Beyond Bouma's window: How to explain global aspects of crowding?

Adrien Doerig [1*], Alban Bornet [1], Ruth Rosenholtz [2], Gregory Francis [3], Aaron M. Clarke [4], Michael H. Herzog[1]

[1] Laboratory of Psychophysics, Brain Mind Institute, École Polytechnique Fédérale de Lausanne (EPFL), Switzerland
[2] Department of Brain and Cognitive Sciences, Computer Science and Artificial Intelligence Laboratory, MIT, Cambridge, MA, USA
[3] Department of Psychological Sciences, Purdue University, West Lafayette, IN, USA
[4] Laboratory of Computational Vision, Psychology Department, Bilkent University, Ankara, Turkey
[*] Corresponding author
E-mail: adrien.doerig@gmail.com (AD)

Abstract

In crowding, perception of an object deteriorates in the presence of nearby elements. Although crowding is a ubiquitous phenomenon, since elements are rarely seen in isolation, to date there exists no consensus on how to model it. Previous experiments showed that the global configuration of the entire stimulus must be taken into account. These findings rule out simple pooling or substitution models and favor models sensitive to global spatial aspects. In order to investigate how to incorporate global aspects into models, we tested a large number of models with a database of forty stimuli tailored for the global aspects of crowding. Our results show that incorporating grouping like components strongly improves model performance.

Author Summary

Visual crowding highlights interactions between elements in the visual field. For example, an object is more difficult to recognize if it is presented in clutter. Crowding is one of the most fundamental aspects of vision, playing crucial roles in object recognition, reading and visual perception in general, and is therefore an essential tool to understand how the visual system encodes information based on its retinal input. Classic models of crowding have focused only on local interactions between neighboring visual elements. However, abundant experimental evidence argues against local processing, suggesting that the global configuration of visual elements strongly modulates crowding. Here, we tested all available models of crowding that are able to capture global processing across the entire visual field. We tested 12 models including the Texture Tiling Model, a Deep Convolutional Neural Network and the LAMINART neural network with large scale computer simulations. We found that models incorporating a grouping component are best suited to explain the data. Our results suggest that in order to understand vision in general, mid-level, contextual processing is inevitable.

Introduction

When an element is presented in the presence of nearby elements or clutter, it becomes harder to perceive, a well-known effect called crowding. One of the main characteristics of crowding is that the element itself is not invisible, contrary to contrast- and backward-masking; rather its features appear jumbled and distorted (figure 1). Crowding is a ubiquitous phenomenon because elements are rarely encountered in isolation in everyday situations (figure 1c). Thus, understanding crowding is crucial for understanding vision in general.

For about half a century, the consensus was that flankers interfere with a target element only when placed within a spatially restricted window around the target, the so-called Bouma law (figure 1b; [1–4]):

$$\text{Size of Bouma's window} \approx 0.5*\text{eccentricity}$$

Classic models of crowding proposed that early visual areas, such as V1, process the features of stimuli with high precision. Crowding occurs when neural signals are pooled along the visual hierarchy, e.g., when V2 neurons pool neural signals from V1 neurons [5]. Hence, in line with classic hierarchical feedforward processing (figure 2a), crowding may be seen as a natural consequence of object recognition in the visual system. For example, a hypothetical neuron coding for a square might respond to signals from neurons coding for the lines making up the square. In order to achieve translational invariance, the square neuron is sensitive to lines all over its receptive field and pools this information in order to decide whether a square is present. According to this logic, crowding occurs when elements that do not belong to the same object are pooled. In this sense, crowding is an unwanted by-product of object recognition and, for this reason, a bottleneck of vision (for a reviews, see [2,6]). Other models have proposed that performance in crowding deteriorates because features of the target are substituted for features of the flanking elements [4,7]. As mentioned, all these models are local in the sense that crowding is determined by nearby elements only. Based on these two lines of thought, pooling and substitution, researchers have suggested that with more flankers performance deteriorates because more irrelevant features are pooled or substituted.

PLEASE INSERT FIGURE 1 AROUND HERE

Figure 1: Crowding. a. In crowding, the perception of a target element deteriorates in the presence of nearby elements. When fixating the left cross, the target letter V on the right is hard to identify because of the nearby flankers. **b.** The task is easier than in (a), because the flankers are further away from the target letter V. Bouma's law states that crowding occurs only when flankers are sufficiently close to the target, within the so-called Bouma's window. **c.** Crowding is a ubiquitous phenomenon since elements are rarely seen in isolation. For example, when fixating the central red dot, the child on the left is easier to detect because it is not surrounded by nearby flankers, as is the child on the right.

The understanding of crowding has largely changed in the last decade. For example, it has been shown that detailed information can survive crowding [8,9]. Crowding occurs in the fovea and is not restricted to the periphery, contrary to earlier proposals [10,11]. Most importantly for the present discussion, performance depends on elements far outside of Bouma's window. For example, in supercrowding, elements outside of Bouma's window decrease performance beyond the decrement arising from elements within the window [12]. Surprisingly, adding flankers can even reduce crowding, and such uncrowding effects can depend on elements outside of Bouma's window (figure 2; [10,13–17], review: [18]). For example, observers performed a vernier discrimination task. When a surrounding square was added to the vernier, the task became much more difficult: a classic crowding effect. However, adding more flanking squares improved performance gradually, i.e., performance improved the more squares were presented ([19]; figure 2b). The entire line of squares extends over 17 degrees in the right visual field, while the single vernier offset threshold is less than 200'' (figure 2d). Hence, performance is not exclusively determined by local interactions: fine-grained vernier acuity in the range of about 200'' depends on elements as far away as 8.5 degrees - a ratio of two orders of magnitude, extending far beyond Bouma's window. Moreover, performance depends on the overall configuration [20]. For example, in three-by-seven displays of squares and stars (figure 2c), a shift of the central row changes performance strongly (figure 2c, 4th and 5th configurations). Similar effects were found with stimuli other than verniers [21,22], as well as in auditory [23] and haptic crowding [24].

PLEASE INSERT FIGURE 2 AROUND HERE

Figure 2: a. Standard view of visual processing. First, edges are detected by low-level neurons with small receptive fields. Higher level neurons pool signals from lower level neurons in a hierarchical, feedforward manner, creating higher level representations of objects by combining low-level features [25,26]. For example, two low-level edge detectors may be combined to create a "corner" representation. Four such corner detectors can be assembled to create a rectangle representation. Receptive field size naturally increases along this pathway since, for example, a rectangle covers larger parts of the visual field than the lines making up the rectangle. **b. Uncrowding.** Observers performed a vernier discrimination task. The y-axis shows the threshold for which observers correctly discriminate the vernier offset in 75% of trials (so

performance is good when the threshold is low). First, only a vernier is presented, an easy task (performance for this condition is shown as the dashed horizontal line). Then, a flanking square is added making the task much more difficult (leftmost stimulus). This is a classic crowding effect. Importantly, adding more flanking squares improved performance gradually, i.e., performance improved the more squares are presented [19]. We call this effect uncrowding. **c. The global configuration** of the entire stimulus determines crowding. Performance is strongly affected by elements far away from the target as shown in these examples [15]. **d. Performance is not determined by local interactions only**. In this display, fine-grained vernier acuity of about 200'' depends on elements as far away as 8.5 degrees - a difference of two orders of magnitude, extending far beyond Bouma's window.

Because they cannot produce long-range effects, local models cannot explain the global aspects of crowding. Here, we tested which global models, integrating information across large parts of the visual field, can explain global effects on crowding (see figure 3 for a list). We also tested the most prominent local models to verify our hypothesis that local models are inadequate to explain global aspects of crowding.

The models that we tested differ with respect to four criteria:

Spatial extent: Local vs. Global. In a local model, elements far from the target do not exert any effects on the target. By contrast, in a global model, any element in the visual field may potentially interfere with target processing.

Mechanism of interference: Pooling, substitution, or other?

Organisation: Feed-forward (features at a given level are only affected by lower level features) vs. recurrent processing (features at a given level can be affected by lower or higher level features).

Grouping component: Does the model incorporate a grouping component? Certain models explicitly compute grouping-like aspects by determining which low-level elements should belong to the same higher level group. Only elements within a group interfere with each other.

PLEASE INSERT FIGURE 3 AROUND HERE

Figure 3: Tested models and their characteristics. Models may integrate information locally or globally, and the interference mechanism may be pooling, substitution, or other. Models are feed-forward or recurrent, and may or may not compute grouping-like aspects of the stimulus. The aim of the current work is to investigate which models can explain the global effects of crowding.

Methods

To test the models, we used human data from previous work exploring the crowding/uncrowding phenomena [10,11,15,17,19,20]. The stimulus database comprises 40 different stimuli belonging to 11 different categories: circles, Gestalts, hexagons, irregular1, irregular2, lines, octagons, patternIrregular, patternStars, squares and stars. An example of each category is shown in figure 4. Behavioral results can be found in the original papers (listed in figure 14a). In each category, we have the vernier target alone, plus crowding and uncrowding configurations. All the stimuli are shown in figure 14a and behavioural results can be found in the original papers. With a few exceptions (see details in the results section), we ran each model on all stimuli. For some models, we could not use the entire database because computation time was too long (deep convolutional networks, LAMINART, Texture Tiling Model), or because the model was not adapted to accommodate certain kinds of stimuli (Population Coding). Human and model results are summarized in the discussion (figure 14). All the code we used is available online at https://github.com/adriendoerig/beyond-boumas-window-code (except the Texture Tiling Model, which Rosenholtz and colleagues will share in a forthcoming publication). All the results can be found at https://github.com/adriendoerig/beyond-boumas-window-results.

There are two fundamentally different approaches to measure model performance. First, a linking hypothesis may be used to relate model output to performance (both are scalar numbers). For example, template matching computes how similar the model output is to the target image. If they are similar, performance is good. The second, textural approach is used to quantify performance in textural models. The idea is that peripheral vision is ambiguous because information is compressed by summary statistics. If a model uses a proper algorithm for representing these ambiguities, presenting the processed image in the fovea should lead to similar human performance as presenting the original unprocessed image in the periphery [27]. Accordingly, to measure the performance of textural algorithms, the stimuli are fed through a texture synthesis procedure. Then, observers freely examine the output image and report vernier orientation. If this task is easy, performance is good. For each model, we used the linking hypothesis proposed by the original authors when available. When this was not possible (for example for Alexnet, which has never been applied to crowding results before), we detail which linking hypothesis we used in the corresponding section. In the following, we present, first, textural models and, second, models using a linking hypothesis.

An important point is that different readouts lead to different results. Hence, the different methods of model evaluation used here could affect our results. However, we are mainly interested in qualitative rather than quantitative comparisons and the readout functions we used cannot confuse crowding and uncrowding. More specifically, the readout processes we use produce results monotonically linked to the model outputs. Hence, they cannot confuse uncrowding cases (a U-shape function where the vernier alone condition leads to good performance, a single flanker deteriorates performance, and multiple flankers lead again to good performance) with cases that do not show uncrowding (a monotonic function where the vernier alone condition leads to good performance, a single flanker deteriorates performance, and multiple flankers deteriorate performance even more).

Because different models were evaluated differently, it was impossible to come up with one performance measure and to compare models via something like the Akaike Information Criterion. However, despite this variety of performance measures, our results are qualitatively unambiguous: each model either is capable of producing uncrowding, or it is not. We took the parameters directly from the original models whenever possible. Otherwise, we tried our best to search the parameter space (see results). We cannot exclude that other combinations of parameters fit the dataset better. However, we will argue that the models that cannot produce uncrowding fail to do so for principled reasons, and not because of poor parameter choices (see discussion).

PLEASE INSERT FIGURE 4 AROUND HERE

Figure 4: Stimulus categories. We used 40 different stimuli from 11 different categories. The task was always to report the offset direction of the central vernier. This figure shows one example from each category. The stimulus database is tailored to test for global effects such as uncrowding. Human data was taken from previous work [10,11,15,17,19,20]. Human and model results are summarized in the discussion (figure 14 shows the results for all stimuli and models).

Results

Texture-like models:

The following models are based on texture analysis. The outputs are images, and the *texture method* is applied as described in the methods.

Epitomes

In the Epitomes model, described by Jojic et al. [28], large repeating patterns are summarized by small repeated representative image patches. Repeated patterns are substituted with their exemplars. The original image can subsequently be retrieved with good accuracy from the compressed representation, even though neighboring features encoded in the same patch are mingled. Epitomes are effectively a "substitution" model that exploits regularities. Although this model was not proposed as a model of crowding, it embodies many of the key characteristics of local pooling and substitution models.

Using the Jojic et al.'s code available online (http://www.vincentcheung.ca/research/sourcecode.html) we ran the model on all stimuli with the original parameters (designed to optimize image reconstruction accuracy for natural images and texture overlays). To evaluate performance, we used the texture evaluation method with the authors as subjects, analysing the results qualitatively (see methods). In addition, we computed the model threshold as

[redacted]

where leftStim(x,y) is the normalized intensity of pixel (x,y) in the left vernier offset version of the output. Effectively, this equation quantifies how different the normalized output images are for the left and the right vernier offset versions of the stimulus. If they are very different, the task is easy. Consistently across the dataset, the model successfully produces crowding but not uncrowding: performance was always worse when adding more flankers (figure 5). We suggest that the model cannot explain uncrowding because it compresses information from local regions of the image, ignoring global structure.

PLEASE INSERT FIGURE 5 AROUND HERE

Figure 5: Epitomes. a. Illustration of the epitome model. An image (left) is compressed into an epitome (center), a summary of local features. The image on the right is reconstructed from the epitome. **b.** As an example for the classic texture evaluation, we show the stimulus and reconstructed image for the 1- and 7-square conditions. Human vernier offset thresholds are better for the 1-square than the 7-square condition. The model does not produce uncrowding because vernier offset direction in the output is not easier to make out in the 7-square than in the 1-square case (according to the authors' judgment). **c.** Example for our performance measure. Human and model thresholds (see main text for how model threshold was computed) for vernier alone (condition 1), single square (condition 2) and 7 squares (condition 3). The 7-square threshold is higher than the 1- square threshold, in contrast with human performance. Note: the model outputs a number quantifying how different the left and right vernier offset versions of the input are (so the higher this difference, the better the performance). To make comparison

with the human threshold easier, we applied the following monotonic transformation to the output: "threshold-like output" = 1/"raw output". Then, we scaled the result to be in the same range as the human results. This monotonic re-scaling cannot not change the conclusions because monotonic outputs are mapped on monotonic performance and the same is true for U-shaped functions (see methods).

Single Texture Model

Portilla & Simoncelli [29] proposed a set of statistics capable of capturing key aspects of texture appearance to human vision (figure 6a). Balas et al., [27] suggested an explanation of crowding in which peripheral vision might measure these texture statistics in pooling regions that overlap and tile the visual field. The intuition is that summary statistics provide an efficient way of extracting relevant information at low computational cost from natural images. Though Balas et al. proposed a model covering the entire visual field as described in the next subsection, they initially tested the predictions of a single pooling region, since texture synthesis procedures did not exist for multiple overlapping pooling regions. Each of their stimuli fell within a single Bouma-sized patch. They have since suggested that this shortcut of using a single pooling region, which greatly reduces computation time, can often suffice for texture-like stimuli that fall within a single pooling region [30].

Although the model was intended by Balas et al. to be applied only over a Bouma's window-sized patch, here we applied it to the entire stimulus to see if this kind of texture synthesis could capture long-range interactions between the vernier and other elements. The texture statistics are computed from pixel intensities taken from the entire image. Using the code provided online by

Portilla & Simoncelli (https://github.com/LabForComputationalVision/textureSynth), we created textures from all of our stimuli and the authors analyzed the results qualitatively using the texture measure (see figure 6c for two examples). The model produces strong crowding: vernier offsets are harder to discriminate from the textures when flankers are present. However, the model cannot explain uncrowding: consistently across our whole dataset, uncrowded conditions are worse than crowded conditions for this model (figure 6c). More elements always deteriorate performance. In their original contribution, Balas et al. seeded the texture synthesis algorithm using a low-pass, noisy version of the stimulus to reduce position noise. We also ran our stimuli using this method (see results repository online). While the output images became less distorted than without using the seed, it did not change the conclusion, because the target vernier remained much harder to detect in the textures synthesized from the uncrowded 7 flankers stimuli than from the crowded single flanker stimuli – i.e., there was no uncrowding.

Texture Tiling Model (TTM)

The TTM model was first described by Balas et al. [27], with its first full instantiation developed by Freeman & Simoncelli [31]. It computes summary statistics for overlapping local patches of the visual field, mimicking the way V2 receptive fields grow in size with eccentricity (figure 6b). Balas, Rosenholtz and others have studied this model extensively, calling it the Texture Tiling Model (TTM; [32,33]). In a series of papers, this model explained well the local aspects of visual tasks such as crowding and visual search. We ran a selection of stimuli through the TTM model (circles, squares, and irregular1). Similarly to the previous textures, the results were analysed by the authors using the *texture measure*. Crowding was well captured, but uncrowding could not be explained by TTM (figure 6d). The vernier was not better represented as the number of flankers increased.

PLEASE INSERT FIGURE 6 AROUND HERE

Figure 6: Texture Synthesis and Texture Tiling Model. a. A texture (right) synthesized from the input on the left using the Portilla & Simoncelli [29] summary statistics. The output resembles crowding. Pooling- and substitution-like effects occur. **b.** In the TTM, instead of applying the summary statistics process to the whole image at once, only local patches of the image are processed, yielding a local summary statistics model. The local patches are thought to reflect V2 receptive fields. **c.** Whole-field summary statistics. From left to right: stimuli and Portilla & Simoncelli textures for the vernier, 1-square and 7-square conditions. The vernier offset is easy to determine from the texture in the vernier alone condition, and slightly harder in the crowded condition (a right-offset is discernable in the middle top of the display). Across all data, the model consistently produces crowding, but no uncrowding, as exemplified in the right condition in which no offset is present at all. **d.** Texture Tiling model. The left column shows three synthesized examples from the 1-square condition. On the right is the 7-flanking squares case. The model cannot produce uncrowding: since the stimulus on the right is less crowded than the stimulus on the left in the human data, the direction of the vernier should be easier to make out on the right than on the left. However, this is not the case.

We suggest that TTM alone cannot explain uncrowding because it is a sophisticated local mechanism that scrambles together neighboring elements. There is no mechanism allowing elements that do not share a pooling region with the target to directly affect the target representation. Our results suggest that neither pooling summary statistics over the entire stimulus nor pooling over previously tested local regions explain the behavioural results. If the whole field is used, uncrowding cannot occur because more elements mean more interference and thus worse performance. On the other hand, using local regions does not help because far away elements cannot improve performance in cases where humans show uncrowding.

Deep Textures

Gatys and colleagues [34] used deep neural networks to create textures. The algorithm starts with a noise image and iteratively modifies it to match the correlations between neuron activities in a set of layers (figure 7a). This procedure synthesizes textures that are often indistinguishable from the original image, creating true metamers [35]. Deep textures were not intended to be applied to images like our stimuli, nevertheless we were interested in seeing if they could handle them because one could think of deep textures as synthesizing textures based on learned features rather than on the hand-coded features of Portilla & Simoncelli [29]. Perhaps the learned features provide a better representation and thus do a better job of predicting crowding.

Using Gatys et al.'s code with their suggested set of parameters (https://github.com/leongatys/DeepTextures), we created textures of each stimulus in our database (Figure 7b shows a selection of examples). We first evaluated model performance by the *texture measure* performed by the authors. Since the results were much less clear than for the previous texture approaches, we also conducted a psychophysical experiment with naive participants. Five subjects performed the *classic texture measure*: they were first explained the texture synthesizing process and then were shown textures synthesized from our stimuli. They were asked to report if they thought the texture was synthesized from a left- or right-vernier stimulus. We used three categories of stimuli (Gestalts, squares and circles), with ten textures per stimulus (a total of 100 textures). Performance was at chance for all stimuli. Textures for the untested stimulus categories strongly resemble the tested categories (the vernier offset orientation is not visible in the textures, even for the vernier-alone condition). We tried different stimulus sizes, but this did not improve the results. In conclusion, despite its clear success at texture synthesis for natural images, the model in its present form is not suitable to study crowding with our stimuli.

Wallis et al. [36] have proposed a foveated model in which these deep statistics are computed over local image patches, just as the TTM computes Portilla and Simoncelli's statistics over local patches. The code is not yet publicly available, so we did not test it explicitly, however, we believe it will not explain uncrowding for exactly the same reasons that the TTM does not handle uncrowding better than Portilla and Simoncelli's whole field statistics: distant elements that are not in pooling regions around the target cannot affect the target representation.

PLEASE INSERT FIGURE 7 AROUND HERE

Figure 7: Deep textures. a. In the deep textures algorithm, the correlation between a deep neural network's unit activities is used as a summary statistic. Textures are then synthesized to match that statistic. **b.** Original stimuli and textures synthesized from these stimuli using the deep textures algorithm by Gatys et al. [34]. The vernier offset is poorly visible, therefore, despite its clear success at synthesizing textures, the

model in its present form in not suitable to model crowding with our stimuli. We tried different zooms on our stimuli but the results did not change.

Models using a linking hypothesis

The following models all use a linking hypothesis to relate their output (a number) to human performance. Whenever possible, we used the same linking hypothesis as in the original contribution. When no linking hypothesis was available, we specify the method used.

Wilson & Cowan Network with End-Stopped Receptive Fields

Wilson & Cowan [37] proposed a mathematical model of simple cortical (excitatory and inhibitory) neurons interacting through recurrent lateral connexions. Variations of this kind of model have successfully accounted for visual masking data using stimuli similar to our lines category [38]. We used a similar neural network for our crowding stimuli. The model first convolves the input image with an on-center, off-surround receptive field mimicking processing by the LGN. Next, the input activations are fed into both an excitatory and an inhibitory layer of neurons, which are reciprocally connected such that the excitatory units excite the inhibitory units and the inhibitory units inhibit the excitatory units. Details of the model, its filters, and its parameters can be found in [38] and [39]. Although the filters are local, the strength of activity at any given pixel location partly depends on the global pattern of activity across the network because of the feedback connections. More generally, the feedback in the network functions like a discontinuity detector by enhancing discontinuities and suppressing regularities. Clarke, Herzog & Francis [40] applied this model to crowding stimuli, but it performed poorly and produced no uncrowding. For example, there was no difference between the stimuli in the Gestalts category and the length of the bars in the lines category had no effect at all on performance. Here, to improve the model, we replaced the classic receptive fields by end-stopped receptive fields so that each neuron is optimally activated only by stimuli of a specific length. There were three different sizes for the end-stopped receptive-fields, corresponding to the size of a vernier bar, the size of the whole vernier, and the size of the flankers. To measure performance for each stimulus, for each end-stopped receptive field size, we took as output the state of the excitatory layer after stabilization (40 time-steps) and cross-correlated it with the vernier alone output. The cross-correlations for each end-stopped receptive field size were summed to yield a single output number per stimulus. We then fitted a psychometric function on one class of stimuli (training set) and used this function to provide model performance for all other classes of stimuli (testing set). Apart from the end-stopped receptive fields modification, we used the same parameters as in Hermens et al. [38].

We fit the psychometric function based on the model's output for the squares category, i.e., the squares category is the training set, and used this fit to measure performance on all other stimulus categories, i.e., all other categories are the testing set. We also tried to use each of the other categories as the training set; using the squares yielded the best results. The model produces crowding: performance drops in the presence of flankers. It also produces uncrowding but only for the training set (squares) and, to a lesser extent, for the irregular1 category. Indeed, performance is better in the 7 squares than in the single square condition (Figure 8b), and marginally better in the 7 irregular1 than in the single irregular1 condition (Figure 8c). For the other categories, there is no uncrowding (see Figure 8d for an example). The choice of the training and testing sets has a strong influence on the conditions that mimic human performance. Squares and lines are the categories for which size regularity seems to play the most important role. For all other classes, there is no uncrowding, regardless of the training. This poor generalization capability suggests that the model uses idiosyncratic features of its training set rather than capturing general regularities, similar to overfitting.

PLEASE INSERT FIGURE 8 AROUND HERE

Figure 8: Wilson and Cowan network with end-stopped receptive fields: a. Structure of the network in [38] which we augmented with end-stopped receptive fields. An excitatory and an inhibitory layer of neurons are activated by the stimulus and interact with one another. The output of the excitatory layer is cross-correlated with a vernier template to measure performance. **b.** Output for the squares category (with psychometric function fitted on the squares category). In accordance with human results, performance is better in the 7 squares than in the 1 square case. **c.** Output for the irregular category (with psychometric function fitted on the squares category). Performance is marginally better in the 7 irregular1 than in the 1 irregular1 case. **d.** Output for the stars category (with psychometric function fitted on the squares category). There is no uncrowding for this stimulus. Uncrowding occurs only for specific kinds of stimuli, where element size regularities seem important. Further, performance depends strongly on which data are used for the training set (i.e., for fitting the psychometric function), suggestive of overfitting. **e.** Model output images. Columns are different stimuli: vernier, 1 square and 7 squares. The first row shows the stimuli, and the three subsequent rows show the model output for the short, medium and long end-

stopped receptive fields. The crucial result is that the vernier is better represented in the short and medium populations in the 7 squares than in the 1 square conditions (i.e., uncrowding occurs). As mentioned, uncrowding occurred for very few stimuli categories. In cases that didn't show uncrowding, the vernier representation deteriorated further when flankers were added (see results on the online repository). Note: the model outputs a cross-correlation quantifying how similar the model output is to the model output in the vernier alone condition (so the higher this cross-correlation, the better the performance). To make comparisons with human thresholds easier, we applied the same linking hypothesis as Hermens et al. [38]: we fitted a psychometric function to link model outputs to behavioural results, as explained in the main text.

Zhaoping's V1 Recurrent Model

This recurrent neural network model is described by Li Zhaoping [41]. The network consists of a grid of neurons tuned to 12 orientations that are linked by lateral connections that follow a specific pattern (see figure 9a&b). The connectivity pattern allows the network to reproduce many experimental effects such as pop-out, figure-ground segmentation and border effects. It has also been shown to highlight certain parts of visual displays such as masked verniers [42], and we wondered if it could similarly produce uncrowding. We recoded the network from scratch following the detailed instructions and using the same parameters as in [41] and studied it as another recurrent model of early visual cortex. We ran all our stimuli and assessed performance by cross-correlating each output with the output of the vernier without flankers. The magnitude of the cross-correlation is taken as a measure of vernier offset discrimination performance. The model produces crowding but not uncrowding consistently across the dataset (see figure 9c).

PLEASE INSERT FIGURE 9 AROUND HERE

Figure 9: V1 Segmentation model. a. The input is sampled at each grid position by neurons tuned to 12 orientations, mimicking V1 simple cells. **b.** The connectivity pattern between cells depends on their relative position and orientation as shown here. Solid lines indicate excitation and dashed lines indicate inhibition. As shown, each neuron excites aligned neurons and inhibits non-aligned neurons. Each neuron has the same connectivity pattern, suitably rotated and translated. **c.** Output images for the square category. Each small oriented bar shows the maximally active orientation at this grid position. **d.** Results for the squares category. The dashed red bar shows the vernier threshold, which is matched for humans and the model. As shown, uncrowding does not occur in the model, because performance is worse for the 7 squares than the 1 square stimulus. Note: the model outputs a cross-correlation quantifying how similar the model output is to the model output in the vernier alone condition (so the higher this cross-correlation, the better the performance). To make comparison with the human threshold easier, we applied the same procedure as we did for the epitomes, i.e., we applied the following monotonic transformation to the output: "threshold-like output" = 1/"raw output". Then we scaled the result to be in the same range as the human results. This monotonic re-scaling does not change the conclusions – the phenomenon of uncrowding cannot be altered.

A Variation of the LAMINART Model

The LAMINART model by Cao & Grossberg [43] is a neural network capable of computing illusory contours between collinear lines. Francis, Manassi & Herzog [44] augmented it with a segmentation process in which elements linked by illusory contours are grouped together by dedicated neural populations. This dedicated neural processing operates in the same way for all conditions and plays an important role in explaining many other visual phenomena (review: [45]). This model process was intended as an implementation of a two-stage model of crowding, with a strong grouping process: stimuli are first segmented into different groups and, subsequently, elements within a group interfere. After dynamical processing, different groups are represented by distinct neural populations. Performance is determined by template matching. Importantly, crowding is low when the vernier is alone in its group (i.e., when the population representing the vernier does not also represent other elements) and high otherwise.

The segmentation process is started by local selection signals and spreads along connected contours (figure 10). The location of each selection signal follows a Gaussian distribution centred on a given location, with a constant standard deviation. Uncrowding occurs when the selection signals hit a group of flankers without hitting the vernier, rescuing it from the deleterious effects of the flankers. In our simulations, each stimulus is run twenty times, each time drawing a new selection signal location. The final performance is averaged over these twenty trials. Crucially, segmentation becomes easier with more flankers, because a group of many flankers connected by illusory contours produces a larger region for selection (figure 10).

To account for the observers' proclivity to succeed in the vernier discrimination task, the central location of a selection signal is tuned to produce the least amount of crowding for each condition. This assumption follows the idea that an observer does the best job possible in each given situation. Although this added flexibility is not present in other models, it does not constitute an unfair advantage for the LAMINART. Indeed, it is not strictly necessary in order for the model to produce uncrowding. For example, if the segmentation signals' central location followed a uniform distribution over the whole stimulus, it would still hit a large group of flankers (without hitting the target) more easily than a small group of flankers. In summary, whenever the flankers form a wide group that can be easily segregated from the vernier, uncrowding should be produced. Hence, uncrowding is largely independent of the selection signals' distribution.

Many stimuli in the dataset had been simulated by the model in Francis et al. [44]. Here, we improved the model by using more orientations and we ran the model on our dataset, using the template matching measure (some stimuli could not be run for reasons detailed below). Overall, the LAMINART explains the data set well (figure 10).

More precisely, the categories circles, Gestalts, lines, octagons, squares and hexagons are all well explained. Categories irreg1, irreg2 and stars cannot be explained, but they include bars of many different orientations, and the current LAMINART simulation is only capable of handling eight orientations. We did not run the stimuli in the patternStars and patternIrregular categories because they are too large to be processed in realistic time. In general, situations where the model fails tend to be those in which the model groups elements while the data suggests it should not, leading in some cases to no uncrowding, and in other cases to excessive uncrowding. One example is when flankers (e.g., squares and stars) group together when they should not. Another example is when flankers group with the target vernier (e.g., irreg1), suggesting the need to improve the grouping mechanism itself (figure 10).

Across all stimuli and all models, the LAMINART is by far the most successful model in this comparative study because it can explain a wide range of uncrowding results, as well as capture classic crowding effects.

PLEASE INSERT FIGURE 10 AROUND HERE

Figure 10: The LAMINART variation. a: Activity in the LAMINART model. Colors represent the most active orientation (red: vertical, green: horizontal). When a stimulus is presented, segmentation starts to propagate along connected (illusory or actual) contours from two locations marked by attentional selection signals. Visual elements linked together by illusory contours form a group. After dynamic, recurrent processing, the stimulus is represented by three distinct neural populations, one for each group. Crowding is high if other elements are grouped in the same population as the vernier, and low if the vernier is alone. On the left, the flanker is hard to segment because of its proximity to the vernier. Across the trials, the selection signals often overlap with the whole stimulus, considered as a single group. Therefore, the flanker interferes with the vernier in most trials, and crowding is high. On the right, the flankers are linked by illusory contours and form a group that spans a large surface. In this case, segmentation signals can easily hit the flankers group successfully (without hitting the vernier). The vernier thus ends up alone in its group in most trials and crowding is low. **b**: The left row shows human performance with the square flanker stimuli. The right row is the output of the LAMINART model. It fits the data very well. The same holds true for a majority of our stimuli. To compute the LAMINART's output values, we used the same linking hypothesis as in the original description of the model [44]: template matching is used to decide if the target vernier offset is left or right, and this result is monotonically transformed into a threshold-like measure. **c**: Sometimes flankers group together (illusory contours are formed) when they should not, erroneously predicting uncrowding for this condition. **d**: Sometimes flankers group with the vernier when they should

not. Here, weak illusory contours connect the central flanker and the vernier. No uncrowding can be produced for this condition because segmentation always spreads to the vernier, independently of the success of the selection signals.

Alexnet (A Convolutional Neural Network)

Deep Convolutional Neural Networks (CNNs) are local, feedforward, pooling networks. Training involves using feedback signals to adjust weights between neurons in subsequent layers. Once the network has been trained, users typically fix the weights and use the network in a feedforward manner. Given enough time and training samples, CNNs can learn any function by learning adequate weights [46,47]. CNNs fit very nicely in the standard view of vision research, in which basic features, such as edges, are combined in a hierarchical, feedforward manner to create higher-level representations of complex objects (figure 2a). We reasoned that crowding would occur in these networks for exactly the same reason as in classic local pooling models: the target and the flankers' representations at a given layer are pooled within the receptive fields of the subsequent layer, thus, leading to poorer performance. Although CNNs obviously compute groups such as objects or animals, these groups have no effect whatsoever on crowding of lower level features. Indeed, there are no connections from higher to lower level layers. Thus, elements far away from the vernier cannot interact with nearby elements and lead to uncrowding. To test this hypothesis, we processed the square category through Alexnet [48], a CNN trained to classify natural images with high accuracy, using Tensorflow [49]. In order to determine vernier offset discrimination in different layers, we trained classifiers to identify the vernier offset from the activations of different layers of Alexnet (figure 11a). The classifiers had a single hidden layer with 512 units, followed by a softmax layer with two outputs, corresponding to left and right. In the training phase, we ran verniers through the network, and trained classifiers to identify the offset orientation from the different layers' activations (which were normalized to zero mean and unit standard deviation). Each layer had its own classifier. We used all ReLU layers following the convolution layers and the last fully connected layer. A different classifier was trained for each of these layers. During the test phase, we used verniers alone, verniers flanked with a single square (crowded stimuli) and verniers with 7 squares flankers (uncrowded stimuli). Both training and testing stimuli had varying sizes, offsets and positions in the image. Figure 11 shows average performance for each layer over 6 runs. For each run, we trained a new classifier on each layer, using 250000 verniers in the training set. In the testing phase, we ran 3000 verniers, 3000 crowded stimuli and 3000 uncrowded stimuli through Alexnet. Our classifiers predicted vernier orientation from the layer activations for each of these inputs. Interestingly, our classifiers could well retrieve the test vernier orientations with 100% accuracy in all convolutional layers (layers 2, 3, 4 and 5). Adding square flankers deteriorated performance strongly. The single square (crowded) stimuli could be decoded only in the convolutional layers 2, 3 and 4, and in fully connected layer 7, but with much poorer accuracy than the vernier alone. Crucially, unlike in humans, the 7 squares (uncrowded) stimulus performance was always worse or equal to the performance on the single square (crowded) stimulus. Hence, the deep network produced crowding, but not uncrowding. We suggest that the mechanism leading to these results is similar to the classic local pooling account of crowding.

PLEASE INSERT FIGURE 11 AROUND HERE

Figure 11: Alexnet. a. Stimuli consisted of either verniers, verniers surrounded by a single square or verniers with seven squares. The stimuli had varying sizes, vernier offsets and positions. Alexnet's architecture and a classifier are shown on the right (there was a classifier at each layer). The boxes correspond to the input (leftmost box) and activated neuron layers (see [48] for the detailed architecture of Alexnet). We trained softmax classifiers on all ReLU layers following the convolution layers and the last fully connected layer to detect vernier orientation from the layer's activity. **b.** Accuracy of softmax classifiers trained to detect vernier orientation from different layers in the deep neural network Alexnet. Across all layers, the offsets in crowded stimuli (1 square flanker) are always better detected than offsets in uncrowded stimuli (7 square flankers). This runs contrary to human performance. NB. This model only produces percent correct, there is no output image.

Hierarchical Sparse Selection (HSS)

This model was described by Chaney, Fischer & Whitney [50]. In a series of experiments, it was shown that in spite of difficulty identifying a crowded target, crowding does preserve some information about the target, i.e., information is rendered inaccessible but not destroyed (see [8,9] for reviews). For example, a face surrounded by other faces cannot be explicitly identified, but information about its features can nevertheless survive crowding and contribute to the perceived average of a set of faces [51]. To accommodate these results, Chaney et al. proposed that information is not lost along the visual processing hierarchy. Instead, crowding occurs because readout is sparse. Specifically, given a feature map representing a stimulus, only a subset of the neurons from this map can be used to decode the target, which leads to crowding's deleterious effects (figure 12a).

Using the author's code, we tested all our stimuli and found that crowding could be explained, but uncrowding did not occur in the model (figure 12b). Originally, the model was used to detect crosses, triangles and circles. We modified the model's readout layer to classify vernier orientation, which was achieved with 99.13% accuracy (the rest of the model does not need any change to accommodate new stimuli). Then, we dropped 75% of the neurons for the sparse readout, which led to a vernier classification accuracy of 81.48%. We tested all our stimuli by asking the model to classify the vernier orientation, first without dropping any neurons, then with 75% of the neurons dropped for the sparse readout, as we did for the verniers. For all stimuli, performance dropped with the sparse readout. For example, the 1 square condition was classified with 93.35% accuracy when all neurons were used, and this dropped to 75.55% with sparse readout. The 7 squares condition had a similar profile, but classification accuracy was worse than for the 1 square condition (71.73% with all neurons and 59.23% with sparse readout). This pattern of results was found in all stimulus categories: sparse readout impaired performance, and adding more flankers impaired performance too. Thus, there was crowding but no uncrowding. We would like to mention that Chaney et al. argue that uncrowding can in fact be explained, if the target and flanker are represented in different feature maps, which are however not implemented at the moment. In essence, visual stimuli are segmented into different feature maps (this must happen early in the visual pathway to explain the low-level vernier results), and subsequently the HSS model applies within feature maps, on this pre-segmented input.

PLEASE INSERT FIGURE 12 AROUND HERE

Figure 12: Hierarchical Sparse Selection model. a. The model posits that receptive fields along the visual hierarchy are large and dense. This allows for "lossless" transmission of information through the visual system. For instance, the offset of the vernier in this illustration is not corrupted by pooling thanks to the density of the receptive fields (blue and red circles). Crowding occurs because, when we try to access information, only a few sparse receptive fields are used for readout (red circles). Hence, crowding occurs at readout because of sparse sampling of receptive fields. This sparse readout can occur at any stage of visual processing, from low-level features (shown here) to faces. **b.** Uncrowding does not occur in the Hierarchical Sparse Selection model because performance is worse for the model on the 7 squares than the 1 square condition, contrary to human performance. NB. This model only produces a scalar output, there is no output image.

Models tested elsewhere

The following models were not implemented here, but we mention them for completeness.

Saccade-Confounded Summary Statistics

Nandy & Tjan [52] proposed a model linking summary statistics to saccadic eye movements: crowding is proposed to occur because the acquisition of summary statistics in the periphery is confounded by eye-movement artifacts. This leads to inappropriate contextual interactions in the periphery and in this way produces crowding. For the present purposes this is not directly relevant, because foveal and peripheral uncrowding results are qualitatively identical [11], which the saccade-confounded summary statistics model cannot explain since it suggests that crowding can only occur in peripheral regions. Furthermore, it is not clear how uncrowding can occur in this model.

Population Coding

This kind of model was first described by Van den Berg, Roerdink, & Cornelissen [53]. A similar model was proposed by Harrison & Bex [54]. Both models elegantly produce both pooling and substitution behaviour by assuming that an element's orientation is represented by a population code: a probability distribution of its orientation. When many elements are present, the population codes interfere and disturb the target element's representation, which leads to crowding. This interference depends on distance and is usually modeled as a 2D Gaussian. Dayan & Solomon [55] also proposed a model in which elements are represented as probability distributions. They added a Bayesian process to account for the accumulation of evidence over time. Their model captures local crowding effects similarly to Van den Berg et al. and Harrison & Bex's models: the interference comes from the representations of neighbouring elements deleteriously affecting each other. This model and the one by Van den Berg and colleagues cannot handle images as input and thus could not be tested with our stimuli.

We have shown elsewhere that the Harrison & Bex [54] implementation cannot explain uncrowding [56]. Agaoglu & Chung [57] showed that the interaction between elements depends on which of them is considered as the target for report. Hence, the crowding interference between elements in the display depends on the task, which is not easily incorporated in the models without a dedicated process. Van den Berg et al. [53] suggested that elements do not interfere when they are represented in different perceptual groups, similar to the LAMINART model. Similarly, Harisson & Bex [54] have suggested that a preprocessing stage determining which elements interfere is needed.

Fourier Model

The Fourier transform is sensitive to global aspects of spatial configurations because it is based on periodic features. Even if it was never explicitly proposed to explain crowding, it may capture some effects of uncrowding that have to do with regularities in the stimulus. Previously [15,40], we used a Fourier-based model and tested it on the entire dataset. Essentially, this is a texture-like model, assuming that the brain Fourier transforms the visual input. Repetitive structures, such as arrays of squares are more compactly coded in the Fourier space than the 2D space. We restate the results here for comparison with the other models. The model first bandpasses filters the stimuli (passing a small range of frequencies at all orientations), then computes the Fourier transforms of the filtered left- and right-offset cases for each stimulus. Similarly to what was done to measure performance of Zhaoping's recurrent V1 model, these are cross-correlated with the filtered versions of the verniers without any flankers and the magnitude of the cross-correlation is taken as a measure of vernier offset discrimination performance. This process is repeated over all possible pass-bands (which is finite given a fixed image size) until the pass-band yielding performance most similar to humans is found. Across the dataset, this approach failed to reproduce the data (see figure 13), suggesting that such a simple use of global regularities in the display is insufficient to explain crowding. Depending on the set of Gabor filters, uncrowding occurred for certain stimuli, but this was never consistent over several stimulus types, which is suggestive of overfitting. With one set of filters the lines category could be explained, with another the Gestalts category could be explained.

PLEASE INSERT FIGURE 13 AROUND HERE

Figure 13: Fourier model. a. The Fourier model computes Fourier transforms for the left- and right-offset versions of each stimulus. If these transforms are very different, crowding is low because the offset direction is easy to decode in Fourier space [15]. **b.** Output of the Fourier model. The model failed on most stimuli [15]. NB. This model only produces a scalar output, there is no output image.

Discussion:

For decades, crowding was thought to be fully determined by nearby elements. For this reason, target elements were presented only with a few nearby elements, and models were local in nature. However, experiments of the last two decades have shown that elements far beyond Bouma’s window can strongly affect performance. Crowding can become stronger [12] or weaker

[10,13–16] when elements are presented outside Bouma's window. Hence, local models cannot provide a complete account of crowding. In addition to spatial extent, it is the specific stimulus configuration that determines crowding. Configurational effects are not small modulations of crowding but have large effect sizes and, more importantly, can qualitatively change the pattern of results. For example in figure 2b, performance changes in a non-linear U-shaped fashion with best performance for the unflanked target, strong crowding for few flankers, and weaker crowding when flankers make up a regular configuration.

A major question is at which computational level crowding occurs. In local models, only nearby elements interfere with target processing, often due to low level mechanisms such as pooling. In global models, features across the entire visual field are potentially important. Global interactions may be restricted to low level features, such as the orientations of the stimulus elements. At the other extreme, explicitly computing objects (such as the squares in figure 2) may turn out to be necessary. Likewise, face crowding may or may not necessitate the explicit computation of faces [8,51,58,59]. For this reason, some global models explicitly compute grouping-like aspects. Only elements within a group interfere with each other. Classically, models restricting themselves to lower level features are given priority because they offer more parsimonious explanations.

Model comparison

Here, we investigated all available models suited to explain the global aspects of crowding.

All models (leaving aside Deep Textures, which was never proposed to explain crowding with laboratory stimuli) produced crowding comparable to the human data. However, only the

LAMINART model was consistently able to produce *un*crowding (figure 14). The Wilson and Cowan network produced uncrowding only for the squares category (and to a lesser extent for the irregular1 category). The Fourier model produced uncrowding only for the Gestalts and lines stimuli. In both models, uncrowding depended heavily on parameter values, a signature of overfitting. In the Wilson and Cowan network, the end-stopped receptive fields led to grouping elements of similar size, but this did not generalize to explain other global effects.

PLEASE INSERT FIGURE 14a AROUND HERE

PLEASE INSERT FIGURE 14b AROUND HERE

Figure 14: a. Summary of results. Results for all models (columns). In black, the left panel displays all crowding stimuli and the right panel displays all uncrowding stimuli (i.e., better performance when extra elements are added to the crowded condition) as observed in human data (rows). Superscript numbers indicate which publication the results are taken from (1: Sayim, Westheimer & Herzog [17]; 2: Manassi et al. [11]; 3: Manassi, Sayim & Herzog [19]; 4: Manassi et al. [15]). Red indicates that the model predicts crowding, green indicates uncrowding and gray indicates that we did not run the model on the stimulus. A perfect model would have only red in the left half of the table and only green in the right half. Only the LAMINART is capable of producing uncrowding consistently. Fourier and the Wilson-Cowan network produce uncrowding, but suffer from overfitting (see discussion). For these two models, we provide the results for the best parameters. For example, the Wilson and Cowan with different parameters can explain the lines category but then it cannot explain the squares and irregular1 categories. **b. Model comparison.** All models produce crowding, but only the Fourier, Wilson and Cowan and LAMINART models produce

uncrowding. The Fourier and the Wilson and Cowan model overfit and thus do not capture general principles. The LAMINART is the only model that explicitly computes grouping like aspects and segments the image into different layers.

We think there are principled reasons why most models cannot reproduce most of the global uncrowding findings. First, the effects of global configuration (figure 2c) operate on a much higher level than most models can capture. To phrase it this way, we think that human performance is based on global configurations and not on simple hidden sub-regularities, such as repeating patterns or simple summary statistics. Second, as Wallis et al. [36] put it: "Based on our experiments we speculate that the concept of summary statistics cannot fully account for peripheral scene appearance. Pooling in fixed regions will either discard (long-range) structure that should be preserved or preserve (local) structure that could be discarded. Rather, we believe that the size of pooling regions needs to depend on image content". For this reason, we think that performance in crowding cannot be explained simply as a by-product of basic spatial processing, e.g., by summary statistics. In contrast, which elements interfere seems to depend on the global stimulus layout. We propose that the LAMINART can consistently produce uncrowding because it can deal with this requirement by incorporating a grouping-like process: elements linked by illusory contours are grouped together and segmented from elements in other groups. Interference happens only between elements within a group.

Another way to approach the importance of grouping for crowding is that it provides extra information that makes one condition inherently easier than another. Vernier acuity tasks are often thought to be mediated by the responses of one or more feature detectors. Each feature detector might itself look like a vernier offset, or might be similar to an orientation detector such as a Gabor. Regardless, correct performance at the vernier task requires precise placement of the detector; a slightly misplaced detector can easily give the wrong answer, particularly when the vernier is flanked by other stimuli. Crowding induces location uncertainty. Any information that can help correctly place the detector – essentially any cue to the right position – would improve performance. Strong stimulus grouping could be one such cue (Rosenholtz et al., under review). In this case too, it is crucial to understand how the brain groups visual elements across the entire visual field.

The LAMINART model links elements by illusory contours, which is a rather basic grouping mechanism. It remains an open question whether more complex features are necessary to explain crowding/uncrowding such as an explicit computation of objects, e.g. squares, faces etc. For example, can the irregular shapes category be explained with simple contour integration? Likewise, it remains an open question whether face crowding can be explained without the explicit computation of faces.

In the LAMINART model, the grouping and interference processes are separate. Alternatively, grouping and interference may be intimately linked. One possibility is that the groups correspond to optimal statistical representations. For example, elements may form a group when they can be well compressed by summary statistics. In this scenario, grouping is part of the summary statistics process itself. There are probably many other ways in which grouping may play a role.

A major problem with the grouping approach is the lack of a well-defined, objective measure of grouping. If there is no objective measure, groups can be chosen ad hoc to explain experimental results, leading to circular explanations. As a first step towards an objective measure of grouping, subjective measures (i.e., asking observers to report what they feel belongs to a group) can complement studies. Such subjective ratings about perceptual groups have correlated well with psychophysical performance levels [11].

Future Models

As we have shown, none of the current models can fully explain (un)crowding. What would the model of the future look like? What components are crucial?

First, as mentioned earlier, we can rule out local models because elements across large parts of the visual field influence perception of the target.

Second, to explain the complex effects of spatial configurations in crowding, our results suggest that grouping-like, mid or higher level aspects need to be incorporated in a model. However, the exact nature of this process is unknown. For example, it may or may not be that mid-level processing is sufficient. In addition, the incorporation of higher level processes does not exclude the additional use of summary statistics and other lower level components. The grouping stage is difficult to study because of the seemingly infinite number of possible visual configurations. We believe that new tools are needed to help navigate the huge search space effectively. For example, Van der Burg, Olivers, & Cass [60] have proposed a genetic algorithm to find configurational features important for crowding.

Third, we cannot rule out feedforward models. Indeed, it is a mathematical fact that any recurrent model can be “unfolded” into a feed-forward network [61–63]. However, these feedforward models are usually extremely large and computationally expensive. For this reason, we suggest that models with feedback connections are much more likely to be able to explain how complex spatial configurations influence target processing. For example, higher level grouping processing, such as computing the squares and grouping them together, may feed back to lower level processing of the target, i.e., the vernier. Support for this hypothesis comes from the finding that the Alexnet CNN could not produce uncrowding, presumably because high-level features cannot influence low-level processing.

Fourth, the nature of interference remains unclear. One option is that interference occurs during complex spatial processing by an unknown mechanism. Another option is that the classic interference mechanisms operate after complex spatial processing is accomplished. For example, pooling may occur only for grouped elements. In the same line of reasoning, Chaney et al. [50], Van den Berg et al. [53] and Harrison & Bex [64] noted that adding a grouping stage to their interference mechanism may help explain a wider range of results. Combining complex spatial processing with good interference mechanisms may, therefore, allow for a happy marriage between interference- and grouping-based mechanisms leading to a truly unified model of crowding.

Conclusion:

The global stimulus configuration plays a crucial role in crowding, which cannot be captured by local models. For this reason, we propose that models of crowding need to include grouping like processes. While our results show that none of the current models lacking a grouping process can explain the global uncrowding phenomena, they may be good candidates for a potential second, interference stage.

How are basic features of the visual field grouped to form objects? The most successful model we analyzed, the LAMINART variation, suggests that this is done by linking features together by illusory contours. Further work is needed to assess how far this mechanism can go and what alternative or additional components are necessary, such as summary statistics. For example, the groups may correspond to optimal statistical representations (elements that can easily be compressed using summary statistics would form a group).

Most importantly, large scale, configurational effects are not restricted to visual crowding with vernier targets. Uncrowding occurs also for letters and Gabors [65], as well as in audition [23] and haptics [24]. Similar effects are found in backward masking [66] and overlay masking [17,67]. Hence, crowding is a special case of contextual processing. Vision research has largely missed these aspects because of the use of well-controlled stimuli, which are usually presented in isolation or with only a few nearby flankers. Our results suggest that in order to understand vision in general, a mid-level, contextual processing stage is inevitable.

Acknowledgments:

We thank Thomas Wallis, Alexander Ecker and Mauro Manassi for helpful discussions and the reviewers for insightful comments. We also thank the authors who shared their code on the web or by email.

References.

1. Bouma H. Visual interference in the parafoveal recognition of initial and final letters of words. Vision Res. 1973;13: 767–782. doi:10.1016/0042-6989(73)90041-2
2. Levi DM. Crowding-An essential bottleneck for object recognition: A mini-review. Vision Res. 2008;48: 635–654. doi:10.1016/j.visres.2007.12.009
3. Pelli DG, Palomares M, Majaj NJ. Crowding is unlike ordinary masking: Distinguishing feature integration from detection. J Vis. 2004;4: 12–12. doi:10.1167/4.12.12
4. Strasburger H, Harvey LO, Rentschler I. Contrast thresholds for identification of numeric characters in direct and eccentric view. Percept Psychophys. 1991;49: 495–508. doi:10.3758/BF03212183
5. Pelli DG. Crowding: a cortical constraint on object recognition. Curr Opin Neurobiol. 2008;18: 445–451. doi:10.1016/j.conb.2008.09.008
6. Pelli DG, Tillman KA. The uncrowded window of object recognition. Nat Neurosci. 2008;11: 1129–1135.
7. Ester EF, Klee D, Awh E. Visual crowding cannot be wholly explained by feature pooling. J Exp Psychol Hum Percept Perform. 2014;40: 1022.
8. Manassi M, Whitney D. Multi-level Crowding and the Paradox of Object Recognition in Clutter. Curr Biol. 2018;28: R127–R133. doi:10.1016/j.cub.2017.12.051
9. Whitney D, Haberman J, Sweeny TD. From textures to crowds: multiple levels of summary statistical perception. New Vis Neurosci. 2014; 695–710.

10. Malania M, Herzog MH, Westheimer G. Grouping of contextual elements that affect vernier thresholds. J Vis. 2007;7: 1–1. doi:10.1167/7.2.1

11. Manassi M, Sayim B, Herzog MH. Grouping, pooling, and when bigger is better in visual crowding. J Vis. 2012;12: 13–13. doi:10.1167/12.10.13

12. Vickery TJ, Shim WM, Chakravarthi R, Jiang YV, Luedeman R. Supercrowding: Weakly masking a target expands the range of crowding. J Vis. 2009;9: 12–12. doi:10.1167/9.2.12

13. Banks WP, Larson DW, Prinzmetal W. Asymmetry of visual interference. Percept Psychophys. 1979;25: 447–456. doi:10.3758/BF03213822

14. Livne T, Sagi D. Configuration influence on crowding. J Vis. 2007;7: 4–4. doi:10.1167/7.2.4

15. Manassi M, Lonchampt S, Clarke A, Herzog MH. What crowding can tell us about object representations. J Vis. 2016;16: 35–35. doi:10.1167/16.3.35

16. Põder E. Crowding, feature integration, and two kinds of “attention.” J Vis. 2006;6: 7–7. doi:10.1167/6.2.7

17. Sayim B, Westheimer G, Herzog MH. Gestalt factors modulate basic spatial vision. Psychol Sci. 2010;21: 641–644.

18. Herzog MH, Sayim B, Chicherov V, Manassi M. Crowding, grouping, and object recognition: A matter of appearance. J Vis. 2015;15: 5–5. doi:10.1167/15.6.5

19. Manassi M, Sayim B, Herzog MH. When crowding of crowding leads to uncrowding. J Vis. 2013;13: 10–10. doi:10.1167/13.13.10

20. Herzog MH, Manassi M. Uncorking the bottleneck of crowding: a fresh look at object recognition. Curr Opin Behav Sci. 2015;1: 86–93. doi:10.1016/j.cobeha.2014.10.006

21. Chakravarthi R, Pelli DG. The same binding in contour integration and crowding. J Vis. 2011;11: 10–10. doi:10.1167/11.8.10

22. Livne T, Sagi D. Multiple levels of orientation anisotropy in crowding with Gabor flankers. J Vis. 2011;11: 18–18. doi:10.1167/11.13.18

23. Oberfeld D, Stahn P. Sequential grouping modulates the effect of non-simultaneous masking on auditory intensity resolution. PloS One. 2012;7: e48054.

24. Overvliet KE, Sayim B. Perceptual grouping determines haptic contextual modulation. Vision Res. 2016;126: 52–58. doi:10.1016/j.visres.2015.04.016

25. DiCarlo JJ, Zoccolan D, Rust NC. How Does the Brain Solve Visual Object Recognition? Neuron. 2012;73: 415–434. doi:10.1016/j.neuron.2012.01.010

26. Riesenhuber M, Poggio T. Hierarchical models of object recognition in cortex. Nat Neurosci. 1999;2.

27. Balas B, Nakano L, Rosenholtz R. A summary-statistic representation in peripheral vision explains visual crowding. J Vis. 2009;9: 13–13. doi:10.1167/9.12.13

28. Jojic N, Frey BJ, Kannan A. Epitomic analysis of appearance and shape. Proceedings Ninth IEEE International Conference on Computer Vision. 2003. pp. 34–41 vol.1. doi:10.1109/ICCV.2003.1238311

29. Portilla J, Simoncelli EP. A Parametric Texture Model Based on Joint Statistics of Complex Wavelet Coefficients. Int J Comput Vis. 2000;40: 49–70. doi:10.1023/A:1026553619983

30. Zhang X, Huang J, Yigit-Elliott S, Rosenholtz R. Cube search, revisited. J Vis. 2015;15: 9–9.

31. Freeman J, Simoncelli EP. Metamers of the ventral stream. Nat Neurosci. 2011;14: 1195–1201. doi:10.1038/nn.2889

32. Keshvari S, Rosenholtz R. Pooling of continuous features provides a unifying account of crowding. J Vis. 2016;16: 39–39. doi:10.1167/16.3.39

33. Rosenholtz R, Huang J, Raj A, Balas BJ, Ilie L. A summary statistic representation in peripheral vision explains visual search. J Vis. 2012;12: 14–14. doi:10.1167/12.4.14

34. Gatys L, Ecker AS, Bethge M. Texture Synthesis Using Convolutional Neural Networks. In: Cortes C, Lawrence ND, Lee DD, Sugiyama M, Garnett R, editors. Advances in Neural Information Processing Systems 28. Curran Associates, Inc.; 2015. pp. 262–270. Available: http://papers.nips.cc/paper/5633-texture-synthesis-using-convolutional-neural-networks.pdf

35. Wallis TSA, Funke CM, Ecker AS, Gatys LA, Wichmann FA, Bethge M. A parametric texture model based on deep convolutional features closely matches texture appearance for humans. J Vis. 2017;17: 5–5. doi:10.1167/17.12.5

36. Wallis T, Funke C, Ecker A, Gatys L, Wichmann F, Bethge M. Towards matching peripheral appearance for arbitrary natural images using deep features. J Vis. 2017;17: 786–786.

37. Wilson HR, Cowan JD. A mathematical theory of the functional dynamics of cortical and thalamic nervous tissue. Kybernetik. 1973;13: 55–80. doi:10.1007/BF00288786

38. Hermens F, Luksys G, Gerstner W, Herzog MH, Ernst U. Modeling spatial and temporal aspects of visual backward masking. Psychol Rev. 2008;115: 83.

39. Panis S, Hermens F. Time course of spatial contextual interference: Event history analyses of simultaneous masking by nonoverlapping patterns. J Exp Psychol Hum Percept Perform. 2014;40: 129.

40. Clarke AM, Herzog MH, Francis G. Visual crowding illustrates the inadequacy of local vs. global and feedforward vs. feedback distinctions in modeling visual perception. Front Psychol. 2014;5.

41. Li Z. Visual segmentation by contextual influences via intra-cortical interactions in the primary visual cortex. Netw Comput Neural Syst. 1999;10: 187–212.

42. Zhaoping L. V1 mechanisms and some figure–ground and border effects. J Physiol-Paris. 2003;97: 503–515.

43. Cao Y, Grossberg S. A laminar cortical model of stereopsis and 3D surface perception: closure and da Vinci stereopsis. Spat Vis. 2005;18: 515–578. doi:10.1163/156856805774406756

44. Francis G, Manassi M, Herzog MH. Neural dynamics of grouping and segmentation explain properties of visual crowding. 2017;

45. Grossberg S. Towards solving the hard problem of consciousness: The varieties of brain resonances and the conscious experiences that they support. Neural Netw. 2017;87: 38–95.

46. LeCun Y, Bengio Y, Hinton G. Deep learning. Nature. 2015;521: 436–444. doi:10.1038/nature14539

47. Lin HW, Tegmark M, Rolnick D. Why Does Deep and Cheap Learning Work So Well? J Stat Phys. 2017;168: 1223–1247. doi:10.1007/s10955-017-1836-5

48. Krizhevsky A, Sutskever I, Hinton GE. Imagenet classification with deep convolutional neural networks. Advances in neural information processing systems. 2012. pp. 1097–1105.

49. Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, et al. TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems [Internet]. 2015. Available: https://www.tensorflow.org/

50. Chaney W, Fischer J, Whitney D. The hierarchical sparse selection model of visual crowding. Front Integr Neurosci. 2014;8.

51. Fischer J, Whitney D. Object-level visual information gets through the bottleneck of crowding. J Neurophysiol. 2011;106: 1389–1398.

52. Nandy AS, Tjan BS. Saccade-confounded image statistics explain visual crowding. Nat Neurosci. 2012;15: 463–469.

53. Van den Berg R, Roerdink JB, Cornelissen FW. A neurophysiologically plausible population code model for feature integration explains visual crowding. PLoS Comput Biol. 2010;6: e1000646.

54. Harrison WJ, Bex PJ. A Unifying Model of Orientation Crowding in Peripheral Vision. Curr Biol. 2015;25: 3213–3219. doi:10.1016/j.cub.2015.10.052

55. Dayan P, Solomon JA. Selective Bayes: Attentional load and crowding. Vision Res. 2010;50: 2248–2260. doi:10.1016/j.visres.2010.04.014

56. Pachai MV, Doerig AC, Herzog MH. How best to unify crowding? Curr Biol. 2016;26: R352–R353. doi:10.1016/j.cub.2016.03.003

57. Agaoglu MN, Chung ST. Can (should) theories of crowding be unified? J Vis. 2016;16: 10–10.

58. Kalpadakis-Smith A, Goffaux V, Greenwood J. Crowding for faces is determined by visual (not holistic) similarity: Evidence from judgements of eye position. 2017;

59. Sun H-M, Balas B. Face features and face configurations both contribute to visual crowding. Atten Percept Psychophys. 2015;77: 508–519. doi:10.3758/s13414-014-0786-0

60. Van der Burg E, Olivers CN, Cass J. Evolving the keys to visual crowding. J Exp Psychol Hum Percept Perform. 2017;43: 690.

61. Hornik K, Stinchcombe M, White H. Multilayer feedforward networks are universal approximators. Neural Netw. 1989;2: 359–366.

62. Schäfer AM, Zimmermann HG. Recurrent Neural Networks Are Universal Approximators. Artificial Neural Networks – ICANN 2006. Springer, Berlin, Heidelberg; 2006. pp. 632–640. doi:10.1007/11840817_66

63. Werbos PJ. Generalization of backpropagation with application to a recurrent gas market model. Neural Netw. 1988;1: 339–356. doi:10.1016/0893-6080(88)90007-X

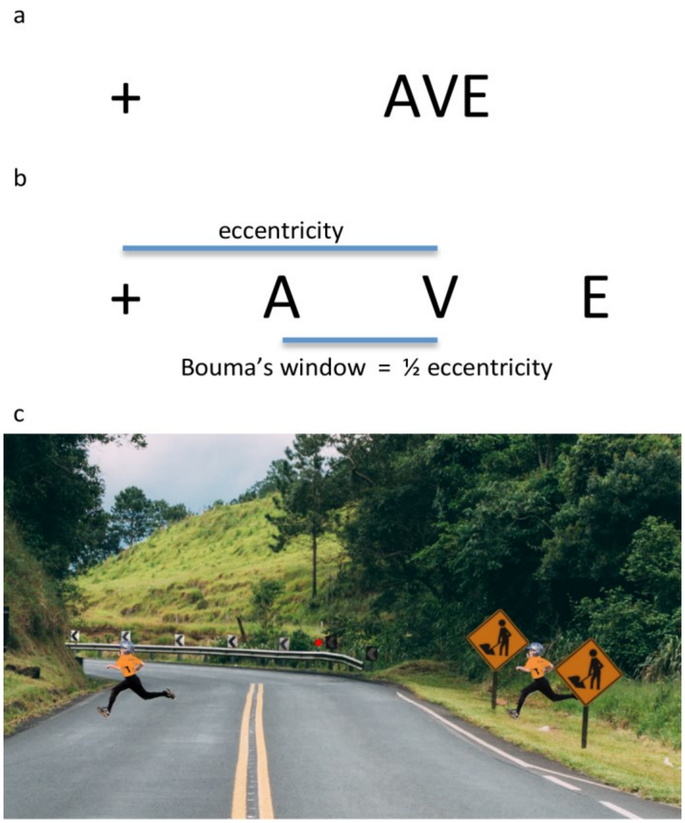
64. Harrison WJ, Bex PJ. Reply to Pachai et al. Curr Biol. 2016;26: R353–R354.

65. Saarela TP, Westheimer G, Herzog MH. The effect of spacing regularity on visual crowding. J Vis. 2010;10: 17–17.

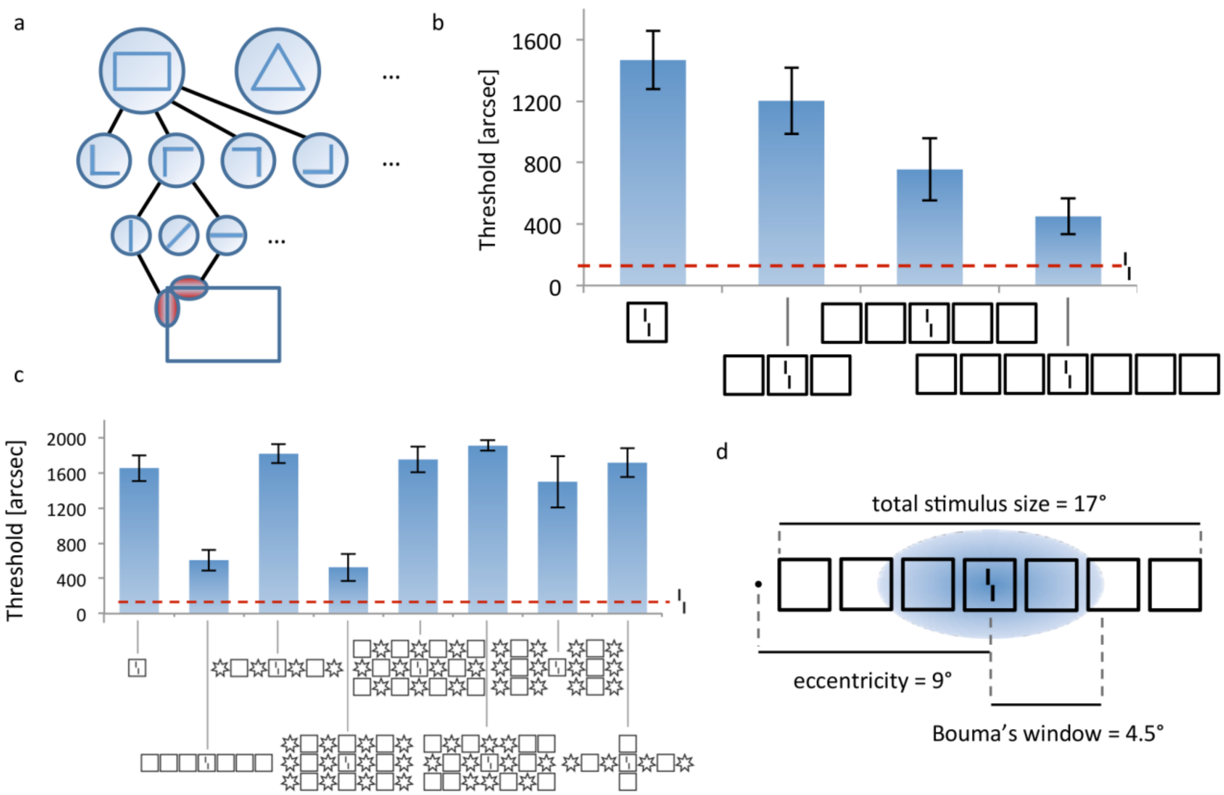
66. Herzog MH, Fahle M. Effects of grouping in contextual modulation. Nature. 2002;415: 433. doi:10.1038/415433a

67. Saarela TP, Sayim B, Westheimer G, Herzog MH. Global stimulus configuration modulates crowding. J Vis. 2009;9: 5–5.

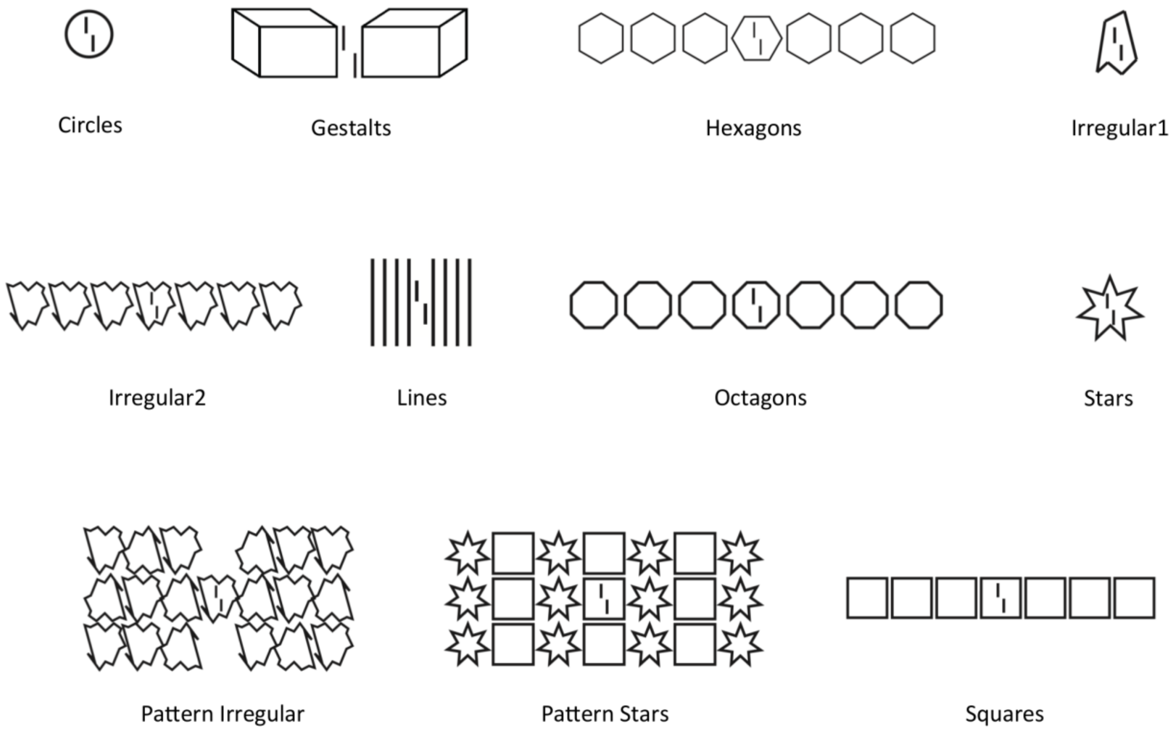
FIGURES:

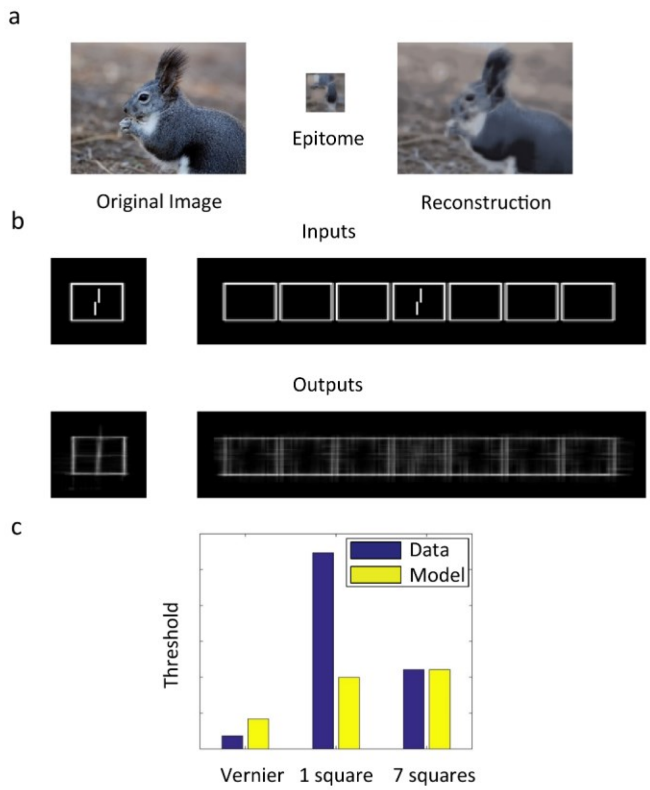


962



a
b
1600
1200
800
400
0
Threshold [arcsec]
c
2000
1600
1200
800
400
0
Threshold [arcsec]
d
total stimulus size = 17°
eccentricity = 9°
Bouma's window = 4.5°


Circles
Gestalts
Hexagons
Irregular1
Irregular2
Lines
Octagons
Stars
Pattern Irregular
Pattern Stars
Squares

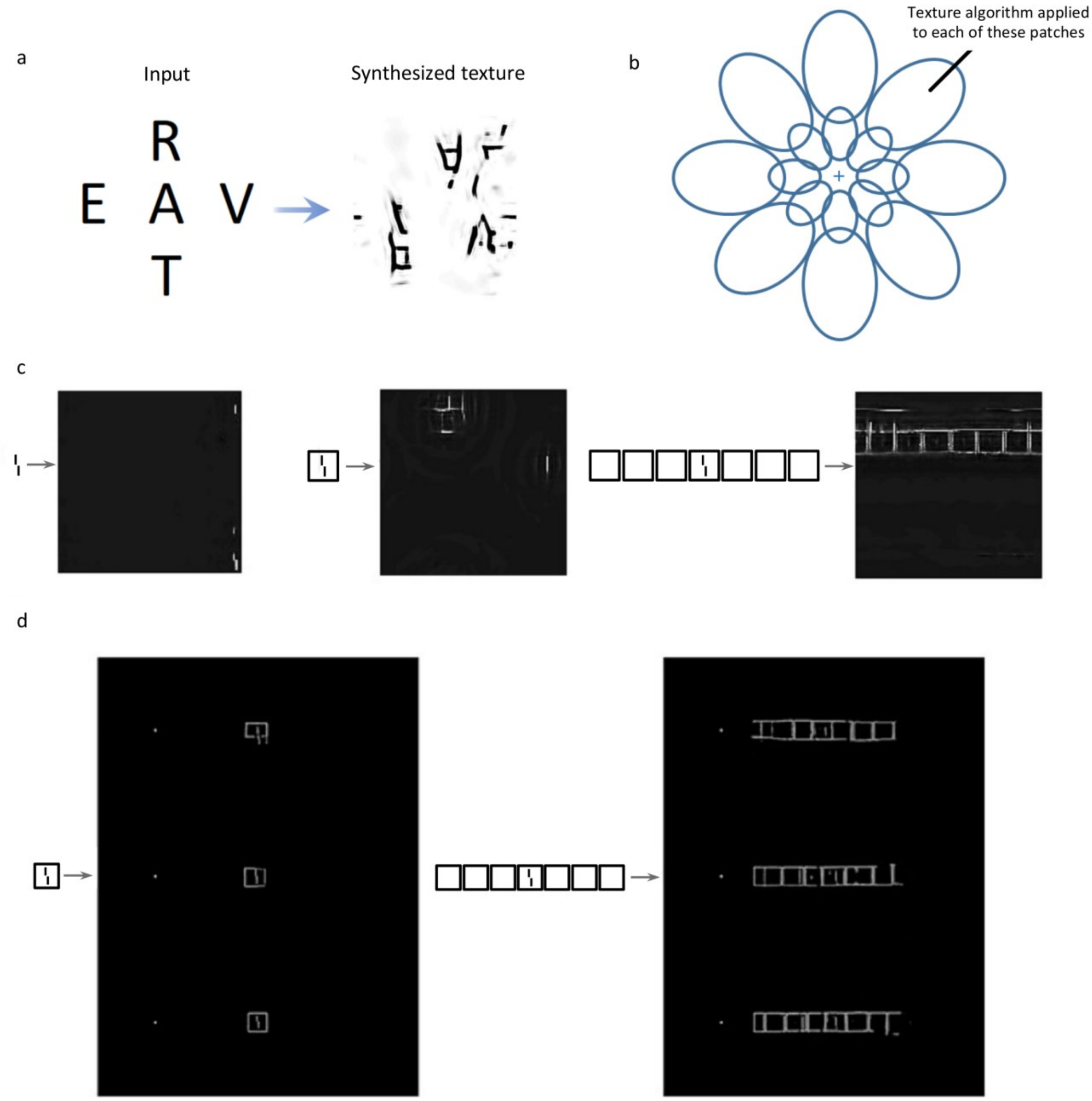

a
Epitome
Original Image
Reconstruction
b
Inputs
Outputs
c
Data
Model
Threshold
Vernier
1 square
7 squares

a

b

c

d

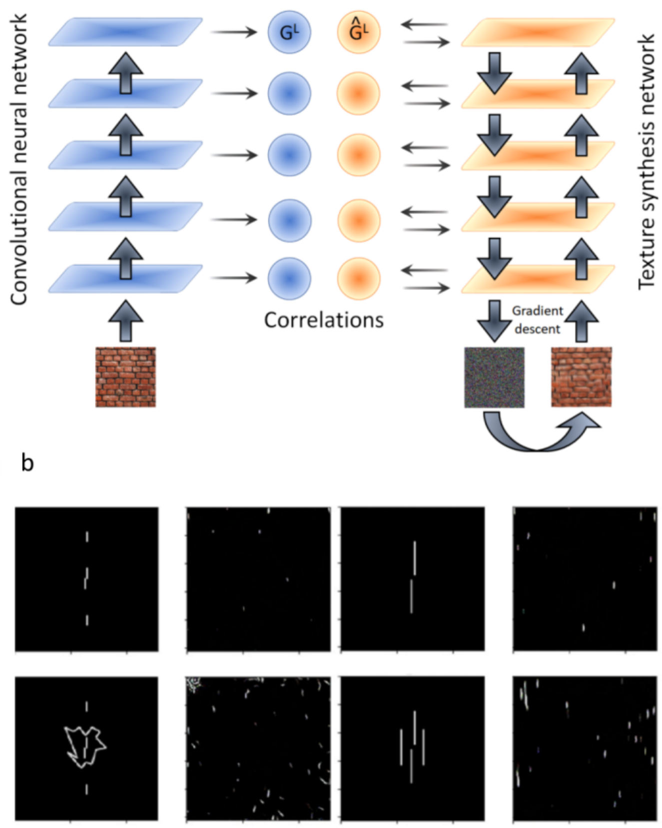


Input
Synthesized texture
R
E A V
T
Texture algorithm applied
to each of these patches

965

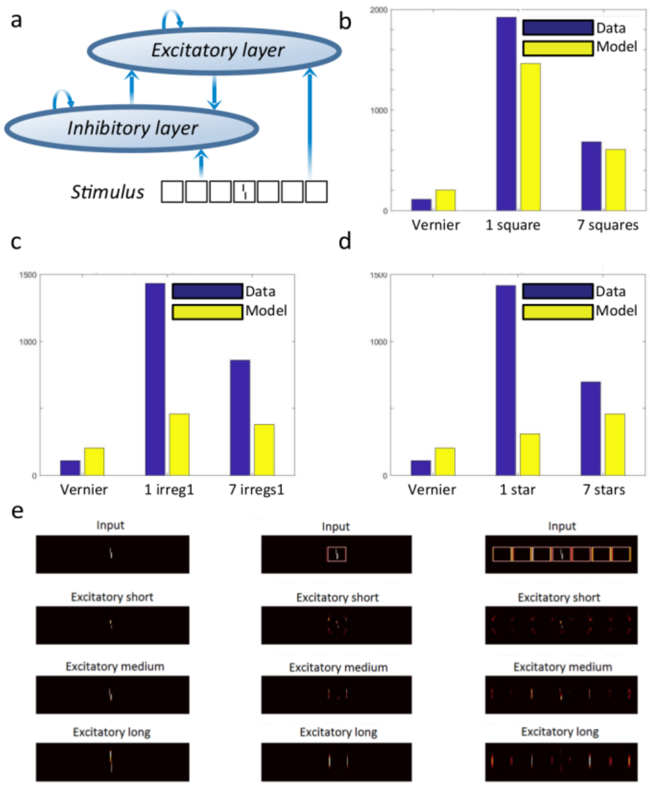
a

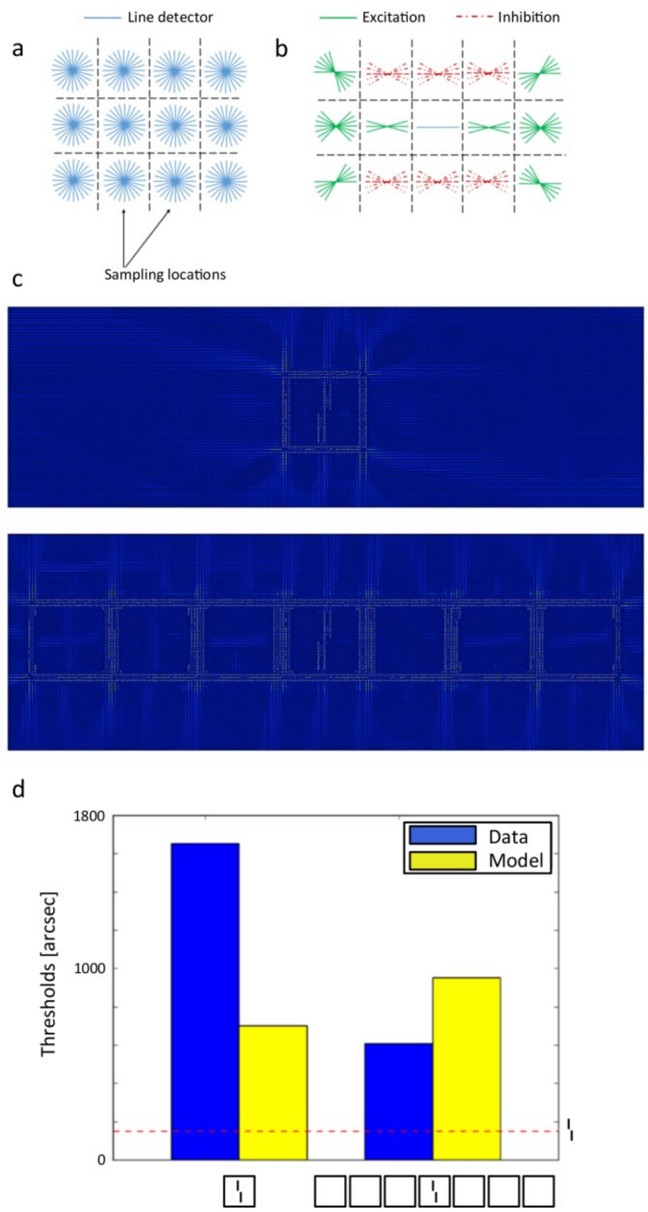
b

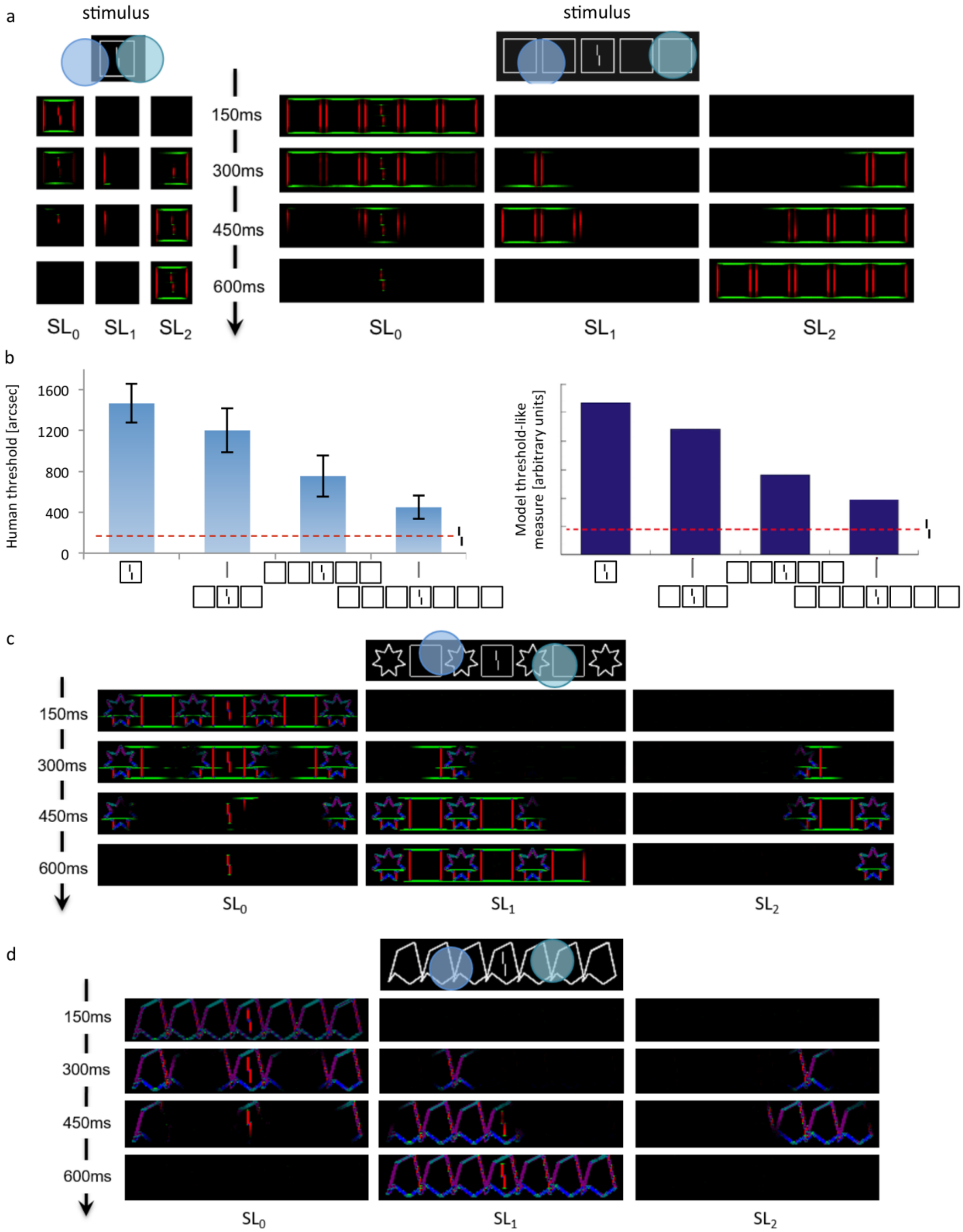


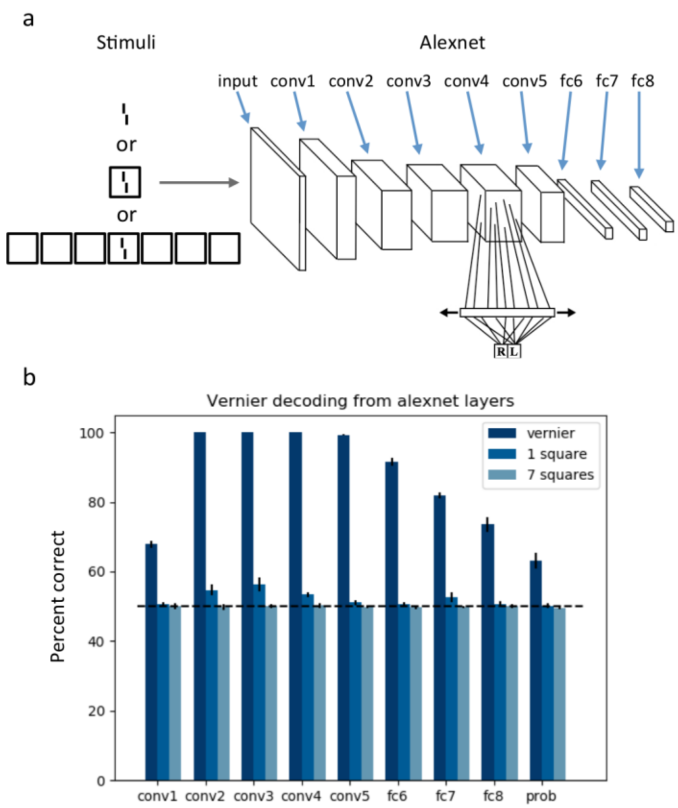
Convolutional neural network
Correlations
Texture synthesis network
Gradient descent

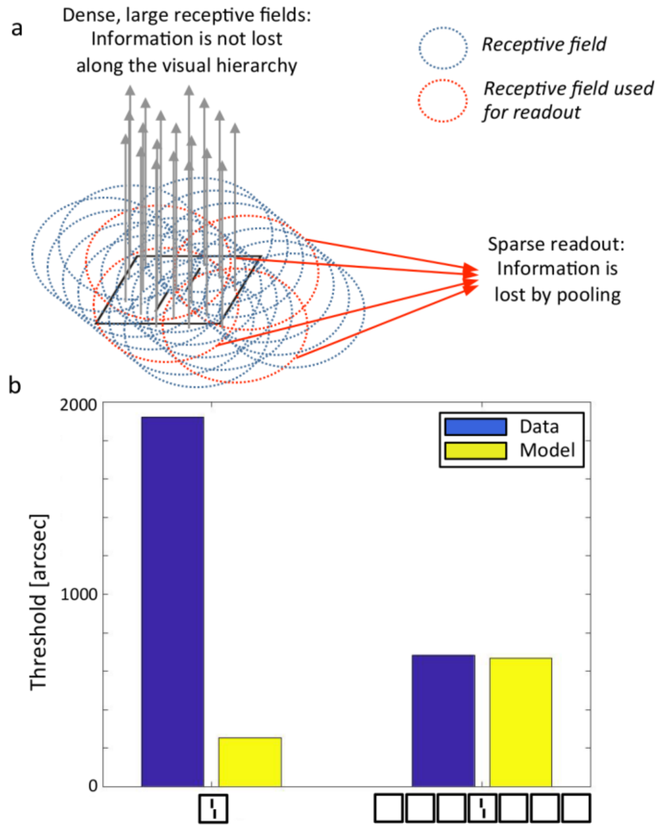
966

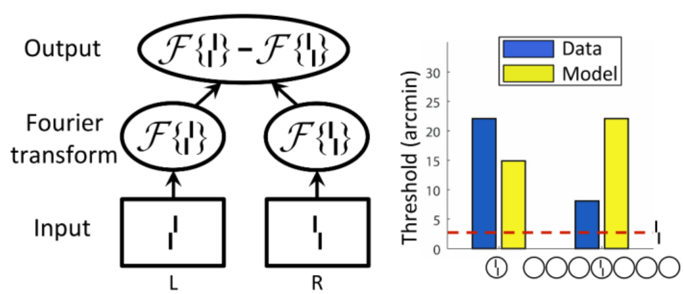

a
Excitatory layer
Inhibitory layer
Stimulus
b
Data
Model
2000
1000
0
Vernier
1 square
7 squares
c
Data
Model
1500
1000
0
Vernier
1 irreg1
7 irregs1
d
Data
Model
1500
1000
0
Vernier
1 star
7 stars
e
Input
Excitatory short
Excitatory medium
Excitatory long

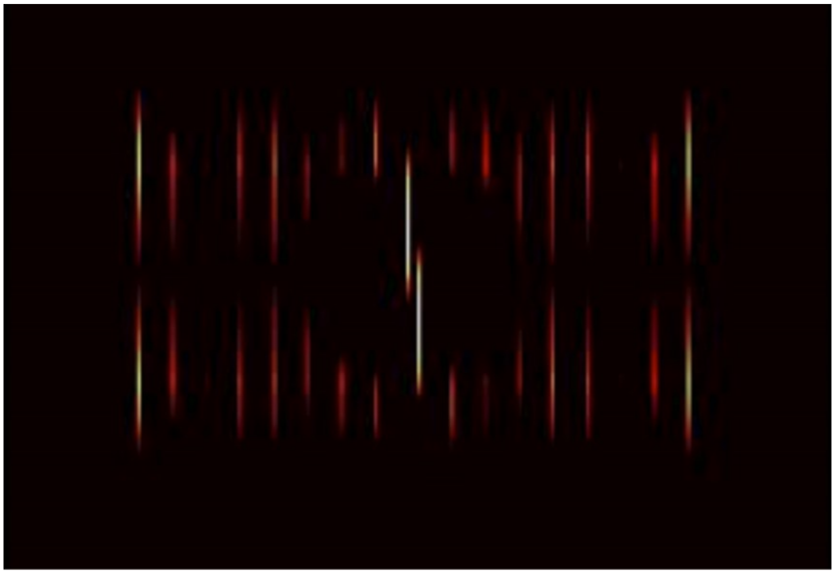

Line detector
Excitation
Inhibition
a
b
Sampling locations
c
d
1800
1000
0
Thresholds [arcsec]
Data
Model

969

a
stimulus
stimulus
150ms
300ms
450ms
600ms
SL0
SL1
SL2
SL0
SL1
SL2
b
Human threshold [arcsec]
1600
1200
800
400
0
Model threshold-like measure [arbitrary units]
c
150ms
300ms
450ms
600ms
SL0
SL1
SL2
d
150ms
300ms
450ms
600ms
SL0
SL1
SL2

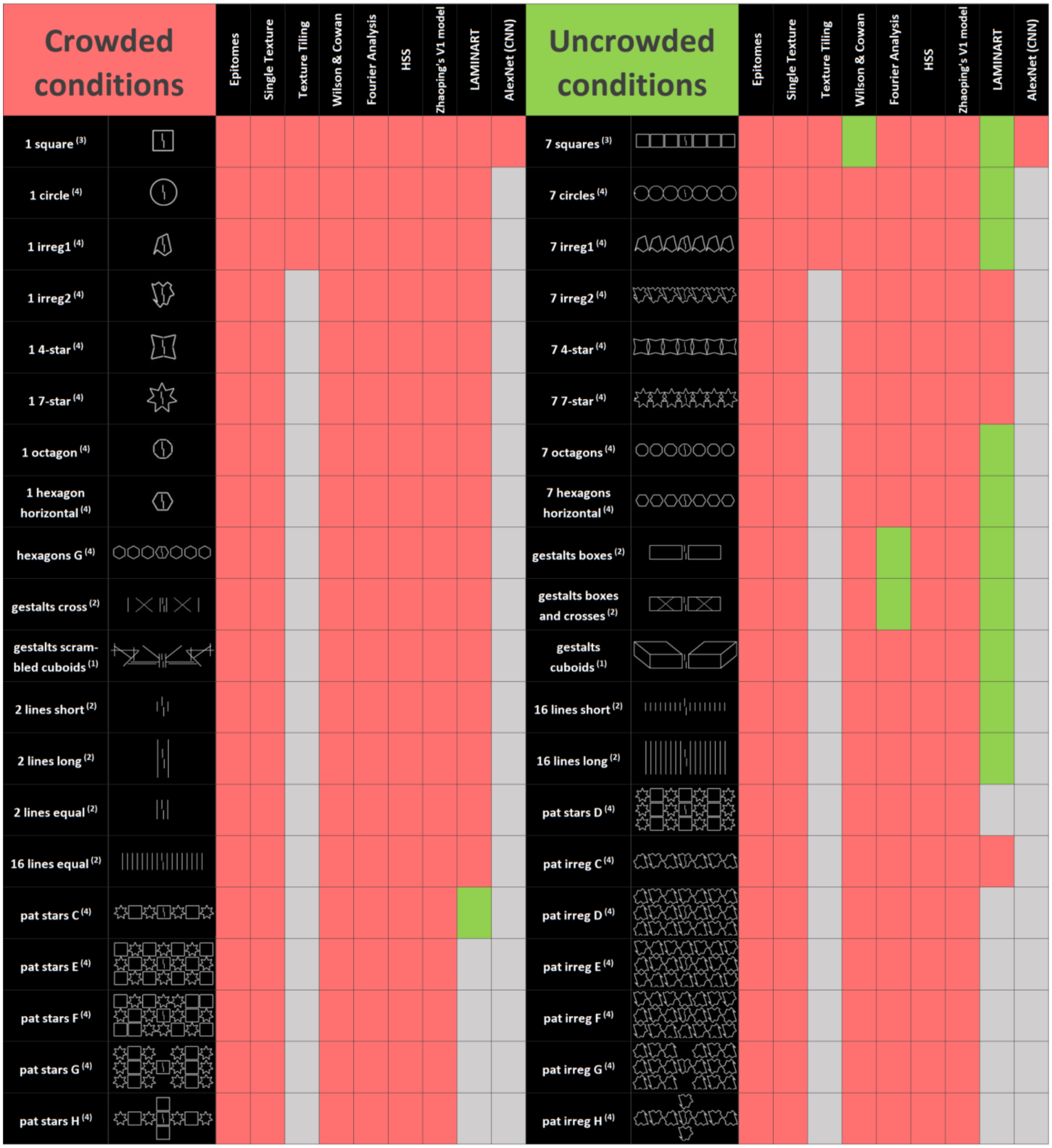

a
Stimuli
Alexnet
input conv1 conv2 conv3 conv4 conv5 fc6 fc7 fc8
or
or
R L
b
Vernier decoding from alexnet layers
vernier
1 square
7 squares
Percent correct
0
20
40
60
80
100
conv1 conv2 conv3 conv4 conv5 fc6 fc7 fc8 prob


a
Dense, large receptive fields:
Information is not lost
along the visual hierarchy
Receptive field
Receptive field used
for readout
Sparse readout:
Information is
lost by pooling
b
Threshold [arcsec]
2000
1000
0
Data
Model


Output
Fourier
transform
Input
L
R
Data
Model
Threshold (arcmin)
30
25
20
15
10
5
0



	Spatial extent		Mechanism			Organisa-tion		
	Local	Global	Pooling	Substitution	Other	Feed-forward	Recurrent	Grouping
Classic pooling	✓		✓			✓		
Classic substitution	✓			✓		✓		
Population coding	✓				✓	✓		
Epitomes	✓			✓		✓		
Single Texture Model		✓	✓			✓		
Texture Tiling	✓		✓			✓		
Deep Textures		✓	✓			✓		
Saccade-confounded Statistics		✓	✓			✓		
Wilson & Cowan-type net		✓	✓				✓	
Fourier Analysis		✓			✓	✓		
Zhaoping's V1 recurrent model		✓	✓				✓	
LAMINART		✓	✓				✓	✓
Hierarchical Sparse Selection	✓		✓				✓	
Alexnet (Convolutional neural network)	✓		✓			✓		



Crowded conditions
Epitomes
Single Texture
Texture Tiling
Wilson & Cowan
Fourier Analysis
HSS
Zhaoping's V1 model
LAMINART
AlexNet (CNN)
Uncrowded conditions
1 square (3)
1 circle (4)
1 irreg1 (4)
1 irreg2 (4)
1 4-star (4)
1 7-star (4)
1 octagon (4)
1 hexagon horizontal (4)
hexagons G (4)
gestalts cross (2)
gestalts scram-bled cuboids (1)
2 lines short (2)
2 lines long (2)
2 lines equal (2)
16 lines equal (2)
pat stars C (4)
pat stars E (4)
pat stars F (4)
pat stars G (4)
pat stars H (4)
7 squares (3)
7 circles (4)
7 irreg1 (4)
7 irreg2 (4)
7 4-star (4)
7 7-star (4)
7 octagons (4)
7 hexagons horizontal (4)
gestalts boxes (2)
gestalts boxes and crosses (2)
gestalts cuboids (1)
16 lines short (2)
16 lines long (2)
pat stars D (4)
pat irreg C (4)
pat irreg D (4)
pat irreg E (4)
pat irreg F (4)
pat irreg G (4)
pat irreg H (4)

	Captures effects of		
Model	Crowding	Uncrowding	Explicit grouping stage
Local Substitution (Strasburger et al., 1991; Ester et al., 2014)	Yes	No	No
Local pooling (Parks et al., 2001; Pelli et al., 2004)	Yes	No	No
Epitomes (Jojic et al., 2003)	Yes	No	No
Single Texture Model (Portilla & Simoncelli, 2000)	Yes	No	No
Texture Tiling Model (Freeman & Simoncelli, 2011; Rosenholtz et al., 2012)	Yes	No	No
Deep Textures (Gatys, Ecker & Bethge, 2015)	Yes	No	No
Wilson & Cowan Network (Wilson & Cowan, 1973; Clarke, unpublished)	Yes	Yes	No
Zhaoping's V1 recurrent model (Zhaoping, 1999)	Yes	No	No
LAMINART (Francis et al., 2017)	Yes	Yes	Yes
AlexNet (Convolutional Neural Network)	Yes	No	No
Hierarchical Sparse Selection (Chaney, Fischer & Whitney, 2014)	Yes	No	No
Saccade Confounded Summary Statistics (Nandy & Tjan, 2012)	Yes	No	No
Population coding (van den Berg et al., 2010, Dayan & Solomon, 2010, Harrison & Bex, 2015)	Yes	No	No
Fourier model (Manassi et al., 2015)	Yes	Yes	No