Challenges to pooling models of crowding: Implications for visual mechanisms

Ruth Rosenholtz*[1,2] rruth@mit.edu

Dian Yu[1] dyumit17@gmail.com

Shaiyan O. Keshvari[1,2] shaiyan.o.keshvari@gmail.com

[1]CSAIL, Massachusetts Institute of Technology, Cambridge, MA USA

[2]Department of Brain & Cognitive Sciences, Massachusetts Institute of Technology, Cambridge, MA USA

Keywords: peripheral vision, crowding, pooling mechanism, high-dimensional pooling models

* Corresponding author

Challenges to pooling models of crowding: Implications for visual mechanisms

Abstract

A set of phenomena, known as crowding, reveal peripheral vision's vulnerability in the face of clutter. Crowding is important both because of its ubiquity, making it relevant for many real-world tasks and stimuli, and because of the window it provides onto mechanisms of visual processing. Here we focus on models of the underlying mechanisms. This review centers on a popular class of models known as pooling models, as well as the phenomenology that appears to challenge a pooling account. Using a candidate high-dimensional pooling model, we gain intuitions about whether a pooling model suffices, and reexamine the logic behind the pooling challenges. We show that pooling mechanisms can yield substitution phenomenon and therefore predict better performance judging the properties of a set versus of a particular item. Pooling models can also exhibit some similarity effects without requiring mechanisms that pool at multiple levels of processing, and without constraining pooling to a particular perceptual group. Moreover, we argue that other similarity effects may in part be due to non-crowding influences like cueing. Unlike low-dimensional straw man pooling models, high-dimensional pooling preserves rich information about the stimulus, which may be sufficient to support high-level processing. To gain insights into the implications for pooling mechanisms, one needs a candidate high-dimensional pooling model, and cannot rely on intuitions from low-dimensional models. Furthermore, to uncover the mechanisms of crowding, experiments need to separate encoding from decision effects. While future work must quantitatively examine all of the challenges to a high-dimensional pooling account, insights from a candidate model allow us to conclude that a high-dimensional pooling mechanism remains viable as a model of the loss of information leading to crowding.

Introduction

The Puzzle of Visual Crowding

In the fovea (i.e., the central rod-free area of the retina, approximately 1.7° diameter), recognition of visual forms is relatively robust and effortless. This is not the case for the 99% of the visual field outside the fovea. It is well known that the visual system has trouble recognizing peripheral objects in the presence of nearby flanking stimuli, a phenomenon known as *crowding* (Levi D. M., 2008; Pelli & Tillman, 2008; Whitney & Levi, 2011). A classic demonstration can be seen in Figure 1. Fixating the upper cross, one can likely easily identify the isolated 'A' on the left but not the one flanked by additional letters. An observer might see that there is an 'A' in the string but not at its correct location, e.g. see the ‘A’ as to the right of ‘R’. They might not see an ‘A’ at all, or might see strange letter-like shapes made up of a mixture of parts from several letters (Lettvin, 1976). Move the neighboring letters – the “flankers” – farther from the target ‘A’, and at a certain *critical spacing* recognition is restored. The critical spacing is approximately 0.4 to 0.5 times the *eccentricity* (the distance from the center of fixation to the target) for a wide range of stimuli and tasks (Bouma, 1970; Pelli, et al., 2009; Pelli, Palomares, & Majaj, 2004). Pelli and Tillman (2008) have dubbed this *Bouma’s Law*. This roughly linear dependence on eccentricity means that moving the display closer or farther from the eyes has little effect on the critical spacing of crowding, over a wide range of viewing distances, which the reader can observe with Figure 1.

Crowding phenomena cannot be attributed to the peripheral loss of acuity (Bouma, 1970). Rather, they highlight peripheral vision’s vulnerability to the presence of clutter. Even the earliest descriptions of crowding noted unique and interesting features distinct from loss of acuity. Korte (1923) described that under conditions of crowding, firm localization of detail becomes extremely difficult. Lettvin (1976) remarked that “it is not as if these things go out of focus – but rather it’s as if somehow they lose the quality of ‘form’.” A peripherally viewed word “only seems to have a ‘statistical’ existence… [preserving] every property save that of the spatial order that would confer shape” (Lettvin, 1976).

Crowding affects many real-world visual stimuli and tasks. It is not only relevant for recognition of arrays of items such as letters. Rather, “self-crowding” can also occur, in which a single object can be sufficiently complex to be “cluttered” on its own, impairing recognition even without the presence of nominal “flankers” (Martelli, Majaj, & Pelli, 2005; Ehinger & Rosenholtz, 2016). Crowding has a far greater impact on perception than the peripheral loss of acuity or color vision, and it is the dominant difference between foveal and peripheral vision (Rosenholtz, 2016). It impacts visual search, object recognition, scene perception, perceptual grouping, shape perception, and reading (e.g. (Rosenholtz, Huang, Raj, Balas, & Ilie, 2012; Pelli & Tillman, 2008; Rosenholtz, Huang, & Ehinger, 2012; Pelli, et al., 2007)). The information that survives crowding must suffice to guide eye movements and give us a

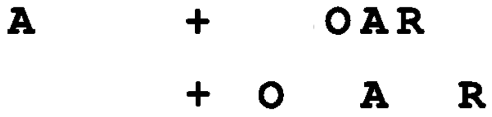

Figure 1. Visual crowding. (top) Fixate the ‘+’, and try to identify the ‘A’s on the left and right. (bottom) At a critical spacing, identification of the 'A' improves.

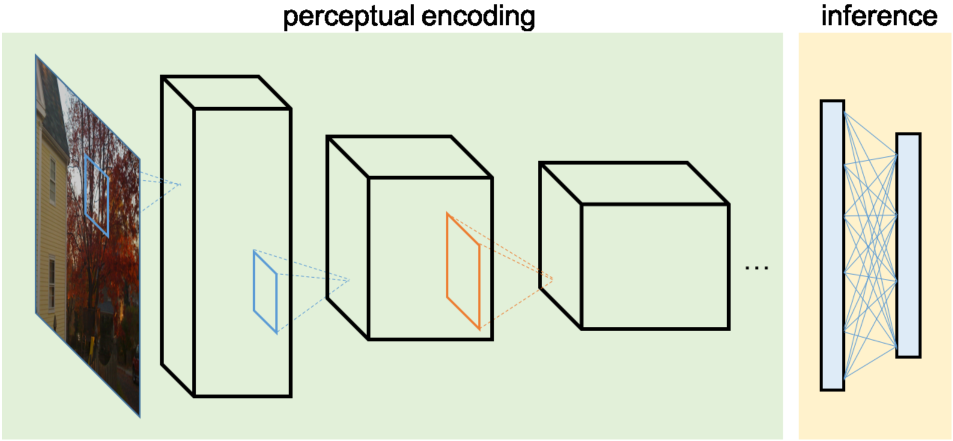

Figure 2. A candidate architecture for object recognition processing. Each visual input proceeds through a series of encoding stages that gradually advance understanding of its contents. The encoding likely favors easy performance of ecologically relevant tasks at the expense of performance at other tasks (DiCarlo & Cox, 2007). Each stage may also lose information, perhaps because of limited resources. Finally, the organism makes inferences about the visual world. The observer may have more information for some decisions than others, making some tasks inherently easier. Object recognition could be difficult because of any of these stages of processing. The goal of understanding crowding is to uncover the mechanisms particular to crowding phenomena. Standard pooling models of crowding presume that crowding results from losses at a single stage of processing, e.g. as indicated in orange.

coherent view of the visual world (Rolfs, Jonikatis, Deubel, & Cavanagh, 2011). Its pervasive effects mean that we cannot hope to understand much of vision without understanding, controlling for, or otherwise accounting for the mechanisms of visual crowding.

A challenge in understanding the mechanisms underlying crowding is distinguishing those mechanisms from the rest of object recognition processing. In crowding experiments, we present a stimulus to the experimental subject, and observe the outcome of the entire processing pipeline (Figure 2). A given condition could be difficult because of any of these stages of processing. We do not aim to elucidate the entire process for recognizing crowded objects (Tyler & Likova, 2007), nor for performing visual tasks more generally, but rather to model an important bottleneck in visual processing, and thus understand what information survives, and how that influences decision-making and predicts the difficulty of visual tasks.

A Dominant Theory of Crowding: Pooling Models

Crowding phenomenology – the jumbling, loss of location information, and the seemingly "statistical" nature of the perceived stimulus – have pointed a number of researchers towards one particular class of crowding mechanisms. Crowding has been attributed to "excessive or faulty feature integration", "compulsory averaging", or "forced texture processing" (Lettvin, 1976; Parkes, Lund, Angelucci, Solomon, & Morgan, 2001; Levi D. M., 2008; Pelli & Tillman, 2008; Balas, Nakano, & Rosenholtz, 2009), resulting from "pooling" over local regions (Parkes, Lund, Angelucci, Solomon, & Morgan, 2001; Pelli, Palomares, & Majaj, 2004; Balas, Nakano, & Rosenholtz, 2009). Pooling has typically been taken to mean averaging (Parkes, Lund, Angelucci, Solomon, & Morgan, 2001) or otherwise computing summary statistics (Lettvin, 1976; Balas, Nakano, & Rosenholtz, 2009) of features within the local region. Despite differences in terminology, these descriptions appear to refer to similar theories: "Excessive" integration – over a region beyond the bounds of the target object – can be thought of as averaging or pooling over a sizeable area of the visual field, and the operations involved in computing summary statistics are similar to mechanisms proposed to account for texture perception (see (Rosenholtz, 2014), for a review). This class of crowding model is commonly known as a **pooling model**.

A fair assessment of the current state of the field is that pooling models dominate theories of crowding. These models are often not well specified, but we can infer a few critical attributes both from implemented pooling models and from research that claims to challenge a straightforward pooling account: First, pooling regions subtend sizeable areas of the visual field and grow linearly with eccentricity (Bouma, 1970). Second, in straightforward versions of pooling models, pooling occurs on a **single processing level** with **pooling regions that are fixed in size**, rather than changing with the stimulus or task. Although peripheral object recognition no doubt requires processing at multiple levels of a visual processing pipeline, crowding models both from our own lab (e.g. (Balas, Nakano, & Rosenholtz, 2009; Rosenholtz, Huang, & Ehinger, 2012)) and from Freeman and Simoncelli (2011) explain crowding phenomena with fixed pooling at a single level (see also (Pelli, Palomares, & Majaj, 2004)). Furthermore, arguments against a straightforward pooling account have explicitly criticized its assumption of fixed pooling regions at a single processing level (e.g. (Kimchi & Pirkner, 2015; Louie, Bressler, & Whitney, 2007; Malania, Herzog, & Westheimer, 2007)). Third, **pooling regions overlap, and sparsely tile the visual field**. In other words, neighboring pooling regions of a particular class – i.e. that pool the same feature -- do not exist at every possible spatial location. Rather, while neighboring regions overlap, their centers are separated by some distance (Balas, Nakano, & Rosenholtz, 2009; Freeman & Simoncelli, 2011). Pooling over sparse, sizeable regions loses information, meaning one cannot generally reconstruct the visual input. Without sparseness and the resulting loss of information, pooling models would predict no crowding. Finally, we assume that after pooling, visual processing continues with whatever information remains.

If pooling occurs at a single level of processing, it is natural to ask at what level. Researchers have found evidence from adaptation studies that this level lies beyond V1 ((He & Cavanagh, 1996; Liu, Jiang, Sun, & He, 2009) although see (Nandy & Tjan, 2012)), and have argued that pooling occurs

shortly after early feature detection, in some sort of "feature integration" stage (Pelli, Palomares, & Majaj, 2004; Pelli & Tillman, 2008). A number of implemented pooling models of crowding either explicitly or implicitly (through their choice of mechanisms) assume that pooling occurs after V1 (Balas, Nakano, & Rosenholtz, 2009; Freeman & Simoncelli, 2011; Parkes, Lund, Angelucci, Solomon, & Morgan, 2001; van den Berg, Johnson, Anton, Schepers, & Cornelissen, 2012).

Crowding impairs many visual tasks, and yet peripheral vision supports a rich percept of the visual world. In order for a pooling model to be viable it must predict both the limitations and the capabilities of visual perception. As a result, Rosenholtz (2014) has argued that we must make two additional assumptions. First, a pooling model must **pool a large number of features**, meaning the mechanism must involve a large number of populations of "receptive fields" (say, on the order of 1000), with each population pooling a different feature. Second, Rosenholtz argues (and for the purposes of this paper we assume), that **the mechanism pools image features**. In other words, it pools the outputs of filtering operations plus non-linearities, as opposed to averaging the features of individuated *items*. An item-based model might, for example, extract the orientation of each bar in an array, and average those orientations. Using object features can simplify modeling; e.g. one can more easily construct an ideal observer for observations consisting of a discrete set of item features (Parkes, Lund, Angelucci, Solomon, & Morgan, 2001; van den Berg, Johnson, Anton, Schepers, & Cornelissen, 2012), than for continuous outputs of image processing operations. However, while one can certainly make interesting progress by studying object-based models (van den Berg, Johnson, Anton, Schepers, & Cornelissen, 2012), ultimately there are limits to the generalizability of such models (Keshvari & Rosenholtz, 2016). To distinguish the class of models that pool large numbers of image features from a pervasive "simple pooling model" (e.g. Parkes et al. (2001); Greenwood et al. (2009); Greenwood et al. (2012); Levi and Carney (2009)) that pools small numbers of features and/or item features, we call the former an **HD (high-dimensional) pooling model**.

Work from Rosenholtz and colleagues has developed and tested an HD pooling model that we call the Texture Tiling Model (TTM).[2] The model consists of two stages. In the first stage, TTM implements a V1-like representation consisting of responses to oriented, multiscale feature detectors. In the second stage, the model computes a large set of second-order correlations from the responses of that first stage, taking the average over local "pooling regions" (TTM also computes more basic 1st-order summary statistics within each color band) (Balas, Nakano, & Rosenholtz, 2009). These pooling regions grow linearly with eccentricity, in accord with Bouma's Law, and overlap and tile the visual field. The information encoded in the second stage, where pooling happens, has been associated with the information encoded

[2] Details about the Texture Tiling Model and relevant discussion can be found in Appendix A.

physiologically, post-V1 (e.g. (Freeman, et al., 2013; Yamins & DiCarlo, 2016)). In addition, standard models of hierarchical visual processing (e.g. (Fukushima, 1980; Riesenhuber & Poggio, 1999), often have as a second stage the computation of co-occurrence of combinations of features from the first stage; 2nd-order correlations are merely co-occurrence computations pooled over significantly larger regions. The set of statistics we measure are those identified by Portilla and Simoncelli (2000), because that set has been successful at capturing the appearance of textures for human perception. Specifically, textures synthesized using this set of statistics are often difficult to discriminate from the original (Balas, 2006). Mounting evidence supports TTM as a good candidate HD pooling model for the peripheral encoding underlying crowding. We have shown that it predicts performance at a range of peripheral recognition tasks involving arrays of letters and other symbols (Balas, Nakano, & Rosenholtz, 2009; Rosenholtz, Huang, & Ehinger, 2012; Rosenholtz, Huang, Raj, Balas, & Ilie, 2012; Keshvari & Rosenholtz, 2016). The same model predicts the difficulty getting the gist of a scene when fixating – i.e. when forced to use extrafoveal vision – compared to when free-viewing that scene (Ehinger & Rosenholtz, 2016). With the same image statistics though somewhat different arrangement of pooling regions, Freeman and Simoncelli (2011) have predicted the critical spacing of crowding. They also showed that equating those local summary statistics creates synthetic "metamer" images difficult to distinguish one from another when viewed with the same fixation as used by the model (though see (Wallis, Bethge, & Wichmann, 2016)). While in all of these studies there has remained variance unexplained by the model, and thus room for improvement, these HD pooling models have so far proven quite powerful at capturing crowding and related visual phenomena.

Challenges to a Pooling Account of Crowding

In spite of the success of HD pooling models, however, questions remain. Behavioral researchers have made considerable progress understanding crowding in the last 1.5 decades. They have substantially expanded crowding phenomenology to a wider array of stimuli and tasks; moving well beyond arrays of letters and Gabors, to include stimuli with higher-level grouping effects, and tasks with complex naturalistic stimuli. As researchers have studied a wider range of stimuli and tasks, a complex pattern of results has emerged. A number of challenges have arisen to the relatively simple pooling model, and it has seemed that a single unifying explanation might not suffice. Researchers have called into question virtually every feature of pooling models highlighted above, and instead proposed that more complex, and often more dynamic models may be necessary. This review paper centers on these challenges to pooling models.

Some behavioral results have seemed to favor a different type of mechanism entirely, for example the substitution mechanisms described in more detail in the next section (Strasburger, 2005; van den

Berg, Johnson, Anton, Schepers, & Cornelissen, 2012). Other results have seemed to suggest that information is not lost, as it would be by a pooling mechanism, but rather rendered unavailable for object recognition (Chaney, Fischer, & Whitney, 2014; Yeh, He, & Cavanagh, 2012). Finally, other results have appeared to point to an attentional rather than a pooling mechanism (Intriligator & Cavanagh, 2001).

Other challenges have pointed to pooling operating at a different level of processing. For instance, (Levi & Carney, 2009) suggested that pooling might follow segmentation into objects, perhaps also implying the pooling of object features rather than image features. Or perhaps multiple bottlenecks limit peripheral processing, rather than a single bottleneck. Researchers have suggested that crowding mechanisms might operate at multiple levels of processing, rather than at a single feature-integration stage (Louie, Bressler, & Whitney, 2007; Farzin, Rivera, & Whitney, 2009; Ikeda, Watanabe, & Cavanagh, 2013; Kimchi & Pirkner, 2015).

Finally, other challenges have suggested that pooling regions, rather than being fixed, instead vary with the stimulus and task (Banks & White, 1984; Bernard & Chung, 2011; Kimchi & Pirkner, 2015; Livne & Sagi, 2007; M. Manassi, Sayim, & Herzog, 2012; Mauro Manassi, Lonchampt, Clarke, & Herzog, 2016; Rosen & Pelli, 2015). In the most popular version of this suggestion, pooling occurs only within perceptual groups (Banks & White, 1984; M. Manassi et al., 2012).

We refer to the set of phenomena challenging a unified pooling model account the *model challenges*. We discuss each of these in more detail in the following sections. It is natural to think that this complex set of phenomena rules out a unifying pooling model explanation. However, upon closer consideration, this may not be the case. We will argue that pooling models remain viable, in spite of numerous challenges.

CHALLENGE 1: The Mechanism of Crowding is Not Pooling, But Substitution

An observer, faced with an array of items such as letters, and asked to report the identity of a target, may instead report one of its flankers. Such "substitution" phenomena are well known and well documented (Huckauf & Heller, 2002; Strasburger, 2005; Poder & Wagemans, 2007). These phenomena at first glance appear to challenge a simple pooling model. Why, if one encodes the average feature, would one report the features of the flanker rather than the target? Relatedly, other researchers have observed that, "…none of our participants ever spontaneously reported seeing [the mean]. This argues against… averaging and [in favor of] an inability to accurately localize features" (van den Berg, Roerdink, & Cornelissen, 2007). While in subsequent work (van den Berg, Roerdink, & Cornelissen, 2010; van den Berg, Johnson, Anton, Schepers, & Cornelissen, 2012) these authors are firmly in favor of pooling models (which they refer to as "integration" models), below we revisit this observation in order to clarify intuitions about what pooling models predict.

Some researchers have further suggested that substitution *phenomena* might arise from a substitution *mechanism*. For example, the visual system might measure the features and possibly even the identities of both target and flankers, but either not encode their locations at all or encode them in a noisy way (Chung & Legge, 2009; Strasburger & Malania, 2013; van den Berg, Johnson, Anton, Schepers, & Cornelissen, 2012). The loss of location information would predict substitution errors, as the observer would accidentally report the incorrect item.

A substitution mechanism would immediately have consequences for set perception. If peripheral vision preserves the identities of the display items, but not their locations, this would make reporting set properties such as the mean orientation easier than recognizing the features of a particular target item. This prediction agrees with behavioral results (Parkes, Lund, Angelucci, Solomon, & Morgan, 2001; Fischer & Whitney, 2011). Fisher and Whitney (2011), for instance, show subjects a peripheral array of faces, and ask for both the facial expression of the central target and for the mean expression of the set. They found that even though subjects have trouble reporting the target facial expression, that expression nonetheless contributes to judgments of the mean expression. They argue that these results are incompatible with a pooling mechanism, reasoning that if pooling loses information about that central target, the information cannot also be available to contribute to perception of the mean.

An HD Pooling Model Can Predict Substitution Behavior and Good Set Perception

We argue that these apparent challenges to a pooling model arise from misunderstanding high dimensional pooling. It is difficult to reason about an HD pooling model, particularly one that measures image rather than object features. Researchers have instead attempted to gain intuitions from lower-dimensional models. However, an HD pooling model will behave fundamentally differently from its low-dimensional brethren.

To get intuitions about the information preserved and lost by an HD pooling model, we can generate members of the equivalence class of the model, i.e. images that are confusable with the original, according to the model. Rosenholtz and colleagues have called these images “mongrels”. Consider the example in Figure 3. Loosely speaking, information that appears clear and unambiguous in these mongrels corresponds to information that survives HD pooling. Tasks that appear easy to perform with these visualizations are predicted by an HD pooling model to be easy tasks. For all of the examples in this paper we generated at least 10 mongrels, and present 1-2 “typical” ones. For example, if the original task were to judge the orientation of a crowded peripheral target, we rank-ordered the mongrels according to our subjective assessment of the quality of the information available to perform that task, and selected mongrels of median quality. To get a sense of the variability among the mongrels for a single input image, one can view the full set of mongrels generated for Figures 3, 4, 7, and 16 in the

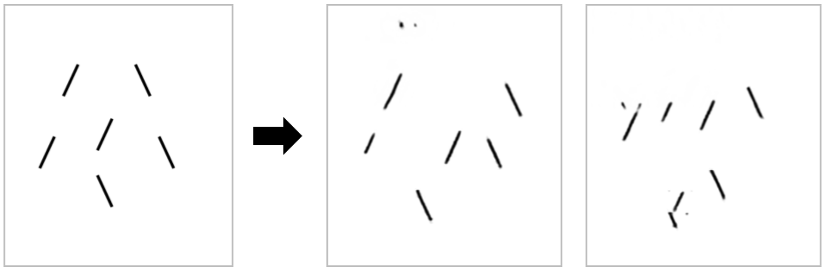

Figure 3. Original image (left) and two members of the equivalence class of our candidate high-dimensional pooling model, TTM (right). The array is 3.6 deg in diameter, each item is 1 deg in length, and the fixation (not shown) is modeled at 10 deg to the right of the central target. An HD pooling model encodes a great deal of information about the stimulus, yet also loses information, introducing uncertainty about the location and number of items.

supplementary material at https://dspace.mit.edu/handle/1721.1/121152. One can also find at that link the code for generating mongrels.

First and foremost, note that the encoding captures a great deal of information about the appearance of the stimulus. Sufficient information survives pooling to determine that the input consists of black lines against a white background. The model was not told anything about oriented black lines, but enough information survives pooling for later processes to get that gist.

Second, note the loss of location information. If the task were to discriminate the orientation of the central bar, observers might have trouble reporting that orientation rather than the orientation of one of the flankers. We can immediately see that a pooling model can predict substitution phenomena, at least qualitatively (Keshvari & Rosenholtz, 2016; Harrison & Bex, 2015).

To what extent swap errors occur in practice likely depends in part on the task, with different answers possible even with the same set of stimuli. Harrison and Bex (2015) found a low number of swap errors, and suggested those errors could be predicted by their population code model. (Note that HD pooling models can also be considered population code models.) Agaoglu and Chung (2016) found that with the same stimuli (concentric "C"s), observers made considerably more swap errors, i.e. reporting the gap location from one C when asked about the gap location for a different C. The high number of swap errors in Agaoglu and Chung (2016) likely arise from their complicated task, which requires observers to report both ring orientations, with the order of the report varying from trial to trial and post-cued. Harrison and Bex's experiments did not have this additional source of uncertainty. Additionally, for these stimuli it remains unclear how many swap errors TTM predicts. Agaoglu and Chung tested an early version of TTM, which used only a single pooling region and seeded synthesis with a blurry version of the original image. They showed that the "texture synthesis [sic[3]] model", rarely

[3] Models like TTM are often referred to as "texture synthesis" models, but this is a misnomer. We do not assume that the visual system "synthesizes" any textures. The texture synthesis is used only to generate members of the equivalence class of the model. It is not part of the model, per se.

produces such substitution errors and concluded that TTM cannot predict the magnitude of substitution effects. However, it is worth noting that we originally used the blurry seed to *reduce* location uncertainty, as the single pooling region version of the model preserves *no absolute location information* (Balas, Nakano, & Rosenholtz, 2009). We no longer use this technique in the full version of TTM with multiple pooling regions. It is an open question whether TTM predicts more swap errors for these stimuli.

The third thing we can note from Figure 3 is that, for this simple display, pooling preserves enough information to determine the distribution of orientations fairly accurately. An observer, asked to report the orientation of the central target, would have no reason to report it as vertical, as the representation makes clear that the stimulus contains no vertical lines. If this were the information available in the periphery, observers should be good at reporting all sorts of set properties. Once pooling loses location information, set perception becomes inherently easier than reporting a particular item. The visual system has information about the set, but information about a given item becomes “inaccessible” in the sense of it being difficult to determine which item is the desired target.

This general logic generalizes to other kinds of set perception, such as the mean facial expression (Figure 4). HD pooling has no difficulty predicting that an incorrectly identified target can contribute to the perceived mean. For such complex stimuli as those of Fisher and Whitney (2011), the more pressing question is whether our particular HD pooling model preserves enough information to predict judgments of mean facial expression at human accuracy levels. These mongrels suggest that HD pooling "render[s] further object processing" difficult, but not "impossible" (Fischer & Whitney, 2011). Clearly set perception performance, according to an HD pooling model, depends upon the complexity of the individual items and of the display.

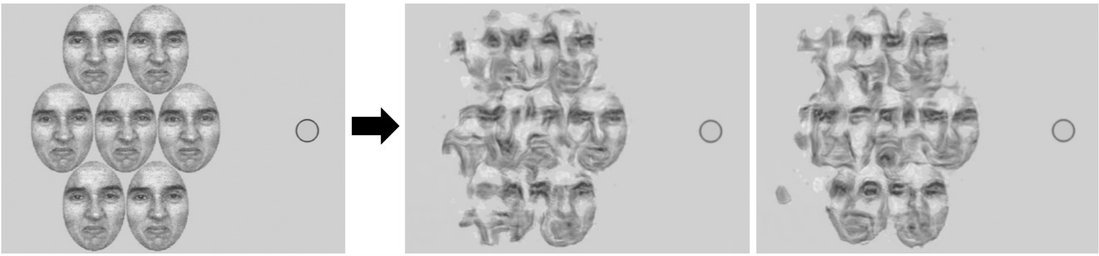

Figure 4. (left) original image from Fisher and Whitney (2011), reprinted with permission, and (right) two typical mongrels. Fixation and a 1 degree “fovea” indicated by the circle. The faces used in the emotional morph were drawn from Ekman’s Pictures of Facial Affect (POFA) collection and are reprinted in modified form with permission from the Paul Ekman Group.

Substitution phenomena do not eliminate pooling models. One might ask why, then, studies ever find averaging effects, e.g. (Parkes, Lund, Angelucci, Solomon, & Morgan, 2001; Greenwood, Bex, & Dakin, 2009). There are two reasons. First, when target and flanker features are sufficiently similar, representation of those features can become poorer. See, for example, representation of the similar orientations in Figure 7. The orientation of the central target appears to be a mix of that of the target and the flankers. Under such conditions, one might imagine that the observer would report something like the mean feature, and they do (Greenwood, Bex, & Dakin, 2009).

Second, observers may base their decision upon the average because doing so is a good strategy for a given task. In their seminal paper, Parkes et al (2001) showed observers one or more "target" Gabors, tilted clockwise or counterclockwise, and observers had to report the direction. All targets on a given trial had the same tilt. In one condition, *n* targets were present with no distractors. In the other condition, nine items appeared, including *n* targets and *9-n* distractor Gabors with zero tilt. Parkes et al. found that they could better fit the results with a model in which observers based their decision on the mean of the noisy orientation observations than by a "signal detection theory" model in which the observer retained observations for every Gabor. One might be tempted to conclude that the observer only has access to the mean ("compulsory averaging"). However, it turns out that responding based upon the mean orientation yields better results than the supposed signal detection theory (SDT) strategy described by

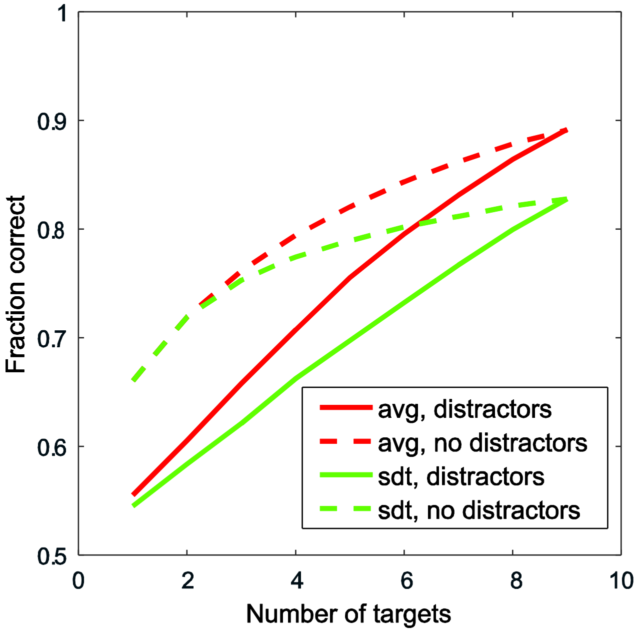

Figure 5. Performance of "compulsory averaging" and SDT models with the same internal noise. Solid curves show predicted performance for the Parkes et al. (2001) condition in which stimuli contained both tilted targets and non-tilted distractors. Dashed curves show predicted performance for the Parkes et al. condition in which stimuli contained only tilted targets. In both conditions the target tilt was 12.5 degrees, and the internal noise was Gaussian distributed with a standard deviation of 30 degrees. Performance is always the same or better for the averaging model (red curves) than for the SDT model (green curves). Rather than compulsory, averaging may be a good strategy for these particular conditions.

Morgan and Solomon (2005)[4]. Their SDT strategy chooses the tilt direction by finding the maximum observation, M, (the most clockwise observation), and the minimum observation, m. The model chooses the clockwise response if |M| > |m|, counterclockwise if not. This strategy is not ideal for a threshold experiment in which multiple targets deviate by the same amount (Ma & Huang, 2009). Intuitively, as the number of targets increases, the SDT model cannot make use of the additional information to estimate the tilt. On the other hand, the averaging model, which chooses "clockwise" if the average of the noisy observations is greater than 0, does make use of information from additional targets. Figure 5 shows for both models the predicted fraction correct as a function of the number of number of targets, given the same internal noise. The pattern of results is the same regardless of the internal noise (excluding extremes in which predicted performance is at ceiling or at floor for both models). An observer following the SDT strategy never performs better than one making their decision based on the average, for the Parkes et al. tasks. An optimal model, with access to the full but noisy distribution of orientations, would behave as if it had access only to the mean.

Much has been made of the apparent dichotomy between substitution and averaging behavior (Ester, Klee, & Awh, 2014; Freeman, Chakravarthi, & Pelli, 2012; Greenwood, Bex, & Dakin, 2009). However, pooling models can produce substitution phenomena, and models capable of producing substitution phenomena, in turn, should sometimes produce averaging behavior.

An HD Pooling Model Does Not Behave Like a Slots Model

One way of implementing a substitution model would be to have *n* slots, one for each item in the stimulus and its features, akin to slot models of short-term memory (Zhang & Luck, 2008). Entire items might swap between slots, leading to classic substitution effects, or features might swap between slots, leading to more complex sorts of confusions.

The previous subsection discussed the fact that a pooling model can produce substitution-like confusions. However, it should also be clear that a pooling model does not behave like a slots model. As we can see in Figure 3, pooling does not simply make items or their features swap positions with each other, but rather makes features and their locations ambiguous and confusable. Pooling can even produce ambiguity about the number of items present. Put another way, an HD pooling model operates on image features, not a list of items and their features; nor is the information available at the output of a pooling model simply a list of items and their features.

Intriligator and Cavanagh (2001) attempted to distinguish between a pooling vs. an attentional selection mechanism for crowding. We discuss their work as part of the substitution-related challenges

[4] The SDT model in (Morgan & Solomon, 2005) corrects an important error in the original (Parkes, Lund, Angelucci, Solomon, & Morgan, 2001) report. Correcting this modeling error does not appreciably change the conclusions of Parkes et al (2001).

because they presume that a pooling model acts like a slots-based substitution model. They cued one of a number of identical disks (Figure 6A), then instructed the observer to move the focus of their attention from item to item in a prescribed way ("left-right-right-left-right…"). Then they asked the observer to identify the item indicated by the cue plus series of instructions. They varied item spacing until observers reached threshold performance levels. The resulting critical spacing was similar to that of crowding. Thus, they suggested that crowding arises from inability to selectively attend to the target. Of interest here is their claim that "mixing of adjacent features… cannot contribute to the critical spacing in our [study] where target features and identity are irrelevant." Their reasoning seems to presume that vision has some number of slots, each containing the features of one item; feature pooling mixes up the contents of the slots – irrelevant in this case since the slots contain identical items – but not the slots themselves. As a result, they reason, their results cannot be due to pooling.

In Figure 6B we have overlaid a pair of mongrels for Figure 6A, with red indicating regions that were black in one mongrel but not the other. The difference between the two mongrels demonstrates the position uncertainty inherent to the model. One might imagine that this degree of uncertainty would make attentionally tracking the target difficult. Because an HD pooling model does predict negative effects of crowding even when the display items are identical, one cannot rule it out as an explanation for the results of Intriligator and Cavanagh (2001).

CHALLENGE 2: Crowding Arises From Multilevel Pooling

Straightforward versions of pooling models presume that crowding arises from pooling at a single critical level of processing. Some recent theories have suggested instead that crowding might involve pooling at multiple levels of processing. In addition to pooling soon after early feature detection, there might also exist crowding mechanisms at the part processing level or shape processing level (Kimchi & Pirkner, 2015), the face processing level (Louie, Bressler, & Whitney, 2007; Farzin, Rivera, &

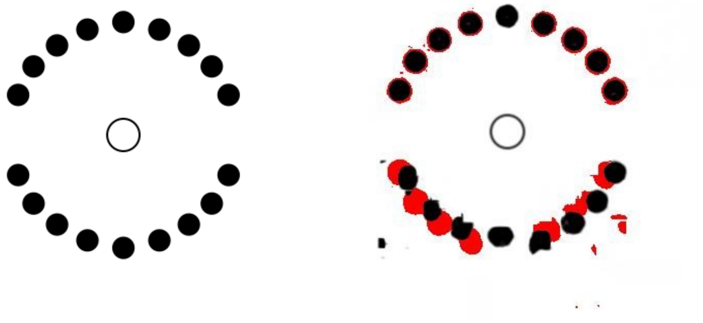

Figure 6. (left) original image based on stimuli from Intriligator and Cavanagh (2001), and (right) two typical mongrels overlaid, with red indicating regions that were black in one mongrel but not the other. Fixation and a 1 degree "fovea" indicated by the circle.

Whitney, 2009), and the point-light-walker processing level (Ikeda, Watanabe, & Cavanagh, 2013). This is an attractive and rather intuitive idea, at first glance, because it parallels standard models of object recognition, e.g. (Fukushima, 1980; Riesenhuber & Poggio, 1999). Standard hierarchical models of vision do involve multiple levels of processing, with some sort of pooling or integration at each level. However while object recognition models do usually alternate between filtering and pooling stages, the pooling typically occurs over an area not much larger than the size of the filter (Riesenhuber & Poggio, 1999; Krizhevsky, Sutskever, & Hinton, 2012), and thus might not cause crowding, per se.

There is an immediate reason to be concerned about the suggestion that models of crowding need to incorporate multiple levels of pooling: Our candidate HD pooling model already loses a lot of information through pooling at a single level, and yet quantitatively predicts a number of phenomena. Our modeling thus far has not suggested the need to pool at additional levels, in spite of testing a relatively large variety of tasks and stimuli. If one were to pool at additional levels of processing, additional information would be lost, perhaps reducing the predictive power of the model. (Note that this particular criticism would not apply to the hierarchical model of Chaney et al. (2014). Rather than losing information at each level of processing, their model predominantly loses information at the decision stage, through a mechanism that can only access a sparse sampling of receptive fields in order to make a decision.) Here we reassess the multilevel pooling claims both on a theoretical level and using intuitions from our HD pooling model.

Many claims of multilevel pooling involve similarity effects, in which it is easier to identify a target flanked by dissimilar items than when it is flanked by similar items. Such effects are prevalent in the crowding literature. Identifying a target letter is easier when it “pops out” from the flanker letters due to a difference in color (Scolari, Kohnen, Barton, & Awh, 2007), and similar effects have been found for large target-flanker differences in orientation (Andriessen & Bouma, 1976), contrast polarity, shape, and binocular disparity, but not eye of origin (Kooi, Toet, Tripathy, & Levi, 1994).

One theory about why these similarity effects occur is that pooling operates only within a feature band. (Alternatively, *inhibition* might occur only within a band (Andriessen & Bouma, 1976; Levi, Hariharan, & Klein, 2002; Kooi, Toet, Tripathy, & Levi, 1994). This distinction is not critical for the present discussion.) According to this theory, if the stimulus has similar flankers, pooling mixes them in with the target, leading to crowding. Dissimilar flankers do not mix with the target, leading to a release from crowding. If this theory is correct, then a seductive corollary would seem to be that one can figure out where in visual processing pooling (i.e. crowding) occurs by looking at what kind of similarity effects one finds. If one finds a shape similarity effect, then pooling must be at the shape processing stage. This suggestion that one might uncover brain mechanisms through simple psychophysical experiments, coupled with standard models of hierarchical processing, makes multi-level crowding doubly attractive as a theory.

There are problems, however, with interpreting similarity effects in terms of the level of pooling. First, the theory presumes that recognition operates by having a receptive field tuned to the target object, reminiscent of "grandmother cells", and that crowding arises because flankers lie within that receptive field, disrupting identification. But, particularly at higher levels of processing, the encoding is likely more distributed. In a distributed encoding scheme, target identification arises through combining information from multiple feature detectors, rather than from the response of a single "band" sensitive only to the target. Even in low-level vision, information from multiple receptive fields in V1, tuned to different orientations, *combines* to identify the underlying orientation.

Second, before using a similarity effect to reason about the level at which pooling occurs, one needs to confirm the level of (dis)similarity of target and flankers. Confirming that it is high-level rather than low-level, or, equivalently, controlling low-level similarity while varying high-level similarity, is notoriously difficult. Higher-level classifications typically derive at least in part because of shared lower-level features.

Both of these theoretical arguments suggest that there may be a mismatch between the apparent level of similarity between target and distractors and the pooling level that produces the effect. In fact, we see evidence of just that. Let us look at some mongrels to get intuitions about what an HD pooling model predicts. Our candidate HD pooling model largely pools pairwise combinations of responses of V1-like orientation detectors. (Notable exceptions are its computation of marginal statistics of luminance and color.) One might think of it as pooling at the "junction-processing" level. Figure 7 shows a similar (top) and dissimilar orientation condition. The slightly tilted target line in the center is the same in both cases. On the right are two mongrels for each condition. What one should look for is how well these mongrels preserve the tilted target, and in particular how well one could judge its

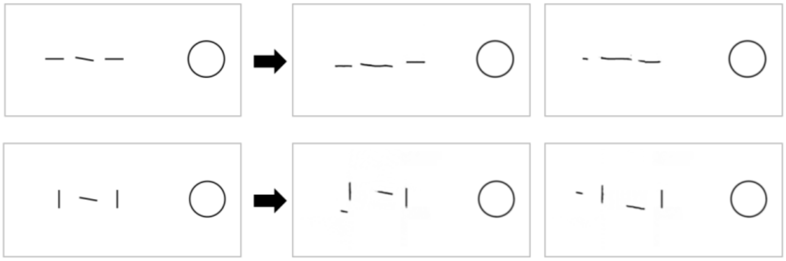

Figure 7. Orientation similarity effect. Determining the tilt of the central target is easier when flankers have dissimilar orientation (bottom) vs. similar (top). In each row, the two images on the right show two mongrels, visualizations of the information available according to our HD pooling model. The target and its orientation are clearer in the two mongrels for the dissimilar condition; HD pooling better encodes the target in that condition. HD pooling predicts this orientation similarity effect without pooling at an orientation processing stage.

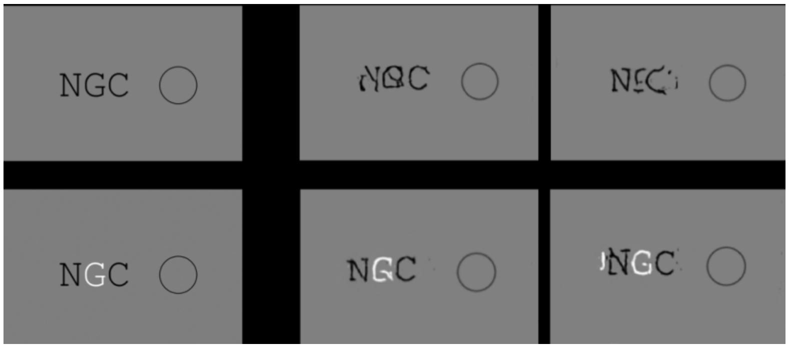

Figure 8. Sign-of-contrast similarity effect. Identifying the central letter is easier when flankers have opposite sign of contrast (bottom). In each row, the two images on the right show two mongrels, visualizations of the information available according to our HD pooling model. The target's identity is clearer in the two mongrels for the dissimilar condition. HD pooling better encodes the target in that condition. HD pooling predicts a sign-of-contrast similarity effect without pooling at a sign-of-contrast processing stage.

orientation. One can easily observe the more faithful representation of the target in the condition with dissimilar flankers. Our HD pooling model can at least qualitatively predict an orientation similarity effect (and in some cases quantitatively; see (Keshvari & Rosenholtz, 2016)), and yet the model has no pooling at the "orientation processing" stage. There is a mismatch between level of similarity and level of pooling.

For another example, consider the similarity effect in Figure 8, based on sign of contrast. It is easier to recognize the target (G) when it has a different sign of contrast than the flankers (Kooi, Toet, Tripathy, & Levi, 1994). Looking at the mongrels, the G shape is quite well preserved in the dissimilar condition. Our HD pooling model predicts, at least qualitatively, a sign-of-contrast similarity effect, and yet the model does not pool at a sign-of-contrast stage. There is a mismatch between the level of similarity and the level of pooling that produces the observed similarity effect.

Finally, consider the shape similarity effect from Kimchi and Pirkner (2015). They flanked a target square composed of L-junctions with a variety of flankers. Figure 9 gives two key examples. In the first condition, flankers have the same overall shape as the target, but consist of nominally different "parts": straight lines instead of L-junctions. In the second condition, flankers consist of the same L "parts" as the target, but those parts form different shapes. The observer indicated the orientation of the target, i.e. whether it appeared as a diamond or a square. At the eccentricity shown, it is easier to recognize the square target when flanked by dissimilar shapes than when flanked by similar shapes. Again, the mongrels indicate that our HD pooling model better encodes the target in the dissimilar condition than in the similar condition; in other words, the HD pooling model predicts the shape similarity effect

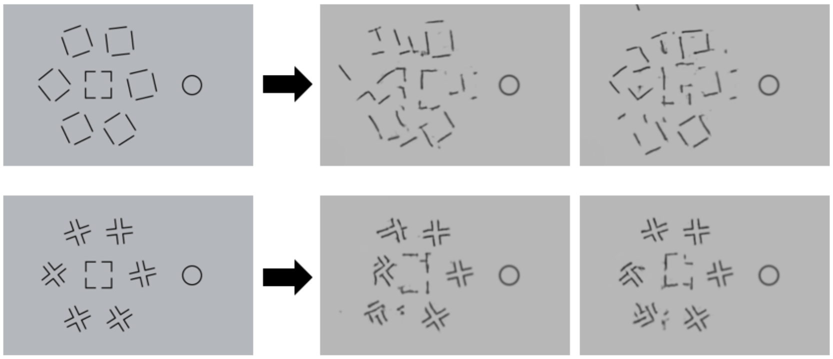

Figure 9. Kimchi and Pirkner (2015) found crowding both when a central square formed of L-junctions was flanked by squares formed without L-junctions (same shape, different parts, top left) and when it was flanked by L-junctions that did not form a square (same parts, different shape, bottom left). However, this does not imply that crowding must happen at both the parts and shape levels of processing. Mongrels on the right of the arrows show signs of crowding for both conditions. Crowding seems worse for the same shape flankers, in agreement with experimental results for the modeled eccentricity of 5 degrees. Original stimuli reprinted with permission.

without any pooling at a shape-processing stage. Again, there is a mismatch between level of similarity and level of pooling.

Several papers from Whitney and colleagues made a somewhat more complicated argument for crowding at a holistic face processing level. Louie, Bressler, and Whitney (2007) asked observers whether a target face appeared on the left or right side of the display, or not at all (Figure 10 shows a single side of the display). They found greater crowding when upright faces flanked the target than when inverted faces did. On the other hand, they found no effect of upright vs. inverted flankers when the task was instead to detect a target house among house flankers (Figure 10). They argued that the difference between faces and houses may derive from holistic processing of faces, versus part-based processing of houses. These results by themselves could arise from a relatively low-level similarity effect. Faces may show an effect of upright vs. inverted flankers, and houses not show such an asymmetry, simply because there is less difference between an upright and inverted house, compared to an upright and inverted face. Cropping the house stimuli further increases the similarity between upright and inverted houses, by eliminating distinguishing roof features. On the other hand, it does not seem obvious that TTM, per se, can predict the similarity effect. From the example mongrels in Figure 10, it appears difficult to identify the face or house in either condition, whereas in the original study observers had a d' in the range of 2.5 to 3 for the upright face targets, and d' around 3.5 for upright house targets. More quantitative study is needed, but we may find that TTM lacks necessary features to predict this relatively good performance. However the additional features that may be required are not obviously at the holistic face processing level.

Louie et al. (2007), however, additionally argue for involvement of holistic face processing based on the results of repeating the two face conditions, but with the entire display inverted. The target faces then appeared upside down. They found that this manipulation eliminated the asymmetry between upright and inverted flankers. They argued that flipping the stimuli should not change low-level similarity. They concluded that their results arise from holistic face processing, and that crowding can occur at the face-processing stage.

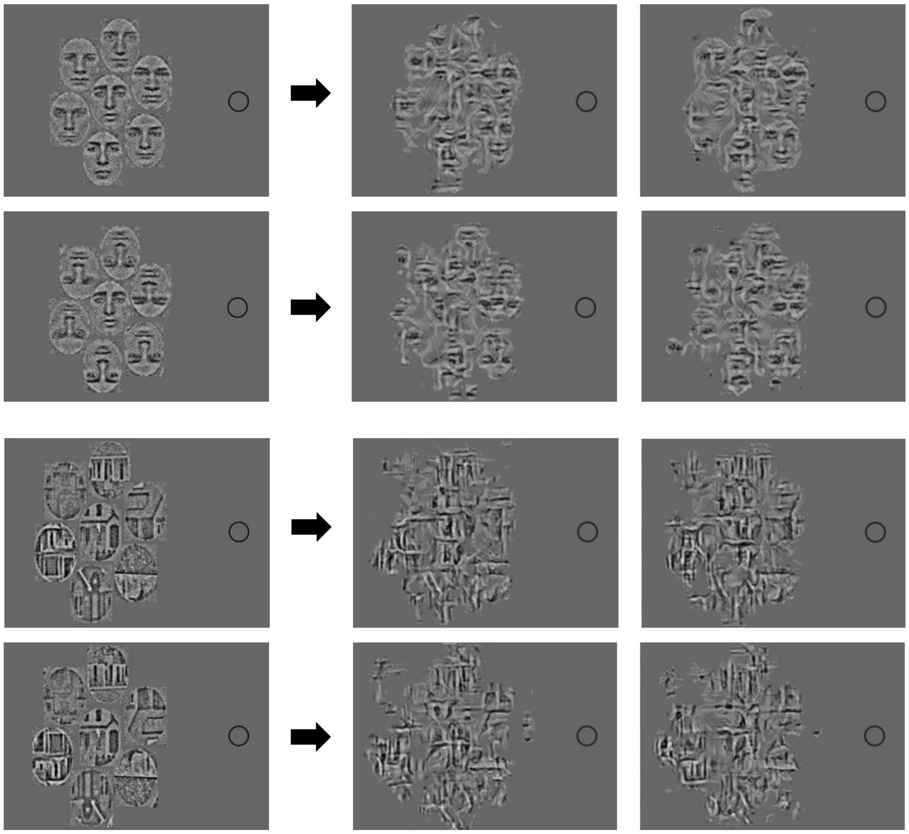

Figure 10. Original stimuli (left of the arrows) and their mongrels (right of the arrows) from Louie et. al. (2007). The first two rows show target face surrounded by upright (1st row) versus inverted (2nd row) flankers. The last two rows show target house surrounded by upright (3rd row) versus inverted (4th row) flankers. Original stimuli adapted from Louie et al. (2007), reprinted with permission.

We agree that inverting the entire display should have minimal effect on low-level similarity, but dispute that their results necessitate that crowding operates at the holistic face processing level. First, it is notable that Sun and Balas (2015) did not replicate the Louie et al. (2007) effect of inverting the entire display. Sun and Balas asked observers to categorize the gender of a target face. They found crowding for a target flanked by face-like stimuli like line drawings and U.S. electrical sockets. Unlike Louie et al., Sun and Balas did find that the upright/inverted flanker asymmetry reversed when they inverted the target face, consistent with a low-level similarity effect.

Second, arguably Louie et al. found no difference between upright and inverted flankers in their inverted face conditions because performance was near floor. Along these lines, Kalpadakis-Smith, Goffaux, and Greenwood (2018) systematically investigated the influence of task difficulty on face crowding. They asked observers to identify the horizontal separation between the eyes. They found a similarity effect for both an upright and inverted target face when the task was easy (large differences in interocular distance). However, when the task was difficult (small differences in interocular distance), they found no similarity effect for either upright or inverted targets. Whitney and colleagues have argued, however, that the lack of a holistic processing pattern of results may arise from using a non-holistic face task (Manassi & Whitney, 2018). However, the same group of researchers showed that discriminating small differences in the horizontal separation between the eyes of a single face was indeed easier when the face was upright compared to inverted, following the pattern of a holistic task. Nevertheless, they found no difference between upright and inverted flankers when observers performed this task on a crowded face in their periphery (Kalpadakis-Smith, Goffaux, & Greenwood, 2018).

Holistic processing may well be involved at a later stage; it operates on the information that survives crowding to produce better performance identifying an upright face than an inverted face. The lack of holistic processing for Louie et al.'s (2007) inverted face conditions likely led to the near-floor performance for inverted targets. However, we would not call this a "crowding" mechanism, per se, as performance is better for upright faces even in the fovea. (Interestingly, visual search for a cube among differently lit cubes also has an asymmetry that does not persist when inverting the entire display. In that case, we similarly argued for a later loss of information due to estimating 3D shape while discounting illumination (Zhang, Huang, Yigit-Elliot, & Rosenholtz, 2015). Again, we would not call this additional loss of information "crowding", and in fact we saw evidence of this loss even when fixating individual, uncrowded cubes.)

Farzin, Rivera, and Whitney (2009) also argued for holistic face crowding, using somewhat different logic. They asked observers to perform a number of tasks with Mooney faces (e.g. judging the orientation or gender of the target), and found classic crowding effects (Figure 11). They also found a similarity effect: greater crowding of an upright Mooney face by upright flankers than by inverted. They did not test for an inversion effect, leaving open the possibility of a low-level similarity effect, and low-level mechanisms. Rather, they argued against low-level crowding on the basis that Mooney face tasks require holistic processing (Kanwisher, Tong, & Nakayama, 1998). Certainly, as illustrated in Figure 11, it is not obvious that TTM predicts the similarity effect. However, we argue that holistic face crowding does not logically follow. Processing Mooney faces requires processing at multiple levels prior to the supposed holistic processing. If recognition of Mooney faces fails under conditions

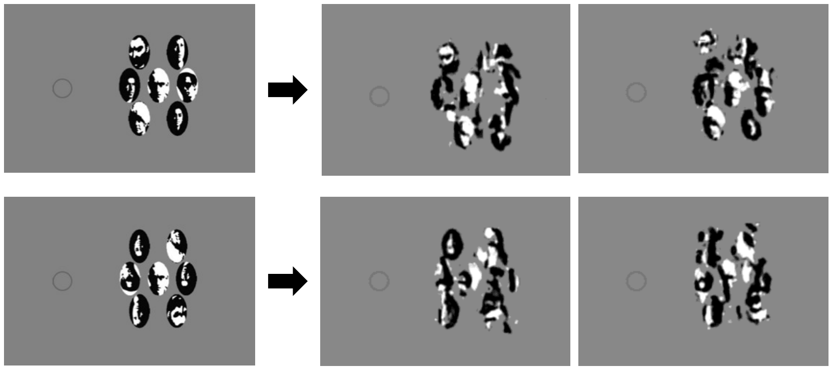

Figure 11. Original stimuli (left of the arrows) and their mongrels (right) from Farzin et al. (2009). Target faces are hard to identify when they are crowded by upright face flankers (top row) compared to inverted flankers (bottom row). Our model, without implementing a holistic face processing mechanism, preserves some information for face tasks, as suggested by the mongrels. Arguably it is also easier to guess the target face identity from the mongrels in the bottom row compared to the top, but further work with a larger set of stimuli would be necessary to quantify this. Original stimuli reprinted with permission.

of crowding, something must have gone wrong with one of those processing stages, but the fault does not obviously lie with the holistic processing stage. For an extreme example, just to make the point: If you closed your eyes and failed to tell the gender of a Mooney face, you would not blame the holistic processing stage.

In summary, one should be careful not to confuse phenomena with mechanisms. Just because crowding occurs when, for example, flankers have similar parts to the target, does not mean that crowding occurs at the parts level of processing. Dakin et al. (2010) similarly argue that the "object-level" crowding effects they find may have a low-level explanation. One cannot easily reason from similarity effects to the stage at which pooling occurs. Nor can we rule out, at this time, that crowding is due to a single level of pooling.

CHALLENGE 3: Flexible pooling regions vary with the stimulus

A third challenge to the pooling account of crowding suggests that pooling regions, rather than being static, might instead vary with the stimulus (Sayim, Westheimer, & Herzog, 2010; Manassi, Sayim, & Herzog, 2012; Manassi, Lonchampt, Clarke, & Herzog, 2016). This challenge is based on an alternative theory of what causes similarity effects. According to this theory, the visual system pools only within a perceptual group. When the target and flankers group together (Figure 12, left column), the visual system pools over both target and flankers, leading to crowding and poor performance identifying the target. When, on the other hand, the target segments from the flankers (Figure 12, right column), the visual system pools over the target alone and flankers alone, leading to a reduction of crowding. As in the previous section, this theory presumes that relief from crowding arises from

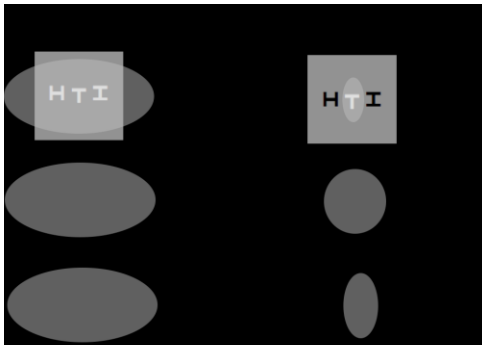

Figure 12. The flexible pooling region hypothesis. According to this hypothesis, when the target and flankers group together, they are pooled together (gray ellipses), resulting in crowding (left column). But when target and flankers are not grouped, they are pooled separately, resulting in less crowding (right column).

mechanisms pooling over the target and not the flankers. However, unlike the previous theory, the relief comes from dynamic adjustment of the region over which pooling occurs, rather than from narrow feature bands.

Once again, there is something attractive about the idea that pooling regions might adapt to the stimulus. What are grouping processes for, after all, if the visual system does not use them to intelligently process the stimulus?

As with our discussion of multilevel crowding mechanisms, one must ask whether an HD pooling model can predict the effects without requiring a more complex mechanism (flexible pooling). We have already demonstrated that it can in some conditions, such as the sign-of-contrast example in Figure 8. Similarly, Keshvari and Rosenholtz (2016) showed that an HD pooling model can predict letter similarity effects without the need for flexible pooling. As a demonstration, consider the stimuli in Figure 13. The target is the letter N in both cases, but the flankers are similar in the top condition, and

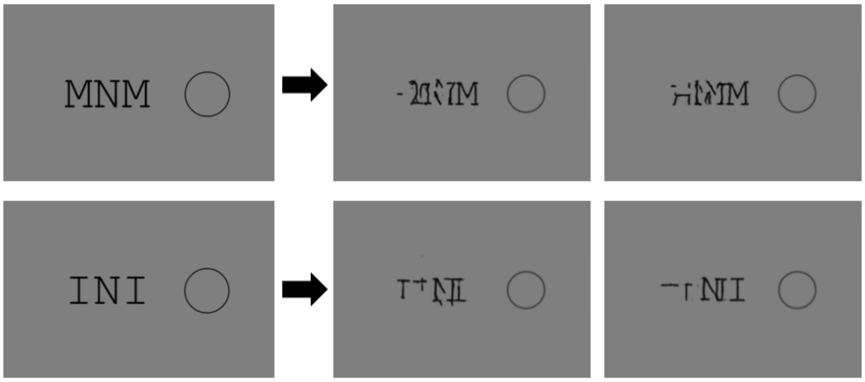

Figure 13. Similarity effect for letter shape. Identifying the central letter ‘N’ is easier when it is flanked by dissimilar flankers (‘I’) compared to similar flankers ‘M’. HD pooling better encodes the target in the dissimilar condition, predicting the letter similarity effect without requiring a flexible pooling mechanism that adjusts to pool over only the target.

dissimilar in the bottom. The target N is better represented in the dissimilar case, suggesting that it will be easier to recognize, even with a fixed pooling mechanism.

These demonstrations suggest that an HD pooling model may at least partially explain grouping effects. In addition, some grouping effects may arise in part from non-crowding mechanisms. Many crowding experiments investigating grouping effects have a potential confound; because crowding leads to location ambiguity, we need to worry about cueing effects. For example, observers asked whether the target line tilts up or down in the dissimilar-orientation condition in Figure 7 may make use of a 100% valid cue that the line with the oddball orientation is the target. HD pooling preserves the feature dissimilarity between target and flankers that would allow the observer to make use of this cue. This cue is not available in the similar-orientation condition. An observer's noisy internal representation of the two arrays may look something like the cartoon in Figure 14 . It should be obvious that the dissimilar condition is inherently easier. The observer in that condition knows to ignore the near-vertical observations, regardless of their noisy observed locations, and respond based on the one near-horizontal observation. In the similar condition, the observer lacks this information, and as a result is strongly impacted by the noise in all three observations. Another way to conceptualize this asymmetry is that an ideal observer also predicts the dissimilar condition will be easier. Grouping effects in crowding may at least in part arise due to generic cueing, i.e. decision-making effects, and not due to mechanisms specific to crowding or peripheral vision. In this toy example, crowding mainly plays a role in introducing location uncertainty, which in turn makes the oddball cue useful.

Along these lines, Rosen and Pelli (2015) first replicated a sign-of-contrast similarity effect (Figure 15, left). The sign of contrast of the target was random on each trial, so the observer did not know to report, say, the white letter, but did know to report the letter with the unique sign-of-contrast. Rosen

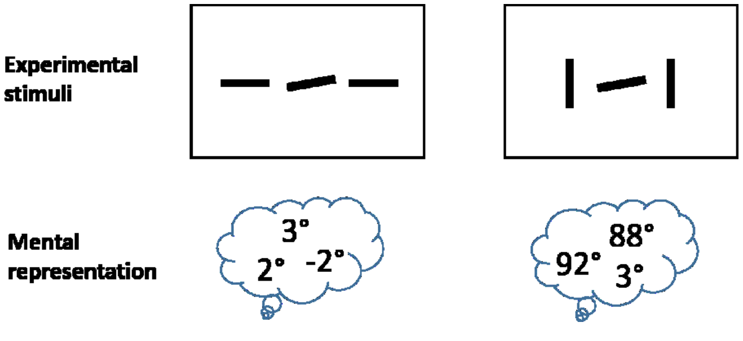

Figure 14. Under conditions of crowding, a pooling model might preserve information necessary to obtain noisy observations of the item features, but not preserve enough location information to tell which observation goes with which item. This does not matter for the dissimilar condition (right), because the target is almost certainly the one with the 3° observation, in this example. This cue is not available in the similar condition, making it inherently more difficult. Any cue that helps reduce that uncertainty could make that difficult condition easier.

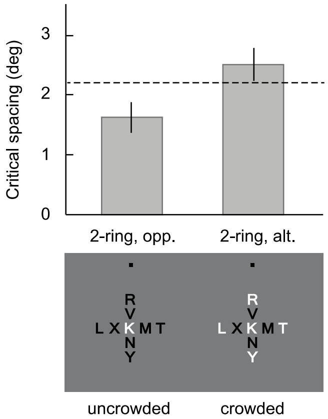

Figure 15. Adapted from Rosen and Pelli (2015). When an odd sign-of-contrast no longer cues the target (right), performance drops, as indicated by an increase in critical spacing.

and Pelli then made the cue less useful, by introducing additional rings of letters with the same sign of contrast as the target. Performance suffered.

In related work, Levi and Carney (2009) varied flanker size, number, spacing, distance to closest point of the flanker, etc. They asked observers to identify the orientation of a Gabor target flanked by a variety of windowed gratings. Figure 16 shows example stimuli from a manipulation that varies the outer diameter of the window while keeping fixed the inner diameter. Over a range of conditions, they found that crowding strength depended upon the spacing between the target and flanker centroids, not on the amount of blank space between them. If crowding mechanisms are sensitive to object centroids, this implies that they operate on objects, or at least after object segmentation has occurred. If so, this would challenge typical pooling accounts, as it suggests that the critical mechanisms operate later than the presumed “feature integration” stage.

However, like the similarity effects discussed above, Levi and Carney’s experimental conditions may be subject to cuing effects. In their critical experiment, the authors varied the size of the flankers while keeping the inner diameter of the flankers fixed. Performance improved as the flankers got larger. The authors attributed this result to the improved encoding of the target Gabor due to the increase in the center-to-center distance between the flanker and the target. However, larger flankers also look considerably less like the target (Figure 16, left). If peripheral vision preserves this information, then the observer might use it to counteract location uncertainty – they should report of the orientation of the small item – leading to improved performance. The mongrels in Figure 16 show that an HD pooling model can capture the difference in flanker size between the two conditions, suggesting peripheral vision may preserve the information necessary to provide a cue to the target. The complex pattern of results in Levi and Carney (2009) may arise from a mix of classic crowding (poorer encoding when target and flankers lie closer

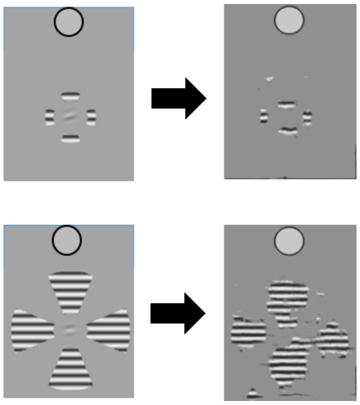

Figure 16. Orientation thresholds for discriminating the orientation of the target are lower when the flankers are large wedges compared to small wedges, for fixed inner diameter of the flankers. Stimuli from (Levi & Carney, 2009) are to the left of the arrows. Mongrels for these two conditions (to the right of the arrows) preserve the distinction between large and small flankers, but obscure the target. This suggests both a potential cueing effect and room for improvement in the model. Original stimuli reprinted with permission.

together), and cueing effects. Determining whether cueing effects influence these results requires experimental verification. It would be premature to claim an object-based crowding mechanism without ruling out this confound.

The mongrels in Figure 16, however, certainly suggest room for improvement in our candidate HD pooling model (or at least in the optimization process that generates the mongrel images). The representation appears to lose the information necessary to report the target orientation. The model may, for instance, need to better mimic contrast sensitivity mechanisms so as to better represent the low contrast target.

Vernier acuity tasks, from Herzog and colleagues, also have the flavor of cueing effects. Vernier acuity requires a decision based on precisely placed feature detectors. In the example in Figure 17, the green detector gives the right answer, whereas the somewhat misplaced orange detectors would both give the wrong answer. The Manassi et al (2012) experiments provided an explicit location cue (the lines above and below the Vernier), but peripheral pooling can disrupt the location of that cue.

Let us further examine the three conditions from Manassi et al (2012) as an example (Figure 18). In the bottom condition, the Vernier pair looks quite different from the long flankers, and the observer can

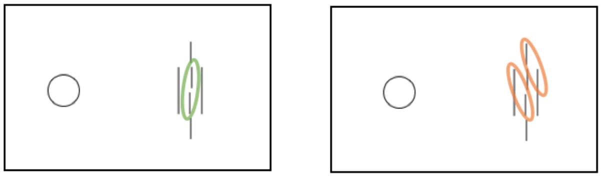

Figure 17. Vernier acuity task require the correct placement of feature detectors. Placing the feature detector like the left panel gives observers the right answer, while placing detectors like the right panel leads to wrong answers.

use this difference to reduce uncertainty; they know to respond to the apparent "tilt" of the short item. This condition should always be easy, regardless of the number of flankers, as found by Manassi et al. In the middle condition, the Vernier pair looks like the flankers, removing a length cue. That condition should always be hard, regardless of the number of flankers, which it is. In the condition on the top, more flankers may form a better group, which in turn may provide a better cue to help localize the Vernier. This would suggest better performance with more flankers, as found by Manassi et al.

In many ways this explanation of the Manassi et al. effects parallels their own interpretation. They demonstrated that grouping strength plays a large role in task performance, and suggested that crowding mechanisms might either operate later in visual processing than grouping mechanisms, or that information from grouping mechanisms feeds back to crowding mechanisms, dynamically adapting those mechanisms. We agree that grouping plays a big role, but attribute that role to providing a cue, rather than to dynamically adapting the mechanisms of crowding. In fact, several recent studies have modeled the crowded Vernier task results either almost solely with a decision mechanism (Zhang, Song,

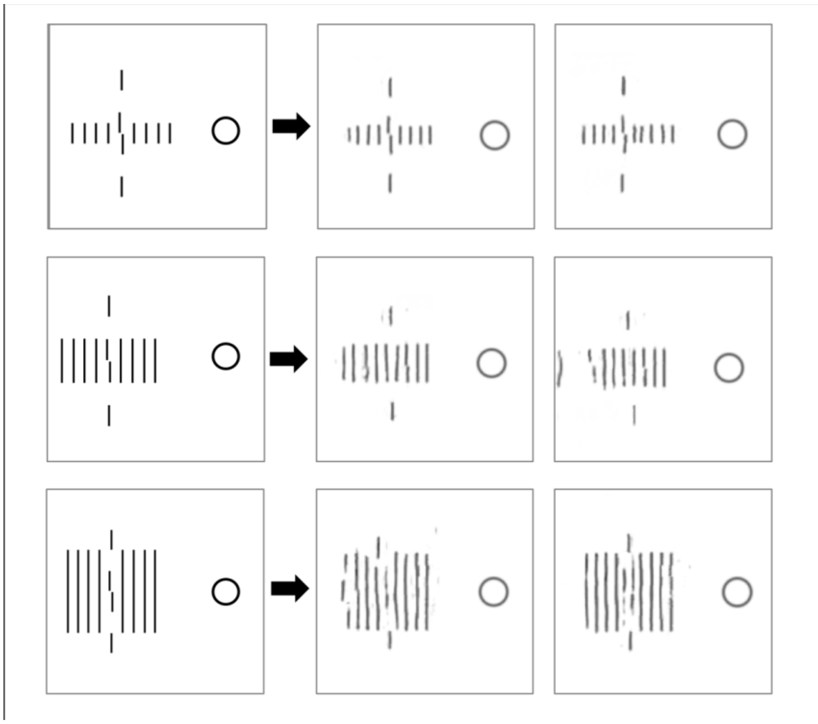

Figure 18. Three conditions from Manassi et al., (2012), (stimuli to the left of the arrows, mongrels to the right). The observer's task is to discriminate the direction of offset of the Vernier. The bottom condition is easy, and the middle condition difficult, and in both performance is independent of the number of flankers. Increasing the number of flankers in the top condition improves performance, perhaps because more flankers leads them to group into an extended object clearly distinct from the target Vernier pair. The mongrels show that sufficient information survives pooling both to distinguish the direction of Vernier offset and to provide a cue distinguishing flankers from Vernier in the easier conditions.

& Yu, 2015), or with grouping processes alone and no special peripheral processing (Francis, Manassi, & Herzog, 2016). It is notable that the original work of (Malania, Herzog, & Westheimer, 2007) demonstrated similar effects in both fovea and periphery; the observed effects may be only minimally due to crowding, per se.

Figure 18 shows a pair of typical mongrels for each condition. Here the question is not whether the target is more poorly represented in the difficult conditions – though that may be true in some cases – but rather whether the representation preserves enough information about both the Vernier offset and the grouping structure to support the use of grouping as a cue. Both seem to be true. However, we note that more recent work has found that for some of the Vernier stimuli TTM seems unable to predict the grouping effects; more flankers leads to worse representation of the target (Doerig, 2019).

Having discussed the potential for cueing confounds, we should revisit the work of (Intriligator & Cavanagh, 2001). Recall that they varied the spacing of a number of disks until observers reached threshold performance tracking the verbally cued item. When the disks were arrayed in an isoeccentric circle about fixation, the critical spacing closely matched that of crowding. However, when Intriligator and Cavanagh (2001) instead asked observers to track the indicated disk among others arranged radially (Figure 19), they found a critical spacing that was smaller than in a traditional crowding task. Importantly, the disks in the radial task varied in size with eccentricity; an attempt to control for cortical magnification. This variation in size, however, likely provided an additional cue that observers could use to keep track of the attended item. If so, this would explain the smaller-than-expected critical spacing in the radial tracking task.

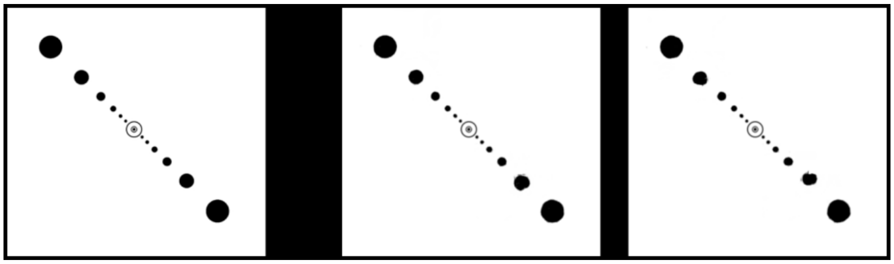

Figure 19. Radial tracking task from Intriligator and Cavanagh (2001). Original image, based on their stimuli, on left. On right, two typical mongrels. White circle indicates 1 degree at fixation. Note that pooling preserves the progressive increase in disk size fairly well.

In summary, it appears plausible that a static HD pooling mechanism could explain some of the grouping effects previously attributed to a flexible pooling mechanism. Some grouping effects, in addition, may arise from cueing confounds rather than from crowding, per se. An HD pooling mechanism loses some information, and maintains other information. Processing then continues, acting on the available information. This includes later grouping processes, as well as a decision stage that makes use of both perceptual organization as well as top-down knowledge to disambiguate the target and perform the task. More work needs to be done first to control for potential cueing confounds, and then to quantitatively test static HD pooling mechanisms on a wide range of phenomena before ruling out such mechanisms in favor of more complicated flexible pooling.

CHALLENGE 4: High-level information survives crowding

In standard pooling models of crowding, the pooling supposedly occurs over fairly low-level features. For example, in our candidate HD pooling model, many of the statistics computed involve pooling at something like a "junction-processing" stage. Empirical studies have nonetheless found evidence that higher-level information can survive crowding. Such results might seem at first glance to preclude the possibility that crowding derives from a low-level pooling mechanism.

For instance, Yeh, He, and Cavanagh (2012) find that Chinese characters that cannot be identified under conditions of crowding (Figure 20) nonetheless can prime a word/non-word lexical decision task when the word has a meaning related to that of the crowded character. The unidentified character can speed responses by about 50 ms compared to trials on which the characters are not semantically related. Surprisingly, they find no significant difference in the magnitude of the priming effect for a crowded vs. uncrowded character, suggesting significant semantic information about the crowded word survives, despite inability to identify it.

It is true that a *low-dimensional* pooling model would degrade the available information so as to render higher level processing nearly impossible. Consider for the sake of argument a model that encodes the visual input using only the mean of a single feature. The loss of information would be *profound*. Such a model could clearly predict difficulty identifying a crowded target. But it would also predict poor performance at nearly all peripheral tasks. How would an observer ever identify anything, even an isolated single letter? However, an HD pooling model behaves fundamentally differently. High-dimensional pooling preserves far more information about the stimulus, and can support many tasks. The mongrels shown in this paper demonstrate that sufficient information survives for higher-level processes to make rough estimates of the number (Figure 3) and size (Figure 16) of objects presented, detect feature pop-out (Figure 7 and Figure 8), construct perceptual groups (Figure 18), and form coarse representations of shapes (Figure 9), and letters (Figure 13). Previous work has shown that the available information suffices for some symbol identification tasks, scene perception and visual search (Balas, Nakano, & Rosenholtz, 2009; Ehinger & Rosenholtz, 2016; Keshvari & Rosenholtz, 2016; Rosenholtz, Huang, Raj, Balas, & Ilie, 2012). Losses from an HD pooling model do not rule out later processing.

Can a pooling model preserve enough information about the target to prime a lexical decision task, while still leading to poor identification performance? While this certainly seems a challenge, it is not out of the question. Identifying a Chinese character requires accessing sufficient information to distinguish it

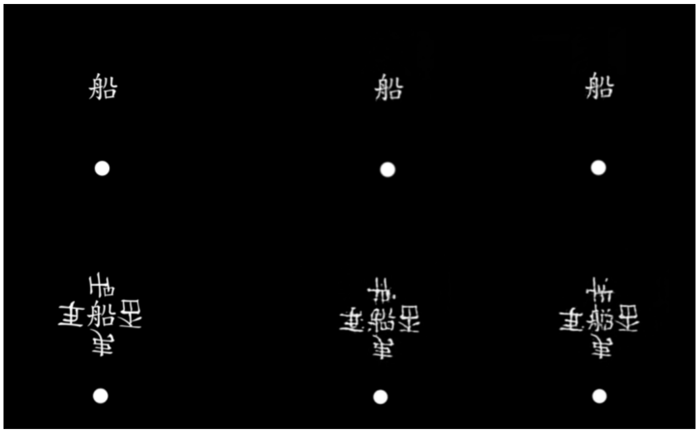

Figure 20. Stimuli from Yeh, He, and Cavanagh (2012) to the left side of the arrow, and mongrels for those original stimuli. HD pooling preserves single characters very well. Encoding of crowded characters is ambiguous, but preserves some information, e.g. some radicals. Thanks to Su-Ling Yeh for providing example stimuli.

from a large number of alternatives. On the other hand, we do not know how much information is required to obtain a priming effect. Plausibly portions of a crowded character might survive pooling; not enough to produce correct identification, but sufficient to provide some information about the meaning. Figure 20 shows that mongrels preserve a fair number of details about the crowded character, for example the radical "口". There also may be enough information to identify the radical "舟". Radicals are characters that either carry meaning when they appear by themselves, or are common subparts in a family of characters that may share similar meaning. For example, the target character in Figure 20 means boat, and its radical, "舟" is commonly associated with other characters related to boats as well. Seeing this radical along might be able to elicit representations associated with boats, thus allowing for a priming effect. Yeh et al. (2012) intentionally picked characters for which the radical did not match the meaning of the word in the lexical decision task, thus avoiding the most obvious confound. Nonetheless, the perception of such radicals may elicit some semantic processing. Observers were approximately 25% correct at the crowded character recognition task, whereas chance performance would be far lower. Though Yeh et al. examined priming only on trials on which object recognition failed, we should not assume that observers had no information about the target on those trials. Rather, 25% correct performance may imply that the observer could narrow the answer to 4 possibilities. Perhaps all 4 choices activated some semantic information, leading to priming while prohibiting correct performance at the harder character identification task. While it seems hard to imagine that the magnitude of priming would be as great as with correct identification of an uncrowded target, it would nonetheless be interesting to examine these conditions to ask what information about Chinese characters survives crowding, and whether TTM preserves that information.

More generally, given our presumption that processing continues after pooling, we expect task to matter. In an earlier section, we discussed this in regards to set perception. Lacking the information for identifying a target in a crowded array does not mean one lacks all information about that target, as if one has failed to select the target and therefore failed to process it. Low-level crowding may permit some higher-level information to "get through the bottleneck of crowding" (Fischer & Whitney, 2011). One may, for example, have sufficient information about the target for it to influence perception of the mean. Nor does ability to perform a task imply that the observer has full information about the stimulus – i.e. that no crowding has occurred. One could perform a target/non-target face task in the periphery (Louie, Bressler, & Whitney, 2007), and yet not preserve sufficient information to support a more fine-grained identification of that face among 100 possible choices. Performance can differ on two tasks because they require different information and/or have different inherent difficulty.

Conclusions

Pooling models of crowding have been popular under a number of different names, from faulty integration, through compulsory averaging, to forced texture perception. Recent empirical results have appeared, on first examination, to challenge a pooling account of crowding. This suggestion, however, has arisen in large part from attempts to gain intuitions about an inherently high-dimensional representation by extrapolating from simple low-dimensional models. Intuitions about low-dimensional models notoriously do not generalize well to higher dimensions. One of our goals in this paper is to provide better intuitions about HD pooling models than one can acquire from introspecting based on low-dimensional straw man models, and based on those intuitions to re-evaluate the seriousness of the challenges to pooling models. High-dimensional pooling preserves enough information to recognize features of individual elements, construct shape and perceptual groups and make judgments about objects and scenes. This calls into question a number of the model challenges. Other apparent model challenges may arise at least in part from decision effects that are not specific to crowding. Further, more quantitative examination is required. Nonetheless, we have suggested that reasonable doubt remains as to whether the challenges truly eliminate pooling models. It appears that pooling remains viable as an explanation of peripheral crowding.

At a higher level, two cautionary lessons emerge. First, one should be careful not to confuse a phenomenon with a mechanism. Substitution phenomena do not imply a substitution mechanism. Crowding by similar orientation, sign-of-contrast, parts, shapes, faces or point-light-walkers does not imply that difficulty identifying the crowded peripheral target arises from mechanisms operating at the corresponding processing stage. Second, one should avoid claiming that a model cannot explain a set of results without actually specifying and testing the model.

Examples in both this paper and in earlier publications clearly point to imperfections in our candidate HD pooling model, TTM. The model lacks some obvious 2nd order statistics that would better capture contour integration, calling into question whether it can explain the effects such as those of (Livne & Sagi, 2007), in which crowding of a target Gabor is relieved if the flanking Gabors align and form a smooth contour. The model captures a good deal of endstopping, and yet lacks explicit endstopping features. Such features may be important for tasks such as distinguishing an array of Os from one of Os and Cs. Lastly, our readers might remember that the target was altogether "lost" in the mongrels for stimuli from Levi and Carney (2009), suggesting at minimum a failure to handle low-contrast information for these stimuli (though whether the fault lies in the encoding itself or in the optimization procedure that generates the mongrels remains to be seen).

Eliminating pooling models as a class is difficult. Choosing what features to pool provides a powerful and flexible way of varying the information lost and maintained by a given model. Vary the number or complexity of features, or the areas over which the model pools them, and the information available can change in profound ways.

One might instead go so far as to think of an HD pooling model in terms of Occam's razor. It provides the (relatively) simple explanation for a range of phenomena, and as such serves as a useful check for whether more complex mechanisms are required, and if so, which ones. The Texture Tiling Model has been particularly successful in this regard, not only explaining a range of crowding results, but calling into question more complex explanations for difficult visual search, change blindness, set perception, and easy scene perception (Rosenholtz, Huang, & Ehinger, 2012; Rosenholtz, 2016). Future work is needed to show quantitatively how much TTM does or does not account for the experimental findings associated with the model challenges. Nonetheless, the mongrels, which provide us with intuitive visualizations of what information is preserved by the model, show promise.

Explaining crowding phenomena may well prove to require more complicated mechanisms. The challenge, for alternative models of crowding, lies in capturing the range of phenomena already explained by pooling models. TTM has, to date, been tested on over 70 conditions. Pooling models have been sufficiently successful that competitors must demonstrate that they have similar or better explanatory power for a wide range of empirical results. Gone are the days in which we can consider a new model to be viable when it explains the results of a single experiment. The model challenges reviewed in this paper provide a useful test set for distinguishing between models as we move forward.

Acknowledgments

The project was funded by NSF/NIH/BMBF IIS-1607486 and NSF BCS-1826757 grants to RR.

References

Agaoglu, M. N., & Chung, S. T. (2016). Can (should) theories of crowding be unified? *Journal of Vision, 16*(15), 10, 1-22.

Andriessen, J. J., & Bouma, H. (1976). Eccentric vision: Adverse interactions between line segments. *Vision Research, 16*(1), 71-78.

Balas, B. J. (2006). Texture synthesis and perception: Using computational models to study texture. *Vision Research, 46*(3), 299-309. doi:10.1016/j.visres.2005.

Balas, B. J., Nakano, L., & Rosenholtz, R. (2009). A summary-statistic representation in peripheral vision explains visual crowding . *Journal of Vision, 9*(12), 13.

Bell, A. J., & Sejnowski, T. J. (1995). An information maximisation approach to blind separation and blind deconvolution. *Neural Computation, 7*(6), 1129-1159.

Bouma, H. (1970). Interactional effects in parafoveal letter recognition. *Nature, 226*, 177-178.

Chaney, W., Fischer, J., & Whitney, D. (2014). The hierarchical sparse selection model of visual crowding. *Frontiers in Integrative Neuroscience, 8*(73), 1-11. doi:10.3389/fnint.2014.00073

Chung, S. T., & Legge, G. E. (2009). Precision of position signals for letters. *Vision Research, 49*, 1948-1960.

Dakin, S. C., Cass, J., Greenwood, J. A., & Bex, P. J. (2010). Probabilistic, positional averaging predicts object-level crowding effects with letter-like stimuli. *Journal of Vision, 10*(10), 1-16.

DiCarlo, J. J., & Cox, D. D. (2007). Untangling invariant object recognition. *Trends in Cognitive Sciences, 11*(8), 333-341.

Doerig, A. B. (2019). Beyond Bouma's window: How to explain global aspects of crowding? *PLOS Computational Biology, 15*(5), e1006580. doi:pcbi.1006580

Ehinger, K. A., & Rosenholtz, R. (2016). A general account of peripheral encoding also predicts scene perception performance. *Journal of Vision, 16*(2), 13.

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.

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

Fischer, J., & Whitney, D. (2011). Object-level visual information gets through the bottleneck of crowding. Journal of Neurophysiology. *Journal of Neurophysiology, 106*(3), 1389-1398.

Francis, G., Manassi, M., & Herzog, M. H. (2016). Cortical dynamics of perceptual grouping and segmentation: Crowding (Abstract). *Vision Sciences Society Annual Meeting. 16*, p. 1114. St. Petersburg, FL: Journal of Vision.

Freeman, J., & Simoncelli, E. P. (2011). Metamers of the ventral stream. *Nature Neuroscience, 14*(9), 1195-1201.

Freeman, J., Chakravarthi, R., & Pelli, D. G. (2012). Substitution and pooling in crowding. *Atten. Percept. Psychophys., 74*(2), 379-396.

Freeman, J., Ziemba, C. M., Heeger, J., D., Simoncelli, E. P., & Movshon, J. A. (2013). A functional and perceptual signature of the second visual area in primates. *Nature neuroscience*, 974–981.

Fukushima, K. (1980). Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. *Biological Cybernetics, 36*, 193-202.

Greenwood, J. A., Bex, P. J., & Dakin, S. C. (2009). Positional averaging explains crowding with letter-like stimuli. *PNAS, 106*, 13130-13135. doi:10.1073/pnas.0901352106

Greenwood, J. A., Bex, P. J., & Dakin, S. C. (2012). Crowding follows the binding of relative position and orientation. *Journal of Vision, 12*(3), 18.

Harrison, W. J., & Bex, P. J. (2015). A unifying model of orientation crowding in peripheral vision. *Current Biology, 25*, 3213-3219.

He, S., & Cavanagh, P. I. (1996). Attentional resolution and the locus of awareness. *Nature, 383*, 334-338.

Huckauf, A., & Heller, D. (2002). Spatial selection in peripheral letter recognition: In search of boundary conditions. *Acta Psychologica, 111*(1), 101-123.

Ikeda, H., Watanabe, K., & Cavanagh, P. (2013). Crowding of biological motion stimuli. *Journal of Vision, 13*, 20.

Intriligator, J., & Cavanagh, P. (2001). The spatial resolution of visual attention. *Cogn. Psychol., 43*, 171-216.

Kalpadakis-Smith, A., Goffaux, V., & Greenwood, J. (2018). Crowding for faces is determined by visual (not holistic) similarity: Evidence from judgments of eye position. *Scientific Reports, 8*(12556), 1-14.

Kanwisher, N., Tong, F., & Nakayama, K. (1998). The effect of face inversion on the human fusiform face area. *Cognition, 68*, B1-B11.

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

Kimchi, R., & Pirkner, Y. (2015). Multiple level crowding: Crowding at the object parts level and at the object configural level. *Perception, 44*(11), 1275-1292.

Kooi, F. L., Toet, A., Tripathy, S. P., & Levi, D. M. (1994). The effect of similarity and duration on spatial interaction in peripheral vision. *Spatial Vision, 8*(2), 255-279.

Korte, W. (1923). U"ber die Gestaltauffassung im indirekten Sehen. *Z. Psych., 98*, 17-82.

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). ImageNet classification with deep convolutional neural networks. *Proc. Neural Inf. Process. Syst. 2012*, (pp. 1097-1105). Lake Tahoe, NV.

Lettvin, J. Y. (1976). On seeing sidelong. *The Sciences, 16*(4), 10-20.

Levi, D. M. (2008). Crowding--An essential bottleneck for object recognition: A mini review. *Vision Research, 48*, 635-654.

Levi, D. M., & Carney, T. (2009). Crowding in peripheral vision: Why bigger is better. *Current Biology, 19*, 1988-1993.

Levi, D., Hariharan, S., & Klein, S. (2002). Supressive and facilitatory spatial interactions in peripheral vision: Peripheral crowding is neither size invariant nor simple contrast masking. *Journal of Vision, 2*, 167-177. doi:10.1167/2.2.3

Liu, T., Jiang, Y., Sun, X., & He, S. (2009). Reduction of the crowding effect in spatially adjacent but cortically remote visual stimuli. *Current Biology, 19*(2), 127-132.

Livne, T., & Sagi, D. (2007). Livne, T., & Sagi, D. (2007). Configuration influence on crowding. *Journal of Vision, 7*(2), 4, 1-12. doi:10.1167/7.2.4

Louie, E. G., Bressler, D. W., & Whitney, D. (2007). Holistic crowding: Selective interference between configural representations of faces in crowded scenes. *Journal of Vision, 7*(2), 24, 1-11. doi:10.1167/7.2.24

Ma, W. J., & Huang, W. (2009). No capacity limit in attentional tracking: Evidence for probabilistic inference under a resource constraint. *Journal of Vision, 9*, 3.

Malania, M., Herzog, M. H., & Westheimer, G. (2007). Grouping of contextual elements that affect vernier thresholds. *Journal of Vision, 7*((2):1), 1-7.

Manassi, M., & Whitney, D. (2018). Multi-level crowding and the paradox of object recognition in clutter. *Current Biology, 28*, R127-R133.

Manassi, M., Lonchampt, S., Clarke, A., & Herzog, M. H. (2016). What crowding can tell us about object representations. *Journal of Vision, 16*(3), 35, 1-13.

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

Martelli, M., Majaj, N. J., & Pelli, D. G. (2005). Are faces processed like words? A diagnostic test for recognition by parts. *Journal of Vision, 5*(1), 6.

Morgan, M. J., & Solomon, J. A. (2005). Capacity limits for spatial discrimination. In L. Itti, G. Rees, & J. K. Tsotsos (Eds.), *Neurobiology of Attention* (pp. 8-10). Burlington: Academic Press.

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

Parkes, L., Lund, J., Angelucci, A., Solomon, J. A., & Morgan, J. (2001). Compulsory averaging of crowded orientation signals in human vision. *Nature Neuroscience, 4*, 739-744.

Pelli, D. G., & Tillman, K. A. (2008). The uncrowded window of object recognition. *Nature Neuroscience, 11*, 1129-1135.

Pelli, D. G., Majaj, N. J., Raizman, N., Christian, C. J., Kim, E., & Palomares, M. C. (2009). Grouping in object recognition: the role of a Gestalt law in letter identification . *Cognitive Neuropsychology, 26*(1), 36-49. doi:10.1080/13546800802550134

Pelli, D. G., Palomares, M., & Majaj, N. (2004). Crowding is unlike ordinary masking: Distinguishing feature integration from detection. *Journal of Vision, 4*(112), 12.

Pelli, D. G., Tillman, K. A., Freeman, J., Su, M., Berger, T., & Majaj, N. J. (2007). Crowding and eccentricity determine reading rate. *Journal of Vision, 7*(2), 20, 1-36.

Poder, E., & Wagemans, J. (2007). Crowding with conjunctions of simple features. *Journal of Vision, 7*(2), 23.

Portilla, J., & Simoncelli, E. P. (2000). A parametric texture model based on joint statistics of complex wavelet coefficients. *Int. J. Comput. Vis., 40*(1), 49-71.

Riesenhuber, M., & Poggio, T. (1999). Hierarchical models of object recognition in cortex. *Nature Neuroscience, 2*(11), 1019-1025.

Rolfs, M., Jonikatis, D., Deubel, H., & Cavanagh, P. (2011). Predictive remapping of attention across eye movements. *Nature Neuroscience, 14*(2), 262-256.

Rosen, S., & Pelli, D. G. (2015). Crowding by a repeating pattern. *Journal of Vision, 15*(6), 10, 1-9.

Rosenholtz, R. (2014). Texture perception. In J. Wagemans (Ed.), *Oxford Handbook of Perceptual Organization* (pp. 167-186). Oxford, UK: Oxford University Press. doi:10.1093/oxfordhb/9780199686858.013.058

Rosenholtz, R. (2016). Capabilities and limitations of peripheral vision. *Annual Rev. of Vision Sci., 2*(1), 437-457.

Rosenholtz, R., Huang, J., & Ehinger, K. A. (2012). Rethinking the role of top-down attention in vision: effects attributable to a lossy representation in peripheral vision. *Frontiers in Psychology, 3:13*. doi:doi: 10.3389/fpsyg.2012.00013

Rosenholtz, R., Huang, J., Raj, A., Balas, B., & Ilie, L. (2012). A summary statistic representation in peripheral vision explains visual search. *Journal of Vision, 12(4):14*, 1-17.

Sayim, B., Westheimer, G., & Herzog, M. H. (2010). Gestalt factors modulate basic spatial vision. *Psychological Science, 21*(5), 641-644.

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.

Strasburger, H. (2005). Unfocused spatial attention underlies the crowding effect in indirect form vision. *Journal of Vision, 5*(11), 8, 1024-1037. doi:10.1167/5.11.8

Strasburger, H., & Malania, M. (2013). Source confusion is a major source of crowding. *Journal of Vision, 13*(1), 24. doi:10.1167/13.1.24

Sun, H.-M., & Balas, B. (2015). Face features and face configurations both contribute to visual crowding. *Attention, Perception, & Psychophysics, 77*(2), 508-519.

Toet, A., & Levi, D. M. (1992). The two-dimensional shape of spatial interaction zones in the parafovea. *Vision Research, 32*, 1349-1357.

Tyler, C. W., & Likova, L. T. (2007). Crowding: A neuroanalytic approach. *Journal of Vision, 7*(2), 16, 1-9.

van den Berg, R., Johnson, A., Anton, A. M., Schepers, A. L., & Cornelissen, F. W. (2012). Comparing crowding in human and ideal observers. *Journal of Vision, 12*(8), 13, 1-15.

van den Berg, R., Roerdink, J. B., & Cornelissen, F. W. (2007). On the generality of crowding: Visual crowding in size, saturation, and hue compared to orientation. *Journal of Vision, 7*(2), 1-11. Retrieved from http://journalofvision.org/7/2/14/

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*, e1000646.

Wallis, T. S., Bethge, M., & Wichmann, F. A. (2016). Testing models of peripheral encoding using metamerism in an oddity paradigm. *Journal of Vision, 16*(2), 4. doi:10.1167/16.2.4

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.

Yamins, D. L., & DiCarlo, J. J. (2016). Using goal-driven deep learning models to understand sensory cortex. *Nature Neuroscience, 19*(3), 356-365.

Yeh, S. L., He, S., & Cavanagh, P. (2012). Semantic priming from crowded words. *Psychological Science, 23*(6), 608-616.

Zhang, S., Song, M., & Yu, A. J. (2015). Bayesian hierarchical model of local-global processing: Visual crowding as a case-study. *Proc. Cognitive Science Society Conference.*

Zhang, X., Huang, J., Yigit-Elliot, S., & Rosenholtz, R. (2015). Cube search, revisited. *Journal of Vision, 15*(3), 9.

Appendix A: The Texture Tiling Model

In the examples in this paper, we use the Texture Tiling Model as our candidate HD pooling model. This Appendix gives a more detailed explanation of the model, along with discussion of various decisions made in creating the model. In particular, this section describes the statistics measured, the pooling regions used, and gives a brief description of the algorithm by which we generate visualizations of the information lost and preserved in peripheral vision. We have provided the MATLAB software at https://dspace.mit.edu/handle/1721.1/121152.

This paper utilizes what we refer to as the "full-field" version of TTM. While some of our papers made predictions based on a "single pooling region" version – to a first approximation equivalent to Portilla and Simoncelli's (2000) texture analysis/synthesis algorithm, modified to work robustly on psychophysical displays with large blank regions – the present model utilizes information from multiple pooling regions across the visual field. Within each pooling region, TTM runs this modified Portilla and Simoncelli algorithm, with the following parameters: Number of scales = 4; number of orientations = 4; and, for these examples, Na = 7. The parameter Na (also referred to as "M" in the original Portilla and Simoncelli (2000) paper), specifies the number of central samples of the autocorrelation used as constraints. In the past we have also used Na = 9, which gives similar results for most stimuli tested. This algorithm computes the following summary statistics: marginal statistics of luminance and color; autocorrelation; correlations of responses of V1-like cells across location, orientation, and scale; and phase correlation across scales. Presuming that the visual system computes local summary statistics as hypothesized by pooling models, further investigation will be required to pinpoint exactly what statistics are involved. Though previous work has suggested that the aforementioned statistics provide a good initial guess, we would be surprised if this initial set of statistics proved to be correct. The summary statistics might involve more biologically plausible computations, or might derive from features learned for ecological tasks such as object recognition. Additional statistics might be needed to better capture contour integration behavior, or to more explicitly compute endstopping. In addition, one would of course expect to include some statistics based on motion and on binocular disparity.

Given a fixation point, TTM tiles the image with overlapping pooling regions. For the examples in this paper, we have used square pooling regions. Though we have previously implemented elongated elliptical pooling regions, the square regions allowed for faster processing and led to fewer synthesis artifacts in our hands, while otherwise not greatly differing in their results from elliptical pooling regions. Square "receptive fields" are of course completely biological implausible. They also appear in opposition to behavioral work, which finds roughly twice the critical spacing in the radial direction than in the tangential (Toet & Levi, 1992), though such differences in critical spacing could also derive from the pattern of overlap of the pooling regions (Rosenholtz, 2016). Further work is required to determine the number, location, size, and degree of overlap of these regions necessary to best predict human behavior. The pooling regions used here have width equal to 0.5 × the eccentricity at their center. The pooling regions have a radial overlap of 45%; for a pair of pooling regions arranged in a radial direction, this specifies what percent of the width of the inner pooling region overlaps with the outer pooling region. Tangential overlap is determined by the size of the pooling regions plus the number of pooling regions in the tangential direction. The examples here use 36 pooling regions tangentially, i.e. with their centers every 10 degrees around a circle centered at fixation.

The algorithm performs two preprocessing steps. First, if one simply lays down pooling regions across the entire image, some pooling regions will land partially outside the original image, necessitating image padding. Second, prior to measurement of statistics, TTM blurs the input image to approximately mimic the loss of acuity with eccentricity.

Synthesis is initiated by assuming that a "foveal region" (which we conceptualize as a small 1-2 degree circle about fixation) is reconstructed perfectly. For the purposes of this paper, we used a default size for this "fovea" of 32 pixels in radius. One might loosely think of this as the number of pixels per degree. Then, moving in an outward sweep, each subsequent pooling region is synthesized using the previous partial synthesis result as the seed for the texture synthesis process (plus, in the first iteration, noise in regions not yet synthesized). When replacing each synthesized patch, we blend it with overlapping regions using a Gaussian distance-weighted average. A second sweep reconstructs the pooling regions from the outer rings inward, ending by reapplying the "fovea". This process iterates a number of times over the entire image. We use a coarse-to-fine strategy, starting with the coarsest scale and adding one scale at a time, to speed convergence. At each stage of this coarse-to-fine procedure we run 10 iterations of synthesis of the entire image. Within each such iteration, we run 3 iterations of each pooling region.

Little previous work speaks to what color statistics the model should compute. It seems likely that the visual system computes summary statistics in several color channels, and perhaps also computes some sort of correlations between those channels. More research is required. Here we first use independent component analysis (Bell & Sejnowski, 1995) to split the image into three color bands. This is somewhat unrealistic in that it suggests the computed statistics change with the contents of the stimulus, but in practice it works well. We measure statistics in each of these bands independently. Within each local pooling region we also measure the covariance between the three color bands, and apply that constraint after synthesizing the three channels of each local pooling region.

Additional details are documented with the code. This includes details involved in the functioning of the code, and recommendations for running the algorithm, including choices of the number of pixels per degree and other adjustments to the parameters of the algorithm.