Spatial Attention Enhances Crowded Stimulus Encoding Across Modeled Receptive Fields by Increasing Redundancy of Feature Representations

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

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

1 Introduction

The cerebral cortex is composed of a hierarchy of processing areas, each containing overlapping neuronal receptive fields (RFs) that tile the visual field

Neural Computation **34, 190–218** (2022)
https://doi.org/10.1162/neco_a_01447

at different spatial scales. The visual systems of humans and other animals use spatial attention to dynamically reconfigure the size and density of RFs (Klein, Harvey, & Dumoulin, 2014; Womelsdorf, Anton-Erxleben, Pieper, & Treue, 2006) to enhance sampling of stimuli (Anton-Erxleben & Carrasco, 2013) and perception (Carrasco, 2011) at attended locations.

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

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

Reconfiguration of RFs by spatial attention is perhaps more relevant to stimulus encoding in the visual periphery, where RFs are larger and less densely arranged compared to foveal RFs (Gattass, Gross, & Sandell, 1981; Gattass, Sousa, & Gross, 1988). As such, limits on the size and density of RFs have been theorized to contribute to the perceptual phenomenon known as visual crowding (Levi, 2008; Whitney & Levi, 2011; Rosenholtz, 2016), in which target objects in the visual periphery that are easily identified in isolation are more difficult to identify when flanked by similar nearby objects. Interestingly, flanking stimuli that are presented more peripherally, relative to a target stimulus location, crowd more than those that are presented more foveally (Banks, Bachrach, & Larson, 1977; Petrov & Meleshkevich, 2011), which suggests that target and flanker features encoded in larger RFs may be spatially overintegrated. Indeed, visual crowding has been modeled as a pooling mechanism in which relative spatial information of features is discarded (Balas, Nakano, & Rosenholtz, 2009; Freeman & Simoncelli, 2011; Van den Berg, Roerdink, & Cornelissen, 2010; Keshvari & Rosenholtz, 2016). However, there are additional aspects of crowding that cannot be explained by a simple pooling model, such as substitution errors in which subjects report one of the flankers instead of the target (Hanus & Vul, 2013; Ester, Klee, & Awh, 2014; Coates, Bernard, & Chung, 2019), categorical target/flanker effects (Reuther & Chakravarthi, 2014), global/contextual effects (Manassi, Sayim, & Herzog, 2012; Herzog, Sayim, Chicherov, & Manassi, 2015), and holistic effects (Farzin, Rivera, & Whitney, 2009).

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

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

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

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

2 Materials and Methods

2.1 Model Description.

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

2.1.2 Convolutional Neural Network Model. We trained a three-layer fully convolutional feedforward neural network to classify grayscale handwritten digits (28 × 28 pixels; MNIST: LeCun, Bottou, Bengio, & Haffner, 1998). Each convolutional layer in the neural network takes an image (or stack of images) as input and decomposes it into a set of feature maps, with each pixel in the feature map indicating the relative presence or absence of that feature. These feature maps are then passed through a nonlinear activation function (rectified linear unit (ReLU) or softmax; see Table 1). Finally, a pooling operation is applied to reduce the image size of the feature maps. Typically, this involves taking the maximum value within subsets of pixels (e.g., 2 × 2 subsets of pixels to reduce image height and width by 50%).

Table 1: Model Architecture Used for Training.

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

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

We trained our model for 10 epochs (10 full passes through a training set of 60,000 images), with a mini-batch size of 10, using supervised learning for digit classification with backpropagation (stochastic gradient descent with a learning rate of 0.001 and momentum of 0.9). The trained model achieved a test set error rate of 0.96% (100 − classification accuracy) on a held-out test set of 10,000 images. Table 1 shows the number of channels, activation functions, and pooling operations for each layer used during training.

The trained model was then used to extract features to be studied in crowding experiments in which multiple digits are simultaneously presented. In order to simulate peripheral vision for these crowding experiments and therefore provide the model with macroscopic realism (Kording et al., 2020), we replaced the max-pooling function in the second layer with an RF pooling array (see Figure 1, described below). We chose the second layer for this because the weights in this layer are more likely to represent unique fragments of the target digit that are shared across different digits, compared to the first-layer weights (which convolve over an area much smaller than a digit) and the third-layer weights (which convolve over an entire digit). Therefore, the second-layer weights better reflect competition between features within RFs. By training the model on individual 28 × 28 pixel digits without the RF pooling array, we ensured that only the size and density of RFs would affect stimulus encoding during the crowding experiments.

2.1.3 Receptive Field Pooling. Unlike a typical max-pooling layer, RF pooling occurs within RFs of variable size. As shown in Figure 1 for an example

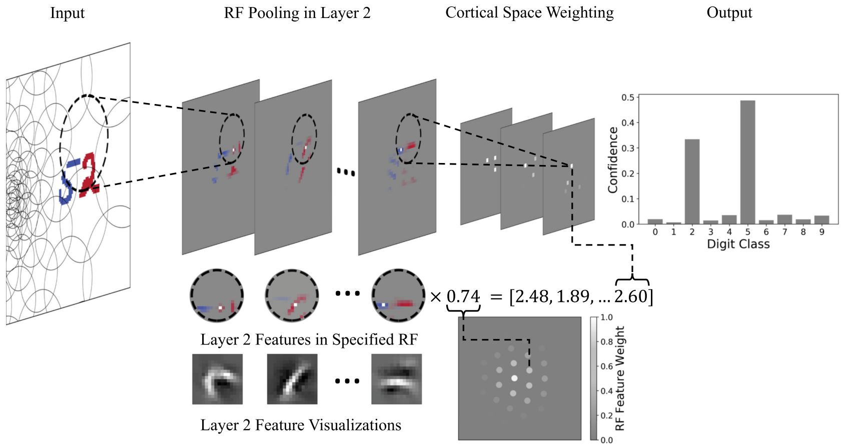

Figure 1: The three-layer fully convolutional neural network was trained to classify handwritten digits, with the softmaxed values in the output vector signifying the confidence of the classification for each digit. Flanking digits (red) were presented at various locations around the central target digit (blue). To model peripheral visual processing, we implemented a pooling operation on an RF array that simulates the eccentricity dependence of RFs in human visual cortex (here, eccentricity increases from left to right, with the fovea centered on the left edge of the input image). Feature maps in the second layer were spatially pooled within each RF separately (as shown for the highlighted example RF). Target (blue) and flanker (red) features compete within each RF, and the maximum value (shown here in white) in each masked feature map is retained while all other pixels were set to zero. In order to classify the target, the second-layer features in each RF were proportionately weighted based on the RF's cortical distance from the target (see equation 2.3), simulating 2D gaussian connections to a third-layer population of neurons that is centered on the target digit. Values within the brackets for the given RF indicate estimates of the relative presence or absence of the respective feature representations shown in the figure.

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

Using this approach, we maintain the spatial organization of the feature maps while pooling information with variable spatial resolution across the image. This allows us to separately examine the outputs across RFs (to assess redundancy of stimulus encoding) as well as the interactions within individual RFs (to assess fidelity of stimulus encoding). Finally, since each RF is defined by its μ and σ values, the RF array can be dynamically updated by allocation of attention to change the center positions and/or sizes of each RF using equation 2.4 (described in section 2.2).

2.1.4 Spatial Organization of the RF Pooling Array. The RF pooling array is organized into concentric rings that expand from a central point (fovea; left edge of Input image in Figure 1), with the circular RFs in each ring increasing exponentially in size as a function of eccentricity. Each RF center μ and size σ is determined by the following equations:

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

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

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

2.1.5 Weighting of RF Features for Digit Classification. In order to simulate peripheral vision while maintaining spatial relationships among features in image space, we weighted RF features based on their respective locations in "cortical space" (see Figure 1). This weighting procedure simulates a selection mechanism in which pooled features in one location are enhanced relative to pooled features from other locations in the image, allowing the model to selectively classify a target object among flanking objects. To convert from eccentricity e_{image} and polar angle θ_{image} values in image space to eccentricity e_{cortical} and polar angle θ_{cortical} values in cortical space, we used the following relationships, which are derived from the inversion of the exponential expansion in equation 2.1:

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

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

With this approach, we assume that the third convolutional layer represents a population of neurons centered on the target digit location and has 2D gaussian connections (in cortical space) to the second-layer RF outputs. Specifically, we computed digit classification by first passing a stimulus image through the first two convolutional layers of the model and the RF pooling array. We then weighted the outputs of each RF by a 2D gaussian in cortical space (see Figure 1):

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

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

2.1.6 Model Summary. In summary, a fully convolutional neural network (see Table 1) was first trained to classify 28 × 28 individual MNIST digits, and the learned weights were then fixed for all experiments. To simulate peripheral vision, we created an eccentricity-dependent RF pooling array (see Figure 1) with an eccentricity scaling factor of 0.2 and a horizontal or vertical offset. For all experiments, the RF pooling array replaced the max-pooling operation after the second convolutional layer (see Table 1). Although the exponential nature of the RF pooling array is important for accurately simulating peripheral vision, the specific values of the eccentricity scaling factor and the pooling operation are not important factors for studying the effects of crowding on task performance in our model.

In order to simulate a target-identification task in the periphery (e.g., Input in Figure 1), we used a 2D gaussian weighting function in cortical space as a selection mechanism to classify the target among flanking digits. The use of cortical weighting in our model is similar to asking a human participant to report the identity of the central digit as opposed to the flankers. Therefore, we weighted values pooled by RFs inversely proportional to the cortical distance from the target (i.e., RFs closer to the target had greater weights than RFs farther from the target). Importantly, we used the same weights for all experiments (i.e., we did not recalculate the weights following attentional modulation of RF properties) to ensure that any changes in the model's ability to classify a target digit were driven primarily by the structural properties of the RF pooling array. The weighted features were then passed through the final convolutional layer, and we computed target classification by selecting the feature class with the greatest value across the image space in the output layer.

2.2 Experimental Design and Statistical Analyses.

2.2.1 Visual Crowding Experiment. Inspired by stimuli used in perceptual experiments on visual crowding, we employed a classification task in which the target object is closely surrounded by flanking objects. We constructed crowded stimuli from a balanced test set of 10,000 MNIST digits that were not used during training. We randomly chose target digits and placed them at the center of the stimulus image, and we randomly chose flankers from nontarget classes. Target/flanker spacing was measured center-to-center. Figure 2 illustrates the four configurations we used in this study (outlined by colored boxes). In this example, the RF pooling array is offset horizontally. The inner (yellow), outer (blue), and radial (green) configurations have flankers at different eccentricities than the target, and the tangential (red) configuration has flankers at approximately the same eccentricity as the

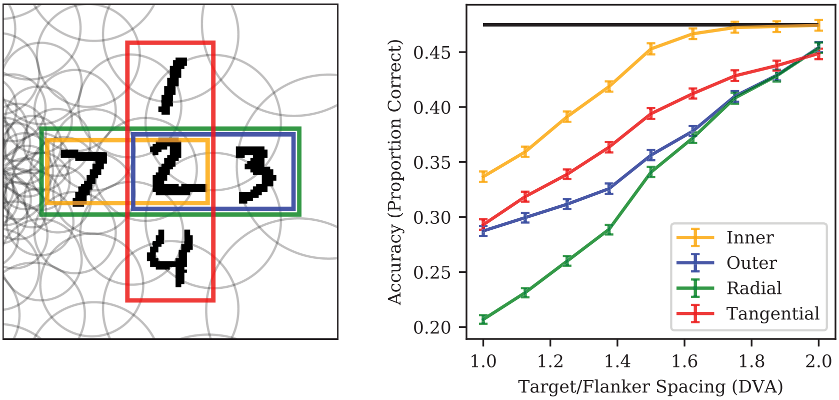

Figure 2: Left: Examples of crowded stimuli (target/flanker spacing = 1.5 DVA). The fixation point in this example is three DVA to the left of the central target digit (i.e., the left edge of the image). Gray circles show locations and sizes of individual RFs, and colored boxes outline the four unique configurations. Right: Target classification accuracy as a function of target/flanker spacing for each configuration. Line and symbol colors correspond to the box colors on the left. The black line indicates accuracy for targets presented without flankers. Chance performance is 0.1 (1 out of 10 possible digits). Error bars are bootstrapped 95% confidence intervals.

target. Note that the RFs that sample the peripheral flanking digit are larger and the RFs that sample the foveal flanking digit are smaller.

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

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

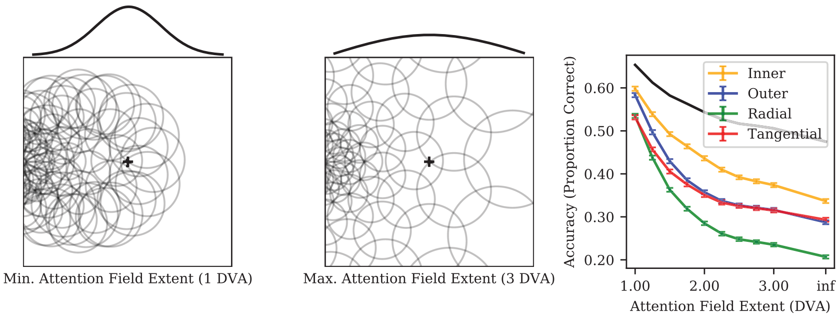

Figure 3: Left: RF pooling array at the minimum attention field extent. Middle: RF pooling array at the maximum attention field extent. Black cross indicates the attended target location. A 1D slice of the gaussian attention field is displayed above the RF array for both 1 and 3 DVA examples. Right: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

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

where *AF* and *RF* index the attention field and RF kernels, respectively. Decreases in the spatial extent of the attention field (i.e., smaller values of σ_{AF}) cause RFs to become smaller and more densely packed around the center of the attention field (see the left and center panels of Figure 3). To study the effects of this simulated attentional allocation, we empirically varied the size of σ_{AF} and computed new values of σ_{RF} and μ_{RF} for each RF in the pooling array via the gaussian multiplication described above. However, we do not assume the range or scale of σ_{AF} used in our experiments has a one-to-one relationship with the full range of attentional modulation in humans.

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

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

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

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

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

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

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

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

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

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

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

3 Results

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

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

We manipulated spacing for four unique target/flanker configurations (inner, outer, radial, and tangential; see Figure 2). In humans, crowding is influenced by target/flanker configuration: a single inner flanker presented foveally to the target causes less crowding than the same outer flanker presented peripherally to the target (Banks et al., 1977). Additionally, crowding is anisotropic: flankers presented on either side of the target along a radial axis emanating from the fovea cause more crowding than flankers presented along a tangential axis that is perpendicular to the radial axis (Toet & Levi, 1992; Chen et al., 2014). In the current experiment, we measured target classification accuracy in each of these target/flanker configurations to determine if our simple (relative to previous models, e.g., Nandy & Tjan, 2012; Chaney, Fischer, & Whitney, 2014) model of RF pooling could reproduce the effects observed in the literature that are described above.

Figure 2 shows that for all four configurations, target classification accuracy increased as a function of target/flanker spacing and that, at large target/flanker spacings, all four configurations approached accuracy levels observed in the target-alone condition (black line). Generally crowding is greatest for target/flanker spacings that are less than one-half of the target eccentricity (in our model, 3 DVA; Bouma, 1970). We also found that accuracy was lower for the radial configuration than for the tangential configuration for spacings at or below 1.5 DVA (green and red lines; bootstrapped p-value [1000 samples] = 0.), consistent with previously reported anisotropies of crowding in human subjects (Toet & Levi, 1992). Moreover, accuracy was lower for the outer configuration compared to the inner configuration for spacings at or below 1.5 DVA (blue and yellow lines; bootstrapped p-value [1000 samples] = 0.), again consistent with asymmetries that have been reported in human subjects (Banks et al., 1977).

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

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

As expected, more precise attention (smaller spatial extent) centered at the target location resulted in greater target classification accuracy for every target/flanker configuration (see Figure 3). Moreover, the effect of increasing attention field extent on target classification decayed exponentially. Furthermore, the relationships among the four configurations remained the same as observed in Figure 2, with outer/radial having lower target classification accuracy than inner/tangential configurations.

3.1.2 Substitution Errors Occur at Above-Chance Levels When Crowding Is Strong. Both increasing target/flanker spacing (see Figure 2) and decreasing the attention field extent (see Figure 3) had positive effects on our model's ability to correctly classify the target digit. These increases in performance are consistent with what has been shown in previous human studies. However, target classification is not the only metric that has been used to study visual crowding in human subjects. Substitution errors—the phenomenon of incorrectly reporting the flanker's identity instead of the target's at an above-chance rate—is an additional metric used to characterize target/flanker interactions in crowding (Ester et al., 2014; Hanus & Vul, 2013; Coates et al., 2019). In this experiment, we analyzed the results from the same target/flanker spacings and attention field extents as before. However, instead of reporting target classification accuracy, we present the number of flanker responses for each configuration as a proportion of incorrect trials (i.e., trials in which the target was not reported). Under strong crowding conditions, RFs that contain both target and flanker features will exhibit competition and therefore have feature interference. This interference should lead to the identities of the flankers being reported at above-chance levels on incorrect trials, compared to all other nontarget digits.

Figure 4 shows that under the strongest crowding conditions (left: 1 DVA spacing; right: 3 DVA extent), the proportion of trials in which the flankers were identified for each configuration was significantly above chance (black lines). Furthermore, the rate of incorrectly reporting the flanker decreased as the target/flanker spacing increased and the attention field extent decreased. Interestingly, the outer flanker was reported more often than the inner flanker across the majority of target/flanker spacings and attention field extents, and this asymmetry was observed both when the inner/outer flankers were presented as a single flanker with the target (solid yellow and blue lines, respectively) as well as when they were presented as pairs of flankers in the inner radial/outer radial conditions (dashed yellow and blue lines, respectively). These results suggest that when there is substantial crowding, representations of the identities of the specific flankers are stronger than those of the identities of all other nontarget classes. Furthermore, these findings indicate that in our model, crowding is due to competition between representations of target and flanker features.

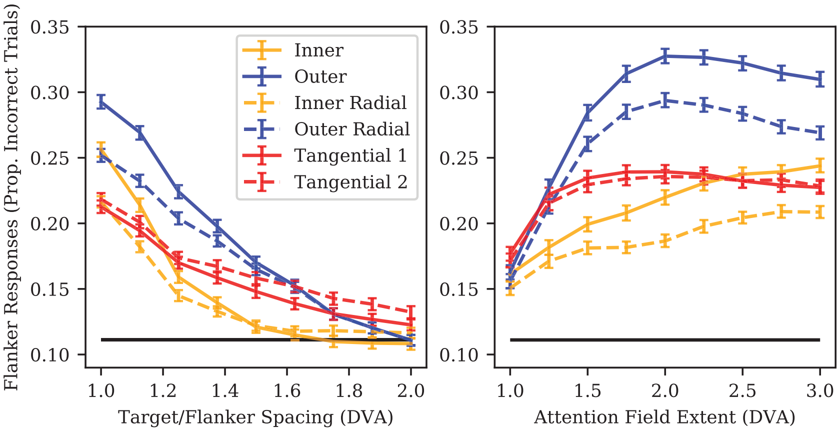

Figure 4: Proportion of incorrect trials for which the flanker digit was erroneously reported as a function of (left) target/flanker spacing and (right) attention field extent. The same four configurations were used as before. The radial/tangential configurations each had two possible flanker choices (Inner/Outer Radial and Tangential 1/2, respectively). Tangential 1/2 correspond to the perpendicular flankers placed below and above the radial axis in Figure 2, respectively. Black lines indicate chance probability for incorrectly reporting a nontarget digit (one out of nine possible digits). Error bars are bootstrapped 95% confidence intervals.

3.1.3 Increases in Target Classification Accuracy Depend Largely on RF Position Shifts. We have shown that reconfiguration of the RF pooling array by attention modifies both RF locations and sizes in our model (see Figure 3). In this experiment, we limited the effects of attention to changes in either the positions or the sizes of the RFs in our pooling array by separately applying updates to either μ or σ from equation 2.4, respectively. Previous fMRI research in humans indicates that shifts in RF position by attention are more important than changes in RF size for population-level encoding of fine spatial information (Vo et al., 2017). This suggests that shifting RFs in our pooling array toward the attended target location, without changing their size should increase target classification accuracy more than decreasing the sizes of RFs without changing their positions.

We employed the same target/flanker configurations and range of attention field extents as before (see Figure 3), but here we applied attention effects separately for RF position and size. Target/flanker spacing was fixed at 1 DVA for this experiment. Figure 5 shows an example RF pooling array for updated position (top left) and size (top right). As expected, shifts in fixed-size RFs toward the target location with attention increased target classification accuracy (see Figure 5; the black solid line indicates the

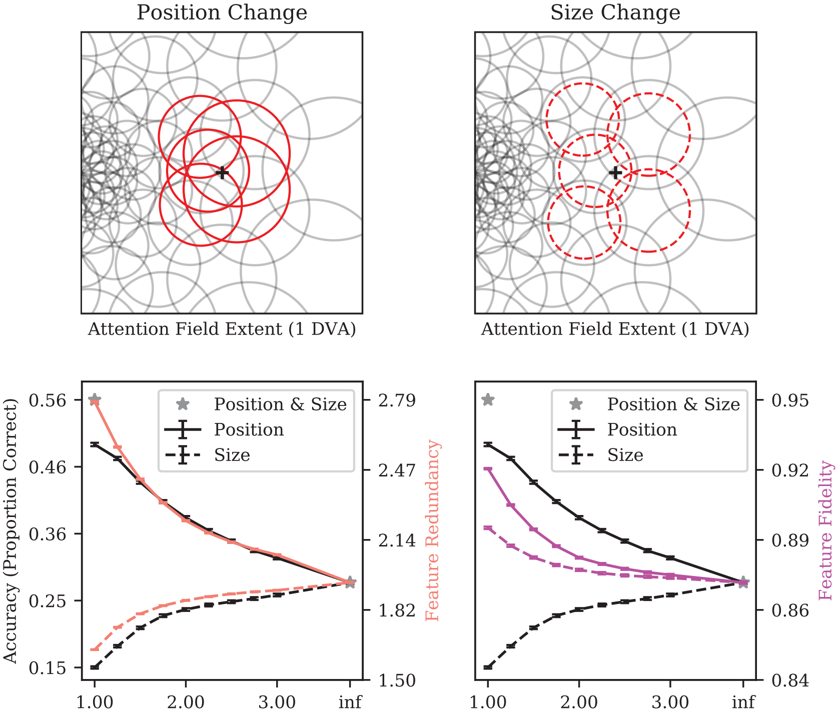

Figure 5: Changes in RF position with more focused spatial attention increase the density of RFs at the attended target location, indicated by the black cross (top left), whereas changes in RF size alone decrease RF density (top right). Target classification accuracy and feature redundancy (see equation 2.7) both increase with attention-related position changes but decrease with size changes (bottom left). In contrast, the fidelity of feature representations (as measured by cosine similarity; see equation 2.5) increases with more focused attention for both RF location and size changes (bottom right). Note that each of the y-axes has been scaled so that the corresponding metric is plotted relative to the value obtained for that metric following changes in both RF position and size with an attention field extent of 1 DVA (left gray star in each bottom panel) and infinity (i.e., "no attention"; right gray star in each bottom panel). All metrics depicted were averaged across the four target/flanker configurations (see Figure 2). Error bars are bootstrapped 95% confidence intervals.

average across configurations). Interestingly, decreasing the size of stationary RFs with attention decreased target classification accuracy (black dashed line). Note that allowing attention to affect both RF position and size together resulted in greater target classification accuracy (i.e., the value of 0.56 indicated by the gray stars at 1 DVA in the bottom panels of Figure 5) than either position or size changes alone.

In the second part of this experiment, we characterized the effects of shifts in RF position and size by attention on the redundancy and fidelity of feature representations. As described in section 2.2.3, we define redundancy as the average number of RFs that represent an activated target feature when corrupted by the flankers and fidelity as a measure (cosine similarity) of how corrupted the target features are by the flanker features. In order to visualize the relationships among these variables with each other and with target classification accuracy, we plotted each metric in Figure 5 relative to the same metric obtained for changes in both RF size and position for an attention field extent at infinity ("no attention") and at 1 DVA. Each metric is therefore relative to these matched points, which are shown as gray stars in the bottom panels of Figure 5.

Shifts in the positions of RFs toward the target location increased the density of target-containing RFs (and therefore the redundancy of feature representations; see Figure 5, top left panel), whereas reductions in the size of RFs decreased redundancy (see Figure 5, top right panel). We found that feature redundancy (salmon lines) was tightly coupled with target classification accuracy (black lines) for both RF position and size changes (see Figure 5, bottom left panel) across a range of attention field extents, suggesting that RF density at the target location (i.e., feature redundancy) is strongly related to downstream effects on target classification accuracy.

Fidelity of feature representations (magenta lines) increased both when the positions of fixed-size RFs were shifted toward the target location and when stationary RFs shrank with attention (see Figure 5, bottom right panel). Decreasing RF size results in less competition for processing between the target and flankers within a single RF, and this is reflected by increased feature fidelity values for smaller spatial extents of attention (magenta dashed line). However, attention field size has a very different relationship with feature fidelity than it has with target classification accuracy, which is worse for smaller attention field size (and therefore for smaller RFs; black dashed line). Together, these results suggest that target classification is more closely related to feature redundancy than it is to the fidelity of feature representations.

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

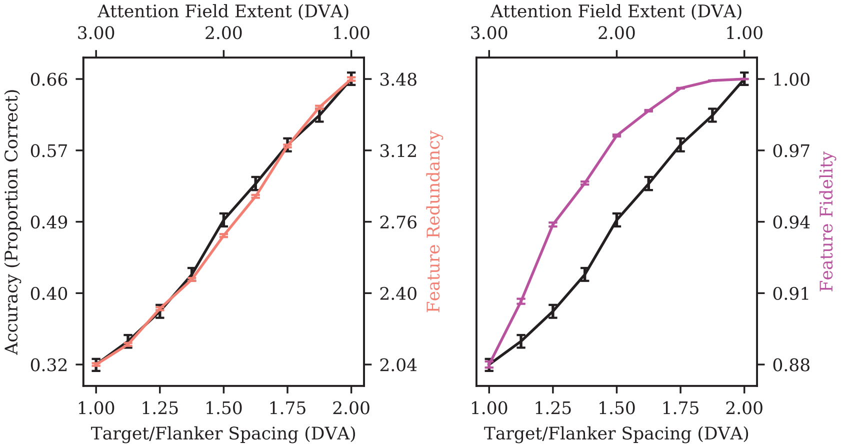

Figure 6: Target classification accuracy plotted with feature redundancy (left; equation 2.7) or feature fidelity (right; see equation 2.5) for a diagonal cross-section of the combined factors of attention field extent and target/flanker spacing. Target classification accuracy is much more closely related to feature redundancy than it is to feature fidelity. All metrics plotted here were averaged across the four target/flanker configurations. Error bars are bootstrapped 95% confidence intervals.

Although these results indicate a stronger relationship between target classification accuracy and feature redundancy compared to the relationship with feature fidelity (see Figure 5, bottom panels), interactions between features are dependent on both RF sampling and the relative distance between the target and flanker. In the previous experiment, all results were obtained with a target/flanker spacing of 1 DVA. We therefore conducted an additional experiment to more fully characterize the effects of feature redundancy and fidelity on target classification accuracy.

We selected a subsample of 1000 test images (from the original 10,000) for each combination of attention field extent and target/flanker spacing values used in the previous experiments. This enabled us to characterize the effects of both of these factors on the full range of observed variation in target classification accuracy that we studied. As shown in Figure 6, both redundancy and fidelity were highly correlated with target classification accuracy ($R^2 = 0.96$, $R^2 = 0.69$, respectively). We then computed the unique variance explained by each factor with multiple linear regression and found that the unique variance explained by redundancy was substantially greater than the unique variance explained by fidelity ($R^2 = 0.28$ versus $R^2 = 4.96 \times 10^{-3}$, bootstrapped p-value of the difference in explained variance [1000 samples] = 0). These results indicate that redundancy of target feature representations is likely to be more important than fidelity for target classification in visual crowding.

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

4 Discussion

Inspired by the normalization model of attention (Reynolds & Heeger, 2009), we constructed a model with a dynamic pooling array of RFs that were modulated by spatial attention in order to characterize how feature redundancy and fidelity relate to downstream target classification during a visual crowding task. Our model reproduced patterns of target classification for different target/flanker spacings and configurations that have been reported in psychophysical visual crowding experiments (Whitney & Levi, 2011). Next, by separately manipulating the effects of spatial attention on RF size and location, we demonstrated a plausible mechanism by which visual crowding is relieved by position shifts in RFs that increase their density at the attended target location. Finally, by varying target/flanker spacing and the spatial extent of attention, we revealed that feature redundancy explained far more unique variance in target classification accuracy than was explained by feature fidelity (see Figure 6).

4.1 A Model of Spatial Attention Effects on Downstream Processing and Perception. In our model, spatial attention increases RF density at the attended target location, resulting in an increase in feature redundancy across populations of RFs that improves target classification in crowded stimuli (see Figure 5, bottom left panel). Our model does not explicitly contain a metric of response amplitude per se but instead quantifies feature representations in individual RFs. Therefore, we did not explore the effects of attention on response gain in our study. However, our model is conceptually compatible with literature demonstrating gain modulation by spatial attention (Moran & Desimone, 1985). The RF pooling operation in our model encodes information in a lossy manner relative to the total information available in the second-layer feature maps. However, more information is preserved with the smaller and more densely organized RFs that are produced by attention, demonstrating an increase in information gain with more precise attention. This is similar to the effect of attention on feature fidelity observed in the bottom right panel of Figure 5, in which spatial attention directed toward the target digit increased the fidelity of the encoded target signal.

Similar to Vo, Sprague, and Serences (2017) and Baruch and Yeshurun (2014), we found that shifts in RF position with attention are more important than changes in RF size for improving feature representations. Interestingly, we found that at the minimum attention field extent, target classification accuracy based only on changes in RF position was considerably lower than what would be predicted by its relationship with feature redundancy (see Figure 5, bottom left panel). This discrepancy may be explained by differences in the effects of feature redundancy measured across partially versus completely overlapping RFs. In our model, as RFs approach complete spatial overlap, they are more likely to represent the same pixel locations for a given feature, which does not provide any benefits for target classification. Indeed, Nigam, Pojoga, and Dragoi (2019) demonstrated that synergistic connections within a cortical column in V1 (i.e., connections between nearby neurons sharing very similar RFs) allow for greater decoding of stimulus information than do redundant connections. This physiological result is consistent with our interpretation of our modeling results that feature redundancy across partially overlapping RFs is more beneficial for perception than redundancy within highly overlapping RFs.

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

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

More recently, Lonnqvist, Clarke, and Chakravarthi (2020) reported a study of visual crowding in deep neural networks. Although the authors observed striking differences between the pattern of visual crowding observed in CNNs and what has typically been observed in human studies, there are important differences between their study and ours. Lonnqvist et al. (2020) logarithmically downsampled images in order to simulate peripheral vision, whereas our model used eccentricity-dependent RF pooling of feature maps. However, downsampling the image simulates peripheral visual input rather than peripheral visual processing, and it is inconsistent with the interpretation of visual crowding as a high-contrast mixing of stimulus features. Additionally, Lonnqvist et al. (2020) did not incorporate a selection mechanism for classifying target objects separate from flankers but instead trained their model to classify a single object at a target location, followed by testing with both target and flanking objects. It is possible that their inability to observe increased performance as a function of target/flanker spacing (see our Figure 2, right panel) was due to overfitting during the target-alone training procedure in their model. These differences highlight the importance of eccentricity-dependent pooling and selection mechanisms for successfully modeling visual crowding.

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

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

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

There are also several models in which spatial attention has been implemented through enhanced responses (Olshausen, Anderson, & Van Essen, 1993; Mozer & Sitton, 1998; Hamker, 2004). For instance, Deco and Lee (2002) used a set of gaussian weights similar to our cortical weighting mechanism (see section 2.1.5) to enhance responses within an attended region. However, our model uses cortical weighting as a method for selecting target features for classification, not for gain modulation.

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

Acknowledgments

This work was supported by National Eye Institute R01 025278 to M.A.S. We thank Devavrat Vartak, Julie Self, Kelly Byrne, Elizabeth Lawler, and Arjun Mukerji for valuable feedback on this letter.

References

Albonico, A., Martelli, M., Bricolo, E., Frasson, E., & Daini, R. (2018). Focusing and orienting spatial attention differently modulate crowding in central and peripheral vision. *Journal of Vision*, *18*(3), 4, 1–17. 29677319

Anton-Erxleben, K., & Carrasco, M. (2013). Attentional enhancement of spatial resolution: Linking behavioural and neurophysiological evidence. *Nature Reviews Neuroscience*, *14*(3), 188–200. 10.1038/nrn3443, PubMed: 23422910

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. 10.1093/cercor/bhp002, PubMed: 19211660

Balas, B., Nakano, L., & Rosenholtz, R. (2009). A summary-statistic representation in peripheral vision explains visual crowding. *Journal of Vision*, *9*(12), 13.1–13.18.

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

Barbot, A., & Carrasco, M. (2017). Attention modifies spatial resolution according to task demands. *Psychological Science*, *28*(3), 285–296. 10.1177/0956797616679634, PubMed: 28118103

Baruch, O., & Yeshurun, Y. (2014). Attentional attraction of receptive fields can explain spatial and temporal effects of attention. *Visual Cognition*, *22*(5), 704–736. 10.1080/13506285.2014.911235

Bouma, H. (1970). Interaction effects in parafoveal letter recognition. *Nature*, *226*(5241), 177–178. 10.1038/226177a0, PubMed: 5437004

Carrasco, M. (2011). Visual attention: The past 25 years. *Vision Research*, *51*(13), 1484–1525. 10.1016/j.visres.2011.04.012, PubMed: 21549742

Chaney, W., Fischer, J., & Whitney, D. (2014). The hierarchical sparse selection model of visual crowding. *Frontiers in Integrative Neuroscience*, *8*, 73.1–73.11. 10.3389/fnint.2014.00073, PubMed: 25309360

Chen, F. X., Roig, G., Isik, L., Boix, X., & Poggio, T. (2017). Eccentricity dependent deep neural networks: Modeling invariance in human vision. In *AAAI Spring Symposium Series* (pp. 541–546).

Chen, J., He, Y., Zhu, Z., Zhou, T., Peng, Y., Zhang, X., & Fang, F. (2014). Attention-dependent early cortical suppression contributes to crowding. *Journal of Neuroscience*, *34*(32), 10465–10474. 10.1523/JNEUROSCI.1140-14.2014, PubMed: 25100582

Chen, Y., Geisler, W. S., & Seidemann, E. (2006). Optimal decoding of correlated neural population responses in the primate visual cortex. *Nature Neuroscience*, *9*(11), 1412–1420. 10.1038/nn1792, PubMed: 17057706

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

Coates, D. R., Bernard, J.-B., & Chung, S. T. (2019). Feature contingencies when reading letter strings. *Vision Research*, *156*, 84–95. 10.1016/j.visres.2019.01.005, PubMed: 30660632

Deco, G., & Lee, T. S. (2002). A unified model of spatial and object attention based on inter-cortical biased competition. *Neurocomputing*, *44*, 775–781. 10.1016/S0925-2312(02)00471-X

Desimone, R., & Duncan, J. (1995). Neural mechanisms of selective visual attention. *Annual Review of Neuroscience*, *18*(1), 193–222. 10.1146/annurev.ne.18.030195.001205, PubMed: 7605061

Deza, A., & Eckstein, M. (2016). Can peripheral representations improve clutter metrics on complex scenes? In D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, & R. Garnett (Eds.), *Advances in neural information processing systems*, *29*, 2847–2855. Red Hook, NY: Curran.

Deza, A., Jonnalagadda, A., & Eckstein, M. P. (2019). Towards metamerism via foveated style transfer. In *Proceedings of the International Conference on Learning Representations*.

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

Ester, E. F., Klee, D., & Awh, E. (2014). Visual crowding cannot be wholly explained by feature pooling. *Journal of Experimental Psychology: Human Perception and Performance*, *40*(3), 1022–1033. 10.1037/a0035377, PubMed: 24364703

Farzin, F., Rivera, S. M., & Whitney, D. (2009). Holistic crowding of Mooney faces. *Journal of Vision*, *9*(6), 18, 1–15. 10.1167/9.6.18, PubMed: 19761309

Freeman, J., & Simoncelli, E. P. (2011). Metamers of the ventral stream. *Nature neuroscience*, *14*(9), 1195–1201. 10.1038/nn.2889, PubMed: 21841776

French, R. M. (1999). Catastrophic forgetting in connectionist networks. *Trends in Cognitive Sciences*, *3*(4), 128–135. 10.1016/S1364-6613(99)01294-2, PubMed: 10322466

Gattass, R., Gross, C., & Sandell, J. (1981). Visual topography of V2 in the macaque. *Journal of Comparative Neurology*, *201*(4), 519–539. 10.1002/cne.902010405

Gattass, R., Sousa, A., & Gross, C. (1988). Visuotopic organization and extent of V3 and V4 of the macaque. *Journal of Neuroscience*, *8*(6), 1831–1845. 10.1523/JNEUROSCI.08-06-01831.1988, PubMed: 3385477

Hamker, F. H. (2004). A dynamic model of how feature cues guide spatial attention. *Vision Research*, *44*(5), 501–521. 10.1016/j.visres.2003.09.033, PubMed: 14680776

Hanus, D., & Vul, E. (2013). Quantifying error distributions in crowding. *Journal of Vision*, *13*(4), 17, 1–27. 10.1167/13.4.17, PubMed: 23525133

He, D., Wang, Y., & Fang, F. (2019). The critical role of V2 population receptive fields in visual orientation crowding. *Current Biology*, *29*(13), 2229–2236. 10.1016/j.cub.2019.05.068, PubMed: 31231052

Herzog, M. H., Sayim, B., Chicherov, V., & Manassi, M. (2015). Crowding, grouping, and object recognition: A matter of appearance. *Journal of Vision*, *15*(6), 5, 1–18.

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

Jia, Y., Huang, C., & Darrell, T. (2012). Beyond spatial pyramids: Receptive field learning for pooled image features. In *Proceedings of the 2012 IEEE Conference on Computer Vision and Pattern Recognition* (pp. 3370–3377). Piscataway, NJ: IEEE.

Keshvari, S., & Rosenholtz, R. (2016). Pooling of continuous features provides a unifying account of crowding. *Journal of Vision*, *16*(3), 39, 1–15. 26928055

Klein, B. P., Harvey, B. M., & Dumoulin, S. O. (2014). Attraction of position preference by spatial attention throughout human visual cortex. *Neuron*, *84*(1), 227–237. 10.1016/j.neuron.2014.08.047, PubMed: 25242220

Kording, K. P., Blohm, G., Schrater, P., & Kay, K. (2020). Appreciating the variety of goals in computational neuroscience. *Neurons, Behavior, Data Analysis, and Theory*, *3*(6), 1–12.

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

Levi, D. M. (2008). Crowding an essential bottleneck for object recognition: A mini-review. *Vision Research*, *48*(5), 635–654. 10.1016/j.visres.2007.12.009, PubMed: 18226828

Lonnqvist, B., Clarke, A. D., & Chakravarthi, R. (2020). Crowding in humans is unlike that in convolutional neural networks. *Neural Networks*, *126*, 262–274. 10.1016/j.neunet.2020.03.021, PubMed: 32272430

Manassi, M., Sayim, B., & Herzog, M. H. (2012). Grouping, pooling, and when bigger is better in visual crowding. *Journal of Vision*, *12*(10), 13, 1–14. 10.1167/12.10.13, PubMed: 23019118

Manassi, M., & Whitney, D. (2018). Multi-level crowding and the paradox of object recognition in clutter. *Current Biology*, *28*(3), R127–R133. 10.1016/j.cub.2017.12.051, PubMed: 29408262

McAdams, C. J., & Maunsell, J. H. (1999). Effects of attention on orientation-tuning functions of single neurons in macaque cortical area V4. *Journal of Neuroscience*, *19*(1), 431–441. 10.1523/JNEUROSCI.19-01-00431.1999, PubMed: 9870971

Moran, J., & Desimone, R. (1985). Selective attention gates visual processing in the extrastriate cortex. *Science*, *229*(4715), 782–784. 10.1126/science.4023713, PubMed: 4023713

Motter, B. C. (2009). Central V4 receptive fields are scaled by the V1 cortical magnification and correspond to a constant-sized sampling of the V1 surface. *Journal of Neuroscience*, *29*(18), 5749–5757. 10.1523/JNEUROSCI.4496-08.2009, PubMed: 19420243

Mozer, M. C., & Sitton, M. (1998). Computational modeling of spatial attention. *Attention*, *9*, 341–393.

Nandy, A. S., & Tjan, B. S. (2012). Saccade-confounded image statistics explain visual crowding. *Nature Neuroscience*, *15*(3), 463–469. 10.1038/nn.3021, PubMed: 22231425

Nigam, S., Pojoga, S., & Dragoi, V. (2019). Synergistic coding of visual information in columnar networks. *Neuron*, *104*(2), 402–411. 10.1016/j.neuron.2019.07.006, PubMed: 31399280

Olshausen, B. A., Anderson, C. H., & Van Essen, D. C. (1993). A neurobiological model of visual attention and invariant pattern recognition based on dynamic routing of information. *Journal of Neuroscience*, *13*(11), 4700–4719. 10.1523/JNEUROSCI.13-11-04700.1993, PubMed: 8229193

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

Petrov, Y., & Meleshkevich, O. (2011). Asymmetries and idiosyncratic hot spots in crowding. *Vision Research*, *51*(10), 1117–1123. 10.1016/j.visres.2011.03.001, PubMed: 21439309

Ren, S., He, K., Girshick, R., & Sun, J. (2015). Faster r-CNN: Towards real-time object detection with region proposal networks. In C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, & R. Garnett (Eds.), *Advances in neural information processing systems*, *28* (pp. 91–99). Red Hook, NY: Curran.

Reuther, J., & Chakravarthi, R. (2014). Categorical membership modulates crowding: Evidence from characters. *Journal of Vision*, *14*(6), 5, 1–13. 10.1167/14.6.5, PubMed: 25325783

Reynolds, J. H., Chelazzi, L., & Desimone, R. (1999). Competitive mechanisms subserve attention in macaque areas V2 and V4. *Journal of Neuroscience*, *19*(5), 1736–1753. 10.1523/JNEUROSCI.19-05-01736.1999, PubMed: 10024360

Reynolds, J. H., & Heeger, D. J. (2009). The normalization model of attention. *Neuron*, *61*(2), 168–185. 10.1016/j.neuron.2009.01.002, PubMed: 19186161

Rosenholtz, R. (2016). Capabilities and limitations of peripheral vision. *Annual Review of Vision Science*, *2*, 437–457. 10.1146/annurev-vision-082114-035733, PubMed: 28532349

Sabour, S., Frosst, N., & Hinton, G. E. (2017). Dynamic routing between capsules. In I. Guyon, Y. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, & R. Garnett (Eds.), *Advances in neural information processing systems*, *30* (pp. 3856–3866). Red Hook, NY: Curran.

Scolari, M., Kohnen, A., Barton, B., & Awh, E. (2007). Spatial attention, preview, and popout: Which factors influence critical spacing in crowded displays? *Journal of Vision*, *7*(2), 7.1–7.23. 10.1167/7.2.7

Simoncelli, E. P., & Olshausen, B. A. (2001). Natural image statistics and neural representation. *Annual Review of Neuroscience*, *24*(1), 1193–1216. 10.1146/annurev.neuro.24.1.1193, PubMed: 11520932

Somers, D. C., Dale, A. M., Seiffert, A. E., & Tootell, R. B. (1999). Functional MRI reveals spatially specific attentional modulation in human primary visual cortex. *Proceedings of the National Academy of Sciences*, *96*(4), 1663–1668. 10.1073/pnas.96.4.1663

Sun, G. J., Chung, S. T., & Tjan, B. S. (2010). Ideal observer analysis of crowding and the reduction of crowding through learning. *Journal of Vision*, *10*(5), 16.1–16.14. 10.1167/10.7.161

Toet, A., & Levi, D. M. (1992). The two-dimensional shape of spatial interaction zones in the parafovea. *Vision Research*, *32*(7), 1349–1357. 10.1016/0042-6989(92)90227-A, PubMed: 1455707

Van den Berg, R., Roerdink, J. B., & Cornelissen, F. W. (2010). A neurophysiologically plausible population code model for feature integration explains visual crowding. *PLOS Comput. Biol.*, *6*(1), e1000646. 10.1371/journal.pcbi.1000646, PubMed: 20098499

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. In I. Guyon, Y. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, & R. Garnett, (Eds.), *Advances in neural information processing systems*, *30* (pp. 5998–6008). Red Hook, NY: Curran.

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. 10.1523/JNEUROSCI.3484-16.2017, PubMed: 28242794

Volokitin, A., Roig, G., & Poggio, T. A. (2017). Do deep neural networks suffer from crowding? In *Advances in neural information processing systems*, *30* (pp. 5628–5638). Red Hook, NY: Curran.

Wandell, B. A., & Winawer, J. (2015). Computational neuroimaging and population receptive fields. *Trends in Cognitive Sciences*, *19*(6), 349–357. 10.1016/j.tics.2015.03.009, PubMed: 25850730

Whitney, D., & Levi, D. M. (2011). Visual crowding: A fundamental limit on conscious perception and object recognition. *Trends in Cognitive Sciences*, *15*(4), 160–168. 10.1016/j.tics.2011.02.005, PubMed: 21420894

Womelsdorf, T., Anton-Erxleben, K., Pieper, F., & Treue, S. (2006). Dynamic shifts of visual receptive fields in cortical area MT by spatial attention. *Nature neuroscience*, *9*(9), 1156–1160. 10.1038/nn1748, PubMed: 16906153

Yeshurun, Y., & Carrasco, M. (1998). Attention improves or impairs visual performance by enhancing spatial resolution. *Nature*, *396*(6706), 72–75. 10.1038/23936, PubMed: 9817201

Yeshurun, Y., & Carrasco, M. (2008). The effects of transient attention on spatial resolution and the size of the attentional cue. *Perception and Psychophysics*, *70*(1), 104–113. 10.3758/PP.70.1.104, PubMed: 18306965

Yeshurun, Y., Montagna, B., & Carrasco, M. (2008). On the flexibility of sustained attention and its effects on a texture segmentation task. *Vision Research*, *48*(1), 80–95. 10.1016/j.visres.2007.10.015, PubMed: 18076966

Yeshurun, Y., & Rashal, E. (2010). Precueing attention to the target location diminishes crowding and reduces the critical distance. *Journal of Vision*, *10*(10), 16.1–16.12. 10.1167/10.10.16

Received January 28, 2021; accepted July 1, 2021.