bioRxiv preprint doi: https://doi.org/10.1101/2022.03.04.483026; this version posted March 7, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Behavioral benefits of spatial attention explained by multiplicative gain, not receptive field shifts, in a neural network model

Kai J. Fox*[1,3], Daniel Birman*[2,3], and Justin L. Gardner[1]

[1]Department of Psychology, Stanford University, Stanford, CA 94305, USA

[2]Department of Biological Structure, University of Washington, Seattle, WA 98195, USA

[3]Correspondence to: kaifox@stanford.edu or dbirman@uw.edu

*Authors contributed equally to this work

March 4, 2022

Abstract

Attention allows us to focus sensory processing on behaviorally relevant aspects of the visual world. Directing attention has been associated with a number of changes in sensory representation including multiplicative gain as well as shifts in the size and location of neuron receptive fields in early visual cortex. But, which, if any, of these physiological effects can account for the behavioral benefits of attention? Here we use a large scale computational model of primate visual cortex to perform a set of experiments in which we decouple changes in spatial tuning from changes in gain. We show that increased gain at cued locations in a neural network observer model mimics the improvement of human subjects on an attentional task with a spatial cue. Increasing gain resulted in changes in receptive field size and location similar to physiological effects, yet when we forced the model to use only these spatial tuning changes the model failed to produce any behavioral benefit. Instead, we found that gain alone was both necessary and sufficient to explain behavioral improvement during attention. Our results suggest that receptive field shifts are a result of the signal gain that boosts behavioral performance rather than the core mechanism of spatial attention.

bioRxiv preprint doi: https://doi.org/10.1101/2022.03.04.483026; this version posted March 7, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Introduction

Deploying goal-directed spatial attention towards important visual locations allows observers to detect targets with higher accuracy (Hawkins et al., 1990), faster reaction times (Posner, 1980), and higher sensitivity (Sagi & Julesz, 1986) providing humans and non-human primates with a mechanism to select and prioritize spatial visual information (Carrasco, 2011). At the same time as behavioral responses are enhanced, sensory responses near attended locations are amplified (Connor et al., 1996; McAdams & Maunsell, 1999) and the receptive fields of neural populations change shape and size, typically shrinking and shifting towards the target of attention (Anton-Erxleben et al., 2009; Ben Hamed et al., 2002; Kay et al., 2015; Klein et al., 2014; van Es et al., 2018; Vo et al., 2017; Womelsdorf et al., 2006). These changes in neural representation are thought to contribute to behavioral enhancement, but because gain effects and changes in spatial properties of receptive fields co-occur in biological systems, it is not possible to disentangle which of these changes gives rise to improved behavior. Computational models of the visual system allow us to design experiments to independently examine the effects of such changes (Eckstein et al., 2000; Lindsay & Miller, 2018; Pelli, 1985).

Shrinkage and shift of receptive fields toward attended targets has been observed in single unit recordings (Anton-Erxleben et al., 2009; Womelsdorf et al., 2006) and in population activity measured with functional imaging (Fischer & Whitney, 2009; Klein et al., 2014; van Es et al., 2018; Vo et al., 2017). These physiological changes could cause behavioral enhancement through a variety of possible mechanisms (Anton-Erxleben & Carrasco, 2013): receptive field changes might magnify the cortical representation of attended regions (Moran & Desimone, 1985), select for relevant information (Anton-Erxleben et al., 2009; Sprague & Serences, 2013), reduce uncertainty about spatial position (Vo et al., 2017), increase spatial discriminability (Fischer & Whitney, 2009; Kay et al., 2015), or change estimates of perceptual size (Anton-Erxleben et al., 2007). Compression of visual space is also observed just prior to saccades and thought to shift receptive fields towards the saccade location (C. L. Colby & Goldberg, 1999; Merriam et al., 2007; Zirnsak et al., 2014) and maintain a stable representation of visual space (C. Colby, Goldberg, et al., 1992; Kusunoki & Goldberg, 2003; Ross et al., 1997; Tolias et al., 2001). It is also possible that shift and shrinkage of receptive fields occur as a side effect of amplifying neural responses in an asymmetrical way across a receptive field (Klein et al., 2014), raising the question of how these two effects combine to enhance perception.

We took a modeling approach to address whether receptive field shift and shrinkage are responsible for the behavioral enhancement of spatial attention or a side-effect of neural gain. We modified a convolutional neural network (CNN) to test various hypotheses by implementing them as elements of the model architecture. CNN architectures can be designed to closely mimic the primate visual hierarchy (Kubilius et al., 2018; Yamins et

bioRxiv preprint doi: https://doi.org/10.1101/2022.03.04.483026; this version posted March 7, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

al., 2014). Training "units" in these networks to categorize images leads to visual filters that show a striking qualitative resemblance to the filters observed in early visual cortex (Krizhevsky et al., 2012) and the pattern of activity of these units when presented with natural images is sufficient to capture a large portion of the variance in neural activity in the retina (McIntosh et al., 2016), in early visual cortex (Cadena et al., 2019), and in later areas (Cichy et al., 2016; Eickenberg et al., 2017; Güçlü & van Gerven, 2015; Khaligh-Razavi & Kriegeskorte, 2014; Yamins et al., 2014). Cortical responses and neural network activity also share a correlation structure across natural image categories (Storrs et al., 2020). These properties of CNNs make them a useful tool which we can use to indirectly study visual cortex, probing activity and behavior in ways that are impractical in humans and non-human primates (Lindsay & Miller, 2018).

Using simulations based on a CNN observer model we found that the benefits of spatial attention are explained by multiplicative gain alone. We designed a simple cued object-detection task and measured improved human performance on trials with focal attention. Using a CNN whose architecture was designed to maximize similarity with the primate visual stream we measured a similar improvement in detection performance when a Gaussian gain augmented inputs coming from a "cued" location. We found that the network mirrored the physiology of human and non-human primates: units shifted their center-of-mass toward the locus of attention and shrank in size, all in a gain-dependent manner. We isolated each of these physiological changes to determine which, if any, could account for the behavioral effects. A model with only gain reproduced the behavioral benefits of attention while models with only receptive field shifts or only changes in receptive field sensitivity were unable to provide any behavioral benefit. We conclude that gain changes alone are both necessary and sufficient to account for the behavioral benefits of spatial attention.

Results

We characterized the ability of human observers to detect objects in a grid of four images, with or without prior information about the object's possible location (Fig. 1). Observers were given a written category label, e.g. "ferris wheel", and shown five exemplar images of that category (Category intro, 1a). This was followed by a block of 80 trials in which observers tried to detect the presence or absence of the target category among the four images in the grid (Each trial, 1a). Half of the 80 trials had focal cues and 50% of the focal (and distributed) trials included a target image. On focal trials a cue indicated with 100% validity the grid quadrant that could contain a target while on distributed trials no information was given as to where an image of the target category could appear. Distractor images were randomly sampled from the nineteen non-target image categories. Stimulus

bioRxiv preprint doi: https://doi.org/10.1101/2022.03.04.483026; this version posted March 7, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

durations were sampled uniformly from 1 (0.008 s), 2 (0.016), 4 (0.033), 8 (0.066), 16 (0.133), or 32 (0.267) frames (Stimulus, 8.33 ms per frame, 1a). Image grids were masked before and after stimulus presentation by shuffling the pixel locations in the stimulus images, ensuring that the luminance during each trial remained constant. Observers had 2 s to make a response and each trial was followed by a 0.25 s inter-trial interval. Observers completed one training block on an unused category prior to data collection.

Human observers improved their performance on this detection task when given a focal cue indicating the potential location of a target (Fig. 1b). At a stimulus duration of 8 ms (one frame) observers were near chance performance regardless of cueing condition. On distributed trials observers exceeded threshold performance ($d' = 1$) at a stimulus duration of 155 ms, 95% CI [135, 197]. For focal trials, the same threshold was reached with only a 38 ms [32, 43] stimulus duration, demonstrating a substantial performance benefit of focal cueing. We characterized this performance benefit by fitting a logarithmic function to the data, scaled by a parameter α for the focal condition. We found that d' in the focal condition was higher than in the distributed condition, average increase across observers $\alpha = 1.67\times$ [1.57, 1.74]. Across all observers the d' function was best fit as:

$$d'(ms) = \alpha \log(163.588ms + 1) \tag{1}$$

Using a drift diffusion model we found that the majority of this performance benefit came from the focal cue, rather than speed-accuracy trade off. We assessed this by fitting a drift diffusion model to the reaction time and choice data (Wagenmakers et al., 2007). Drift diffusion models assume that responses are generated by a diffusion process in which evidence accumulates over time toward a bound. We used the equations in Wagenmakers et al. (2007) to transform each observer's percent correct, mean reaction time, and reaction time variance for the twenty categories and two focal conditions into drift rate, bound separation, and non-decision time. The drift rate parameter is designed to isolate the effect of external input, the non-decision time reflects the fastest responses an observer makes, and the bound separation is a proxy for how conservative observers are. Comparing the drift rate parameter we observed a similar effect to what was described above for d': the average drift rate across observers in the focal condition was 1.61$\times$, 95% CI [1.39, 1.77] the drift rate in the distributed condition. This suggests that the majority of the performance gain observed in the d' parameter came from increased stimulus information. We did find that the other parameters of the drift diffusion model were also sensitive to duration and condition, but in opposite directions. We found larger bound separation at longer stimulus durations and on focal trials (focal bound-separation 1.57$\times$ distributed [1.37, 1.75]), consistent with observers being more conservative on trials where more information was available. But this increase in cautiousness was offset by a shorter non-decision time on focal trials (0.26 s) compared to distributed (0.38, [0.34, 0.41]).

bioRxiv preprint doi: https://doi.org/10.1101/2022.03.04.483026; this version posted March 7, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

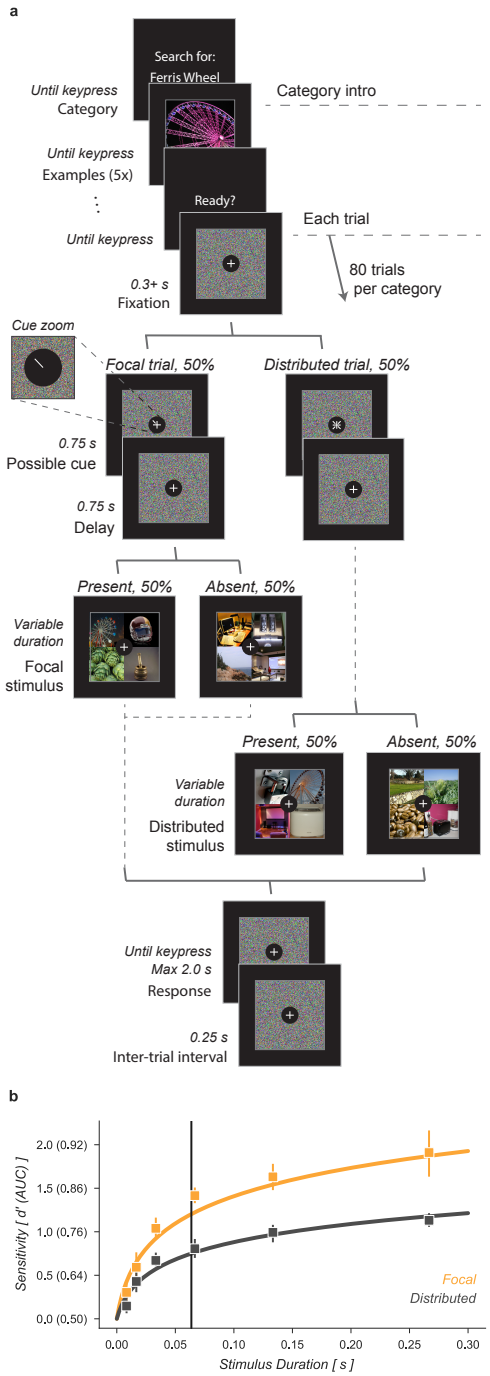

Figure 1: Cued object detection task. (a) Observers were asked to perform object detection with or without a spatial cue. At the start of a block, observers were shown five examples of the target category. This was followed by 80 trials: 40 with a spatial cue indicating the possible target quadrant and 40 with no prior information. Stimulus presentation was pre and post-masked. The stimuli consisted of a composite image of four individual object exemplars. The target category was present in 50% of trials and always in the cued location on focal trials. Human observers used a keyboard to make a fast button response to indicate the target presence before moving on to the next trial. (b) Human observers showed a substantial improvement in performance when given a focal cue indicating the quadrant at which the target might appear. Vertical line at 64 ms indicates the duration at which the best-fit d' curve for the Distributed condition matched CNN model performance without gain. Markers indicate the median and error bars the 95% confidence intervals.

bioRxiv preprint doi: https://doi.org/10.1101/2022.03.04.483026; this version posted March 7, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Having shown that a spatial cue provides human observers with increased stimulus information in this task, we next sought to show that a neural network model of the human visual stream could replicate this behavior under similar conditions. We used a convolutional neural network (CNN) model, CORnet-z (Kubilius et al., 2018), a neural network designed to mimic primate V1, V2, V4, and IT and optimized to perform object recognition for images at a similar scale to our task. We added to this model a set of output layers to predict the presence of the twenty object categories, thus creating a neural network observer model, i.e. a model designed to idealize the computations performed by human observers performing the 4-quadrant object detection task. We applied the observer model to a task analogous to the one human observers performed (Fig. 2). The prediction layers added to the end of the model provided independent readouts for the presence or absence of the different target categories (Linear classifier, Fig. 2c). These output layers were trained on a held out set of full-size images from each category.

To examine the computational mechanisms that could underlie the performance benefit of the focal cue we added a multiplicative Gaussian gain centered at the location of the cued image (Fig. 2b). We applied this gain at the first layer of the model, analogous to a gain signal modulating responses in primate V1, and tested various strengths of gain.

To align the human and model performance for this task we took the performance of the model in the distributed condition (Distributed, Fig. 2a) and found the stimulus duration at which subjects in the distributed condition of the human data matched this performance level (64 ms, Fig. 1b). We then scaled up the amplitude of the Gaussian gain incrementally and found that we could mimic the performance enhancement of attending from the human data by setting the maximum of the Gaussian gain field to approximately $4\times$. The model with this level of gain had a median AUC across categories of 0.80, 95% CI [0.77, 0.82] compared to 0.71 [0.67, 0.72] without gain and a median AUC improvement of 0.09 [0.08, 0.12] within each category.

The gain strengths necessary to induce the behavioral effect in the neural network observer model were relatively large compared to the gain due to directed attention observed in measurements of single unit (Luck et al., 1997; Treue & Trujillo, 1999) and population (Birman & Gardner, 2019) activity. We attribute this difference to the lack of any non-linear "winner-take-all" type of activation in the CNN. In the primate visual system, it is thought that non-linearities such as exponentiation and normalization can accentuate response differences (Carandini & Heeger, 2012) and act as a selection mechanism for sensory signals (Pestilli et al., 2011). We tested whether similar non-linear mechanisms would allow for smaller gain strengths to be amplified to the range needed by our model by raising the activations of units by an exponent before re-normalizing the activation of all units at the output of each layer (see Methods for details). This has the effect of amplifying active units and further

bioRxiv preprint doi: https://doi.org/10.1101/2022.03.04.483026; this version posted March 7, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

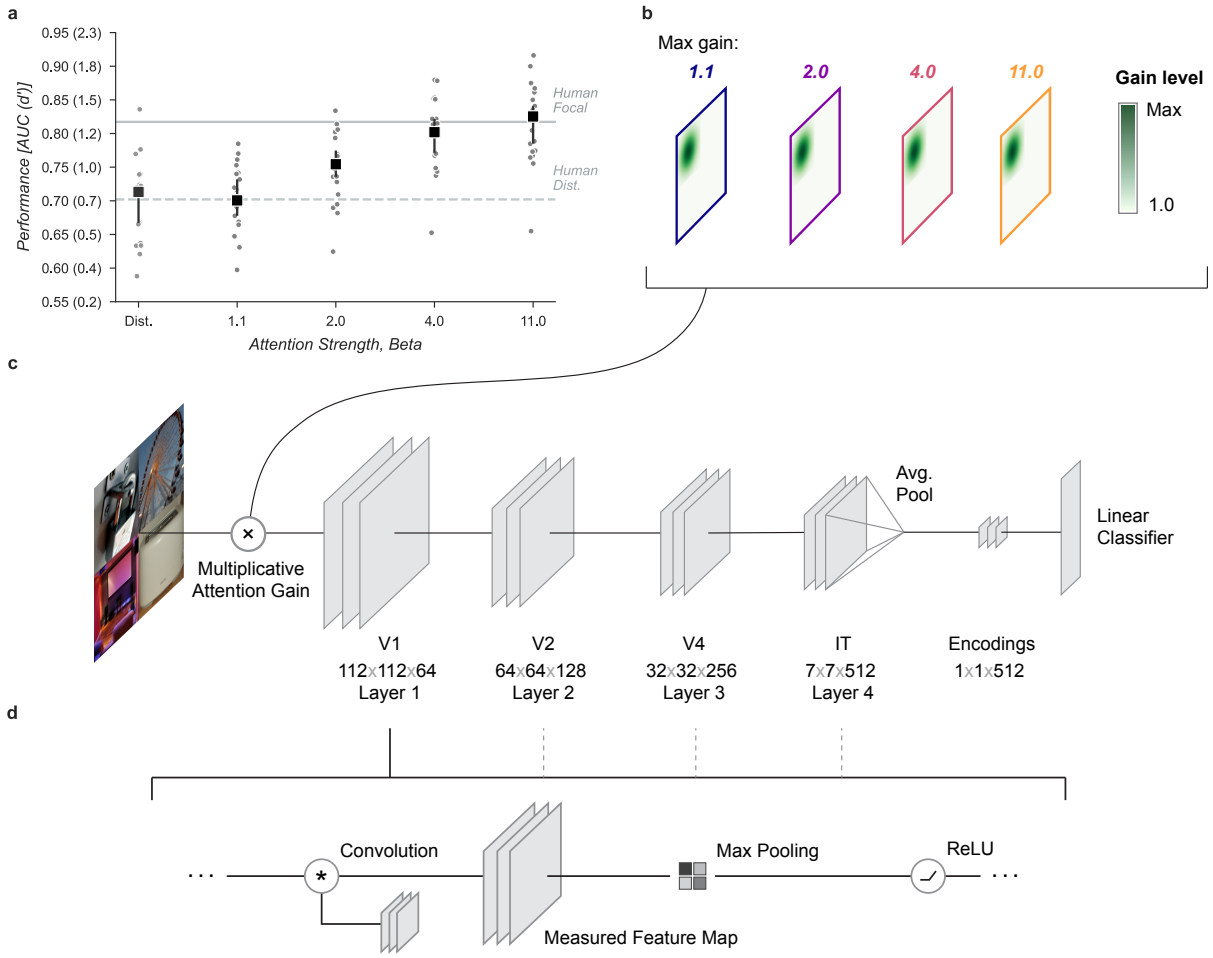

Figure 2: Neural network observer model. (a) Using a Gaussian gain the neural network observer was able to replicate the behavioral benefit of spatial attention for human observers. Human performance is shown at a stimulus duration of 64 s which provided the closest match to the convolutional neural network (CNN) performance without gain. Markers indicate the median and error bars the 95% confidence intervals. (b) The Gaussian gain was implemented by varying the maximum strength of a multiplicative gain map applied to the "cued" quadrant. (c) The gain was applied prior to the first layer of the convolution neural network (CNN). The neural network observer model consisted of a four layer CNN with linear classifiers applied to the output layer. Individual classifiers were trained on examples of each object category. (d) Each of the four convolutional layers consisted of a convolution operation followed by max pooling and a rectified linear unit. Unit activations were measured after the convolution, prior to the max pooling step.

bioRxiv preprint doi: https://doi.org/10.1101/2022.03.04.483026; this version posted March 7, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

suppressing inactive ones. Using this approach we found that a relatively small gain of 1.25× combined with an exponent of 3.8 led to a much larger effective gain of 2.09× after just one layer (Fig. 3j). This form of non-linearity is consistent with the finding that static output non-linearities in single units range from about 2 to 4 (Albrecht & Hamilton, 1982; Gardner et al., 1999; Heeger, 1992; Sclar et al., 1990) and thus this simulation suggests a plausible physiological mechanism by which the larger gains predicted by our model could be implemented. Repeated use of exponentiation and normalization in successive layers of the visual system could produce an even larger effective gain. To avoid training a new convolutional neural network and possibly violate the close relationship between the primate visual system and the CNN we studied, we continued our analysis without introducing an exponentiation and normalization step.

The Gaussian gain could have its effect on the neural network observer model's performance by increasing the activation strength of units with receptive fields near the locus of attention. These changes in activation strength might directly affect behavior, or work indirectly through mechanisms such as changes in receptive field size, location, or spatial tuning. We observed all of these effects in our model (Fig. 3). To measure receptive fields we computed the derivative of each unit with respect to the input image and then fit these with a 2D Gaussian (see Methods for details). We found that the gain caused receptive fields to shift and shrink toward the locus of attention (Fig. 3a,b). The receptive field shift and shrinkage were magnified in deeper layers of the model (Fig. 3d,e) consistent with physiological observations (Klein et al., 2014). The gain in activation strength propagated through the network without modification (Fig. 3f). To measure the effective gain experienced by the layer four units (Fig. 3i) we computed the ratio of the standard deviations of unit activations after the nonlinear ReLU function (Fig. 2d) with and without gain applied. The gain also produced a non-linear change in the information represented by units late in the model (Fig. 3c), reflecting a change in units' spatial tuning rather than simply a propagation of gain. We measured this by comparing the correlation of layer 4 unit activations with and without the Gaussian gain. All three observed effects: receptive field shift, shrinkage and expansion, and effective gain were directly related to the gain strength at the input layer (Fig. 3g-i). All of these changes have been proposed as mechanisms that could account for the behavioral benefits of attention. We designed models to try to isolate these effects with the goal of testing their independent contributions to behavior.

We next sought to test whether receptive field shifts alone could account for the behavioral benefits of the neural network observer model. To do this, we built a model variant that could shift receptive fields without introducing gain. To develop an intuition for how this could affect perceptual reports, consider a CNN with just four units in a 2×2 grid with each unit having its receptive field centered on one image in the composite. When shown a composite grid of four images, a logistic regression using the output of these four units would receive one quarter the information it expects from being trained on full size images. Shifting the receptive fields of the

bioRxiv preprint doi: https://doi.org/10.1101/2022.03.04.483026; this version posted March 7, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

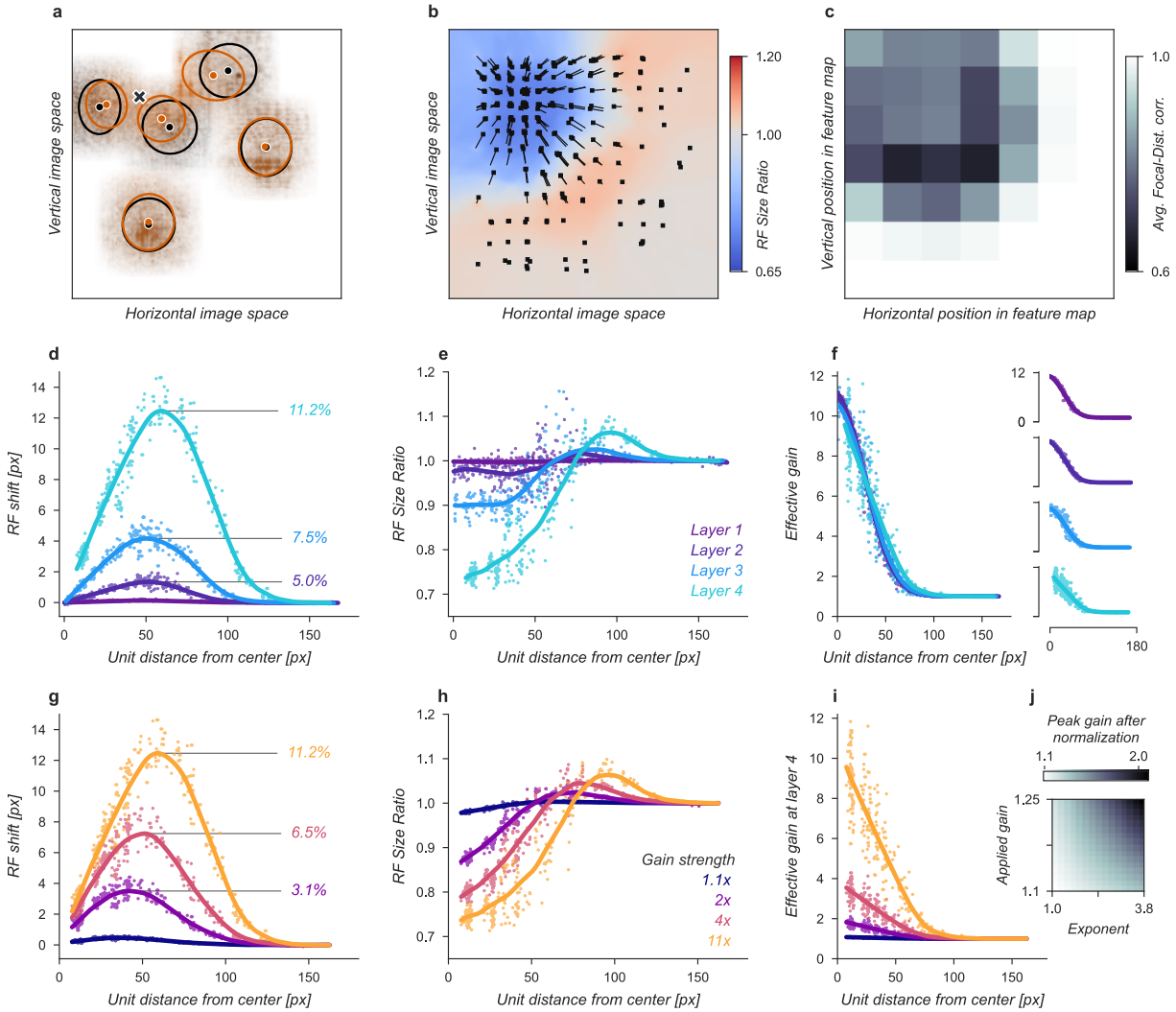

Figure 3: Effects of Gaussian gain on neural network units. (a) The Gaussian gain applied to Layer 1 units caused the measured receptive field (RF) of units in Layer 4 to shift (black ellipse, original; brown ellipse, with gain) toward the locus of attention (black $\times$). (b) A 2D spatial map demonstrates the effects of Gaussian gain in Layer 4: shift of RF center position (black arrows), shrinking RF size near the attended locus (blue colors) and an expansion of size near the gain boundaries (red colors). (c) A spatial correlation map of the $7\times7\times512$ output layer between the focal and distributed conditions demonstrates that the information content of the output layer has been affected by the Gaussian gain in ways beyond simple scaling. (d,e) Scatter plots demonstrate that each layer magnifies the effect of the gain on RF shift and RF size. The RF shift percentages are the ratio of pixel shift at the peak of the curve relative to the average receptive field size, measured as the full-width at half-maximum. (f) Later layers do not magnify the effective gain (shown for an $11\times$ gain), which stays constant across layers. Gain strength does influence the size of RF position shifts (g), RF size (h), and effective gain (i). (j) Adding an additional non-linear normalizing exponent at the output of each layer allows for much smaller gains to be magnified across layers.

bioRxiv preprint doi: https://doi.org/10.1101/2022.03.04.483026; this version posted March 7, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

three non-target units to overlap more with the cued image could add additional task-relevant information to the output.

We designed a variant of our model that could be used to test the hypothesis that receptive field shifts alone are responsible for the behavioral enhancement (Fig. 4). In this model we re-wired the units in the first layer to reproduce the effect of Gaussian gain on receptive fields in the fourth layer based on our measurements of receptive field shift in the Gaussian gain model, as reported in Figure 3g. To mimic those shifts, we changed the connections between the input image pixels and layer one (Fig. 4a). This manipulation worked as designed and changed the receptive field locations and size (Fig. 4b-d) but since no gain was added to the model, the overall responsiveness of units remained constant (Fig. 4e). Because receptive field shifts due to gain are not the result of actual rewiring it is unsurprising that the shift and shrinkage in this model variant are only qualitatively matched to those caused by the original Gaussian gain. Note that the effective gain in layer four *did* change, a result of the units receiving different inputs, but the average change across images was zero.

We found that the model with receptive field shifts but no gain had no effect on task performance, strong evidence against the hypothesis that receptive field shifts are the key to understanding the effects of spatial attention (Fig. 4f). The model imitating shifts from $4\times$ Gaussian gain had a median AUC across categories of 0.71, 95% CI [0.66, 0.73] compared to 0.71 [0.67, 0.72] with no attention and a median change in AUC of -0.01 [-0.02, 0.01] within each category.

Another way to understand the possible effect of the Gaussian gain on task performance is to note that the spatial tuning profile of units is "shifted" towards the locus of attention: sensitivity is enhanced closer to the locus of attention, but the receptive field itself has not truly moved in the manner studied by the previous model. If different parts of a receptive field receive asymmetrical gain, as expected for Gaussian gain, then the local structure of the receptive field has been changed (Fig. 5a). We designed another model variant to test the hypothesis that these local sensitivity shifts alone might be sufficient to explain the behavioral effect without inducing receptive field shifts or gain. To implement this model at layer L, we examined the effect of the Gaussian gain on each unit (green differential gain, Fig. 5a). We normalized this differential gain within each unit's receptive field to prevent any overall gain effect and re-scaled the units kernel accordingly. Overall this manipulation of each unit's kernel preserved a portion of the receptive field shift effect but guaranteed that there was no effective gain.

The sensitivity shift model was designed to only change the spatial tuning of individual units without inducing gain, which naturally caused some shifts in the measured receptive field size and location (solid lines and markers, Fig. 5b-d) but these were smaller than the effects observed under Gaussian gain (dashed lines). The normalization prevented the model from introducing any spatial pattern of gain change (Fig. 5e). Note that there were still

bioRxiv preprint doi: https://doi.org/10.1101/2022.03.04.483026; this version posted March 7, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

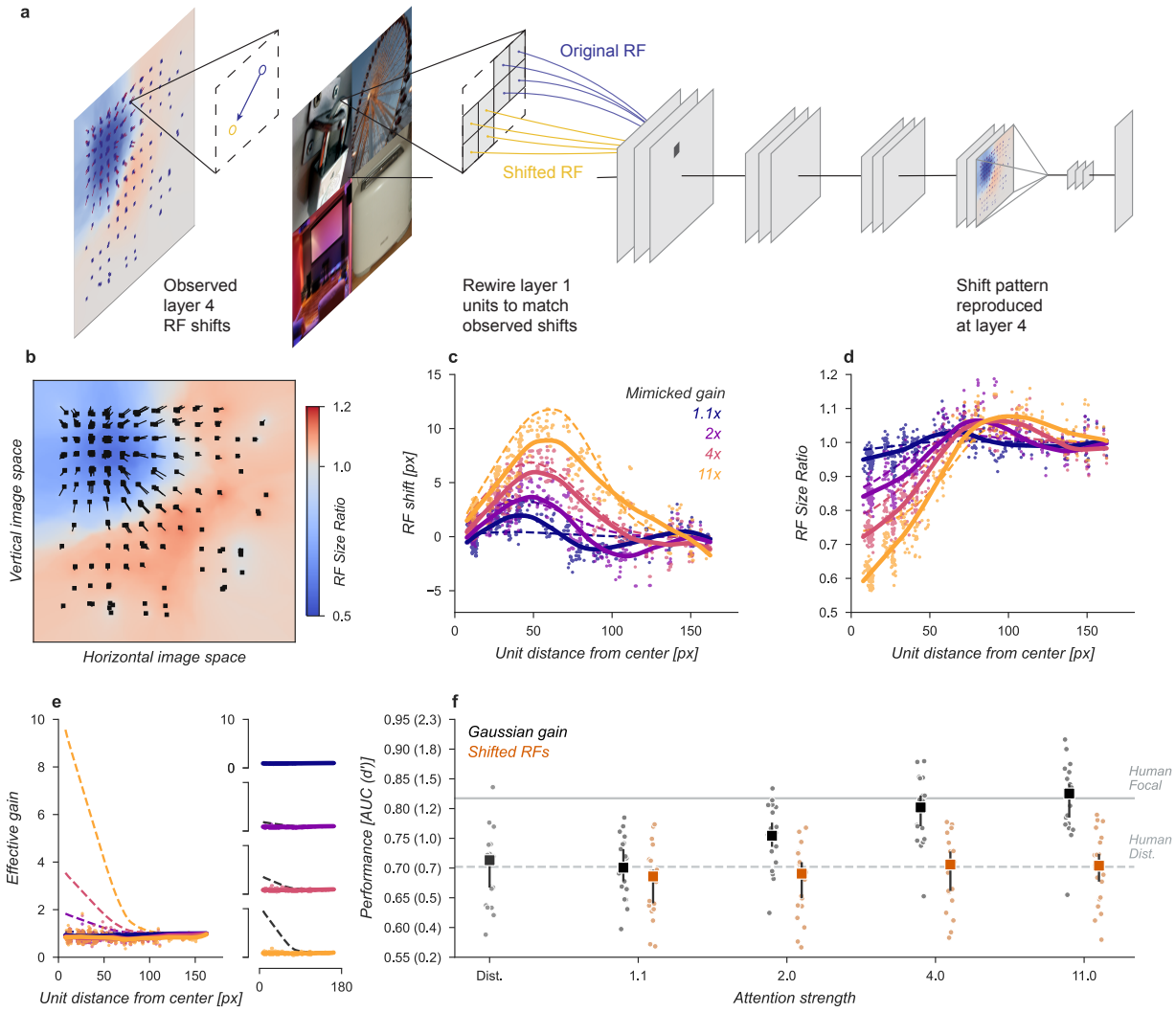

Figure 4: Receptive field shift model. (a) To mimic the effects of the Gaussian gain on receptive field position without inducing gain in the model we re-assigned the inputs to units in Layer 1. This re-assignment was performed so that the pattern of receptive field shift in Layer 4 would match what was observed when the Gaussian gain was applied. (b) The observed pattern of receptive field shifts and shrinkage is shown for a sample of units in layer 4, qualitatively matching the effects of the Gaussian gain. (c) RF shift is shown (solid lines and markers) compared to the effect in the Gaussian gain model (dotted lines). (d) Conventions as in c for the RF size change. (e) Conventions as in c,d for the effective gain of units. (f) The behavioral effect of shifting receptive fields is shown to be null on average across categories when compared to the effect of Gaussian gain. Markers indicate the median and error bars the 95% confidence intervals.

bioRxiv preprint doi: https://doi.org/10.1101/2022.03.04.483026; this version posted March 7, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

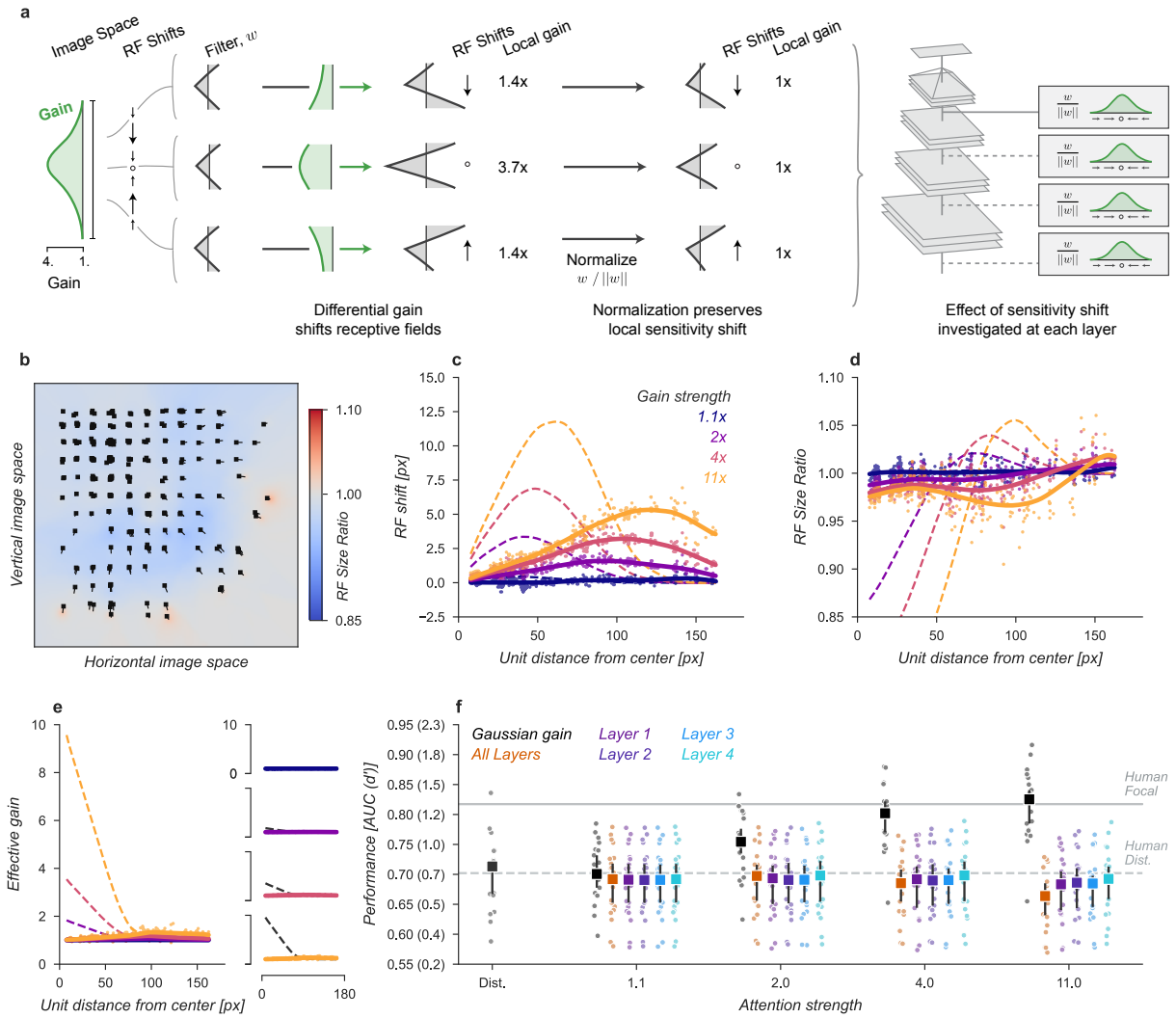

Figure 5: Sensitivity-shift model. (a) We adjusted the kernels of each convolutional neural network unit according to the effect of a Gaussian gain, subtly shifting the the sensitivity within individual units. To avoid inducing a gain change we then normalized each units output such that the sum-of-squares of the weights was held constant, ensuring the local gain at that unit remained at 1×. This model was implemented individually at each layer, replicating the effect of a Gaussian gain of 1.1× to 11× as well as at all layers at once. (b-f) conventions as in Fig. 4.

bioRxiv preprint doi: https://doi.org/10.1101/2022.03.04.483026; this version posted March 7, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

small changes in overall sensitivity of units in this model, for example, the $4\times$ model had an average gain of 1.08, 95% CI [1.07, 1.09] across all units, which we attribute to the fact that inputs to a unit may exhibit correlations due to spatial structure. These receptive field changes and small gain effects were distinct from those observed under Gaussian gain (Fig. 5c-e).

The sensitivity shift model, like the receptive field shift model, was unable to account for the behavioral effects of the Gaussian gain. No matter how deep in the model the sensitivity shift was applied, and even when applied at all layers, the average performance across categories remained flat (Fig. 5f). Compared to the median distributed AUC across categories of 0.71 [0.67, 0.72], the sensitivity model applied to all layers had a median AUC across categories of 0.69 [0.65, 0.72] when imitating gain of $1.1\times$, 0.70 [0.65, 0.72] for $2\times$ gain, 0.69 [0.65, 0.71] for $4\times$ and 0.66 [0.63, 0.69] for $11\times$. Each of these conditions resulted in a median AUC change within category of -0.02 [-0.03, 0.00], -0.01 [-0.03, 0.00], -0.02 [-0.04, -0.01], and -0.04 [-0.05, -0.03], respectively. When applied to early layers we observed a slight drop in performance, which we attribute to how this model directly alters the kernels in the CNN. This breaks the assumption that the CNN kernels at each layer are consistent with those that were optimized when the model weights were trained.

Having ruled out that receptive field shift or changes in spatial tuning could account for the behavioral effect, we next designed a model to amplify signals in the cued quadrant without other effects and found that this model was able to explain the behavioral effects of attention. In all of the models explored so far an asymmetry in gain was created within the receptive fields of the units. To remove this effect we flattened the gain within the cued quadrant (Fig. 6a) by setting the gain at each pixel to the average of the Gaussian gain across the entire quadrant. By itself, this change has the unintended consequence that units centered in an uncued quadrant but with receptive fields overlapping the cued quadrant will still shift in a gain-dependent manner. To remove this effect, we split the CNN feature maps into the four quadrants and computed these separately with padding and concatenated the results. This forces all units in the model to receive information about only a single quadrant. The zero padding at the borders causes receptive field shift, but these are now independent of the gain strength.

Using the gain-only model we were able to reproduce the behavioral effect of the original Gaussian gain (Fig. 6). By design, the model induced the same pattern of receptive field shift and size change at all gain strengths (Fig. 6b-d) and a flat effective gain within the cued quadrant (Fig. 6e). We found that increasing the strength of a flat gain was sufficient to capture the full behavioral effect of the original model (Fig. 6f). The median AUC across categories of the $4\times$ flat gain model was 0.78, 95% CI [0.76, 0.83] compared to 0.80 [0.77, 0.82] for the $4\times$ Gaussian gain model. The confidence intervals in flat gain and Gaussian gain performance overlapped at all gain strengths, with a difference of 0.00 [-0.00, 0.02] at 1.1$\times$ gain, -0.01 [-0.02, 0.00] at 2$\times$ gain, -0.01 [-0.02,

bioRxiv preprint doi: https://doi.org/10.1101/2022.03.04.483026; this version posted March 7, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

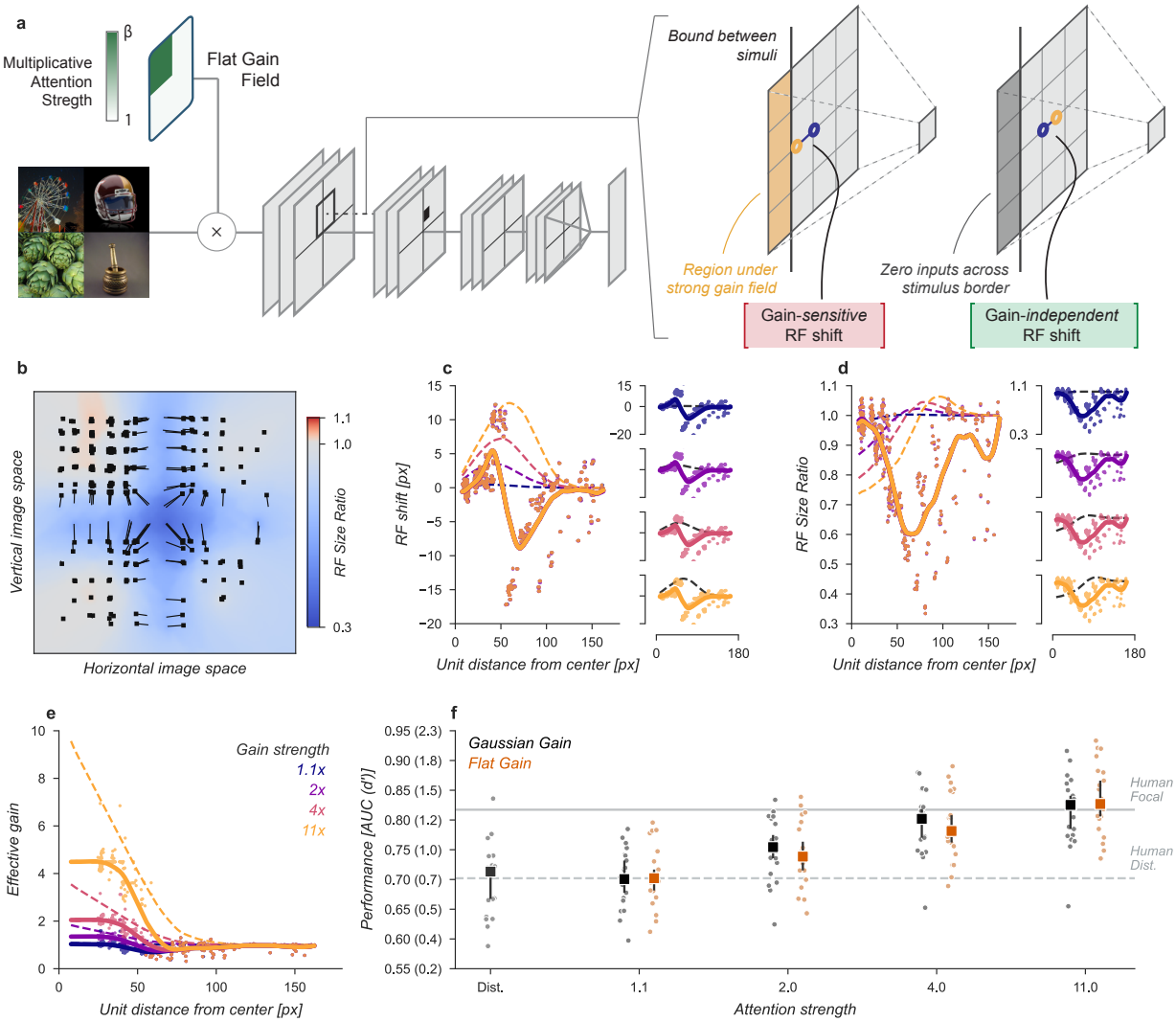

Figure 6: Gain-only model. (a) To create a gain effect without modifying the receptive fields of units we applied a flattened gain field, with the gain set to the average of the original Gaussian gain for each attention strength. The flat gain alone causes units to shift their receptive field at the boundary between the four stimulus quadrants, to modify gain while ensuring shifts were gain-independent we computed the four quadrants separately with zero padding and then concatenated the results. (b-f) conventions as in previous figures.

bioRxiv preprint doi: https://doi.org/10.1101/2022.03.04.483026; this version posted March 7, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

0.00] at 4× gain, and 0.02 [0.00, 0.04] at 11× gain.

Having found that the behavioral enhancement could be explained not by receptive field changes, but instead by the change in the overall activation strength, we asked whether this increased activation strength propagated through the network was both necessary and sufficient to explain behavioral enhancement. To test necessity and sufficiency we ran the task images through the Gaussian gain model (first row, Fig. 7a) and measured the effective gain propagated to units in the final layer output ($7 \times 7 \times 512$, before averaging). We averaged these effective gains over features to obtain a propagated gain map (Layer 4 feature map, 7×7, Fig. 7b). To test the hypothesis that this propagated gain was sufficient to account for the behavioral effect we re-applied it to the output layer of a model with no gain applied.

We found that the propagated gain map, when used to multiply the outputs of a model with no Gaussian gain (Multiply by propagated gain, Fig. 7a) was sufficient to reconstruct the behavioral benefits of Gaussian gain applied to the input (Propagated gain vs. Gaussian gain, Fig. 7c). The median AUC across categories using the propagated gain map was 0.72, 95% CI [0.68, 0.75], compared to 0.71 [0.67, 0.72] in the distributed model. The propagated gain map is not a perfect replacement for the Gaussian gain and we found that within categories there was a median drop in AUC within categories of -0.02 [-0.03, 0.01] when replacing the full Gaussian gain model with the propagated map multiplication.

To test the hypothesis that the propagated gain was necessary to account for the behavioral effect we divided the final layer activations by the propagated gain map (Divide by propagated gain, Fig. 7a). We found that the behavioral effect of an early gain was reversed by this manipulation (Removed gain vs. Distributed, Fig. 7c). The median AUC across categories after dividing out the propagated gain was 0.72, 95% CI [0.68, 0.75], compared to 0.71 [0.67, 0.72] in the distributed condition. Dividing by the propagated map did not perfectly reverse the Gaussian gain, we found a median AUC improvement within categories of 0.02 [0.01, 0.03].

Discussion

Human observers are more accurate when trying to detect objects at a cued location. Our results demonstrate that this behavioral benefit can also be observed in a neural network model of visual cortex when a Gaussian gain is applied over the pixels of a "cued" object. To determine the source of this behavioral benefit we explored different mechanisms of attention in a series of neural network observer models. In the shift-only model we re-wired units to move receptive fields without introducing gain and found that this produced no behavioral benefits. In the

bioRxiv preprint doi: https://doi.org/10.1101/2022.03.04.483026; this version posted March 7, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

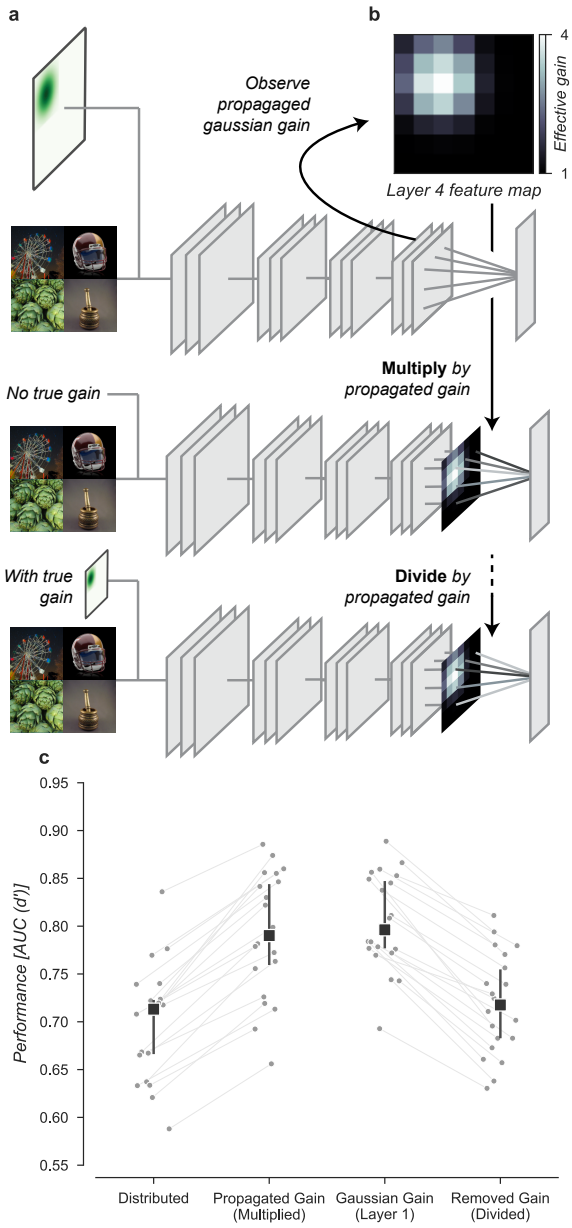

Figure 7: Gain is both necessary and sufficient to explain the behavioral effects of attention. (a) To test necessity and sufficiency of gain on performance we propagated the effect of Gaussian gain through the model and measured the effective gain at the output layer. (b) We averaged the effective gain across features to obtain a "propagated gain map". To test sufficiency we multiplied the output of a model with no true gain by the propagated gain map. To test necessity we divided the output of a model with true gain by the propagated gain map. (c) Multiplying the output by the propagated gain recovered the effect of Gaussian gain, while dividing removed this effect, confirming that gain was both necessary and sufficient to account for the behavior. Markers indicate the median and error bars the 95% confidence intervals.

bioRxiv preprint doi: https://doi.org/10.1101/2022.03.04.483026; this version posted March 7, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

sensitivity-shift model we modified the spatial tuning of individual units to mimic the effect of gain but again found no behavioral benefits. It was only by applying a gain while keeping receptive field properties stable that we were able to reproduce the behavioral benefits of the original Gaussian gain. In line with this, we found that applying gain at the final layer was both necessary and sufficient to account for the behavioral benefits of spatial attention. Our results demonstrate that spatial gain is sufficient to increase the weight of relevant visual information on decision-making. While gain can change receptive field properties, our results suggest that these are secondary effects and only a consequence of applying gain, rather than the cause of the behavioral improvements as others have suggested.

We tested our hypothesis using an image-computable model, which has the benefit that all of the computational steps from sensory input to decision making are constrained by the model architecture. In our case, the advantage of this type of model is that the model architecture encodes the relationship between gain and receptive field shift: any time a gain occurs in an asymmetrical manner across a receptive field, downstream units will show an apparent "shift" in position. Many of the previous studies which have probed the mechanisms of spatial attention have used computational models which linked only the measurements which they made in their experiments, usually without specifying any direct connection between gain and receptive field shift. We know from the large literature exploring the physiology of attention that receptive field shifts are correlated with spatial attention (Anton-Erxleben & Carrasco, 2013; Anton-Erxleben et al., 2007; Anton-Erxleben et al., 2009; Fischer & Whitney, 2009; Kay et al., 2015; Vo et al., 2017; Womelsdorf et al., 2006) and several authors have proposed computational mechanisms to explain how shifts could account for enhanced behavior. For example, receptive field shifts could increase the information capacity of a population of neurons by reducing spatial uncertainty about position (Kay et al., 2015) or enhancing discriminability (Vo et al., 2017). Although these computational models accurately described the visual system they failed to compare different possible model architectures (Gardner & Merriam, 2021), in particular architectures where gain and shift are linked. By using a model designed to mimic the visual system and built from independent data, we believe our analysis is more likely to reveal the true relationship between behavioral performance and receptive field properties. For this task, behavior appears to be solely affected by spatial signal gain.

Our analysis is limited by the accuracy of the neural network observer model. There are several reasons to suggest that the model captures properties of both object recognition and the primate visual system that are relevant to testing mechanisms of attention. We chose to analyze a CNN whose architecture was designed to reflect the primate visual system and that was evaluated by comparing the similarity of CNN unit activity against measurements of single unit activity in the primate visual cortex (Schrimpf et al., 2018). After training, the image features that the CNN units become selective for align closely with those that activate single units in visual cortex

bioRxiv preprint doi: https://doi.org/10.1101/2022.03.04.483026; this version posted March 7, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

(Carter et al., 2019; Yamins et al., 2014). In addition, the designers of the architecture we used (CORnet, Kubilius et al. (2018) optimized for "core object recognition", detecting a dominant object during a viewing duration of natural fixation (100-200 ms) in the central visual field (10 deg). We re-used core object recognition in our human object detection task and projected our test images in a 10 degree square aperture to obtain similar perceptual characteristics. In the analysis of our task we showed that distributed performance was similar for humans and the CNN at around 65 ms, confirming that the intended design of CORnet generalized to the new dataset and task that we used.

One of the main differences between convolutional neural networks and the primate visual cortex comes from the different ways in which units and neurons project their activation between layers. In our analysis, this was most apparent in the scale of the gain necessary to produce human-matched behavioral enhancements during spatial attention. In the Gaussian gain CNN the model passes the gain from layer to layer in a linear manner and we found that a gain of about 4$\times$ was needed to reproduce human behavioral enhancement. For comparison, neural recordings in primates have measured an attentional effect on the order of a 20-40% gain (1.2-1.4$\times$) (Luck et al., 1997; Treue & Trujillo, 1999). When we added a local selection mechanism implemented by divisive gain control to mimic the way that visual cortex selects and amplifies the strongest signals (Kaiser et al., 2016; Pestilli et al., 2011) we found that the CNN observer model required a much smaller gain of 1.25$\times$, which when amplified across four layers could produce the necessary combination of gain and selection required to reproduce the behavioral effect. Such non-linear selection mechanisms are thought to be a critical component of attention (Pestilli et al., 2011) and could be achieved by normalization of signals in visual cortex (Carandini & Heeger, 2012). This analysis does not imply that gain must be applied early in the visual hierarchy. Although physiological measurements have found evidence for this (Luck et al., 1997; McAdams & Maunsell, 1999; Motter, 1993), it is equally possible that the gain is applied at a late stage close to decision making and signal gains early in visual cortex are a result of backward projections to these areas (Buffalo et al., 2010).

Our finding that changing the tuning of units has little effect on behavior implies that the preferred features of units are a poor proxy for determining their influence on decision making (Lindsay & Miller, 2018). In theory, the visual system has the capacity to identify neurons that respond strongly to particular visual features and boost these to improve behavioral performance. In practice, Lindsay and Miller (2018) have shown that this is computationally inferior to magnifying unit activity according to the gradient on the output. In addition, feature sensitivity appears to be weakly correlated with the gradient of a unit on output behavior, with the implication that this is also true in the visual system. This leads to what could be a paradox: if you want to attend to complex features, how do you design a system to target the particular units with strong gradients on downstream decision-making areas? Our results, as well as the finding that attention can sometimes reduce perceptual performance (Yeshurun & Carrasco,

bioRxiv preprint doi: https://doi.org/10.1101/2022.03.04.483026; this version posted March 7, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

1998), suggest that the attention system doesn't solve this problem directly but instead deploys non-specific gain. Instead of over-optimizing the attentional system for every possible visual target, attention seems to operate over a more limited set of features. Visual search tasks suggest that this set of features is limited to a few basic elements, possibly color, orientation, motion, size, and position (Wolfe & Horowitz, 2004). Because these basic feature detectors are combined by the visual system to create complex neurons it is in some sense unsurprising that when we measure high-level features, even those as simple as the receptive fields in late layers, these appear to be the targets of attentional effects. In fact, these apparent effects of attention may simply be epiphenomena as we have demonstrated.

Our finding conceptually shifts experimental and theoretical work on neural mechanisms of attention. The vast literature on receptive field size and location shifts due to attention provide only a suggestion that these effects could underlie behavioral enhancement, but do not test that hypothesis explicitly. By using computational modeling we tested this directly and find strong support for gain as the fundamental mechanism driving behavioral enhancement with attention.

Methods

Human observers

Seven observers were subjects for the experiments (1 female, 6 male, mean age 22 y, range 19-24). All observers except one (who was an author) were naïve to the intent of the experiments. No observers were excluded during the initial training sessions (see eye-tracking below). Observers completed 1600 trials in two 60 minute sessions. Observers wore lenses to correct vision to normal if needed. Procedures were approved in advance by the Stanford Institutional Review Board on human participants research and all observers gave prior written informed consent before participating.

Hardware setup for human observers

Visual stimuli were generated using MATLAB (The Mathworks, Inc.) and MGL (Gardner et al., 2018). Stimuli were displayed at 60 cm viewing distance on a 22.5 inch VIEWPixx LCD display (resolution of 1900x1200, refresh-rate of 120 Hz) and responses collected via keyboard. Experiments were performed in a darkened room where extraneous sources of light were minimized.

bioRxiv preprint doi: https://doi.org/10.1101/2022.03.04.483026; this version posted March 7, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Eye-tracking was performed using an infrared video-based eye-tracker at 500 Hz (Eyelink 1000; SR Research). Calibration was performed at the start of each session to get a validation accuracy of less than 1 degree average offset from expected, using a thirteen-point calibration procedure. During training, trials were initiated by fixating the central cross for 0.5 s and canceled on-line when an observer's eye position moved more than 1.5 degree away from the center of the fixation cross for more than 0.3 s. Observers were excluded prior to data collection if we were unable to calibrate the eye tracker to an error of less than 1 degree of visual angle or if their canceled trial rate did not drop to near zero, all of the observers passed these criteria. During data collection the online cancellation was disabled and trials were excluded if observers made a saccade outside of fixation ($> 1.5\mathrm{deg}$) during the stimulus period.

Experimental Design

We compared the ability of humans and neural networks to detect objects in a grid of four images covering 10 degrees of visual angle (224 px). Given a grid of images, the observers were asked to identify whether or not a particular target category was present. On half of the trials we gave observers prior information telling them which of the four grid locations could contain the object (100% valid cue). This focal condition was compared with a distributed condition, in which no information was provided about which grid location could contain the target object. For humans, the prior in the focal condition was a spatial cue, a visual pointer to one corner of the grid. For the neural network model, the prior for the focal condition was implemented by a mechanistic change in the model architecture, which differed according to the model of attention being tested.

Stimuli

The stimuli presented to both humans and the neural network observer model were composed of four base images arranged in a grid (henceforth a "composite grid"). Each base image contained an exemplar of one of 21 ImageNet (Deng et al., 2009) categories. Composite grids always contained images from four different categories. The base images were cropped to be square, and resized to 122×122 pixels, making each composite grid 224×224 pixels. We pulled 929 images from each of 21 ImageNet categories: analog clock (renamed to "clock"), artichoke, bakery (renamed to "baked goods"), banana, bathtub, bonsai tree (renamed to "tree"), cabbage butterfly, coffee, computer, Ferris wheel, football helmet, garden spider (renamed to "spider"), greenhouse, home theater, long-horned beetle (renamed to "beetle"), mortar, padlock, paintbrush, seashore, stone wall, and toaster. These base images were usually representative of their category. However, many included other distracting elements (people,

bioRxiv preprint doi: https://doi.org/10.1101/2022.03.04.483026; this version posted March 7, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

text, strong reflections, etc). Two authors (KF and DB) selected 100 base images for each category absent of distracting elements (low-distraction base images) to be used for the human task. From these low-distraction base images we set aside 5 to use as exemplars when introducing the category to human participants.

To create the human stimulus set we generated composite grids for each of the 20 target categories. Each category required 80 composite grids: 40 including target objects and 40 without. We therefore needed 40 base images from the target category and 280 ($3 \times 40 + 4 \times 40$) base images from the non-target categories. We sampled all images from the low-distraction base images. Targets were placed 10 times in each of the four corners.

The neural network observer model was trained and tested on an expanded stimulus set. We set aside 50 base images for each category to train the linear classifiers (see Linear Classifiers, below). The approach was otherwise identical to that described above, but 829 composite grids were created with a target and 829 without. Because CNN models are translation invariant we formed all target composites with the target base image in the NW corner, to simplify analysis.

Human task

Human observers performed blocks of trials in which they had to report the presence or absence of a specified category in composite grids. At the start of each block we showed the human observers the words "Search for:" followed by the name of the current target category (Fig. 1a, Category). They were then shown five held-out (i.e. not shown in the task) exemplar base images to gain familiarity with the target category (Fig. 1a, Examples) and advanced through these with a self-paced button click. This was followed by individual trials of the task. At all times a fixation cross (0.5 deg diameter, white) was visible at the center of the screen in front of a black circle (1 deg diameter). This fixation region obscured the center of the composite grid, but made maintaining fixation easier for observers. At the start of each trial the pixels of the current composite grid were scrambled to create a luminance-matched visual mask. This was displayed until an observer maintained fixation for 0.3 s (Fig. 1a, "Fixation"). Once fixation was acquired a cue was shown for 0.75 s, informing the observer about whether the trial was focal (in which case the possible target location was indicated) or distributed (four possible target locations indicated). The focal cue was a 0.25 deg length white line pointing toward the cued corner of the grid. The distributed cue was four 0.25 deg length white lines pointing toward all four corners of the grid. Distributed and focal cues were presented in pseudo-randomized order throughout each block. The cue was followed by a 0.75 s inter-stimulus interval (Fig. 1a, Delay) before the composite grid (10 $\times$ 10 deg) was shown for either 1 (0.008 s), 2 (0.017), 4 (0.033), 8 (0.067), 16 (0.133), or 32 (0.267) video frames (Fig. 1a, Stimulus). The mask

bioRxiv preprint doi: https://doi.org/10.1101/2022.03.04.483026; this version posted March 7, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

then replaced the stimulus and observers were given 2 s to make a response (Fig. 1a, Response), pressing the "1" key for target present or the "2" key for absent. Feedback was given by changing the fixation cross color to green for correct and red for incorrect until the 2 s period elapsed. A 0.25 s inter-trial interval separated trials.

Observers completed one training block (the "tree" category) as practice before data collection began. They then completed each category block (40 focal trials with 20 target present and 20 target absent, and 40 distributed trials with 20 target present and 20 target absent) before moving on to the next category. Block order was pseudo-randomized for each observer. Each block took about five minutes to complete and a break was provided between blocks, as needed. In total the experiment took about two hours, split into two one hour sessions on different days.

Neural network observer model

We modeled the ventral visual pathway using CORnet-Z, a convolutional neural network (CNN) proposed by Kubilius et al. (2018). The model consists of four convolutional layers producing feature maps of decreasing spatial resolution (Table 1). The model which we used was trained on ImageNet, details can be found in Kubilius et al. (2018). At the last convolutional layer we took the average over the spatial dimensions of each feature map to create the neural network's representation (512-dimensional vector) of the input image.

	Layer Type	Kernel Size	Output Shape	FWHM (px, deg)
Input			$224 \times 224 \times 3$	
V1 Block	conv, stride=2	7×7	112 × 112 × 64	11 (0.5)
	max pool	2×2	56 × 56 × 64	
	ReLU		56 × 56 × 64	
V2 Block	conv	3×3	56 × 56 × 128	26.8 (1.21)
	max pool	2×2	28 × 28 × 128	
	ReLU		28 × 28 × 128	
V4 Block	conv	3×3	28 × 28 × 256	55.6 (2.52)
	max pool	2×2	14 × 14 × 256	
	ReLU		14 × 14 × 256	
IT Block	conv	3×3	14 × 14 × 512	111.4 (5.06)
	max pool	2×2	7 × 7 × 512	
	ReLU		7 × 7 × 512	
Encodings	avg. pool		1 × 1 × 512	

Table 1: CORnet-Z structure. Average receptive field (RF) full-width at half-maximum (FWHM) is measured using ellipses fit to the backpropagated gradients of units in a convolutional layer with respect to the input image pixels. 22.4 pixels corresponds to one degree of visual angle (Kubilius et al., 2018).

bioRxiv preprint doi: https://doi.org/10.1101/2022.03.04.483026; this version posted March 7, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Linear classifiers

To allow the neural network observer model to perform an object detection task we trained a set of linear classifiers on the model output to predict the presence or absence of each of the twenty target categories. Each of these fully-connected layers received as input the (512-dimensional) feature output from the CNN and projected these to a scalar output. Weights were fit using logistic regression, using *scikit-learn* and the *LIBLINEAR* package (Pedregosa et al., 2011). We trained the classifiers on a held out set of base images not used to generate the task grids, using 50 images with the target present and 50 images with the target absent.

To test model performance in the detection task the observer model was presented with each of the composite grids in the full image set. We report the model's area under the curve (AUC) as a measure of performance.

Spatial attention: Gaussian gain model

To introduce Gaussian gain as a mechanism for spatial attention we multiplied the pixel intensity of the input image at row r and column c by the magnitude of a 2-dimensional Gaussian, using the following equation:

$$g_{r_0,c_0,\sigma,\beta}(r,c) = (\beta - 1)\exp\left(-\frac{(r-r_0)^2 + (c-c_0)^2}{2\sigma^2}\right) + 1 \tag{2}$$

Where r_0 and c_0 set the row and column location for the center of the gain field and β controls the strength, i.e. the multiplicative factor at the peak of the Gaussian. The Gaussian was centered in the cued quadrant and σ was set to 56 pixels (approx 2.5 degrees). We explored four values of β: 1.1, 2, 4, and 11.

Quantifying the effects of gain on receptive fields and activations

To reduce computational requirements we randomly sampled 300 units per layer (1,200 total units) for receptive field analysis, with higher density near the attended locus.

To determine the location and size of the receptive field of each CNN unit we computed the derivative of their activation with respect to the pixels in the input image. This derivative was taken across a batch of 40 task images evenly distributed across categories. The magnitude of derivatives with respect to the red green and blue channels were summed to create a sensitivity map. Receptive field location and size were estimated by fitting a 2D Gaussian distribution to the sensitivity map. The Gaussian fit was performed by treating the sensitivity map

bioRxiv preprint doi: https://doi.org/10.1101/2022.03.04.483026; this version posted March 7, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

as an unnormalized probability distribution and choosing the Gaussian with the same the mean and covariance matrix as that distribution. Receptive field location was measured as the mean of the Gaussian fit. We report the full-width at half-maximum for the receptive field size.

To measure the effect of gain on the activation of CNN units we computed the effective gain and feature correlation across the sampled units. We defined effective gain as the ratio between the standard deviation of a unit's activity after applying an attention mechanism compared to before. We also measured the feature correlation at each spatial location by taking the Pearson correlation of the sampled unit activation with and without an attentional mechanism applied. We computed the effective gain and correlation measures across all features and all stimuli.

Nonlinear normalization

In order to test the ability of "winner-take-all" normalization to amplify small gains, we isolated the first layer of the CNN, and applied nonlinear normalization with exponent ξ. More precisely, if the output feature map of the first layer had size M rows by N columns by C channels and activations a_{ijc}, we calculated the normalized outputs

$$b_{ijc} = \frac{\sum_{k,l,d=1}^{M,N,C} |a_{kld}|}{\sum_{k,l,d=1}^{M,N,C} |a_{kld}|^{\xi}} a_{ijc}^{\xi}.$$

To measure the resulting amplified gain we applied a small Gaussian gain between $1\times$ and $1.25\times$ to the input image in the same manner as in the full Gaussian gain model. We then measured the ratio of average effective gain for units contained entirely within the gain field against the average effective gain of units entirely outside the attention gain field.

Spatial attention: Shift-only model

In the Gaussian gain model we applied the gain at layer 1 and observed changes in the model's detection performance at the output layers. We took a parallel approach here to design a model that could mimic the receptive field shifts at layer 4 (induced by gain at layer 1) while producing no systematic effect on response gain. To cause the layer 4 units to observe different parts of the input image we shifted the connections between pixels in the input image and first layer. We preserved all other connections, so layer 4 units of the neural network continued to receive information from the same layer 1 units.

To obtain the size of connection shifts we created a "shift map" in input image space by measuring the distance and direction that layer 4 units moved when the Gaussian gain was applied. To make this measurement, we took

bioRxiv preprint doi: https://doi.org/10.1101/2022.03.04.483026; this version posted March 7, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

each input image pixel location (r, c) and calculated the average receptive field shift of the 20 sampled layer 4 units with the closest receptive field centers without attention. Because we used a sampling procedure and not the full set of layer 4 units we weighted the sampled units by their Euclidean distance from the target pixel. To reduce noise in the shift map we applied a Gaussian blur with $\sigma = 8$ pixels. Using the shift map, we then re-assigned the connections from the input image to the layer 1 units so that these would reproduce the shifts observed in layer 4. The simplest way to to implement this involved swapping the activation of each layer 1 unit with the activation of the unit at its shifted location. For example, if unit $(75, 75)$ was shifted by $(-10, -10)$ we assigned it the activation of the unit at $(65, 65)$. To deal with decimal shifts we performed linear interpolation using neighboring units.

Spatial attention: Sensitivity shift by local gain

In the sensitivity shift model we aimed to mimic the spatial tuning changes induced by the Gaussian gain at a particular layer but without changing the effective gain of units. To do to this, we first computed the true gain propagated to the target layer L by scaling the Gaussian gain map to the size of layer $L-1$'s feature map. With this change alone the weights of units closer to the locus of attention are scaled more than the weights farther from the locus, introducing differential gain. To avoid a change in the overall scale of units' weights, we re-scaled the kernel to match the L2-norm (sum-of-squares) of the original kernel weights.

To summarize, suppose that layer $L-1$'s feature map is t times the size of the input image so that a unit at row r and column c of the layer $L-1$ feature map has an effective effective gain of $g_{tr_0,tc_0,t\sigma,\beta}(tr, tc)$ under the Gaussian gain model. Then if $w \in \mathbb{R}^N$ is the original weight vector of a unit in the unraveled convolution at layer L whose input vector $a \in \mathbb{R}^N$ contains the activations of post-ReLU units of layer $L-1$, and if the row-column positions in the $L-1$ feature map of the unit described by a_i is (r_i, c_i), then the replacement weight vector in the sensitivity shift model is given by the vector $w' \in \mathbb{R}^N$, whose entries are:

$$w'_i = \left(\frac{\sum_{i=1}^{N} w_i^2}{\sum_{i=1}^{N} w_i^2 g_{tr_0,tc_0,t\sigma,\beta}(tr_i, tc_i)^2} \right)^{1/2} w_i,$$

Spatial attention: Gain-only model

We designed a model which could effect gain without receptive field shift by flattening the gain in the cued quadrant. Receptive field shifts occur because there is a differential gain across the receptive field of a unit. To

bioRxiv preprint doi: https://doi.org/10.1101/2022.03.04.483026; this version posted March 7, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

get rid of this, you can simply put a flat gain across the cued quadrant. This naive approach has the problem that units that overlap two quadrants will still shift and shrink according to the strength of the gain. To prevent these units from shifting in a manner correlated to the gain we separated the CNN feature maps into four parts corresponding to the four image quadrants, ran the model forward with zero padding around each quadrant, and then concatenated the results back together. This ensured that each unit experienced a flat gain across its inputs and that as gain increased units near the quadrant boundaries did not experience gain-dependent receptive field shift or shrinkage.

Necessary and sufficient test

To obtain a propagated gain map in the final layer output we applied the Gaussian gain to the start of the neural network observer model and measured the average effective gain of the 7×7 layer 4 output units across a representative sample of images. We call this the "propagated gain map", since it represents the effect of the input gain on the output layers. We tested necessity by dividing the network output by the map for a model with gain applied and we tested sufficiency by multiplying the outputs from a no-gain model.

Behavioral analysis

We analyzed the human behavioral data by binning trials according to their duration and computing sensitivity d' from the equation:

$$d' = Z(H) - Z(FA) \tag{3}$$

Where Z is the inverse of the cumulative normal distribution and H and FA are the hit and false alarm rate, respectively. We fit a logarithmic function to the d' data using the equation:

$$d'(t) = \alpha * \log(\kappa t + 1) \tag{4}$$

Where t is the stimulus duration and α and κ are parameters that control the shape of the logarithmic function.

To compare human and model performance we can also convert between d' and the area under the curve (AUC) by the equation:

bioRxiv preprint doi: https://doi.org/10.1101/2022.03.04.483026; this version posted March 7, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

$$d' = \sqrt{2}Z(AUC) \tag{5}$$

Confidence intervals

All error bars are calculated by bootstrapping the given statistic with $n = 1000$ and reported as the 95% confidence interval.

Data and code availability

The images and composite grids used in this study as well as the code necessary to replicate our analyses are available in the Open Science Framework with the identifier 10.17605/OSF.IO/AGHQK.

Acknowledgments

We acknowledge the generous support of Research to Prevent Blindness and Lions Clubs International Foundation, and the Hellman Fellows Fund to JLG as well as the UW Vision Training Grant (NEI T32EY07031) and Washington Research Foundation Postdoctoral Fellowship to DB. We thank Josh Wilson for help with data collection and Eline Kupers and Maggie Henderson for early discussions.

Contributions

All of the authors conceived of and designed research, interpreted results of experiments, edited and revised the manuscript, and approved the final version of the manuscript. KF performed neural network experiments, analyzed data, and prepared figures. DB performed human behavioral experiments, analyzed data, and drafted the manuscript.

References

Albrecht, D. G., & Hamilton, D. B. (1982). Striate cortex of monkey and cat: Contrast response function. *Journal of neurophysiology*, *48*(1), 217–237.

bioRxiv preprint doi: https://doi.org/10.1101/2022.03.04.483026; this version posted March 7, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Anton-Erxleben, K., & Carrasco, M. (2013). Attentional enhancement of spatial resolution: Linking behavioural and neurophysiological evidence. *Nature Reviews Neuroscience*, *14*(3), 188–200.

Anton-Erxleben, K., Henrich, C., & Treue, S. (2007). Attention changes perceived size of moving visual patterns. *Journal of Vision*, *7*(11), 5–5.

Anton-Erxleben, K., Stephan, V. M., & Treue, S. (2009). Attention reshapes center-surround receptive field structure in macaque cortical area mt. *Cerebral cortex*, *19*(10), 2466–2478.

Ben Hamed, S., Duhamel, J.-R., Bremmer, F., & Graf, W. (2002). Visual receptive field modulation in the lateral intraparietal area during attentive fixation and free gaze. *Cerebral cortex*, *12*(3), 234–245.

Birman, D., & Gardner, J. L. (2019). A flexible readout mechanism of human sensory representations. *Nature communications*, *10*(1), 1–13.

Buffalo, E. A., Fries, P., Landman, R., Liang, H., & Desimone, R. (2010). A backward progression of attentional effects in the ventral stream. *Proceedings of the National Academy of Sciences*, *107*(1), 361–365.

Cadena, S. A., Denfield, G. H., Walker, E. Y., Gatys, L. A., Tolias, A. S., Bethge, M., & Ecker, A. S. (2019). Deep convolutional models improve predictions of macaque v1 responses to natural images. *PLoS computational biology*, *15*(4), e1006897.

Carandini, M., & Heeger, D. J. (2012). Normalization as a canonical neural computation. *Nature Reviews Neuroscience*, *13*(1), 51–62.

Carrasco, M. (2011). Visual attention: The past 25 years. *Vision research*, *51*(13), 1484–1525.

Carter, S., Armstrong, Z., Schubert, L., Johnson, I., & Olah, C. (2019). Activation atlas [https://distill.pub/2019/activation-atlas]. *Distill*. https://doi.org/10.23915/distill.00015

Cichy, R. M., Khosla, A., Pantazis, D., Torralba, A., & Oliva, A. (2016). Comparison of deep neural networks to spatio-temporal cortical dynamics of human visual object recognition reveals hierarchical correspondence. *Scientific reports*, *6*(1), 1–13.

Colby, C. L., & Goldberg, M. E. (1999). Space and attention in parietal cortex. *Annual review of neuroscience*, *22*(1), 319–349.

Colby, C., Goldberg, M. Et al. (1992). The updating of the representation of visual space in parietal cortex by intended eye movements. *Science*, *255*(5040), 90–92.

Connor, C. E., Gallant, J. L., Preddie, D. C., & Van Essen, D. C. (1996). Responses in area v4 depend on the spatial relationship between stimulus and attention. *Journal of neurophysiology*, *75*(3), 1306–1308.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., & Fei-Fei, L. (2009). ImageNet: A Large-Scale Hierarchical Image Database, In *Cvpr09*.

bioRxiv preprint doi: https://doi.org/10.1101/2022.03.04.483026; this version posted March 7, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Eckstein, M. P., Thomas, J. P., Palmer, J., & Shimozaki, S. S. (2000). A signal detection model predicts the effects of set size on visual search accuracy for feature, conjunction, triple conjunction, and disjunction displays. *Perception & psychophysics*, *62*(3), 425–451.

Eickenberg, M., Gramfort, A., Varoquaux, G., & Thirion, B. (2017). Seeing it all: Convolutional network layers map the function of the human visual system. *NeuroImage*, *152*, 184–194.

Fischer, J., & Whitney, D. (2009). Attention narrows position tuning of population responses in v1. *Current biology*, *19*(16), 1356–1361.

Gardner, J. L., Anzai, A., Ohzawa, I., & Freeman, R. D. (1999). Linear and nonlinear contributions to orientation tuning of simple cells in the cat's striate cortex. *Visual neuroscience*, *16*(6), 1115–1121.

Gardner, J. L., & Merriam, E. P. (2021). Population models, not analyses, of human neuroscience measurements. *Annual Review of Vision Science*, *7*, 225–255.

Gardner, J. L., Merriam, E. P., Schluppeck, D., & Larsson, J. (2018). MGL: Visual psychophysics stimuli and experimental design package. *Zenodo*.

Güçlü, U., & van Gerven, M. A. (2015). Deep neural networks reveal a gradient in the complexity of neural representations across the ventral stream. *Journal of Neuroscience*, *35*(27), 10005–10014.

Hawkins, H. L., Hillyard, S. A., Luck, S. J., Mouloua, M., Downing, C. J., & Woodward, D. P. (1990). Visual attention modulates signal detectability. *Journal of Experimental Psychology: Human Perception and Performance*, *16*(4), 802.

Heeger, D. J. (1992). Half-squaring in responses of cat striate cells. *Visual neuroscience*, *9*(5), 427 443.

Kaiser, D., Oosterhof, N. N., & Peelen, M. V. (2016). The neural dynamics of attentional selection in natural scenes. *Journal of neuroscience*, *36*(41), 10522–10528.

Kay, K. N., Weiner, K. S., & Grill-Spector, K. (2015). Attention reduces spatial uncertainty in human ventral temporal cortex. *Current Biology*, *25*(5), 595–600.

Khaligh-Razavi, S.-M., & Kriegeskorte, N. (2014). Deep supervised, but not unsupervised, models may explain it cortical representation. *PLoS computational biology*, *10*(11), e1003915.

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

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, *25*, 1097–1105.

Kubilius, J., Schrimpf, M., Nayebi, A., Bear, D., Yamins, D. L., & DiCarlo, J. J. (2018). Cornet: Modeling the neural mechanisms of core object recognition. *BioRxiv*, 408385.

Kusunoki, M., & Goldberg, M. E. (2003). The time course of perisaccadic receptive field shifts in the lateral intraparietal area of the monkey. *Journal of neurophysiology*, *89*(3), 1519–1527.

bioRxiv preprint doi: https://doi.org/10.1101/2022.03.04.483026; this version posted March 7, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

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

Luck, S. J., Chelazzi, L., Hillyard, S. A., & Desimone, R. (1997). Neural mechanisms of spatial selective attention in areas v1, v2, and v4 of macaque visual cortex. *Journal of neurophysiology*, *77*(1), 24–42.

McAdams, C. J., & Maunsell, J. H. (1999). Effects of attention on orientation-tuning functions of single neurons in macaque cortical area v4. *Journal of Neuroscience*, *19*(1), 431–441.

McIntosh, L. T., Maheswaranathan, N., Nayebi, A., Ganguli, S., & Baccus, S. A. (2016). Deep learning models of the retinal response to natural scenes. *Advances in neural information processing systems*, *29*, 1369.

Merriam, E. P., Genovese, C. R., & Colby, C. L. (2007). Remapping in human visual cortex. *Journal of neurophysiology*, *97*(2), 1738–1755.

Moran, J., & Desimone, R. (1985). Selective attention gates visual processing in the extrastriate cortex. *Science*, *229*(4715), 782–784.

Motter, B. C. (1993). Focal attention produces spatially selective processing in visual cortical areas v1, v2, and v4 in the presence of competing stimuli. *Journal of neurophysiology*, *70*(3), 909–919.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Et al. (2011). Scikit-learn: Machine learning in python. *the Journal of machine Learning research*, *12*, 2825–2830.

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

Pestilli, F., Carrasco, M., Heeger, D. J., & Gardner, J. L. (2011). Attentional enhancement via selection and pooling of early sensory responses in human visual cortex. *Neuron*, *72*(5), 832–846.

Posner, M. I. (1980). Orienting of attention. *Quarterly journal of experimental psychology*, *32*(1), 3–25.

Ross, J., Morrone, M. C., & Burr, D. C. (1997). Compression of visual space before saccades. *Nature*, *386*(6625), 598–601.

Sagi, D., & Julesz, B. (1986). Enhanced detection in the aperture of focal attention during simple discrimination tasks. *Nature*, *321*(6071), 693–695.

Schrimpf, M., Kubilius, J., Hong, H., Majaj, N. J., Rajalingham, R., Issa, E. B., Kar, K., Bashivan, P., Prescott-Roy, J., Schmidt, K., Et al. (2018). Brain-score: Which artificial neural network for object recognition is most brain-like? *BioRxiv*, 407007.

Sclar, G., Maunsell, J. H., & Lennie, P. (1990). Coding of image contrast in central visual pathways of the macaque monkey. *Vision research*, *30*(1), 1–10.

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

bioRxiv preprint doi: https://doi.org/10.1101/2022.03.04.483026; this version posted March 7, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Storrs, K. R., Kietzmann, T. C., Walther, A., Mehrer, J., & Kriegeskorte, N. (2020). Diverse deep neural networks all predict human it well, after training and fitting. *bioRxiv*.

Tolias, A. S., Moore, T., Smirnakis, S. M., Tehovnik, E. J., Siapas, A. G., & Schiller, P. H. (2001). Eye movements modulate visual receptive fields of v4 neurons. *Neuron*, *29*(3), 757–767.

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

van Es, D. M., Theeuwes, J., & Knapen, T. (2018). Spatial sampling in human visual cortex is modulated by both spatial and feature-based attention. *Elife*, *7*, e36928.

Vo, V. A., Sprague, T. C., & Serences, J. T. (2017). Spatial tuning shifts increase the discriminability and fidelity of population codes in visual cortex. *Journal of Neuroscience*, *37*(12), 3386–3401.

Wagenmakers, E.-J., Van Der Maas, H. L., & Grasman, R. P. (2007). An ez-diffusion model for response time and accuracy. *Psychonomic bulletin & review*, *14*(1), 3–22.

Wolfe, J. M., & Horowitz, T. S. (2004). What attributes guide the deployment of visual attention and how do they do it? *Nature reviews neuroscience*, *5*(6), 495–501.

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

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

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

Zirnsak, M., Steinmetz, N. A., Noudoost, B., Xu, K. Z., & Moore, T. (2014). Visual space is compressed in prefrontal cortex before eye movements. *Nature*, *507*(7493), 504 507. https://doi.org/10.1038/nature13149