e) and (f) are the same as (c) and (d) but for the flat and dense layer of CNNs with training cue validity of 50%. All results are for the target and cue on the left location but similar results are obtained for the right location.

Figure 5a-d also plot the AUROC for a BIO's decision variables at different stages of its computational hierarchy (Figure 2a): 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 5a-b). Neurons in the dense layer (Figure 5d) 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 5c) show an intermediate stage with more neurons involved in separate cue and target detection and a small proportion of TC 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 toward target-cue integration across layers.

As a control, Figure 5e-f shows analyses for the flat and dense layers of a network trained on a 50% predictive cue (see Supplementary section S.3 and Figure S.2 for Conv1 and Conv2). Results show similar separate subpopulations of target- and cue-tuned neurons in the flat layer (Figure 5e) but a significantly lower number of TC neurons in the dense layer when compared to the CNN trained with an 80% predictive cue (Figure 5f vs. 5d).

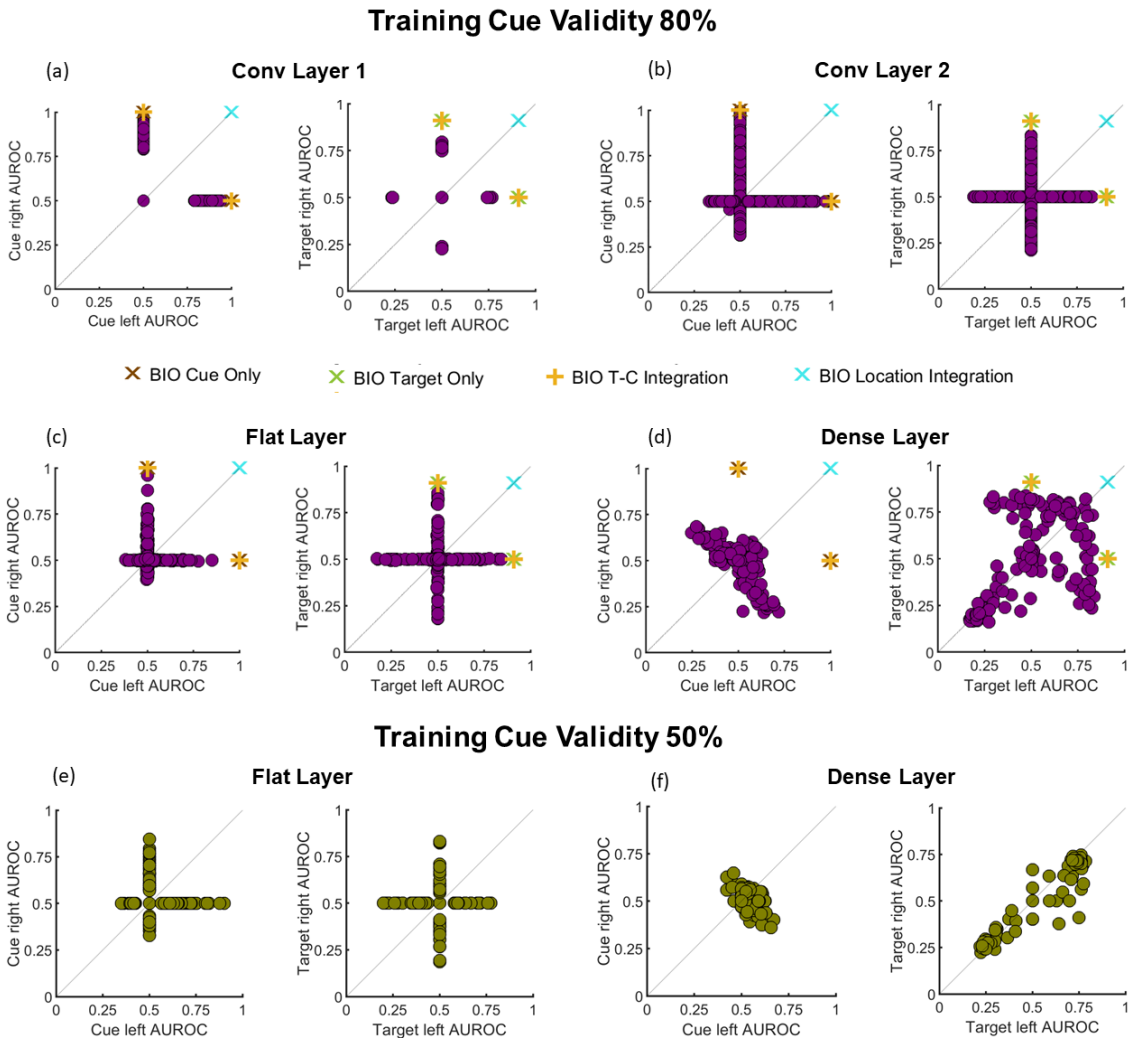


Figure 6|Integration across locations Top row: Scatter plots between neuronal AUROCs for cue detection and target detection at the right and left locations for neurons in the first two convolution layers for CNN with training cue validity of 80%. The AUROCs obtained from the corresponding stage of the 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 only at local spatial locations 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. We assessed the relationship between each neuron's AUROC for

target or cue detection on the left and the right locations (Figure 6). Conv1, Conv2, and flat layers (Figure 6a-c) have target and cue neurons that show no integration across the two locations. In contrast, the dense layer (Figure 6d) showed neurons that integrate across both locations (AUROC ≠ 0.5 for both locations). For the target-only neurons, 38.64 ± 7.30% were target-excitatory or target-inhibitory at both locations (location-summation neurons), and 1.0 ± 0.7% were target-excitatory at one location, and target-inhibitory at the other (location-opponent neurons). Among cue-only neurons, 75.34 ± 10.99% at the dense layer were location-summation and 2.16 ± 0.15 were location-opponent. Of the TC-neurons, 22.08 ± 1.45% were target-summation/cue-summation (TsCs), 18.47 ± 1.46% were target-opponent/cue-opponent (ToCo), 39.34 ± 3.22% were target-summation/cue-opponent (TsCo), and 2.53 ± 0.68% were target-opponent/cue-summation (ToCs) among the dense layer TC-neurons.

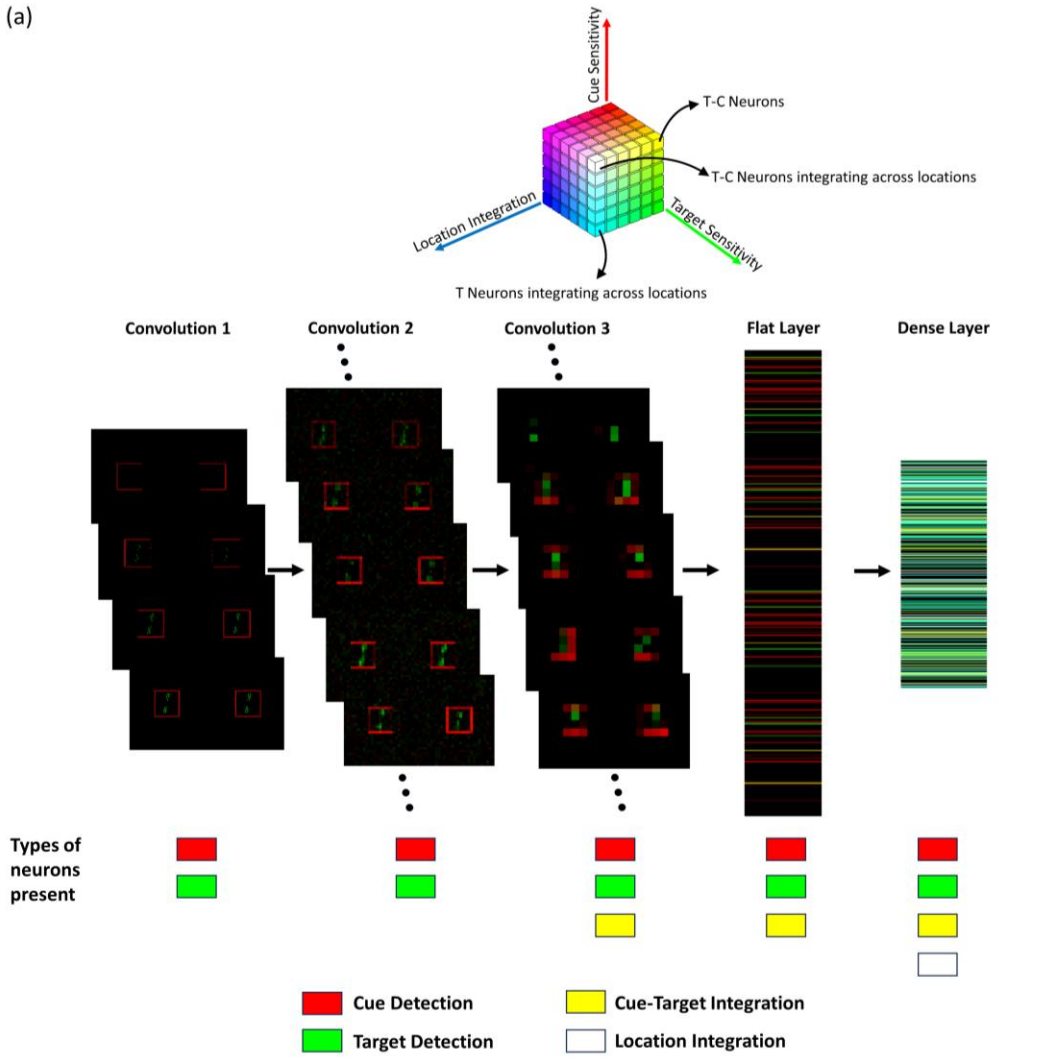
As a comparison, Figure 6e-f shows analyses for the flat and dense layers of a network trained on a 50% predictive cue (see Supplementary section S.3 and Figure S.2 for Conv1 and Conv2). Results show similar separate subpopulations of neurons tuned to each location for the flat layer (Figure 6e) and a dense layer integrating target and cue information across locations. However, compared to the network trained with the 80% predictive cue, the most notable difference is the lower number of location-opponent neurons (Figure 6f vs. 6d).

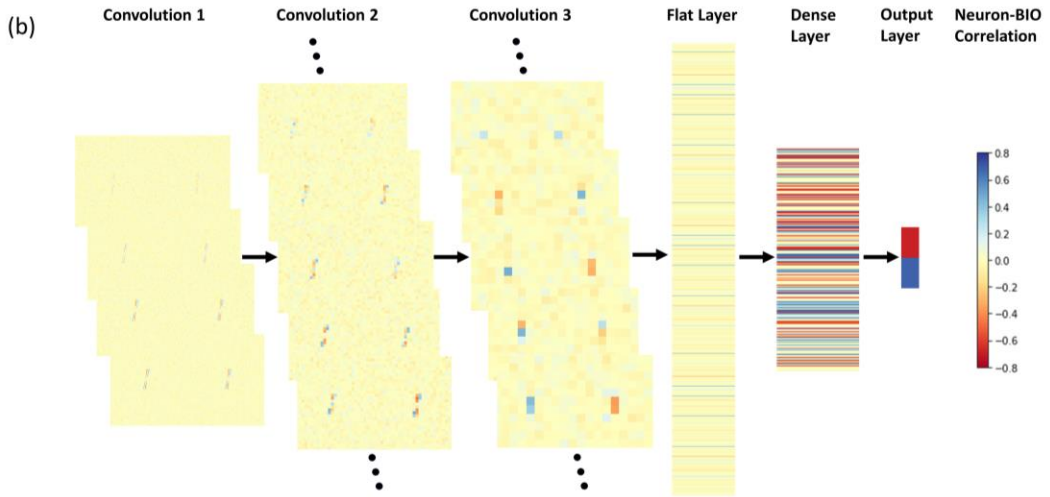
Visualization of neuronal computational stages across network layers

Figure 7a shows a summary visualization of the properties of the neurons at the different layers in terms of their tuning to the target, cue, joint tuning, and location integration. For simplicity, the visualization shows the five activation maps in each layer (out of 8, 24, and 32 activation maps per layer) that showed the highest responsiveness. Below, each layer we indicate corresponding functional computational stages.

CNN and optimal computations measured by trial correlations

To obtain a quantitative measure of the correspondence between neurons' responses and optimal computations we correlated the BIO decision variable (log-likelihood, see methods) and each individual neuron's response across trials. The correlation between the output variables of these two models is 0.68±0.015 across ten models. Figure 7b shows a visualization of layerwise neuron-BIO correlations for one network (absolute values). The convolution layers have similar mean neuron-BIO correlations (Figure 7c) with an increase for the flat, dense layers ($p < 10^{-10}$ across ten networks) and output layers ($p < 10^{-5}$). The max neuron-BIO correlations steadily increase from the first convolution layer, flat layer, and dense layers (Figure 7c, $p < 10^{-10}$ across ten networks) and asymptotes for the output layers ($p = 0.2451$).


(a)
Cue Sensitivity
T-C Neurons
T-C Neurons integrating across locations
Location Integration
Target Sensitivity
T Neurons integrating across locations
Convolution 1
Convolution 2
Convolution 3
Flat Layer
Dense Layer
Types of neurons present
Cue Detection
Target Detection
Cue-Target Integration
Location Integration


(b)
Convolution 1
Convolution 2
Convolution 3
Flat Layer
Dense Layer
Output Layer
Neuron-BIO Correlation
0.8
0.6
0.4
0.2
0.0
−0.2
−0.4
−0.6
−0.8

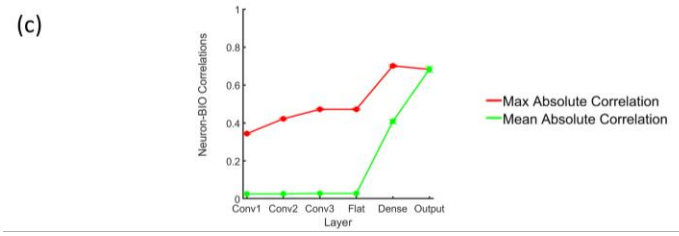

(c)
Neuron-BIO Correlations
1
0.8
0.6
0.4
0.2
0
Conv1
Conv2
Conv3
Flat
Dense
Output
Layer
Max Absolute Correlation
Mean Absolute Correlation

Figure 7|Layerwise Neuron types and Neuron-BIO Correlations. (a) A layerwise map of neuron types present in the network. Each neuron is represented by an RGB triplet, where the Red value depicts the Cue sensitivity (Max{|AUROC Cue Left – 0.5|,|AUROC Cue Right – 0.5|}x 2) of the neuron, the Green value depicts the Target sensitivity (Max{|AUROC Target Left – 0.5|,|AUROC Target Right – 0.5|}x 3) and the Blue value depicts the location integration of the neuron (|AUROC Target Left – 0.5|x|AUROC Target Right – 0.5|x 3). (b) Neuron-BIO correlations. For each neuron, we calculated 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 whose size has been amplified for visualization purposes. (c) Mean and max neuron-BIO correlations (absolute value) for each layer averaged across the ten CNNs.

Influence of cue on mean target response

Fundamental to the understanding of cue-mediated covert attention is 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. We first analyzed the neurons' responses to the target when the cue is valid vs. when the cue is absent (see Figure 8a for example stimuli). The presence of the cue results in larger increases and decreases in the mean response to the target in a subpopulation of neurons (Figure 8b; see tails of distributions) as the processing progresses through the network's layers. Comparison to networks trained with the 50% predictive cue shows that the influence of the predictive (80%) cue arises in the flat and dense layer (The Kolmogorov-Smirnov (KS) test measures the statistical significance of differences in distributions and resulted in: $p=0.9815$ for Conv1, $p=0.8111$ for Conv2, $p= 0.1606$ for flat layer only-T neurons, $p=1.792\times10^{-9}$ for flat layer TC 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 vs. when the cue was invalid (Figure 8c) to assess any potential inhibitory influence of the invalid cue on the response to the target. Comparisons to the 50% predictive cue show that the influence of the 80% predictive cue on the distributions of the neurons' mean responses arises at the dense layer (KS test: $p= 0.9578$ for

Conv1, p= 0.7676 for Conv2, p= 0.1238 for flat layer only-T neurons, p= 0.1244 for flat layer TC 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 location integration neurons. Figure 8c 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 8c). Similarly, the BIO response variables to the target are also inhibited by the invalid cue only at the stage where information is integrated across spatial locations (see dotted line in Figure 8c).

Influence of cue on target detection AUROC

The influence of the cue on the neurons' mean responses to the target, distractor, and their variability (see Supplementary Section S.4. and Figure S.3. for influence on variability) will impact the neurons' sensitivity to the target. We quantified the influence of the cue on sensitivity by comparing AUROC to the target present vs. absent (distractor present) when the valid cue is present (Figure 8a, AUROC valid) vs. when the cue is absent (Figure 8a, AUROC no cue). A comparison of the neuronal cueing (AUROC valid minus AUROC No Cue) distributions (Figure 8d) for networks trained with 80% vs. 50% predictive cues shows that the influence of the predictive cue on target sensitivity arises in the flat and the dense layers of the network (KS test: p= 0.3343 for Conv1, p= 0.2885 for Conv2, p= 0.1140 for flat layer only-T neurons, p= 0.0132 for flat layer TC 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, unlike the CNN, arises only at the stage of integration across locations[34].

We observed a similar trend of the increased cue influence with the processing stage in the network when measuring the percentage of neurons showing a significant change in AUROC for trials with valid cue vs. no cue (Supplementary section S.5).

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% vs. 50% predictive cues 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 only-T neurons, $p= 0.0829$ for flat layer TC neurons and $p= 0.0021$ for the dense layer). The layerwise percentage of neurons showing significant differences in AUROC for no cue vs. invalid cue trials is consistent with these results (Supplementary Section S.5). The results are consistent with the concept that only neurons integrating information across the two spatial locations 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.

ReLU Activation Function and cueing effects 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. 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 0.697 ± 0.007 without the cue. The ReLU

zeroes all negative values, effectively cutting off part of the tails of the target and distractor distributions. This 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 0.684 ± 0.006. The valid cue increases the neuron's mean response to the target and distractor, and a larger part of both response distributions shift to the right above the threshold of zero. Thus, less of both distributions are truncated, resulting in a lower decrease in AUROC for valid cue (Pre-ReLU: AUROC valid cue = 0.684 ± 0.006, AUROC no cue = 0.697 ± 0.007, $p = 0.0753$; Post-ReLU: AUROC valid cue = 0.663 ± 0.005, AUROC no cue = 0.571 ± 0.004, $p < 10^{-5}$; see Figure 8f, right and Supplementary Section S.6 for dense layer).

The ReLU non-linearity has a larger influence at the flat layer TC neurons where the mean absolute neuronal cueing decreased from 0.034 to 0.010 (KS test: $p < 10^{-5}$; Figure 8g) while the mean absolute neuronal cueing in the dense layer TC neurons decreased from 0.032 to 0.028 (KS test: $p = 0.2219$) after removing the ReLU (Figure 8h, left)).

ReLU thresholding vs. location integration: Which causes the behavioral cueing?

We first assessed the functional importance of location integration by removing the weights coming into the dense layer from the flat layer neurons tuned to the uncued location and found that the mean absolute neuronal cueing decreased from 0.032 to 0.011 (KS test: $p < 10^{-5}$; Figure 8h, right), indicating that the integration across locations is required to give rise to the neuronal cueing effects in the dense layer.

To evaluate the functional importance of the ReLU thresholding on behavioral cueing, we trained a network that used an activation function with no thresholding (tanh, hyperbolic tangent function) in the flat layer. This eliminated the effect of thresholding on target sensitivity at the flat layer (not

significantly different from the CNN trained on 50% predictive cue, KS test: $p = 0.5362$; Supplementary Section S.7 and Figure S.4) while still exhibiting a comparable behavioral cueing as with the CNN with ReLU (valid hit rate – invalid hit rate = 0.16 ± 0.018 for tanh vs. 0.17 ± 0.014 for ReLU). These analyses suggest that the location integration at the dense layer rather than the target sensitivity at the individual locations at the flat layer (through the effect of the ReLU thresholding) mediates the CNNs' behavioral cueing effect.

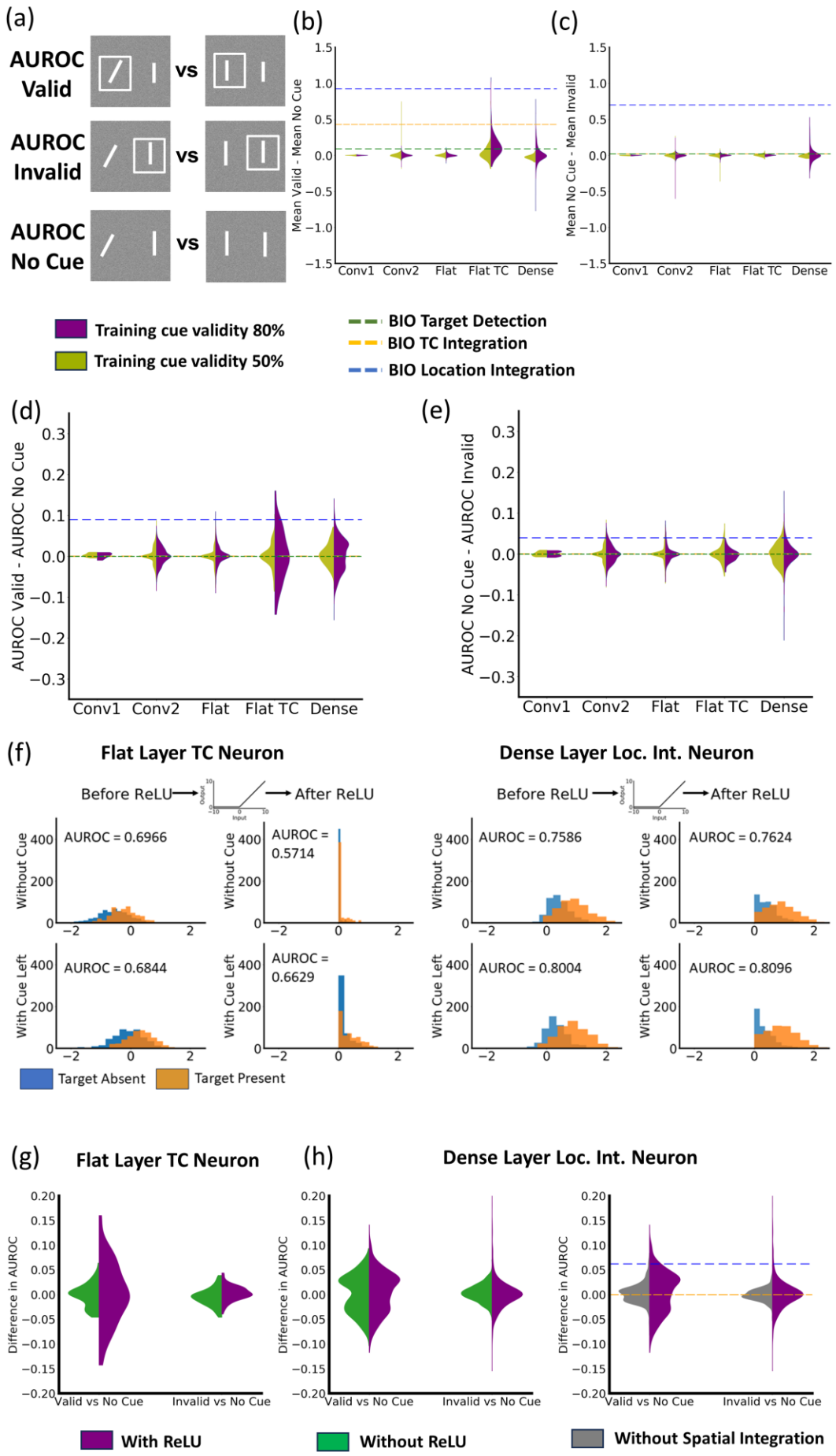

(a)
AUROC Valid
vs
AUROC Invalid
vs
AUROC No Cue
vs
(b)
Mean Valid - Mean No Cue
Conv1 Conv2 Flat Flat TC Dense
(c)
Mean No Cue - Mean Invalid
Conv1 Conv2 Flat Flat TC Dense
Training cue validity 80%
Training cue validity 50%
BIO Target Detection
BIO TC Integration
BIO Location Integration
(d)
AUROC Valid - AUROC No Cue
Conv1 Conv2 Flat Flat TC Dense
(e)
AUROC No Cue - AUROC Invalid
Conv1 Conv2 Flat Flat TC Dense
(f)
Flat Layer TC 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)
Flat Layer TC Neuron
(h)
Dense Layer Loc. Int. Neuron
Difference in AUROC
Valid vs No Cue
Invalid vs No Cue
With ReLU
Without ReLU
Without Spatial Integration

Figure 8|The effect of hierarchy on neuronal cueing. (a) Example figures depicting stimuli that the AUROCs are computed for. (b)-(e) Layerwise distributions of (b) differences in mean neuronal responses between images with valid cue at the left location vs. no cue for networks trained with 50% (green) and 80% valid cues (blue), (c) differences in mean neuronal responses between images with invalid cue at the right location vs. no cue for networks trained with 50% (green) and 80% valid cues (blue), (d) differences in neuronal target presence AUROC between images with valid cue at the left location vs. no cue for networks trained with 50% (green) and 80% valid cues (blue), (e) differences in neuronal target presence AUROC between images with invalid cue at the right location vs. no cue for networks trained with 50% (green) and 80% valid cues (blue). Dotted lines represent the BIO. (f) Example responses before and after ReLU of example flat layer TC 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) Effect of removing ReLU (green) on the cueing of flat layer TC neurons relative to with ReLU (purple). (h) Left: Effect of removing ReLU (green) on the cueing of dense layer location integration neurons relative to with ReLU (purple); Right: Effect of removing weights from the uncued location (gray, 'without spatial integration') on the cueing of dense layer location integration neurons relative to with ReLU and spatial integration. Analyses 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.

Circuits resulting in neurons with a cueing effect

An important goal is to determine how a neuron in the dense layer that shows neuronal cueing arises by combining the inputs of more elementary neurons. We analyzed the response properties of the input neurons and their weights to characterize such neuronal connectivity. Figure 9a-b shows a sample target-summation/cue-summation neuron (T_L+C_L+T_R+C_R+) and a target-opponent/cue-opponent neuron (T_L-C_L-T_R+C_R+) in the dense layer, and how these neurons arise from combining a variety of neurons tuned excitatory and inhibitory to the target, cue, and TC neurons in the flat layer. We also characterized the linear component of these neurons' receptive fields (RF) using a regression analysis of the neuron's response vs. the input noisy image values (see Methods; Figure 9c-f). The summation neuron shows an increased weighting (mean positive and negative weights of the estimated RF) to visual information from the cued location relative to the uncued location (Figure 9c; $p < 0.01$ for each location with and without the cue), a mechanism similar to the CNN's output neurons (Figure 9e) and BIO's location integration stage (Figure 9f). The opponent cell (T_L-C_L-T_R+C_R+) shows increased weighting of the

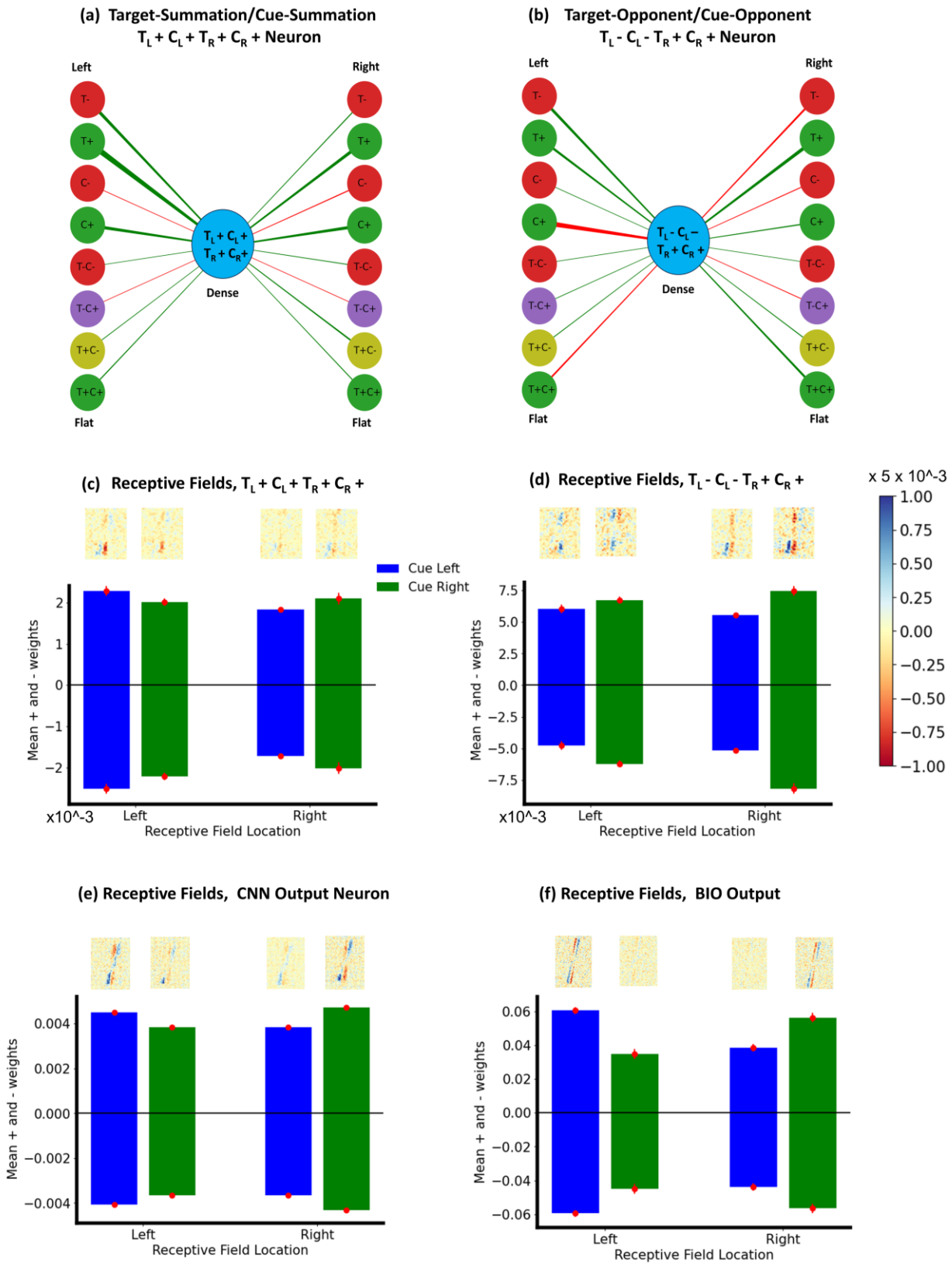
target excitatory RF at the cued location and also increased weighting for the target inhibitory RF at the location opposite to the cue, a distinct computation from any stage in the BIO.

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

To understand the role of the individual neuronal response properties at the dense layer in the final behavioral decision, we computed the correlations between neuronal cueing of dense layer TC neurons and the weights between these neurons and the target presence output neuron. We found that across ten networks, neuronal cueing has a high positive correlation ($r = 0.635$; $p < 10^{-50}$; Supplementary Section S.8 and Figure S.5) with the weights projected onto the target presence output neuron.

Supplementary analyses

The presented analyses are for a specific network architecture. We further validated the generalization of the results for many network or task variations, including the effect of cue validity, regularization, the dense layer size, the overall number of layers, convolution stride size, kernel size, and varying the noise amplitude. The results are discussed in the Supplement (Sections S.9-16 and Figures S.6-S.12).


(a) Target-Summation/Cue-Summation
$T_L + C_L + T_R + C_R$ + Neuron
Left
Right
T-
T+
C-
C+
T-C-
T-C+
T+C-
T+C+
Flat
$T_L + C_L + T_R + C_R$+
Dense
(b) Target-Opponent/Cue-Opponent
$T_L - C_L - T_R + C_R$ + Neuron
$T_L - C_L - T_R + C_R$ +
(c) Receptive Fields, $T_L + C_L + T_R + C_R$ +
Cue Left
Cue Right
Mean + and - weights
x10^-3
Left
Right
Receptive Field Location
(d) Receptive Fields, $T_L - C_L - T_R + C_R$ +
x 5 x 10^-3
(e) Receptive Fields, CNN Output Neuron
(f) Receptive Fields, BIO Output

(g)

CNN Prediction	Empirical Evidence
Cue modulates mean responses to the target (Fig 8b-e).	Maunsell and Cook 2002 [51]
Increasing influence of cue on going deeper into the hierarchy (Fig 8b-e).	Maunsell and Cook 2002 [51], Noudoost et al. 2010 [31]
Higher tuning to the cue at the middle processing stages (Fig 4f).	Wang et al. 2022 [21]
Neuronal modulation is influenced by cue validity (Fig 8b-e).	Basso and Wurtz 1997. [33]
Changes in the cue validity are reflected in deeper and not earlier layers (Fig 8b-e).	Rungratsameetaweemana et al. 2018 [30]
Neurons in deeper regions show correlations with the log-likelihood ratio (Fig 7b).	Yang and Shadlen 2007 [52]
Cue-inhibitory cells (Fig 4c-e).	Wang et al. 2022 [21]
Location-opponent cells (Fig 6d).	Not observed/reported yet.
Cue effects arise after integration across locations (Fig 8h).	Not observed/reported yet.

Figure 9| Example neuron input weight distributions and model predictions. (a)-(b) Weights projected by various flat layer neurons from the two locations (left/right) into two sample neurons in the dense layer. The location summation or opponency and the tuning properties of each dense neuron to the target and cue at both locations are noted at the top (T_L+ and T_R- denote that the neuron is target-excitatory on the left and target-inhibitory on the right, C_L+ and C_R- denote that the neuron is cue-excitatory on the left and cue-inhibitory on the right). The flat layer neuron types are noted on each neuron (colored circles). The color of each connecting line (green/red) denotes whether the sum of weights from that neuron type to the shown dense neuron is positive/negative. The thickness of each connecting line corresponds to the magnitude of the sum of weights between that neuron type and the shown dense neuron. (c)-(d) Receptive fields obtained via linear regression predicting neuronal output response for the neurons shown in (a)-(b) with cue on the left (above blue columns) and the right (above green columns). The mean of positive (depicted by heights of the bars above the black line) and negative (depicted by heights of the bars below the black line) receptive field weights shown above each bar at the left and right locations when the cue is on the left (blue) and on the right (green) for the two neurons shown in (a)-(b). Note that the RF for the left location of the location opponent cell (T_L-) is excitatory to the vertical distractor and inhibitory to the tilted distractor. Error bars indicate standard errors over ten repetitions. (e) Same as (c)-(d) but for the target presence output neuron. Error bars indicate standard errors over ten networks. (f) Same as (c)-(d) but for the BIO log-likelihood ratio. (g) A summary of predictions made by our models, and previous instances of similar results in the literature.

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[1,53–55], model-based mechanisms (Bayesian decision weighting by priors[7,17,34,46,47,56,57], probabilistic population codes[58], divisive normalization[16], neural networks[19,59]), or neuronal response properties (correlations[3], contrast response function[18], tuning curve amplitude[4]), there are no 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 representations can be more easily related to actual neural systems than other modeling approaches[34]. We trained the networks to optimize target detection with predictive cues.

A critical aspect of our CNN approach is that we did not incorporate into the network any specialized attention mechanism[19,59] that operates at the cued location nor any explicit resource limitation[45] that had to be shared across locations. In the training of the network to maximize overall task accuracy (proportion correct) 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 that were similar to that of human observers. The behavioral cueing effect and the associated neuronal properties are emergent properties of task optimization. The systemwide characterization of the network resulted in specific theoretical observations about the computational stages across layers, neuron types, neuron circuits, and mechanisms mediating the influence of the cue on neuron responses. These can be compared to existing physiological studies in the literature. Many of our results are in good agreement with existing findings (see Figure 9g for a summary).

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 and dense layers), the integration of cue and target information at each location emerges first, followed by the integration of information across the two locations (dense layer).

In terms of the influence of the cue, our results showed 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[32,60]. The convolution layers show a small effect, akin to that reported in V1 primary visual cortex single-cell measurements.[4,31] The dense layer shows strong modulation by the cues more similar to the superior colliculus (SC)[31,61] and lateral intraparietal areas (LIP)[31,62]. The flat layer neurons show intermediate effects of the cue, perhaps more similar to those reported in area V4[31]. 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 4f). Cue-induced influences on target responses are causally related to the cue validity. Changes in cue validity showed stronger influences in later layers consistent with EEG results[30].

We also show a correspondence between neuronal properties of the CNN and the stages of the BIO (Figure 7a and Figure 2a) and a higher correlation with the BIO decision variable as processing progresses through the network (Figure 7b), mirroring correlates of BIO log-likelihood ratio in the lateral intraparietal (LIP) region[52]. 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.

Our analyses identified different neuron subtypes that can be compared to single-cell animal studies[4,21,32,61,63] and might serve as theoretical predictors of neurons not yet 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 (T+C+) but also found other combinations: T+C-, T-C+, and T-C-. Arguably, the more commonly studied cell type is the target- and cue-excitatory neuron[4,64], whose mean response increases when the cue is present. However, recent studies show a handful of cue-inhibitory neurons in the mouse superior colliculus (SC; Figure 2a in Wang et al.[21]), although they are not discussed in that paper. In terms of location integration, the CNN resulted in location-opponent neurons that were excited by the cue and/or the target at one location but inhibited at the other location (T_L+C_L+T_R-C_R- where the subscripts left and right refer to the location). We do not know of a report of neurons with such properties. Importantly, we show how TC neurons showing cueing effects arise by combining the inputs of various simpler neurons in the previous layer (Figure 9a-b for two examples).

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 (Figure 9c-d). This is the main mechanism in the BIO model using priors equal to the cue validity (Figure 8b and S.3) before integration across locations. This mechanism leads to larger target AUROC and the cueing effects after location integration, like the BIO[17,34]. In addition, the CNN shows cueing effects arising from location opponent cells that inhibit target and cue influences when the cue is not present. This mechanism is distinct from the BIO. A 3rd mechanism 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 (Figure 8f-g) but we showed that it is not needed to generate cueing effects. However, there might be other important functions for the ReLU cue mechanism that have not been explored in this paper.

There are various shortcomings in our work. Our neuronal analysis pipeline is difficult to scale up to large, deeper CNNs or higher dimensional tasks where multiple features across multiple locations need to be integrated. Our work is restricted to peripheral cues that appeared simultaneously with the target rather than other varieties of cues including pre-cues[2,10] and central cues[1,2]. However, recent work[49] has shown how 3D CNNs also have emergent cueing effects with pre-cues and central cues, indicating that some of our current findings should generalize to those tasks.

We did not add biological constraints such as divisive normalization, which has been used to model neuronal properties of attention[16], or explicit biological attention[19,20], or feedback connections[31,63,65,66], into the networks. Our analysis pipeline provides a testing bed to understand the importance of additional architectural constraints.

Finally, our CNN-based technique provides a novel approach to make predictions about neuronal properties mediating visual information processing. The framework can be of potential use to generate theoretical predictions about neuron subtypes that could guide neurophysiologists in their empirical measurements for a variety of biological organisms[21,27–29] that have been shown to manifest behavioral cueing effects.

The proposed task-based CNN approach results in a distribution of cells that vary in their tuning but, through the network processing, result in decisions at the output stage that are similar to Bayesian optimal computations. Thus, we propose that the approach might provide neuronal predictions that might be closer to what we observe in the brain and a useful complement to the stylized mechanisms and theories conceived by neuroscientists that do not always agree with what is observed in neuronal recordings[67,68].

Methods and Materials

Human Data

We used the human subjects' behavioral data from the Posner task in Srivastava et al., 2024[49]. 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 had 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 adhered to 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. 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. The cue contrast was 0.40.

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:

$$ReLU(x) = \max(0, x) \tag{1}$$

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

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

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:

$$\sum_{n=1}^{32} y_{n1} \log \hat{y}_{n1} + (1 - y_{n2}) \log (1 - \hat{y}_{n2}) \tag{3}$$

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 target-present images with valid cues, and 1000 target-present images with invalid cues 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%) while the number of target present images with valid and invalid cues were both 2500.

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.

The Bayesian Ideal Observer and its computational stages

We used the same BIO formulation that was used in Srivastava et al., 2024[49] which is repeated in Supplement Section S.1. Here we repeat the computational stages of the BIO that is used throughout the paper.

In the first stage, the BIO detects the presence of the cue and target at each of the two locations by computing the likelihood of the data (noisy luminance values G_{Bi} 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 G_i 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=1, target present and j=0 target absent, distractor present). This stage is denoted as cue detection and target detection in Figure 2a.

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:

$$\pi_i . P(G_{Bi}|Cue_i) . P(G_{B!i}|No\ Cue_{!i}) .\ P(G_i|H_{1i})/P(G_i|H_{0i}) \tag{4}$$

Here, !i denotes the location other than the ith location. This calculation is separately computed for each location and denoted by the target-cue integration stage in Figure 2a.

The third stage integrates the evidence across spatial locations by summing joint probabilities across possible target/cue and location configurations and is denoted as location integration in Figure 2a.

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 cue 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. 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 a ROC curve 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 2. 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.*

Computing the neuron-BIO correlations

We extracted the log-likelihood ratio from the BIO for 1000 unseen test images (500 target absent, 400 valid, 100 invalid) and used the same images for the CNN, obtained the response of each neuron to these 1000 images, and correlated each neuron's trial-to-trial responses with the BIO's log-likelihood ratios for those trials. For the CNN's output layer, we used the SoftMax response.

Estimating neuronal receptive fields

We used ten sets of 40k noisy images with the target (14 degrees) and distractor (7 degrees) present at both locations with no angle noise to fit a linear regression model between input image pixels and

neuronal responses. The superposition of the target and the distractor was done to make sure that all the relevant pixels in the image had a baseline input required to cause a variation in the neuron output. The predictor variables were noisy pixels of cropped regions of size 100 by 170 from the input images that contained both locations. The level of additive Gaussian luminance noise that was added to the images was the same as that used in training the networks. Separate regressions were run for the cue left and cue right conditions for each neuron or BIO output depicted in Figure 9c-f. The training R^2 was above 0.9 for all calculated RFs. The out-of-sample R^2 was 0.44 and 0.46 for the summation neuron (Figure 9c), 0.37 and 0.33 for the opponent neuron (Figure 9d), 0.29 and 0.27 for the output neuron (Figure 9e), and 0.39 and 0.40 for the BIO output (Figure 9f) for the cue on the left and right conditions respectively.

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 and will be available publicly upon publication.

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 and will be available publicly upon publication.

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. 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).
9. Müller, N. G., Bartelt, O. A., Donner, T. H., Villringer, A. & Brandt, S. A. A Physiological Correlate of the "Zoom Lens" of Visual Attention. *J. Neurosci.* **23**, 3561–3565 (2003).
10. 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).
11. Li, H.-H., Pan, J. & Carrasco, M. Different computations underlie overt presaccadic and covert spatial attention. *Nature Human Behaviour* **5**, 1418–1431 (2021).
12. Itthipuripat, S., Garcia, J. O., Rungratsameetaweemana, N., Sprague, T. C. & Serences, J. T. Changing the Spatial Scope of Attention Alters Patterns of Neural Gain in Human Cortex. *Journal of Neuroscience* **34**, 112–123 (2014).
13. Dosher, B. A. & Lu, Z.-L. Noise Exclusion in Spatial Attention. *Psychol Sci* **11**, 139–146 (2000).
14. Lu, Z.-L. & Dosher, B. A. External noise distinguishes attention mechanisms. *Vision Research* **38**, 1183–1198 (1998).
15. Burgess, A., Wagner, R., Jennings, R. & Barlow, H. B. Efficiency of human visual signal discrimination. *Science* **214**, 93–94 (1981).
16. Reynolds, J. H. & Heeger, D. J. The Normalization Model of Attention. *Neuron* **61**, 168–185 (2009).

17. 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).
18. Reynolds, J. H. & Chelazzi, L. Attentional modulation of visual processing. *Annu. Rev. Neurosci* **27**, 611–647 (2004).
19. Lindsay, G. W. & Miller, K. D. How biological attention mechanisms improve task performance in a large-scale visual system model. *eLife* **7**, e38105 (2018).
20. Fox, K. J., Birman, D. & Gardner, J. L. Gain, not concomitant changes in spatial receptive field properties, improves task performance in a neural network attention model. *eLife* **12**, e78392 (2023).
21. 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).
22. 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).
23. 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).
24. Maunsell, J. H. R. Neuronal Mechanisms of Visual Attention. *Annual Review of Vision Science* **1**, 373–391 (2015).
25. Rust, N. C. & Cohen, M. R. Priority coding in the visual system. *Nat Rev Neurosci* **23**, 376–388 (2022).
26. Moore, T. & Zirnsak, M. Neural Mechanisms of Selective Visual Attention. *Annual Review of Psychology* **68**, 47–72 (2017).
27. 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.
28. 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).

29. 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).

30. Rungratsameetaweemana, N., Itthipuripat, S., Salazar, A. & Serences, J. T. Expectations Do Not Alter Early Sensory Processing during Perceptual Decision-Making. *Journal of Neuroscience* **38**, 5632–5648 (2018).

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

32. 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).

33. Basso, M. A. & Wurtz, R. H. Modulation of neuronal activity by target uncertainty. *Nature* **389**, 66–69 (1997).

34. 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).

35. Geisler, W. S. *Ideal Observer Analysis*. (MIT Press, Boston, 2003).

36. Barlow, H. B. The absolute efficiency of perceptual decisions. *Philos. Trans. R. Soc. Lond., B, Biol. Sci* **290**, 71–82 (1980).

37. 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).

38. Burge, J. Image-Computable Ideal Observers for Tasks with Natural Stimuli. *Annual Review of Vision Science* **6**, 491–517 (2020).

39. 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).

40. Ma, W. J., Beck, J. M., Latham, P. E. & Pouget, A. Bayesian inference with probabilistic population codes. *Nat Neurosci* **9**, 1432–1438 (2006).

41. Palmer, J., Verghese, P. & Pavel, M. The psychophysics of visual search. *Vision Res* **40**, 1227–1268 (2000).

42. 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).

43. 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).

44. Verghese, P. Visual search and attention: a signal detection theory approach. *Neuron* **31**, 523–535 (2001).

45. 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).

46. Vincent, B. T. Bayesian accounts of covert selective attention: A tutorial review. *Atten Percept Psychophys* **77**, 1013–1032 (2015).

47. Yu, A. J. & Dayan, P. Uncertainty, neuromodulation, and attention. *Neuron* **46**, 681–692 (2005).

48. 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).

49. Srivastava, S., Wang, W. Y. & Eckstein, M. P. Emergent human-like covert attention in feedforward convolutional neural networks. *Current Biology* **0**, (2024).

50. Green, D. M. & Swets, J. A. *Signal Detection Theory and Psychophysics*. (Peninsula Pub, 1989).

51. Maunsell, J. H. R. & Cook, E. P. The role of attention in visual processing. *Philos. Trans. R. Soc. Lond., B, Biol. Sci* **357**, 1063–1072 (2002).

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

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

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

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

56. 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).

57. 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).

58. 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).

59. 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.

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

61. 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).

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

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

64. Reynolds, J. H., Pasternak, T. & Desimone, R. Attention Increases Sensitivity of V4 Neurons. *Neuron* **26**, 703–714 (2000).

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

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

67. Sun, H., Smithson, H. E., Zaidi, Q. & Lee, B. B. Specificity of Cone Inputs to Macaque Retinal Ganglion Cells. *Journal of Neurophysiology* **95**, 837–849 (2006).

68. Chatterjee, S., Ohki, K. & Reid, R. C. Chromatic micromaps in primary visual cortex. *Nat Commun* **12**, 2315 (2021).

Supplementary Information

The Bayesian ideal observer (BIO) derivation shown in this section is the same as used in Srivastava et al., 2024[1]**.** We repeat it here for the sake of completeness.

S.1 Bayesian Ideal Observer Models

General developments of the Bayesian ideal observers for the Posner cueing task can be found in many publications [2–9]. The current treatment differs from previous Bayesian ideal observer treatments [10–17] in incorporating the process of the detection of the cue. Previous model treatments incorporate a prior reflecting the conditional probability of the presence of the target given the cue but implicitly assume the perfect detection of the cue. Here, we develop a theoretically more general model that explicitly calculates the likelihoods of the data given the joint presence of the cue and target, target alone, and cue alone. This is an image-computable Bayesian ideal observer[3,9,18,19]. Luminance noise is added to the image perturbing the luminance values of the target, distractor, and cue. When the cue visibility is high, the model makes identical performance predictions to the BIO model that only uses the priors and assumes perfect detection of the cues.

In the BIO framework, a decision across hypotheses $\boldsymbol{s_i}$ is arrived at by calculating a posterior probability of each hypothesis given the sensory data, $P(\boldsymbol{s_i}|\boldsymbol{g})$[4,20]. Bayes' theorem relates the posterior probability of a hypothesis to the likelihood of the observed sensory data given the hypothesis, $P(\boldsymbol{g}|\boldsymbol{s_i})$ and the prior probability of the hypothesis, $P(\boldsymbol{s_i})$:

$$P(\boldsymbol{s_i}|\boldsymbol{g}) = \frac{P(\boldsymbol{s_i})P(\boldsymbol{g}|\boldsymbol{s_i})}{P(\boldsymbol{g})} \tag{S.1}$$

Typically, the denominator, $P(\boldsymbol{g})$, which is the probability of the data over all possible hypotheses, is a constant across the various hypotheses. Thus, the BIO chooses the hypothesis with the highest product of likelihood and prior.

Calculating Likelihoods

Throughout the following sections, we will make references to the likelihood terms defined here. Suppose we have an n by m image ($\boldsymbol{g}$) which contains either a luminance-defined element, a tilted line with orientation 1 ($\boldsymbol{s_1}$), or a tilted line with orientation 0 ($\boldsymbol{s_0}$) or a cue with additive Gaussian independent noise (with mean μ and standard deviation σ). Thus, each pixel luminance, g_p of $\boldsymbol{g}$ is a corresponding luminance pixel of $s_{i,p}$ (where the first subscript refers to the hypothesis and the second subscript the location in the image) and some additive noise.

Because the noise added is Gaussian, the likelihood of observing a pixel value g_p given that the ith signal is present is given by the normal probability density function:

$$P(g_p|s_{i,p}) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{\left(g_p - s_{i,p}\right)^2}{2\sigma^2}\right) \tag{S.2}$$

where σ is the pixel standard deviation of the luminance noise. We note that the likelihoods are defined with respect to pixel luminance values rather than hypothetical internal responses, feature values, or neural responses.

Because the noise added to each pixel is independent and identical, the joint likelihood of all pixels given the presence of the ith signal, $P(\boldsymbol{g}|\boldsymbol{s}_i)$, reduces to the product of individual pixel likelihoods:

$$
\begin{aligned}
P(\boldsymbol{g}|\boldsymbol{s_i}) = \prod_p P(g_p|s_{i,p}) &= \prod_p \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{\left(g_p - s_{i,p}\right)^2}{2\sigma^2}\right) \\
&\propto \prod_p \exp\left(\frac{-g_p^2 + 2g_p s_{i,p} - s_{i,p}^2}{2\sigma^2}\right) \\
&\propto \exp\left(\frac{-\boldsymbol{g}^T\boldsymbol{g} + 2\,\boldsymbol{s}_i^T\boldsymbol{g} - \boldsymbol{s}_i^T\boldsymbol{s}_i}{2\sigma^2}\right)
\end{aligned}
\tag{S.3}
$$

where $\boldsymbol{g}$ is a 1-D vectorized version of the image and $\boldsymbol{s_i}$ is a vectorized version of the ith signal**.** Eq (S.3) is a basic likelihood calculation and contains a template matching term ($\boldsymbol{s_i}^T\boldsymbol{g}$). Variants of Eq. S.3 are used throughout the development of the BIO for each specific task.

Combining likelihoods across within hypothesis sub-events

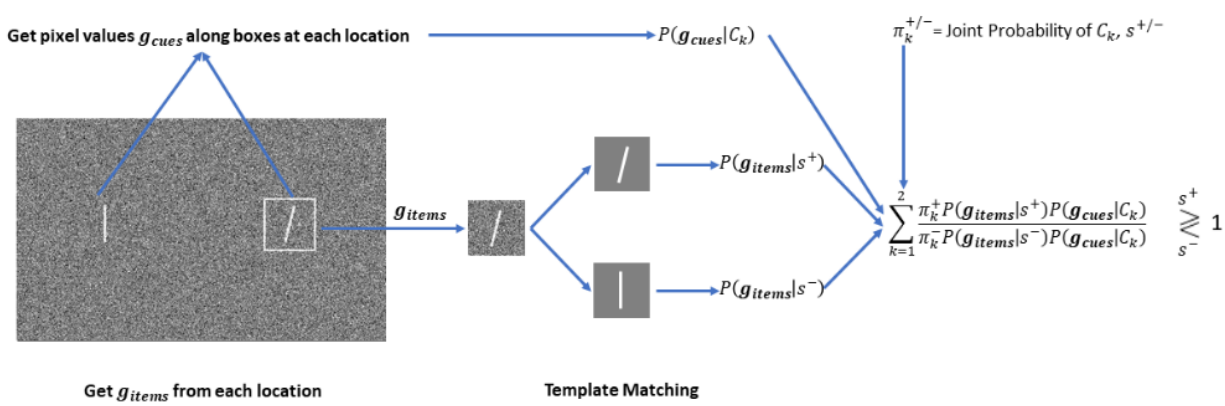
Another important property that we will use repeatedly in our development of the BIO for the various tasks is that of summing likelihoods. In basic probability, the probability of an event A involves summing probabilities across all mutually exclusive events that qualify as A: $P(A) = P(A_1) + P(A_2) + \ldots + P(A_n)$. Similarly, in many tasks, calculating the likelihood of the data given the presence of the target involves considering different locations, orientations, etc., for the target and distractors. The likelihood is calculated for each of the possible events and then summed to obtain a sum of likelihoods of the data given the presence of the target.

$$P(\boldsymbol{g}|\boldsymbol{s_i}) = \sum_{k=1}^{K} P(\boldsymbol{g}|\boldsymbol{s_{i,k}}) \cdot \frac{P(\boldsymbol{s_{i,k}})}{P(\boldsymbol{s_i})} \tag{S.4}$$

where $\frac{P(\boldsymbol{s_{i,k}})}{P(\boldsymbol{s_i})}$ is the prior probability of the kth event or target parameter value. Evaluating Equation S.4 requires the BIO to know apriori all possible values or locations attained by the target and distractors.

Posner Paradigm

In the Posner cueing task (Supplementary Figure S.1), the hypothesis $\boldsymbol{s}^+$ is that the target is present (target line with orientation 1 is at one location and a line with a smaller tilt, orientation 0, is at a second location), and the hypothesis $\boldsymbol{s}^-$ is that the target is absent (both elements have lines with orientation 0). The box cue is present at either location with equal probability. The target appears with the cue with a probability of 80%. The likelihood of the sensory data, $\boldsymbol{g}$, given the presence of the target has to be summed across the mutually exclusive locations of the target.



Supplementary Figure S.1: Flow Chart for the Ideal Observer for Posner Paradigm.

$$P(\boldsymbol{s}^{+}|\boldsymbol{g}) = \sum_{i=1}^{2} P(\boldsymbol{s}_{\boldsymbol{i}}^{+\prime}\,|\boldsymbol{g}) \tag{S.5}$$

Where $\boldsymbol{s}_{\boldsymbol{i}}^{+\prime}$ stands for the target being present at location i, and the target being absent at the other location.

$$P(\boldsymbol{s}_{\boldsymbol{1}}^{+\prime}|\boldsymbol{g}) = P(\boldsymbol{s}_{\boldsymbol{1}}^{+}|\boldsymbol{g}_{\boldsymbol{1}}).P(\boldsymbol{s}_{\boldsymbol{2}}^{-}|\boldsymbol{g}_{\boldsymbol{2}})$$

And (S.6)

$$P(\boldsymbol{s}_{\boldsymbol{2}}^{+\prime}|\boldsymbol{g}) = P(\boldsymbol{s}_{\boldsymbol{1}}^{-}|\boldsymbol{g}_{\boldsymbol{1}}).P(\boldsymbol{s}_{\boldsymbol{2}}^{+}|\boldsymbol{g}_{\boldsymbol{2}})$$

where $\boldsymbol{g}_{\boldsymbol{1}}$ and $\boldsymbol{g}_{\boldsymbol{2}}$ are pixel values from the first and second locations, respectively.

Marginalizing over the two configurations corresponding to the cue appearing at the right or left location, we have

$$P(\boldsymbol{s}_{\boldsymbol{i}}^{+\prime}\,|\boldsymbol{g}) = \sum_{k=1}^{2} P(\boldsymbol{s}_{\boldsymbol{i}}^{+\prime}\,, C_K\;\;|\boldsymbol{g}) \tag{S.7}$$

Where C_1 refers to cue on the Left Location, no box on the right location and C_2 refers to the cue on the right location, no cue on the left location.

Now,

$$P(\boldsymbol{s_i^{+\prime}}, C_K \,|\boldsymbol{g}) \propto P(\boldsymbol{g}|\boldsymbol{s_i^{+\prime}}, C_K).P(C_K, \boldsymbol{s_i^{+\prime}}) \tag{S.8}$$

The likelihood can be separated into the independent contributions of the items (target/distractor) and the cues:

$$P(\boldsymbol{g}|\boldsymbol{s_i^{+\prime}}, C_K) = P(\boldsymbol{g_{items}}, \boldsymbol{g_{cues}}|\boldsymbol{s_i^{+\prime}}, C_K) = P(\boldsymbol{g_{items}}|\boldsymbol{s_i^{+\prime}}).P(\boldsymbol{g_{cues}}|C_K) \tag{S.9}$$

And

$$P(C_K, \boldsymbol{s_i^{+\prime}}) = P(\boldsymbol{s_i^{+\prime}}\,|C_K).P(C_K) \tag{S.10}$$

Thus, (S.9) reduces to:

$$\begin{aligned} P(\boldsymbol{s_i^{+\prime}}, C_K|\boldsymbol{g}) &\propto P(\boldsymbol{g_{items}}|\boldsymbol{s_i^{+\prime}}).P(\boldsymbol{g_{cues}}|C_K).P(\boldsymbol{s_i^{+\prime}}|C_K).P(C_K) \\ &\propto P(\boldsymbol{g_i}|\boldsymbol{s_i^{+}}).P(\boldsymbol{g_{!i}}|\boldsymbol{s_{!i}^{-}}).P(\boldsymbol{g_{cues}}|C_K).P(\boldsymbol{s_i^{+\prime}}\,|C_K).P(C_K) \end{aligned} \tag{S.11}$$

Where $\boldsymbol{g_i}$ indicates pixels from location i, and $\boldsymbol{g_{!i}}$ indicates pixels from the location other than i. $\boldsymbol{s_i^{+}}$ indicates target presence at location i, while $\boldsymbol{s_{!i}^{-}}$ indicates distractor presence at the location other than i.

Now, in our implementation of the Posner Paradigm, the target angle is sampled from a Gaussian distribution with a mean of 15 degrees and standard deviation of 3.5 degrees, while the distractor angle is sampled from a Gaussian distribution with a mean of 7 degrees and a standard deviation of 3.5 degrees. However, since the number of pixels is finite, only a finite number of angles can be drawn on the actual images of the target and distractors. While calculating the likelihood of a target being present, we marginalize across all possible angles along with their prior probabilities.

Thus, using equation (S.4), it follows that:

$$P(\boldsymbol{g_i}|\boldsymbol{s_i^+}) = \sum_{\theta\,\in[0,2\pi)} P(\boldsymbol{g_i}|\boldsymbol{s_{i\theta}^+}) \cdot \frac{\pi_\theta^+}{P(\boldsymbol{s_i^+})}$$

And

$$P(\boldsymbol{g_{!i}}|\boldsymbol{s_{!i}^-}) = \sum_{\theta\,\in[0,2\pi)} P(\boldsymbol{g_{!i}}|\boldsymbol{s_{!i\theta}^-}) \cdot \frac{\pi_\theta^-}{P(\boldsymbol{s_{!i}^-})} \tag{S.12}$$

Where $\boldsymbol{s_{i\theta}^+}$ denotes the presence of the target with an angle θ at location i, $\boldsymbol{s_{!i\theta}^-}$ denotes the presence of the distractor with an angle θ at the location other than i, π_θ^+ is the prior probability of sampling a template with angle θ from the target angle distribution and π_θ^- is the prior probability of sampling a template with angle θ from the distractor angle distribution.

Thus, combining equations (S.5), (S.7), (S.9), and (S.12), we get:

$$P(\boldsymbol{s}^+|\boldsymbol{g}) \propto \sum_{i=1}^{2}\sum_{k=1}^{2}\sum_{\theta_b\in[0,2\pi]}\sum_{\theta_a\,\in[0,2\pi)} P\left(\boldsymbol{g_i}\middle|\boldsymbol{s_{i\theta_a}^+}\right) P\left(\boldsymbol{g_{!i}}\middle|\boldsymbol{s_{!i\theta_b}^-}\right).\, P(\boldsymbol{g_{cues}}|C_K).\, P(\boldsymbol{s_i^{+\prime}}\,|C_K).\, P(C_K) \cdot \frac{\pi_{\theta_a}^+}{P(\boldsymbol{s_i^+})} \cdot \frac{\pi_{\theta_b}^-}{P(\boldsymbol{s_{!i}^-})} \tag{S.13}$$

The first three likelihood terms in the summand here are obtained using (eq (S.3)), and the last four terms are known. The four summations are across target location (i), cue location (k), target angle (θ, and distractor angle (θ).

Likewise, to calculate the posterior probability for the target absence

$$P(\boldsymbol{s}^{-}|\boldsymbol{g}) = P(\boldsymbol{s}_{\boldsymbol{1}}^{-}, \boldsymbol{s}_{\boldsymbol{2}}^{-}|\boldsymbol{g}) = P(\boldsymbol{g}|\boldsymbol{s}_{\boldsymbol{1}}^{-}, \boldsymbol{s}_{\boldsymbol{2}}^{-}).\frac{P(\boldsymbol{s}_{\boldsymbol{1}}^{-}, \boldsymbol{s}_{\boldsymbol{2}}^{-})}{P(\boldsymbol{g})} \propto 0.5\, P(\boldsymbol{g}|\boldsymbol{s}_{\boldsymbol{1}}^{-}, \boldsymbol{s}_{\boldsymbol{2}}^{-})$$

$$\propto 0.5\, P(\boldsymbol{g}_{\boldsymbol{1}}, \boldsymbol{g}_{\boldsymbol{2}}|\boldsymbol{s}_{\boldsymbol{1}}^{-}, \boldsymbol{s}_{\boldsymbol{2}}^{-}) \propto 0.5\, P(\boldsymbol{g}_{\boldsymbol{1}}|\boldsymbol{s}_{\boldsymbol{1}}^{-}).P(\boldsymbol{g}_{\boldsymbol{2}}|\boldsymbol{s}_{\boldsymbol{2}}^{-}) \tag{S.14}$$

where $\boldsymbol{s}_{\boldsymbol{i}}^{-}$ denotes the distractor being present at location i. $P(\boldsymbol{g}_{\boldsymbol{1}}|\boldsymbol{s}_{\boldsymbol{1}}^{-})$ and $P(\boldsymbol{g}_{\boldsymbol{2}}|\boldsymbol{s}_{\boldsymbol{2}}^{-})$ are again marginalized over all possible distractor angles similar to (S.12).

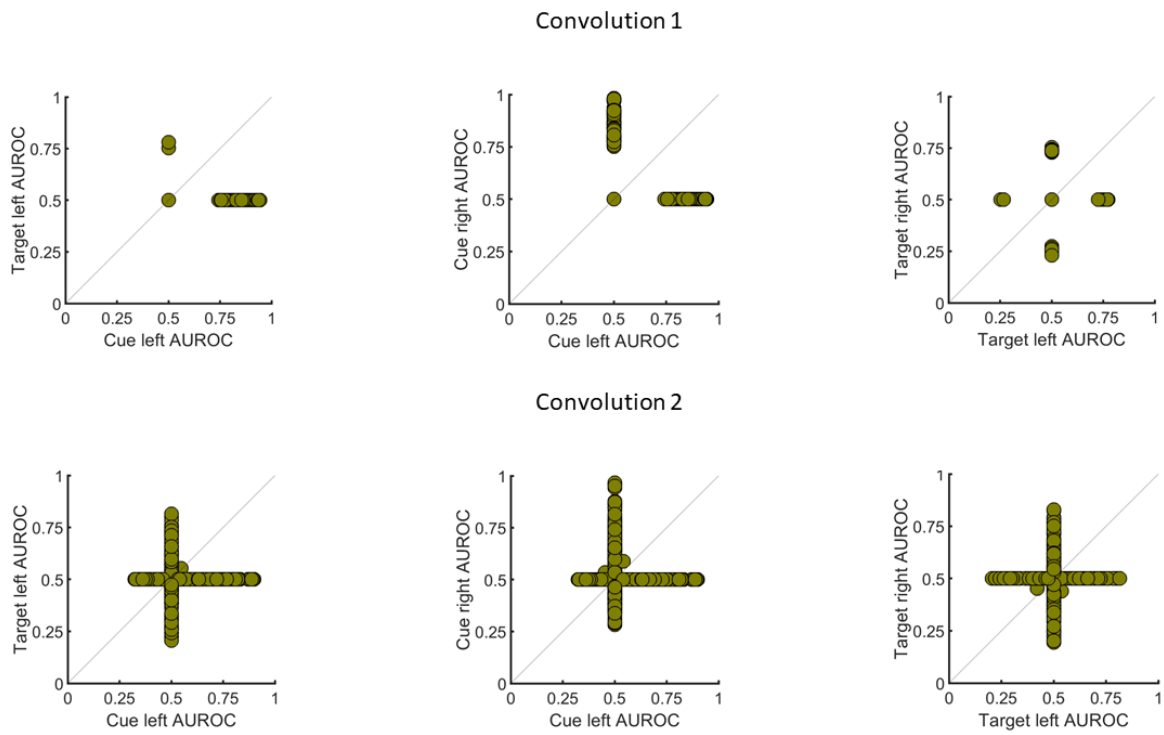
Finally, the BIO makes the decision: target absent, $\boldsymbol{s}^{-}$, if $P(\boldsymbol{s}^{-}|\boldsymbol{g}) > P(\boldsymbol{s}^{+}|\boldsymbol{g})$ and target present, $\boldsymbol{s}^{+}$, if $P(\boldsymbol{s}^{-}|\boldsymbol{g}) < P(\boldsymbol{s}^{+}|\boldsymbol{g})$.

S.2 Statistical Efficiency

The statistical efficiency[2] (SE) is the ratio of the variance (RMS^2) of the stimuli noise shown to the CNNs or humans compared to the BIO. The SE of humans was 0.39 (averaged across 18 observers) and that of the CNNs was 0.90, as reported in Srivastava et al., 2024[1].

S.3 Early layer scatterplots for models trained on 50% predictive cues

The main text presents the scatterplots for target-cue integration (Figure 5) and location integration (Figure 6) only for the flat and dense layers for models trained on 50% predictive cues. Supplementary Figure S.2 presents these scatterplots for the first two convolution layers. We find that similar to the models trained on the 80% predictive cues, the first two convolution layers show no target-cue integration or location integration.

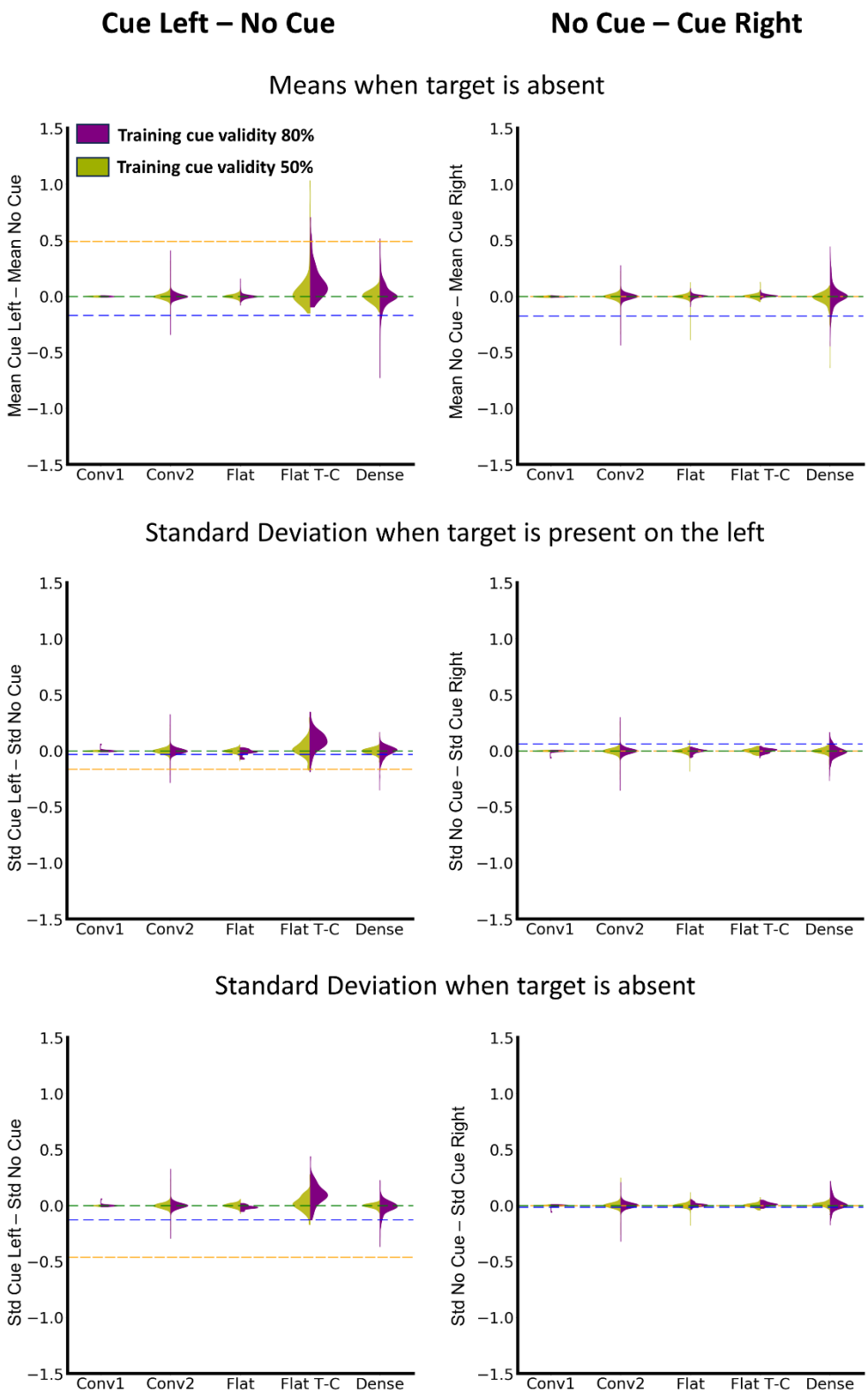


Supplementary Figure S.2| Cue-target and location integration in early layers for models trained on 50% predictive cue. Top row: Convolution layer 1, Bottom row: Convolution layer 2. No evidence for TC-integration or location integration in the early layers, similar to networks trained on 80% predictive cue (main text Figures 5-6). The flat and the dense layers for models trained on 50% predictive cue are plotted in main text Figures 5-6.

S.4 Violin plots for layerwise effect of cue on means and standard deviations

In the main text, we present the effect of the cue on the mean of neuronal responses when the target is present (Figure 8b-c). In Supplementary Figure S.3, we present the effect of the cue on the mean of neuronal responses when the target is absent and the effect of the cue on the standard deviations of neuronal responses. On comparing the distributions of differences in the mean neuronal responses to the cue, we find that Flat T-C and dense layers show a significant difference between networks trained with 80% and 50% predictive cues (KS-test: $p < 10^{-5}$) for cue left vs. no cue and dense layer for cue right vs. no cue (KS-test: $p < 10^{-5}$). On comparing the distributions of differences in the standard deviation of neuronal responses to the cue, we find that Flat T-C and dense layers show a significant difference

between networks trained with 80% and 50% predictive cues (KS-test: $p < 10^{-5}$) for cue left vs. no cue and dense layer for cue right vs. no cue (KS-test: $p < 10^{-5}$) both when the target is present and absent.



Supplementary Figure S.3| Distributions of the effect of the cue on means and standard deviations of responses. Left Column: Comparison between cue left trials and no cue trials. Right Column: Comparison between no cue

trials and cue right trials. Top Row: Layerwise comparison between means of responses with and without cue when the target is absent (See main text Figure 8b-c for means when target present). Middle Row: Layerwise comparison between standard deviations of responses with and without the cue when the target is present. Bottom Row: Same as the middle row but when the target is absent.

S.5 Layerwise percentage of neurons showing significant cueing

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) vs. 1.48% (50% predictive) for Conv1, 15.24% (80% predictive) vs. 5.78% (50% predictive) for Conv2, 11.05% (80% predictive) vs. 11.95% (50% predictive) for flat layer non-TC neurons, 78.95% (80% predictive) vs. 11.72% (50% predictive) for flat layer TC neurons and 65.28% (80% predictive) vs. 39.57% (50% predictive) neurons in the dense layer.

For no cue vs. invalid cue, we find that 2.15% (80% predictive) vs. 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-TC neurons in the flat layer, 8.42% (80% predictive) and 12.49% (50% predictive) for TC neurons in the flat layer and 71.06% (80% predictive) and 38.42% (50% predictive) for neurons in the dense layer show a significant change in AUROC.

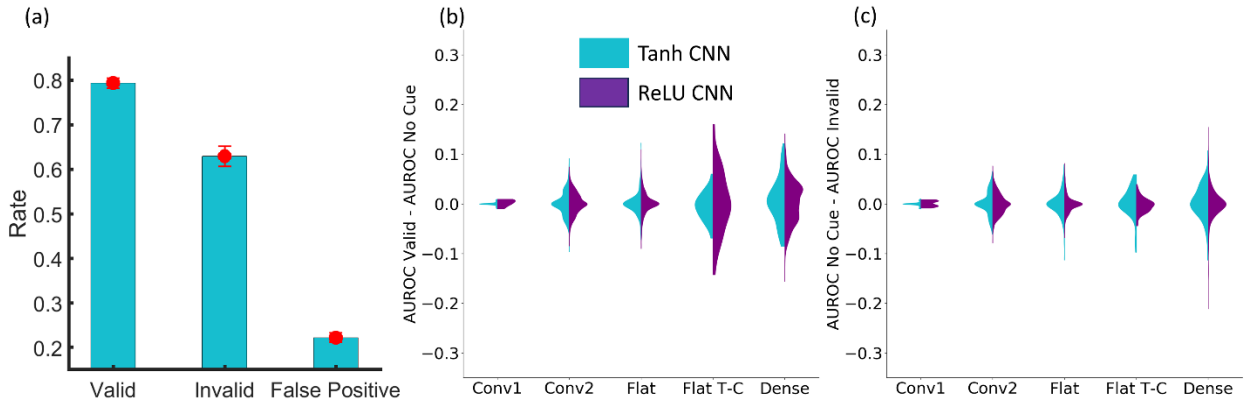
S.6 ReLU is not the primary cause of cueing in the dense layer

Figure 8f right (in the main text) shows pre- and post-ReLU response distributions for a sample dense layer neuron. Dense layer neurons have higher means relative to the flat layer TC 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 ReLU thresholding. 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.

S.7 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 crucial for the cueing effect, we trained a variations of the CNN where 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 (Supplementary Figure S.4) 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 TC integrating neurons in the flat layer with tanh was 0.008 compared to 0.034 in the flat layer with ReLU and the cueing distribution for the flat layer with tanh was not significantly different from that in CNNs trained on 50% predictive cue (KS test: $p = 0.5362$). At the output level, the CNN still exhibited a comparable cueing effect (valid hit rate – invalid hit rate = 0.16) as with the ReLU activation function. This suggests that the cue-induced target sensitivity change in TC 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.

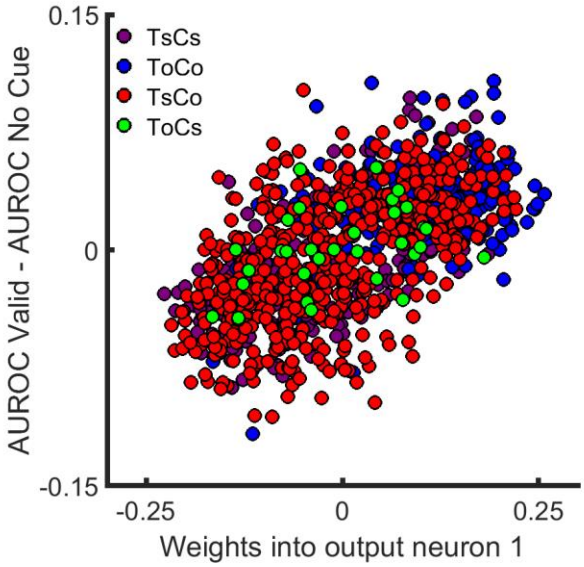


Supplementary Figure S.4| Cueing in Tanh Network: (a) Model performance on valid, invalid, and target-absent trials for networks with Tanh instead of ReLU. Error bars indicate the standard error of the mean across ten

networks. (b) Layerwise effect of valid cue (vs. no cue) on the AUROC for target presence vs. absence in the image for Tanh and ReLU networks. (c) Same as (b) but for the effect of no cue vs. invalid cue.

S.8 The relation between neuronal cueing and weights

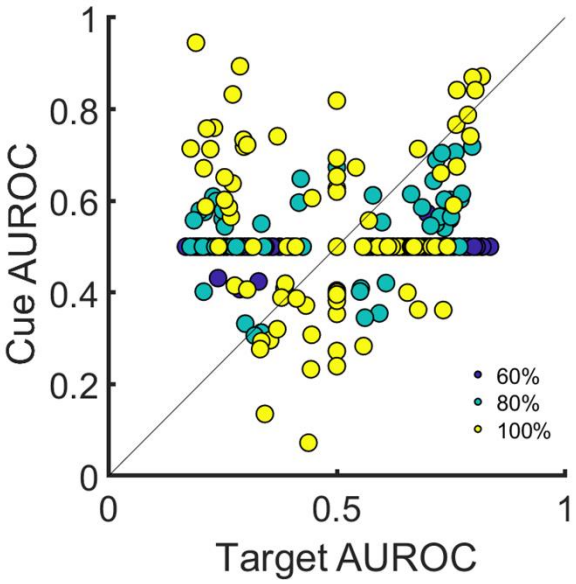
We computed the correlations between neuronal cueing of dense layer TC neurons and the weights between these neurons and the target presence output neuron. Across ten networks, neuronal cueing has a high positive correlation ($r = 0.635$; Supplementary Figure S.5) with the weights projected onto the target presence output neuron. We further divided these neurons by whether they are summation or opponent neurons for the cue and the target. The correlation between neuronal cueing and weights is 0.662 for target-summation/cue-summation neurons (TsCs in Supplementary Figure S.5), 0.397 for target-opponent/cue-opponent neurons (ToCo in Supplementary Figure S.5), 0.551 for target-summation/cue-opponent (TsCo in Supplementary Figure S.5), and 0.378 for target-opponent/cue-summation neurons (ToCs in Supplementary Figure S.5). ToCo neurons had significantly higher weights than the other three neuron types ($p < 10^{-5}$ for all comparisons on unpaired t-tests).



Supplementary Figure S.5| Relation between neuronal cueing and weights. Scatterplot depicting the relationship between neuronal cueing in the dense layer and the weights projected by neurons into output neuron 1 (target presence output neuron). Neurons are separated by their integration type for the cue and the target. TsCo, for example, stands for Target-Summation/Cue-Opponent, indicating that the neuron is either target-inhibitory or target-excitatory on both locations and cue-inhibitory on one and cue-excitatory on the other location. Results are shown across all dense TC neurons across ten networks.

S.9 Effect of changing the cue validity

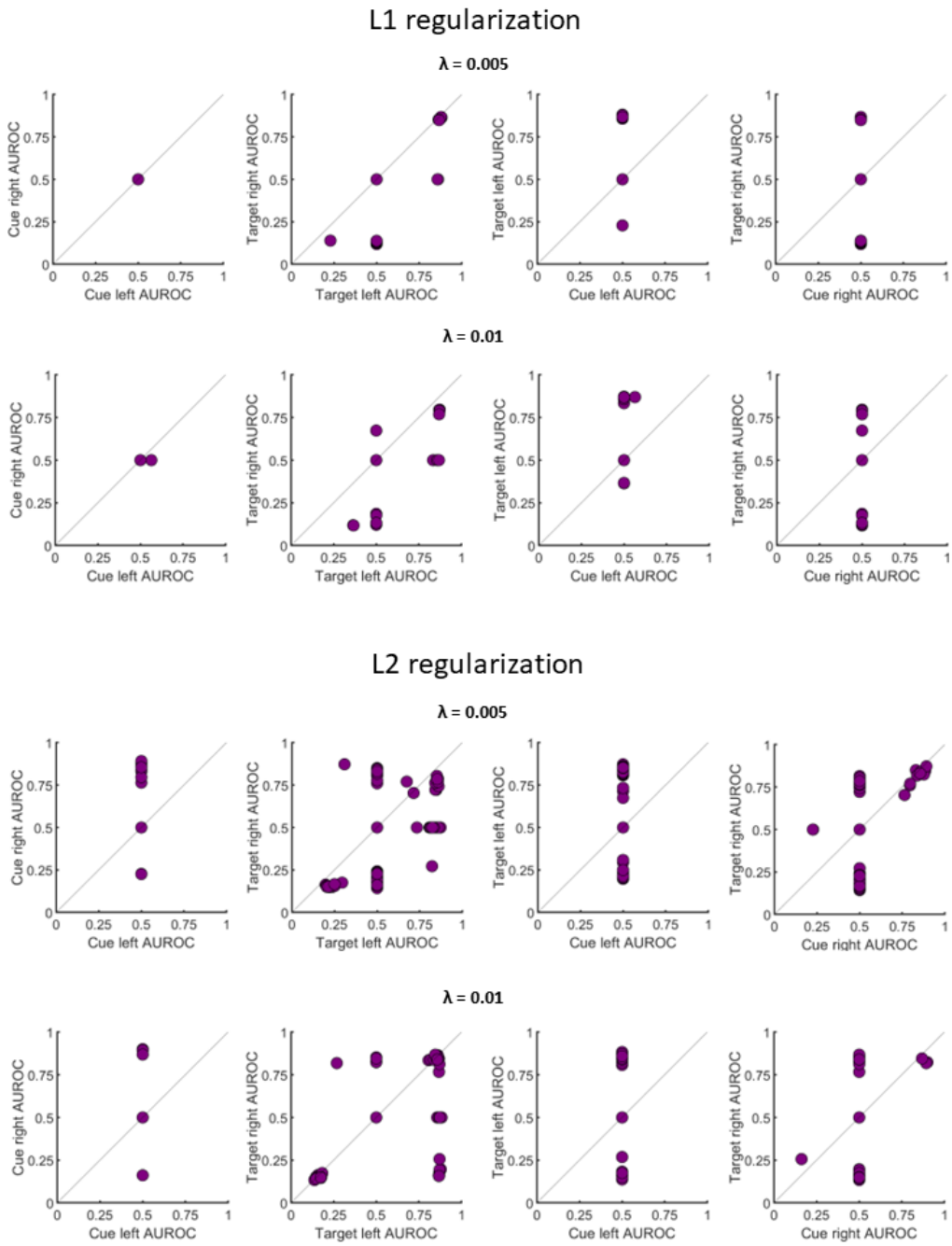
Changing the cue validity was reflected in changes in the dense layer neuron AUROCs, particularly the neurons that are tuned to both the target and the cue (TC neurons). When the cue validity is increased, the dense layer neurons that integrate cue and target information show a larger range of cue AUROC while the range of target AUROCs does not vary significantly (Supplementary Figure S.6). Interestingly, increasing the cue validity also resulted in a higher percentage of cue-excitatory/target-inhibitory (C+T-) neurons (0% for 60% validity, 27.5% for 80% validity and 38% for 100% validity) and a higher percentage of cue-inhibitory/target-inhibitory (C-T-) neurons (20% for 60% validity, 12.5% for 80% validity and 28% for 100% validity).



Supplementary Figure S.6| The effect of changing cue validity on the dense layer neurons. The AUROC scatter plots for cue-target integration in the dense layer are shown for three different cue validities (60%, 80%, and 100%)**.** The bluer colors correspond to lower validity. The range of the cue AUROCs are seen as increasing while that of the target AUROCs is nearly the same. The results shown are for the right location but generalize to the left location.

S.10 The effect of regularization

Regularization is a machine learning technique that forces the network to optimize under constraints on total weights, biases, or activations. We tried L1 and L2 activity regularization. L1 regularization resulted in more sparse representations. Heavy L1 regularization resulted in lower cueing, fewer cue-tuned neurons, and a lower extent of cue-tuning in those neurons, indicating that the network started ignoring the cue instead of the uncued location when trying to minimize resource usage. L2 regularization is less punishing in this sense and the resultant representations show target-cue integration, unlike L1 regularization. The results are shown in Supplementary Figure S.7.

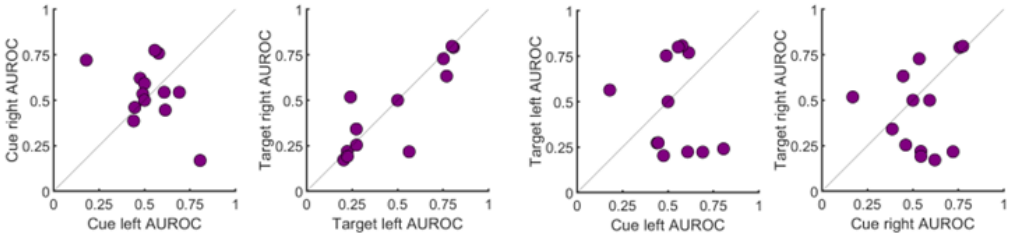


Supplementary Figure S.7| The effect of weight regularization on the dense layer neurons. The regularization was applied to the dense layer. The first two rows depict two levels of L1 regularization, and the last two rows depict the effect of two levels of L2 regularization. L1 regularization results in the network not showing cue-tuning and no target-cue integration. L2 regularization results in cue-tuning at only one location.

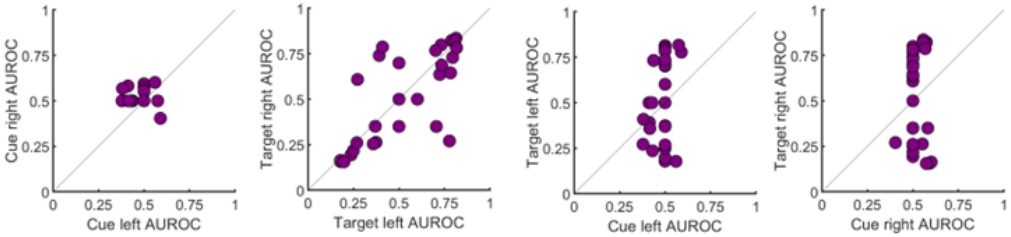
S.11 Effect of changing the size of the dense layer

We tried six different sizes of the dense layer, increasing by multiples of two starting from 32, until 1024. On increasing the size of the dense layer, more untuned neurons emerge but the cue-target scatter plots, the cue on both sides scatter plots, and the target on both sides scatter plots continue to show qualitative consistency (See Supplementary Figure S.8).

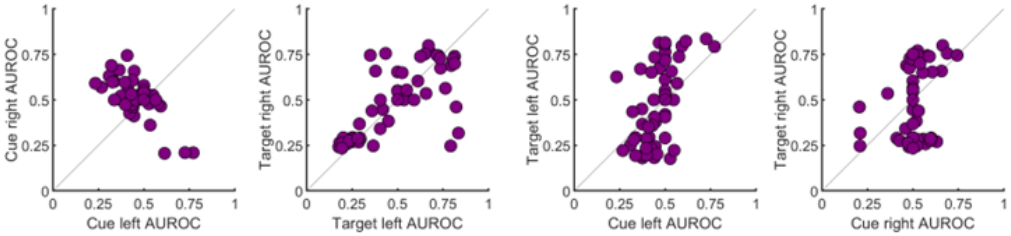
32 dense neurons



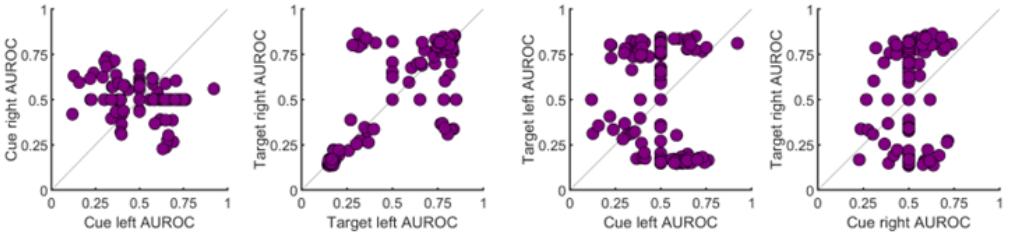
64 dense neurons



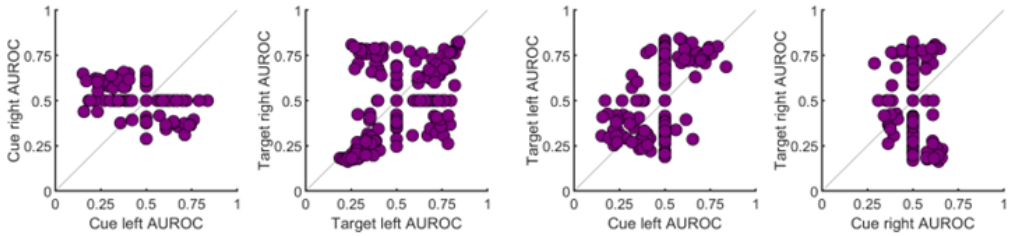
128 dense neurons



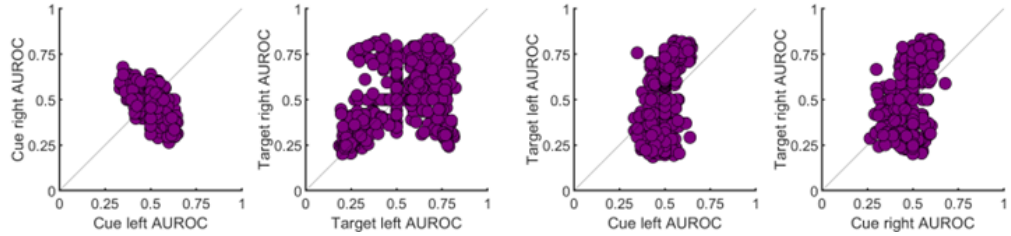
256 dense neurons



512 dense neurons



1024 dense neurons



Supplementary Figure S.8| The effect of changing the size of the dense layer. The rest of the network architecture was kept the same as the network in the main text. Both target-cue integration and location integration are seen across all sizes.

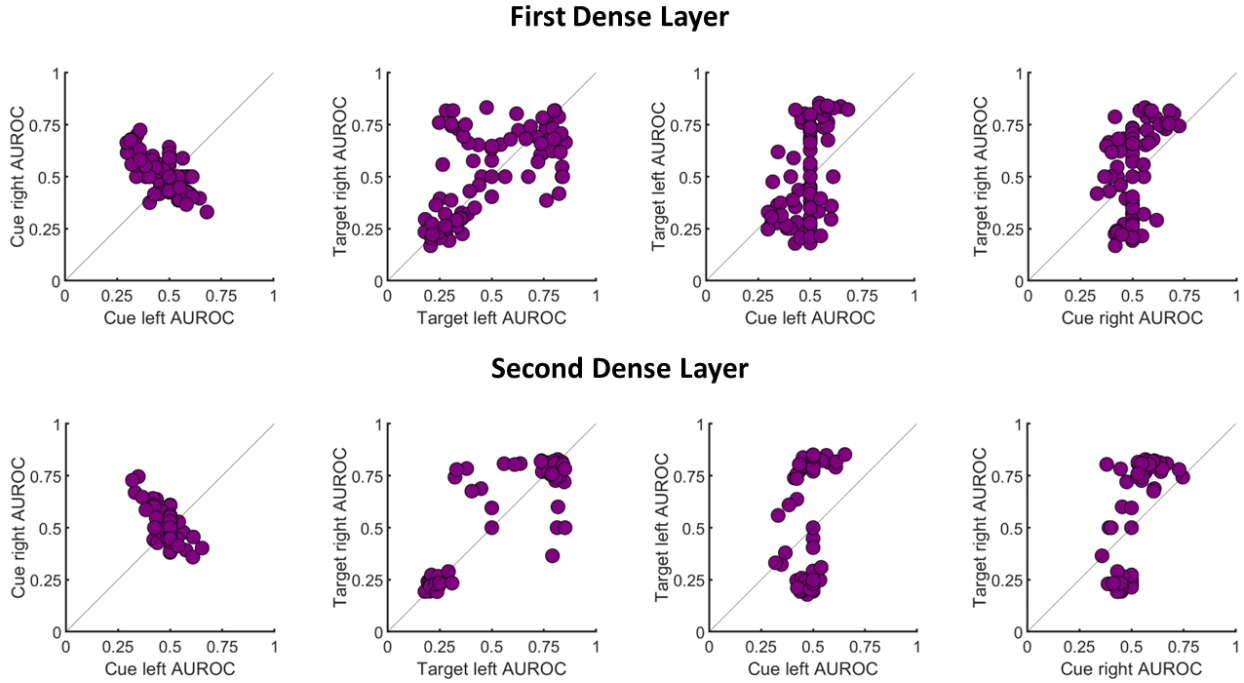
S.12 Effect of increasing the number of layers

We tried two ways of increasing network depths: adding another dense layer or adding one or two more convolutional layers. The neuronal properties depend on which integration stage the neuron is in, and not the depth of the network. The network performance stays in the same range in all cases (between 0.73-0.75).

S.12(a) The effect of depth of the network: adding another dense layer

When an additional dense layer is added, it shows similar integration across locations and of cue and target as the first dense layer (Supplementary Figure S.9).

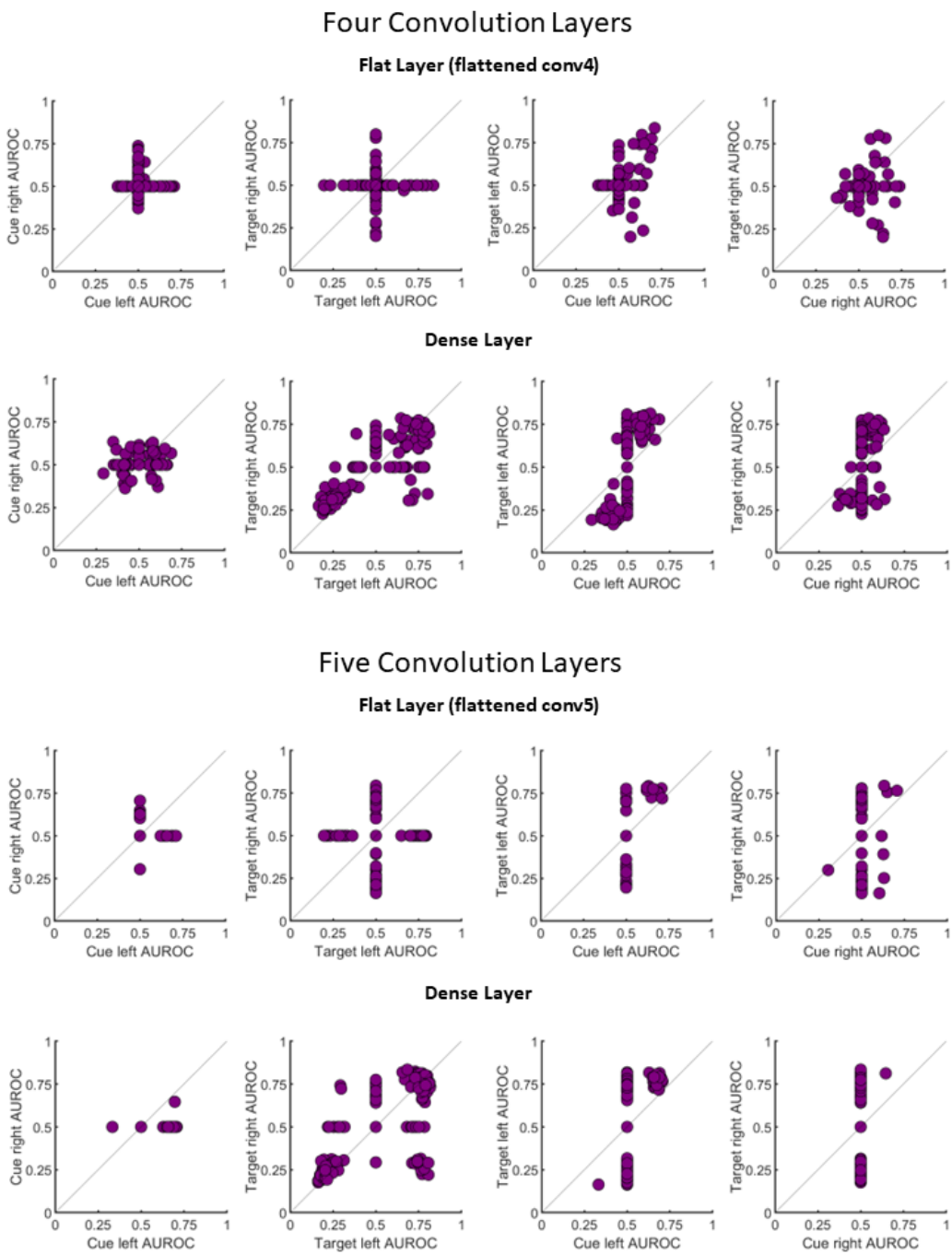
Adding the dense layer did not greatly change the output correlation with the BIO (0.6612 vs 0.6832 noted for the ten networks above).



Supplementary Figure S.9| The effect of adding another dense layer. The rest of the network architecture was kept the same as the network in the main text. The first dense layer has 256 neurons while the second dense layer has 100 neurons. The second dense layer doesn't look qualitatively different than the first dense layer or most other dense layers reported throughout the main text and the supplement.

S.12(b) The effect of depth of the network: adding more convolution layers

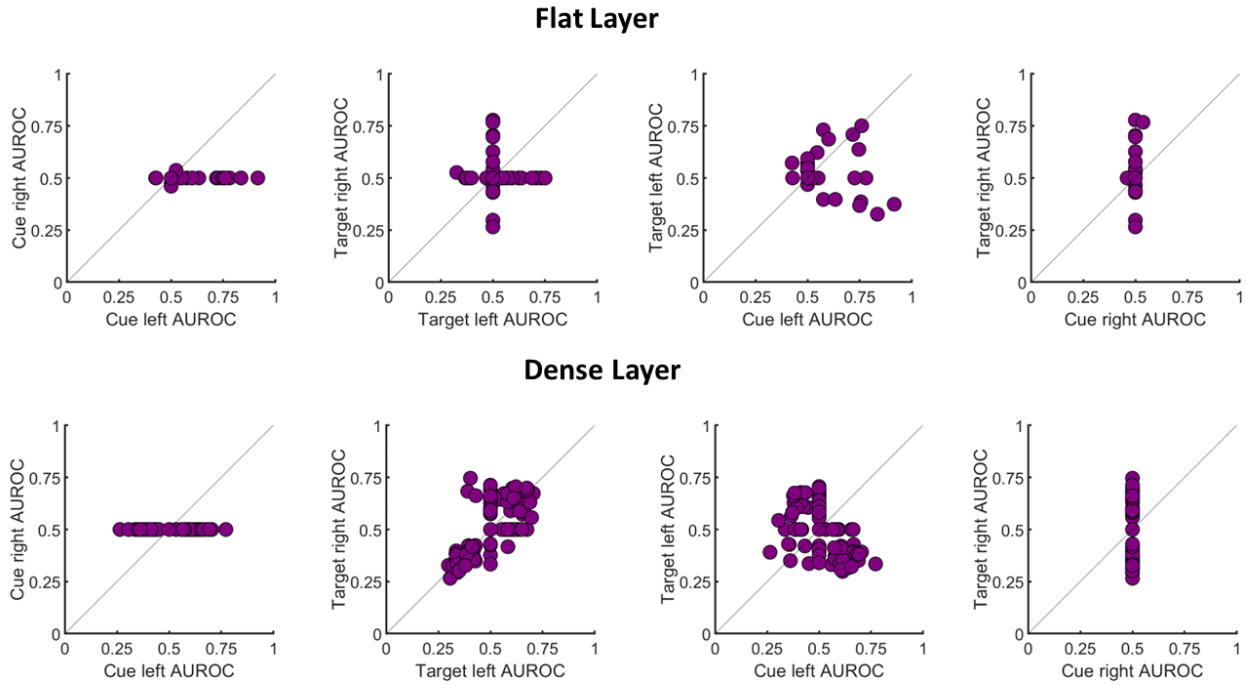
Increasing depth via convolutions does increase the extent of local target-cue integration in the flat layer but not across locations (See Supplementary Figure S.10). Adding more convolution layers changes the output correlation with the BIO (0.7324 for four convolution layers and 0.8110 for five).



Supplementary Figure S.10| The effect of adding another more convolution layers. The rest of the network architecture was kept the same as the network in the main text. The flat layer in both cases refers to the flattened last convolution layer. Increasing the number of flat layers improves local target-cue integration in the flat layers (last two columns; contrast with Figure 6 in the main text), but not the integration across location (first two columns; compare with Figure 7 in the main text).

S.13 Effect of increasing convolution stride size

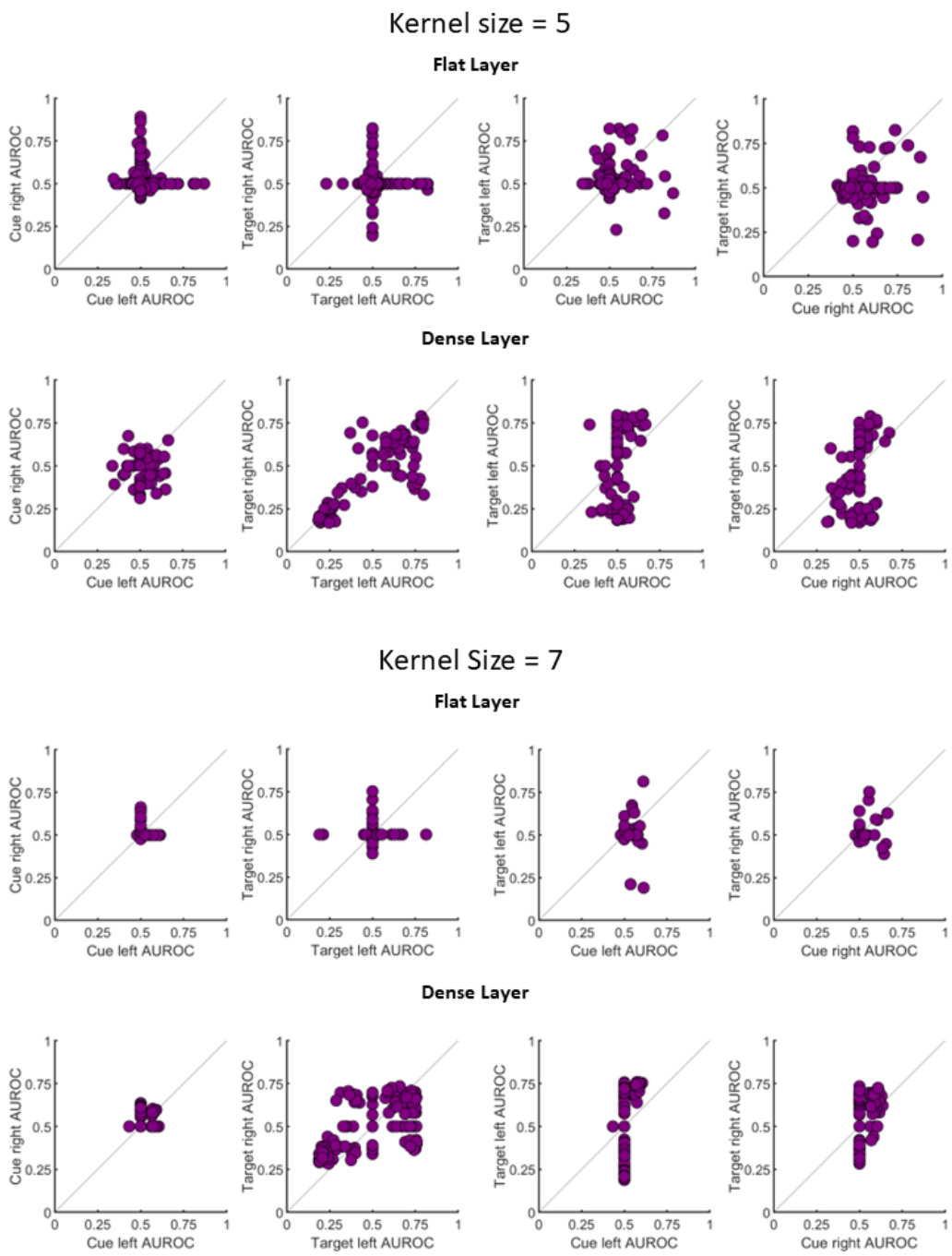
Increasing the stride size of all convolutions from 3 to 5 reduced the network performance to 0.71 while increasing the local target-cue integration on the left location but not the right in the flat layer (Supplementary Figure S.11). The output correlations with the BIO were also reduced to 0.625. These results are perhaps due to the loss of resolution due to the strides.



Supplementary Figure S.11| The effect of increasing the stride of convolution layers. The rest of the network architecture was kept the same as the network in the main text, but the stride of all convolutions was increased from 3 to 5. The local target-cue integration on the left location was improved in the flat layer (third column, first row. Compare with Figure 6 in the main text). However, the cue-integration and the target-cue integration on the right in the dense layer suffer, perhaps due to the loss of resolution due to downsampling (second row, first and last columns. Compare with Figures 6 and 7 in the main text).

S.14 Effect of increasing kernel size

We tried kernel sizes 5 and 7 besides the existing networks that have kernel size 3. In both cases, the network performance increases (0.762 and 0.778, respectively), the amount of local target-cue integration increases in the flat layers but not the integration across locations (Supplementary Figure S.12), and the output correlation with the BIO also increases (0.77 and 0.72 respectively).



Supplementary Figure S.12| The effect of increasing the kernel size of convolution layers. The rest of the network architecture was kept the same as the network in the main text, but the kernel size of all convolutions was increased from 3 to 5 or 7. The local target-cue integration was improved in the flat layers (The last two columns. Compare with Figure 6 in the main text) but not the integration across the two locations.

S.15 What causes higher BIO correlation in the dense layer?

Between the flat and the dense layer, there is a large increase in both the mean and the max absolute correlations. This suggests that the integration across locations is a crucial aspect of having a high correlation with the BIO. The failure of the flat layer to have a comparable correlation is not due to the network depth. When more convolution layers are added, the flat layer (the last convolution layer flattened) still has less correlation than the dense and output layers. With four convolution layers, the flat layer shows a max absolute correlation of 0.44 with a mean absolute of 0.044. In contrast, the dense layer has a max absolute of 0.64 with a mean absolute of 0.3576, and the output correlation is 0.7324. With five convolution layers, the flat layer has a max absolute of 0.4912 with a mean absolute of 0.1754. In comparison, the dense layer has a max absolute of 0.7635 and a mean absolute of 0.5031. The output correlation is 0.8110.

If the depth is increased via the addition of just another dense layer instead of more convolution layers, both the dense layers show high correlations with mean absolutes of 0.39 and 0.51, respectively, and max absolutes of 0.68 and 0.71, respectively.

Increasing the stride or the kernel size does increase local target-cue integration in the flat layer (Supplementary Figures S.11 and S.12), but these methods also do not increase the correlations in the flat layers. For a network with a convolution stride size of 5, the flat layer gets a max absolute correlation of 0.368, while the dense layer gets a max absolute of 0.53. For a network with a kernel size of 5, the flat layer gets a max absolute of 0.49 compared to the dense whose max absolute is 0.75.

When the kernel size is increased to 7, the flat layer gets a max absolute of 0.48, compared to the dense max absolute of 0.72.

Together, these additional controls show that it is not the depth of the network but the spatial integration across the two locations that causes a high correlation with the BIO output. The high correlations at the dense and output levels are also not simply due to the hierarchical structure of the network. A randomly initialized network gets an output correlation of 0.023, a dense max absolute of 0.025, and a mean absolute of 0.05.

The dense layer had some neurons with comparable correlation to the output layer, and the neurons with the highest correlation were neurons that integrate the target across the two locations.

S.16 Effect of noise

While our analysis uses networks that were trained with noise to bring their performance in a comparable range to human subjects in [1], the point of this noise fitting is to make a visual comparison between humans and the models easier. The results, however, do not depend on the noise. The model still shows cueing in the absence of pixel noise (because there is noise in the target and distractor angles – the same angle noise as in the images shown to humans.). When there is no noise, the model achieves an accuracy of 0.78 with a difference of 0.15 in the valid and invalid hit rates. The correlation with BIO is 0.88 at the output level, and in the dense layer, it achieves a mean absolute correlation of 0.62 and a max absolute correlation of 0.86 with the BIO. The max absolute in the flat layer is 0.55, and the mean absolute is 0.19.

Supplementary References

1. Srivastava, S., Wang, W. Y. & Eckstein, M. P. Emergent human-like covert attention in feedforward convolutional neural networks. *Current Biology* **34**, 579-593.e12 (2024).
2. Burgess, A. E., Wagner, R. F., Jennings, R. J. & Barlow, H. B. Efficiency of human visual signal discrimination. *Science* **214**, 93–94 (1981).
3. Burgess, A. E. & Ghandeharian, H. Visual signal detection. II. Signal-location identification. *J Opt Soc Am A* **1**, 906–910 (1984).
4. Geisler, W. S. *Ideal Observer Analysis*. (MIT Press, Boston, 2003).
5. Geisler, W. S. Contributions of ideal observer theory to vision research. *Vision Research* **51**, 771–781 (2011).
6. Landy, M. S., Banks, M. S. & Knill, D. C. Ideal-observer models of cue integration. *Sensory cue integration* 5–29 (2011).
7. Liu, Z., Knill, D. C. & Kersten, D. Object classification for human and ideal observers. *Vision Res* **35**, 549–568 (1995).
8. Shimozaki, S. S., Eckstein, M. P. & Abbey, C. K. Stimulus information contaminates summation tests of independent neural representations of features. *J Vis* **2**, 354–370 (2002).
9. Burge, J. Image-Computable Ideal Observers for Tasks with Natural Stimuli. *Annual Review of Vision Science* **6**, 491–517 (2020).
10. 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).
11. 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).
12. 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).

13. 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).
14. Schoonveld, W., Shimozaki, S. S. & Eckstein, M. P. Optimal observer model of single-fixation oddity search predicts a shallow set-size function. *Journal of Vision* **7**, (2007).
15. Vincent, B. T. Bayesian accounts of covert selective attention: A tutorial review. *Atten Percept Psychophys* **77**, 1013–1032 (2015).
16. Vincent, B. T., Baddeley, R. J., Troscianko, T. & Gilchrist, I. D. Optimal feature integration in visual search. *J Vis* **9**, 15.1–11 (2009).
17. Ma, W. J., Navalpakkam, V., Beck, J. M., Berg, R. van den & Pouget, A. Behavior and neural basis of near-optimal visual search. *Nat. Neurosci* **14**, 783–790 (2011).
18. Eckstein, M. P., Whiting, J. S. & Thomas, J. P. Role of knowledge in human visual temporal integration in spatiotemporal noise. *J Opt Soc Am A Opt Image Sci Vis* **13**, 1960–1968 (1996).
19. Burgess, A. E. & Ghandeharian, H. Visual signal detection. I. Ability to use phase information. *J Opt Soc Am A* **1**, 900–905 (1984).
20. Peterson, W., Birdsall, T. & Fox, W. The theory of signal detectability. *IRE Professional Group on Information Theory* **4**, 171–212 (1954).