Global and high-level effects in crowding cannot be predicted by either high-dimensional pooling or target cueing

Alban Bornet
Laboratory of Psychophysics, Brain Mind Institute,
École Polytechnique Fédérale de Lausanne (EPFL),
Lausanne, Switzerland

Oh-Hyeon Choung
Laboratory of Psychophysics, Brain Mind Institute,
École Polytechnique Fédérale de Lausanne (EPFL),
Lausanne, Switzerland

Adrien Doerig
Laboratory of Psychophysics, Brain Mind Institute,
École Polytechnique Fédérale de Lausanne (EPFL),
Lausanne, Switzerland
Donders Institute for Brain, Cognition and Behaviour,
Nijmegen, Netherlands

David Whitney
Department of Psychology, University of California,
Berkeley, California, USA
Helen Wills Neuroscience Institute,
University of California, Berkeley, California, USA
Vision Science Group, University of California, Berkeley,
California, USA

Michael H. Herzog
Laboratory of Psychophysics, Brain Mind Institute,
École Polytechnique Fédérale de Lausanne (EPFL),
Lausanne, Switzerland

Mauro Manassi
School of Psychology, University of Aberdeen,
King’s College, Aberdeen, UK

In visual crowding, the perception of a target deteriorates in the presence of nearby flankers. Traditionally, target-flanker interactions have been considered as local, mostly deleterious, low-level, and feature specific, occurring when information is pooled along the visual processing hierarchy. Recently, a vast literature of high-level effects in crowding (grouping effects and face-holistic crowding in particular) led to a different understanding of crowding, as a global, complex, and multilevel phenomenon that cannot be captured or explained by simple pooling models. It was recently argued that these high-level effects may still be captured by more sophisticated pooling models, such as the Texture Tiling model (TTM). Unlike simple pooling models, the high-dimensional pooling stage of the TTM preserves rich information about a crowded stimulus and, in principle, this information may be sufficient to drive high-level and global aspects of crowding. In addition, it was proposed that grouping effects in crowding may be explained by post-perceptual target cueing. Here, we extensively tested the predictions of the TTM on the results of six different studies that highlighted high-level effects in crowding. Our results show that the TTM cannot explain any of these high-level effects, and that the behavior of the model is equivalent to a simple pooling model. In addition, we show that grouping effects in crowding cannot be predicted by post-perceptual factors, such as target cueing. Taken together, these results reinforce once more the idea that complex target-flanker interactions determine crowding and that crowding occurs at multiple levels of the visual hierarchy.

Citation: Bornet, A., Choung, O.-H., Doerig, A., Whitney, D., Herzog, M. H., & Manassi, M. (2021). Global and high-level effects in crowding cannot be predicted by either high-dimensional pooling or target cueing. *Journal of Vision*, *21*(12):10, 1–25, https://doi.org/10.1167/jov.21.12.10.

https://doi.org/10.1167/jov.21.12.10 Received February 27, 2021; published November 23, 2021 ISSN 1534-7362 Copyright 2021 The Authors

Introduction

In crowding, perception of a target strongly deteriorates when flanking elements are added (Pelli, 2008; Strasburger, Rentschler, & Jüttner, 2011; Whitney & Levi, 2011). Classically, crowding was explained by pooling or bottleneck models where features of the target and nearby flankers are pooled within receptive fields of low-level neurons (Levi, 2008; Wilkinson, Wilson, & Ellemberg, 1997). In line with this hypothesis, target-flanker interactions in crowding were characterized as (1) locally confined (Bouma's law; Bouma, 1970; Toet & Levi, 1992), (2) deleterious (Parkes, Lund, Angelucci, Solomon, & Morgan, 2001; Wilkinson et al., 1997), and (3) low-level feature specific (Andriessen & Bouma, 1976; Chung, Levi, & Legge, 2001; Levi, Toet, Tripathy, & Kooi, 1994; Levi, Hariharan et al., 2002).

Classic pooling models were seriously challenged by recent results in the last decade, and widely dismissed. First, elements beyond Bouma's window were shown to modulate crowding strength (Harrison, Retell, Remington, & Mattingley, 2013; Malania, Herzog, & Westheimer, 2007; Manassi, Sayim, & Herzog, 2012; Vickery, Shim, Chakravarthi, Jiang, & Luedeman, 2009). Second, it was shown that grouping determines crowding: depending on the stimulus configuration, adding flankers can reduce or increase crowding strength (Livne & Sagi, 2007; Levne & Sagi, 2010; Malania et al., 2007; Saarela, Westheimer, & Herzog, 2010). Third, crowding was shown to occur at multiple levels along the visual hierarchy (e.g., for objects and faces; Kimchi & Pirkner, 2015; Louie, Bressler, & Whitney, 2007; Sun & Balas, 2015; Xia, Manassi, Nakayama, Zipser, & Whitney, 2020). Taken together, target-flanker interactions in crowding are (1) global, (2) complex (i.e., crowding does not simply increase when more flankers are added), and (3) occur at multiple levels of the visual processing (for reviews, see: Herzog, Sayim, Chicherov, & Manassi, 2015; Herzog, Sayim, Manassi, & Chicherov, 2016; Herzog & Manassi, 2015; Manassi & Whitney, 2018; see also Banks, Larson, & Prinzmetal, 1979; Banks & White, 1984; Egeth & Santee, 1981; Huckauf, Heller, & Nazir, 1999; Mason, 1982; Mewhort, Marchetti, & Campbell, 1982; Wolford & Chambers, 1983). As a consequence, simple pooling models do not seem adequate to explain this large body of results (Doerig, Bornet, Rosenholtz, Francis, Clarke, & Herzog, 2019; Rosenholtz, Yu, & Keshvari, 2019).

In response to this line of evidence, Rosenholtz et al. (2019) recently proposed that high-dimensional pooling models (e.g., the Texture Tiling Model [TTM]; Rosenholtz, 2014; Rosenholtz, Huang, & Ehinger, 2012; Rosenholtz, Huang, Raj, Balas, & Ilie, 2012), can explain all these effects. In a first stage, the TTM computes V1-like responses from low-level, multiscale, and oriented feature detectors. In a second stage, the model pools these features locally to generate a large set of second-order correlations (high-dimensional pooling). Contrary to simple pooling models, the high-dimensional pooling stage preserves rich information, which supports a fine-grained representation of the visual input and may, in principle, explain complex crowding effects at a later post-perceptual stage. Still, the TTM shares the characteristics of the simpler pooling models: pooling occurs only in spatially confined regions, is restricted to low-level processing, and occurs at a single processing level. Crucially, if the TTM can predict all of the high-level effects in the recent literature, it means that target-flanker interactions are not as high-level as previously thought.

Rosenholtz et al. (2019) proposed two ways in which grouping might affect the perception of a crowded stimulus, without requiring explicit visual grouping processes. First, what we call grouping might simply be a collateral effect of high-dimensional pooling. For example, the TTM might "group" together elements that can easily be described using summary statistics. Second, what we call "grouping" might reflect processes that happen after high-dimensional pooling. For example, the high-dimensional pooling stage may reduce the position uncertainty of visual elements (cueing). Moreover, Rosenholtz et al. (2019) proposed that the TTM can also reproduce holistic effects in crowding without requiring high-level feature interactions. The rich information preserved by the high-dimensional pooling stage of the TTM may drive holistic processing (e.g., upright and inverted faces being perceived differently), in a post-perceptual stage.

Here, we tested these hypotheses by probing the TTM behavior on a large body of evidence for high-level effects in crowding (Canas-Bajo & Whitney, 2020; Farzin, Rivera, & Whitney, 2009; Manassi et al., 2012; Manassi, Sayim, & Herzog, 2013; Manassi, Hermens, Francis, & Herzog, 2015; Manassi, Lonchampt, Clarke, & Herzog, 2016). First, we show that, in contrast to what Rosenholtz et al. (2019) claimed, the TTM does not reproduce any of the grouping effects in (Manassi et al., 2012; Manassi et al., 2013; Manassi, Hermens, Francis, & Herzog, 2015; Manassi, Lonchampt, Clarke, & Herzog, 2016; section "TTM & Grouping Effects"). Second, we show that the TTM has the same limitations as simple pooling models, strictly dependent on flanker pixel density and blind to high-level configurational aspects (subsection "TTM & prediction power"). Third, as previously mentioned, Rosenholtz et al. (2019) argued that the grouping effects in crowding (Manassi et al., 2012; Manassi et al., 2013; Manassi et al., 2015; Manassi et al., 2016) might arise because different flanker configurations cue the target location in different ways and, thus, may modulate crowding strength in a later post-perceptual stage. We show that cueing plays no real role in crowding (Manassi et al., 2012; Manassi et al., 2013; Manassi et al., 2015; Manassi et al., 2016; subsection "Grouping effects and target cueing"). Fourth, we show that holistic face processing can occur in peripheral vision despite low-level crowding, and that the TTM cannot reproduce this result because low-level information is lost irretrievably at the pooling stage of the model (section "TTM & Face Crowding," single face discrimination task). Fifth, we show that the TTM cannot account for crowding between holistic representations of faces (Farzin et al., 2009; section "TTM & Face Crowding," gender face discrimination task).

General materials and methods

Mongrel generation

To assess TTM performance, we generated mongrels for different stimuli, by using the code shared by Rosenholtz et al. (2019; https://dspace.mit.edu/handle/1721.1/121152). The TTM takes an image as input and outputs several images rather than a performance measure, such as accuracy. The outputted images, called mongrels, share the same pooled statistics as the original input image. The idea is that mongrels, when viewed foveally and for unlimited time, mimic the peripheral perception of the input image (Balas, Nakano, & Rosenholtz, 2009; Rosenholtz et al., 2019).

Stimulus images were taken from (Manassi et al., 2012; Manassi et al., 2013; Manassi et al., 2015; Manassi et al., 2016; Canas-Bajo & Whitney, 2020; Farzin et al., 2009). The layout of the stimuli was identical to the original publications. Every pixel was 1/30 degrees of the stimulus used in the original experiment (i.e., the resolution was 30 pixels per degree). In the original experiment of (Manassi et al., 2012; Manassi et al., 2015), stimuli were displayed on oscilloscopes. Here, we adapted our stimuli to an LCD presentation by having white lines on a black background, as in (Manassi et al., 2013; Manassi et al., 2016).

Model assessment and potential shortcomings

To assess the TTM behavior, following Rosenholtz et al. (2019), we asked participants to perform the original crowding experiments of (Manassi et al., 2012; Manassi et al., 2013; Manassi et al., 2015; Manassi et al., 2016; Canas-Bajo & Whitney, 2020; Farzin et al., 2009), but using the mongrels presented in free viewing conditions. All original experiments were two alternative forced choice (2AFC) target discrimination tasks (more detail in the methods subsection of each experiment). To quantify the TTM performance, we measured target discrimination accuracy for each condition. We attempted to address potential shortcomings of our model assessment method in the following ways.

First, we used the code from the official repository to generate the mongrels. The TTM has a variable parameter that needs to be set, namely the radius of the fovea. The code documentation suggests a value between 16 and 32 pixels. The latter value is what was used in Rosenholtz et al. (2019). Because a value of 32 did not yield sufficiently strong crowding in pilot experiments, which would rule out the TTM as a model of crowding, we used a value of 16 pixels. In order to control for ceiling effects, we repeated some experiments with a radius of 32 pixels (details in the subsection of each experiment).

Second, a single mongrel cannot be regarded as the true output of the TTM but merely as an illustration of its behavior. To have a precise measure of the model output, we generated as many mongrels as we could for each stimulus (10 to 200, depending on the number of conditions we needed to run for each experiment). Moreover, we made all generated mongrels available at https://github.com/albornet/TTM_Verniers_Faces_Mongrels.

Third, humans may have strong individual biases in the perception of the mongrels, which may average out existing effects. For this reason, we also used bias-free algorithms to perform the mongrel discrimination tasks (more details in the Methods subsection of each experiment and in the Discussion section).

Ethics

Participants gave oral consent before the experiment, which was conducted in accordance with the Declaration of Helsinki except for preregistration (World Medical Organization, 2013) and was approved by the local ethics committee (Commission éthique du Canton de Vaud, protocol number: 164/14, title: Aspects fondamentaux de la reconnaissance des objets protocole général).

TTM and grouping effects

Methods

Stimuli

The stimuli that we used to generate the mongrels consisted of a vernier target alone or surrounded by various flanker configurations (Figure 1). The vernier target consisted of two vertical 40 arcmin lines separated by a vertical gap of 4 arcmin. The vernier

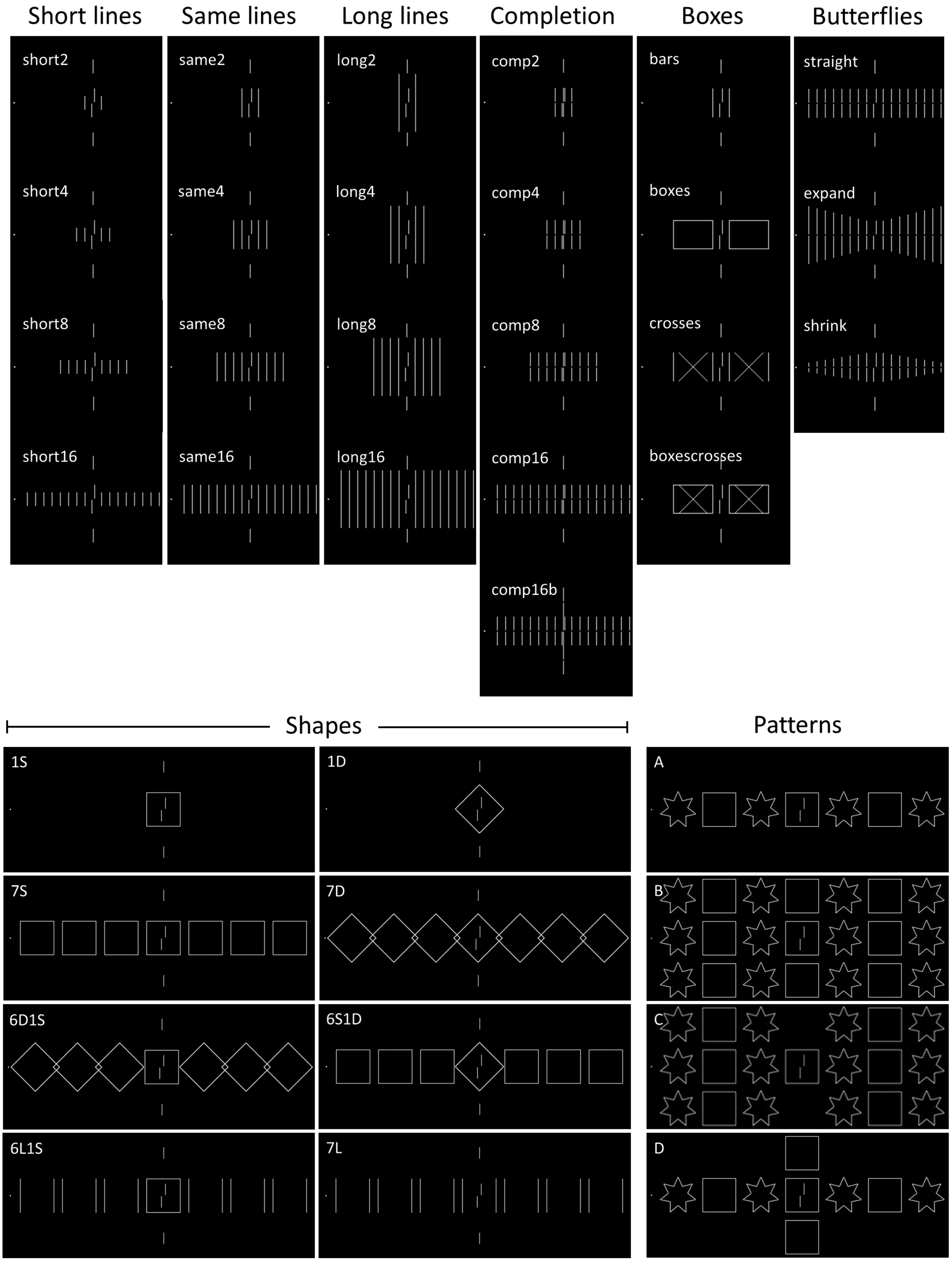

Figure 1. Stimuli used to validate the TTM. In the original experiments, observers were asked to discriminate the offset of a vernier target presented in the right hemifield and in the periphery (here shown in the center of each image), while looking at a fixation dot.

→

←

Different flanker configurations were presented across the studies: “Short/Same/Long lines” and “Boxes” in Manassi et al. (2012); “Completion” and “Butterflies” in Manassi et al. (2015); “Shapes” in Manassi et al. (2013); “Patterns” in Manassi et al. (2016). In the original experiments as well as in the TTM validations, the target eccentricity was 3.88 degrees in the “Lines,” “Boxes,” “Completion,” and “Butterflies” experiments, and 9 degrees in the “Shapes” and “Patterns” experiments. Note that, in all original experiments except “Patterns”, two vertical lines (pointers) were added above and below the vernier target to reduce target location uncertainty.

target was offset either to the left or to the right. The offset size varied according to the eccentricity at which the vernier target was presented (see next paragraph).

Sixteen flanker configurations were taken from Manassi et al. (2012; Figure 1, “Short/Same/Long lines” and “Boxes”) and eight configurations from Manassi et al. (2015; Figure 1, “Completion” and “Butterflies”). For these conditions, each stimulus configuration was presented to the TTM with a vernier target eccentricity of 3.88 degrees and a vernier offset size of 8 arcmin. Eight configurations were taken from Manassi et al. (2013; Figure 1, “Shapes”) and four configurations from Manassi et al. (2016; Figure 1, “Patterns”). For these conditions, each stimulus configuration was presented to the model with a vernier target eccentricity of 9 degrees and a vernier offset size of 14 arcmin. These vernier offsets correspond to approximately five times the thresholds measured in the original experiments for the unflanked conditions (vernier alone).

In all configurations, except the ones in the “Patterns” experiment, two vertical lines (called the “pointers”) were placed above and below the vernier target. In the original experiments, the pointers were used to reduce target location uncertainty (Manassi et al., 2012; Manassi et al., 2013; Manassi et al., 2015). For these configurations, we also generated mongrels using stimuli in which the pointers were removed. In total, 72 different flanker configurations were used (including the vernier alone conditions, at both eccentricities, with and without pointers). For each configuration, 30 different mongrels were generated (split equally between left and right vernier offset), for a total of 2160 unique mongrel samples shown to every participant.

Vernier offset discrimination task

Crowding strength in the TTM was quantified by performing a target discrimination task in free-viewing conditions using the mongrels. We presented the generated mongrel images to observers and asked them to discriminate between left and right vernier offset (2AFC task). The mongrels were shown in a random order (mixed conditions).

In order to familiarize with the task, prior to the experiment, observers were shown 10 examples of the original stimulus images in which only the target was present, followed by 10 original stimulus images in which the target was embedded in different flanker configurations, and finally 10 mongrels. In all these examples, the vernier target (or the part of the mongrel that corresponded to the vernier target) was highlighted and labeled.

Thirteen observers performed this task (6 men and 7 women, 31.8 ± 2.9 years old). For each flanker configuration, we measured the discrimination performance (error rate = 1-accuracy) and computed the corresponding standard error of the mean across observers. Human performance in the vernier offset discrimination task was compared to the human data coming from the corresponding original crowding experiments (Figures 2 to 6).

Vernier offset matching algorithm

To avoid biases introduced by observers using different strategies to perform the mongrel discrimination tasks, we also performed mongrel vernier offset discrimination using a template matching algorithm. The algorithm searched for a target in the mongrels by sliding left- and right-sided vernier target templates over the whole image. For each location in the mongrel, a match value was defined as the sum of the point-wise multiplication between the template and the part of the mongrel image that lay under the target template centered at that location. Each match value was weighted by a function that decreased with the distance of the location of the template to the original position of the target, to help the algorithm focus on the most likely location of the vernier in the mongrel (Equation 1).

$$M^s\left(i, j\right) = \mathrm{e}^{-(D(i,\ j)/\sigma)^2} \cdot \sum\nolimits_{k,l} T^s_{k,l} \cdot I_{i+k,j+l} \qquad (1)$$

$M^s(i, j)$ was the weighted match value of the s-sided vernier template at location (i, j), $T^s_{k,l}$ was the value of the s-sided vernier template at location (k, l) in the template coordinates, I was the mongrel array. $D(i, j)$ was the distance in pixels between the location of the template and the original target position, and σ was the width of the weighting function in pixels. σ was set to 50 pixels. For each mongrel, the algorithm decided for a left or a right vernier as the side of the template that obtained the highest weighted match value.

Results

Lines experiment

In Manassi et al. (2012), crowding was strong when a vernier target was flanked on each side by two short lines or by two lines of the same length as the vernier, but weak when flanked by two longer lines. When increasing the number of flankers, crowding decreased for short flankers, stayed constant with same-length flankers, and slightly decreased with long flankers (see Figure 2, left). Hence, adding flankers can lead to nonmonotonic effects in crowding strength, contrary to what is predicted by simple pooling models.

As with the simple pooling models, in both TTM validation tasks, crowding strength increased when increasing the number or the size of the flankers (see Figure 2, center and right). The TTM performance differs from human data, in which adding flankers reduced crowding strength in certain conditions.

Completion experiment

In Manassi et al. (2015), crowding was strong when a vernier was flanked by 16 same-length straight verniers but decreased when a same-length straight vernier mask was added at target location (see Figure 3, left, straight versus comp16). Crowding was strong for control conditions in which a longer mask was used or using a same-length mask but having only two vernier flankers (see Figure 3, left, comp16b and comp2). Hence, adding a single element can drastically change crowding strength, which cannot be explained by simple pooling models.

In both TTM validation tasks, crowding strength decreased when adding a same-length vernier mask at target location, as in the human data (see Figure 3, center and right, straight versus comp16). However, crowding strength also decreased when using a longer mask or having only two vernier flankers (see Figure 3, center and right, straight versus comp16b and comp2), and gradually increased when adding more flankers (Supplementary Information Figure SA), showing that the configuration played no role.

Boxes and crosses experiment

In Manassi et al. (2012), crowding was strong when the vernier target was flanked by two same-length flankers (see Figure 4, left, bars). Crowding decreased when adding flankers to form boxes or boxes containing a cross (see Figure 4, left, boxes and boxes and crosses), but stayed high when the added flankers were not embedded in box shapes (see Figure 4 left, crosses). These results were taken as evidence that flanker configuration modulates crowding strength.

The TTM failed to reproduce these results. In both TTM validation tasks, weak crowding was observed for the bars, and stronger crowding was observed when adding more flankers (see Figure 4, center and right, bars versus boxes and crosses and boxes and crosses), regardless of the configurations.

Shapes experiment

In Manassi et al. (2013), crowding was strong when the vernier target was flanked by a single square (see

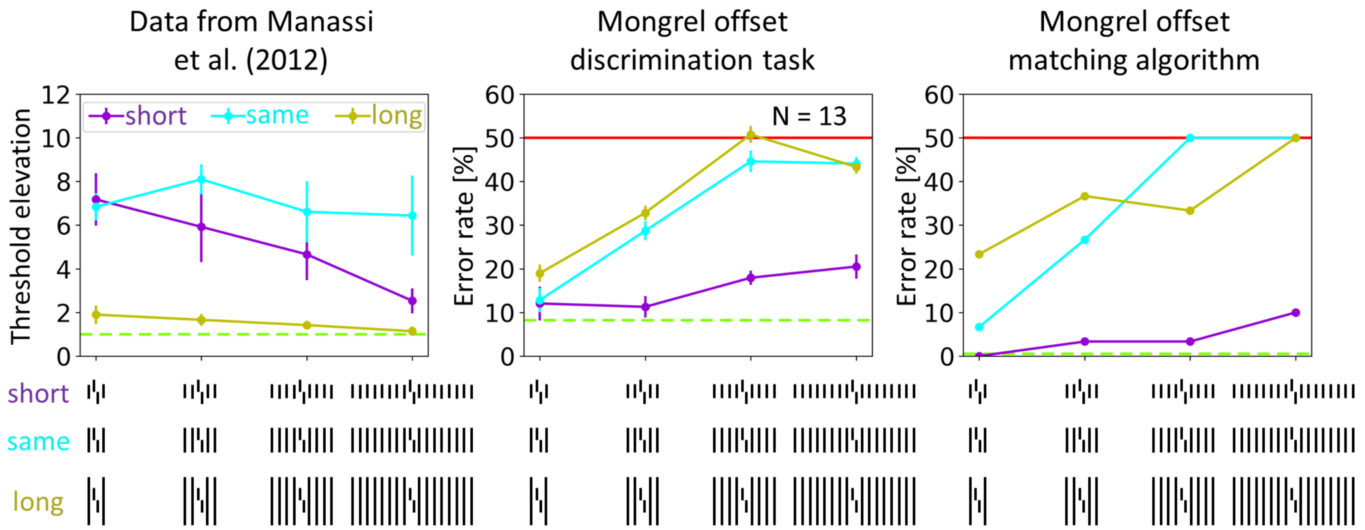

Figure 2. Lines. Left. Data from Manassi et al. (2012). Offset discrimination thresholds were determined for vernier targets presented in the periphery at 4 degrees of eccentricity. Center. TTM validation in which observers discriminate between left and right offset verniers in mongrel images. Right. TTM validation with a template matching algorithm using the same mongrels as in the human experiment. Green dashed lines indicate vernier alone performance. Red lines indicate chance level (50% accuracy). Note that the y-axis labels are different.

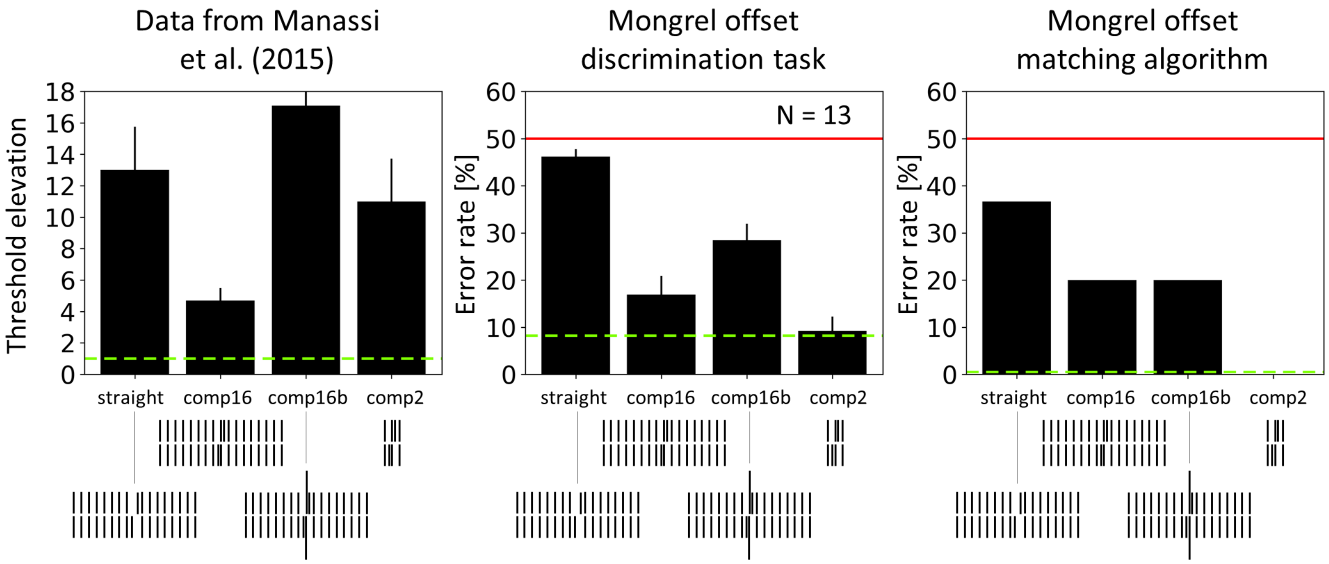

Figure 3. Completion. Left. Data from Manassi et al. (2015). Offset discrimination thresholds were determined for vernier targets presented in the periphery at 4 degrees of eccentricity. Center. TTM validation in which observers discriminate between left and right offset verniers in mongrel images. Right. TTM validation with a template matching algorithm using the same mongrels as in the human experiment. Note that the algorithm made 0% errors for in the comp2 condition (the data is not missing). Green dashed lines indicate vernier alone performance. Red lines indicate chance level (50% accuracy). Note that the y-axis labels are different.

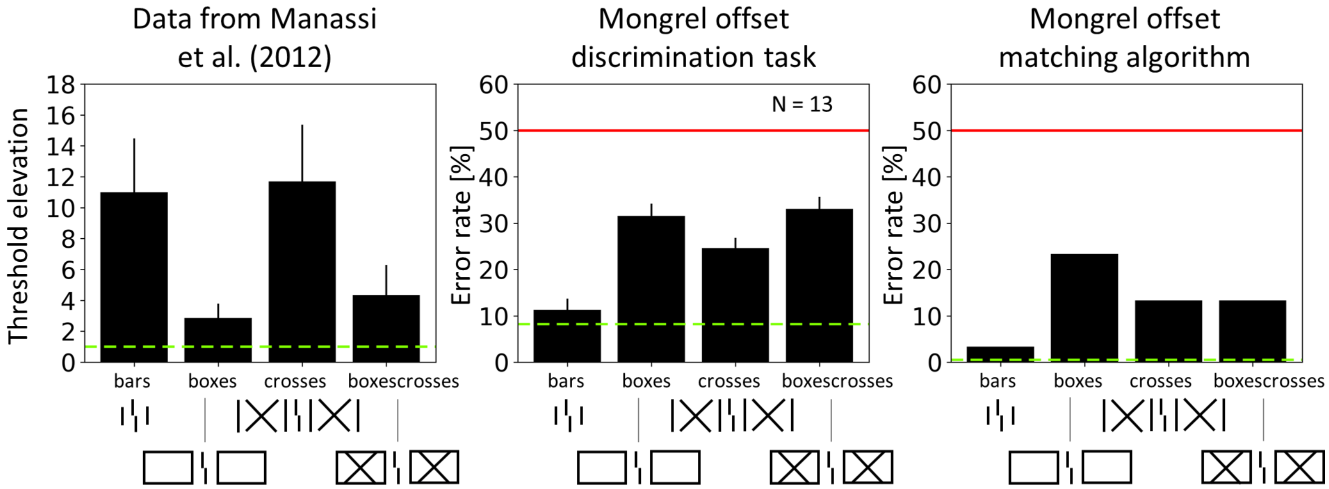

Figure 4. Boxes and crosses. Left. Data from Manassi et al. (2012). Offset discrimination thresholds were determined for vernier targets presented in the periphery at 4 degrees of eccentricity. Center. TTM validation in which observers discriminate between left and right offset verniers in mongrel images. Right. TTM validation with a template matching algorithm using the same mongrels as in the human experiment. Green dashed lines indicate vernier alone performance. Red lines indicate chance level (50% accuracy). Note that the y-axis labels are different.

Figure 5, left, 1S). Crowding decreased when the vernier was flanked by three additional squares on each side but remained strong when the added flankers were diamonds (see Figure 5, left, 7S versus 7D1S). Crowding was strong in control conditions (see Figure 5, left, 7L and 6L1S). The results showed that high-level shape processing can determine low-level vernier acuity.

The TTM did not reproduce this set of results. In both TTM validation tasks, crowding was strong for all tested conditions, independently of shape configuration (see Figure 5, center and right). A similar pattern was found using diamonds instead of squares (Supplementary Information Figure SB).

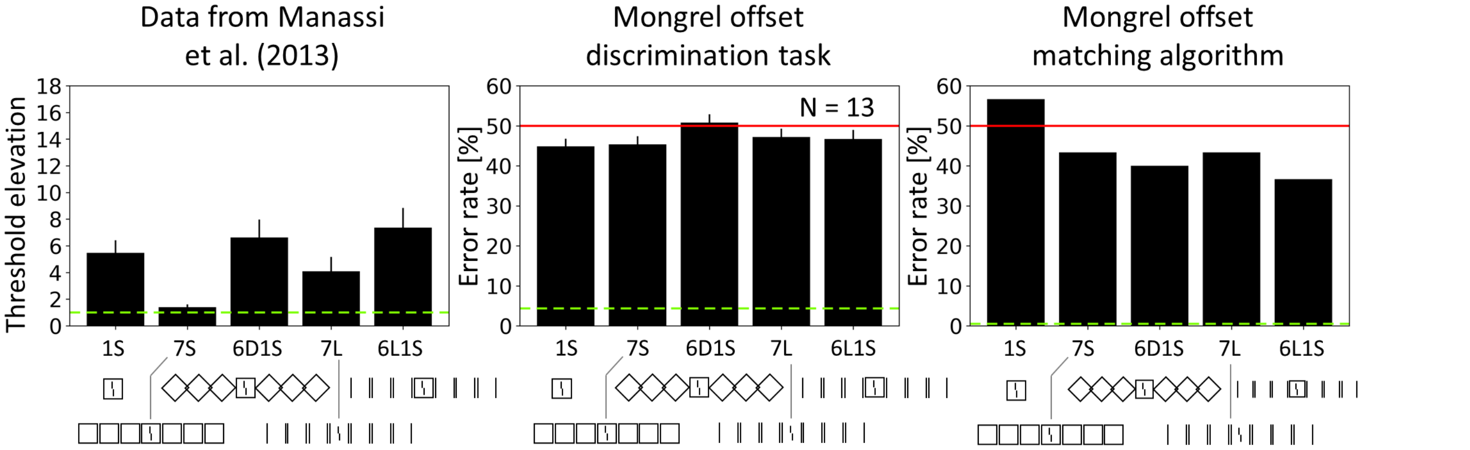

Figure 5. Shapes. Left. Data from Manassi et al. (2013). Offset discrimination thresholds were determined for vernier targets presented in the periphery at 9 degrees of eccentricity. Center. TTM validation in which observers discriminate between left and right offset verniers in mongrel images. Right. TTM validation with a template matching algorithm using the same mongrels as in the human experiment. Green dashed lines indicate vernier alone performance. Red lines indicate chance level (50% accuracy). Note that the y-axis labels are different.

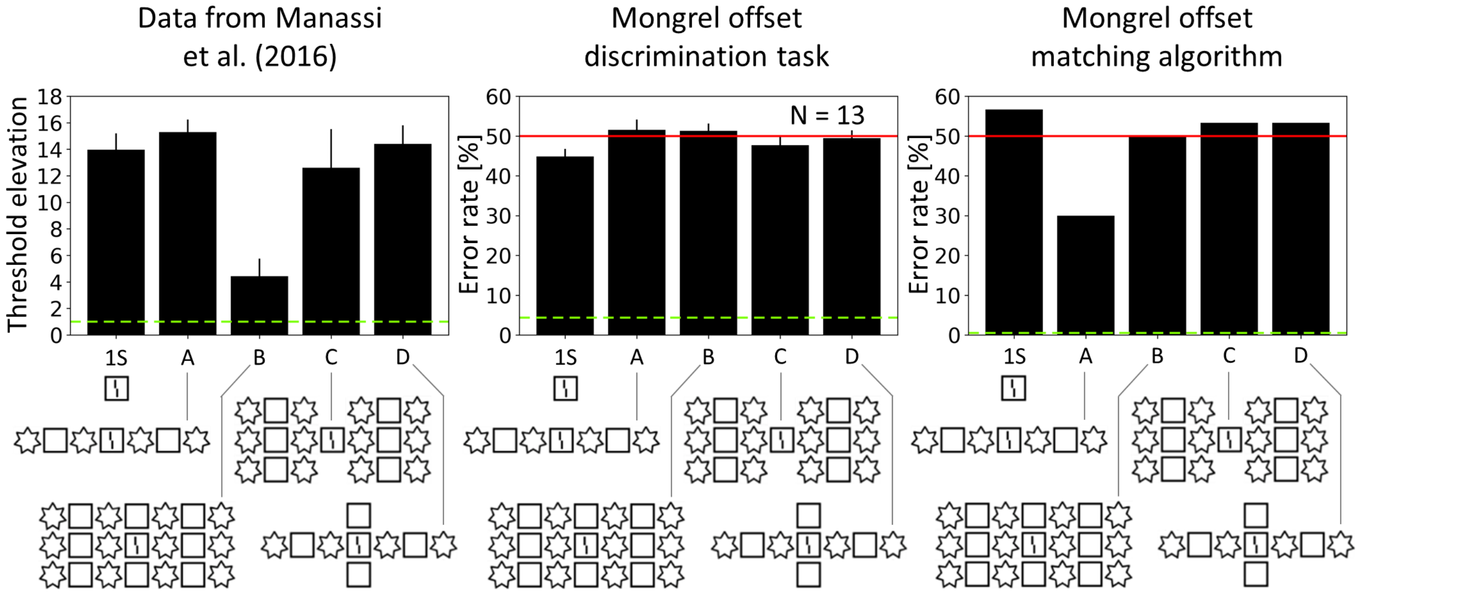

Figure 6. Patterns. Left. Data from Manassi et al. (2016). Offset discrimination thresholds were determined for vernier targets presented in the periphery at 9 degrees of eccentricity. Center. TTM validation in which observers discriminate between left and right offset verniers in mongrel images. Right. TTM validation with a template matching algorithm using the same mongrels as in the human experiment. Green dashed lines indicate vernier alone performance. Red lines indicate chance level (50% accuracy). Note that the y-axis labels are different.

Pattern experiment

In Manassi et al. (2016), crowding was strong when the vernier was embedded in a single square (see Figure 6, left, 1S). Crowding was still strong when the vernier was embedded in an array of alternating squares and stars, but strongly decreased when the vernier was embedded in three identical rows of alternating squares and stars (see Figure 6, left, A versus B). Crowding was strong in both control conditions (see Figure 6, left, C and D). These results showed that the high-level spatial configurations of elements across large parts of the visual field, well beyond the range attributed to local pooling (Bouma, 1970), affect vernier discrimination performance.

Again, the TTM failed to reproduce these results. In both TTM validation tasks, crowding was strong for all tested conditions (see Figure 6, center and right).

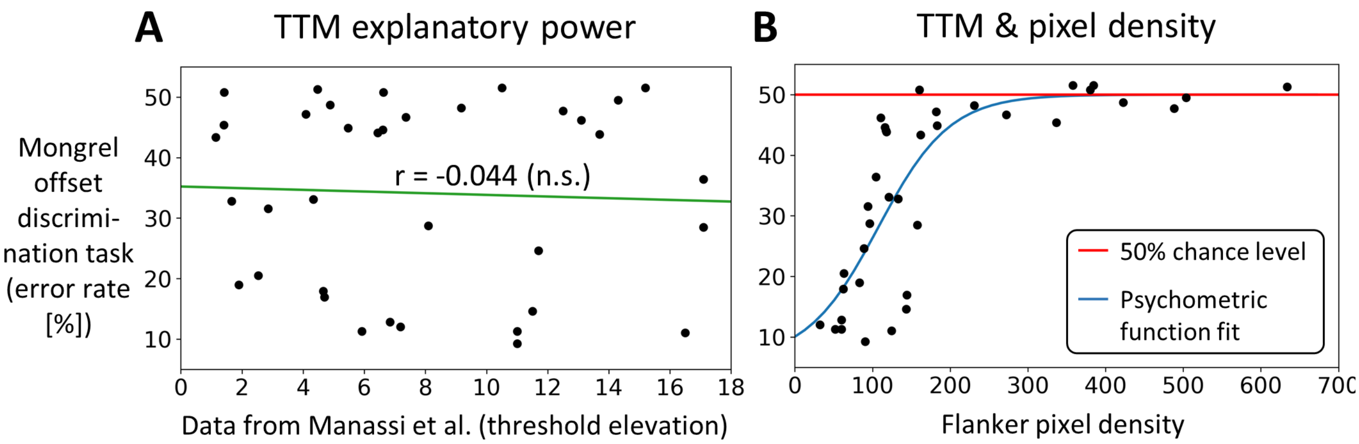

Figure 7. (**A**) TTM performance in the mongrel vernier offset discrimination task showed no correlation (r = −0.044, $p = 0.799$, $BF_{01} = 4.672$; Ly, Verhagen, & Wagenmakers, 2016; Rouder, Speckman, Sun, Morey, & Iverson, 2009) with the original data from (Manassi et al., 2012; Manassi et al., 2013; Manassi et al., 2015; Manassi et al., 2016). (**B**) TTM performance as a function of the sum of the flanker pixels in the corresponding conditions. Each dot indicates a flanking condition in Figure 1. The red line indicates chance level performance. For illustrative reasons, we plotted all tested conditions in a unique graph. Separate plots for all experiments are shown in the supplementary information (Supplementary Information Figure SF). Fitting the data with a psychometric function (see Equation 3 in Supplementary Information SL), we found a strong correlation between the TTM and the fitted performance (r(34) = 0.796, $p < 0.001$, $BF_{10} > 10^6$).

Note that, to avoid ceiling effects in which crowding is too high to show differences between conditions, we also generated mongrels with a larger foveal radius (32 instead of 16 pixels) for all conditions in the Shapes and Patterns experiments (i.e., the ones in Figures 5 and 6, as well as Supplementary Information Figure SB). We also computed the TTM performance for these mongrels, using the template matching algorithm. We obtained lower crowding levels, but a similar qualitative behavior was observed (Supplementary Information Figure SC).

Taken together, the results of the TTM matched none of the results of (Manassi et al., 2012; Manassi et al., 2013; Manassi et al., 2015; Manassi et al., 2016), which showed that: (1) increasing the number of flankers led to nonmonotonic effects (see Figure 2); (2) adding a single element drastically changed crowding behavior (see Figure 3; completion effect); (3) flanker configuration determined crowding (see Figure 4); (4) high-level processing determined low-level processing in crowding (see Figure 5); and (5) adding flankers beyond Bouma's window considerably affected crowding strength (see Figure 6). None of these effects were reproduced by the TTM.

TTM and prediction power

As a global measure of the explanatory power of the TTM for each condition of (Manassi et al., 2012; Manassi et al., 2013; Manassi et al., 2015; Manassi et al., 2016), we plotted the error rates (%) in the mongrel vernier offset discrimination task as a function of the threshold elevation in the original crowding experiments (Figure 7A). The measured correlation was not significantly different from zero (r(34) = −0.044; p value = 0.799), indicating that none of the reported results can be explained by the TTM. A similar correlation was found using the template matching algorithm (Supplementary Information Figure SE).

Second, to assess the TTM behavior, we plotted its performance for each condition as a function of the flanker “density” in the corresponding original stimulus images (Figure 7B). To compute the flanker density, we counted the number of flanker pixels around the target. Each pixel contribution was weighted by a function that decreased with the distance to the target, mimicking Bouma's law (Bouma, 1970). For each condition, the pixel density was defined as the sum of all weighted pixel contributions belonging to the flanker configuration (all details about the methods are given in Supplementary Information SL). The error rate increased with flanker density (see Figure 7B). Fitting the data with a psychometric function (see Equation 3 in Supplementary Information SL), we found a strong correlation between the TTM and the fitted performance (r(34) = 0.796, $p < 0.001$, $BF_{10} > 10^6$). Crucially, this is the exact result that would be expected using a simple pooling model, suggesting that the TTM is blind to complex stimulus configuration and grouping cues, and simply relies on pixel density.

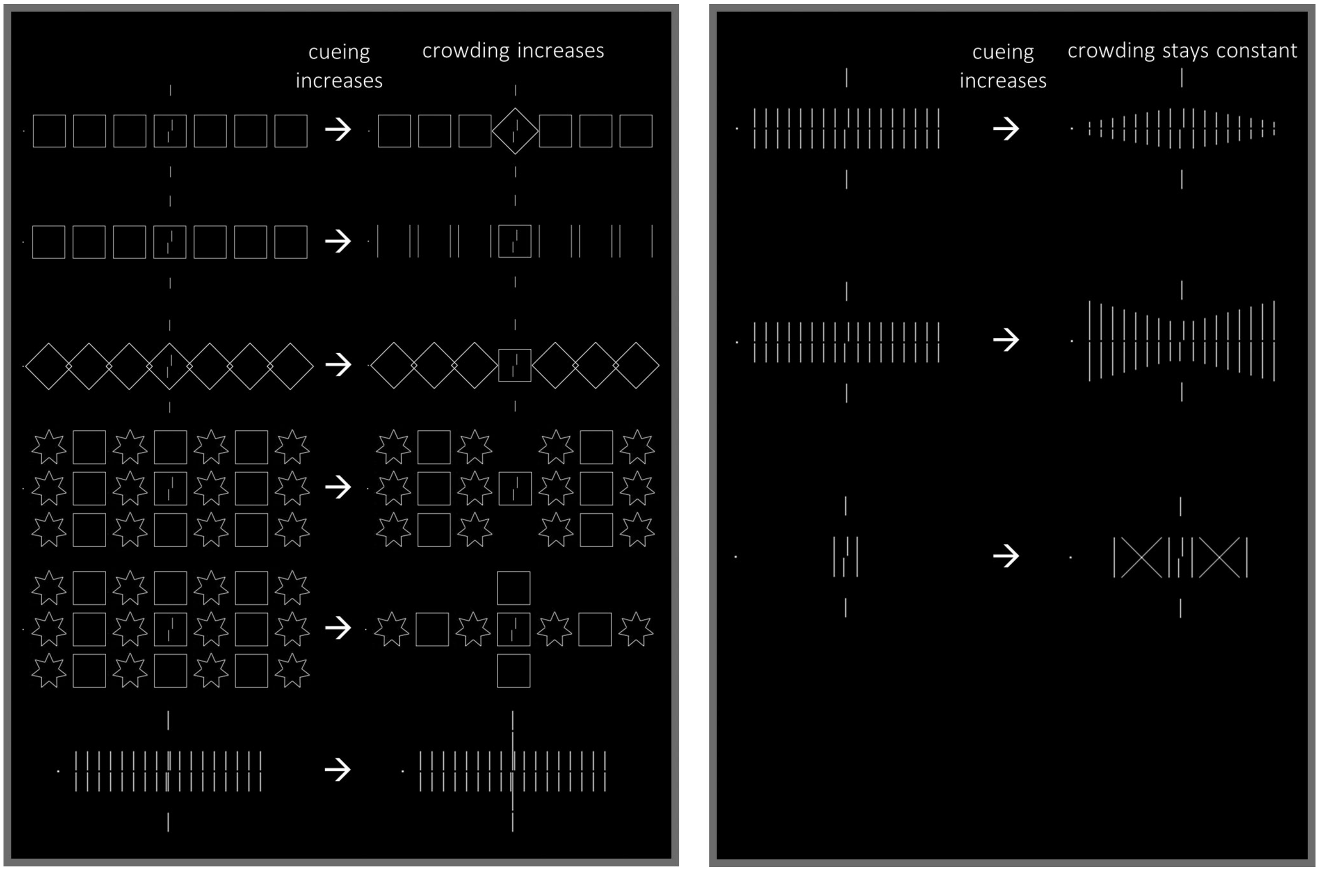

Figure 8. Right column, for both panels. Conditions in which the target location is weakly cued by the flanker configuration. Left Column, for both panels. Conditions in which the target location is strongly cued by the flanker configuration. If cueing had a strong impact on target discrimination performance, crowding would decrease from left to right in all comparisons. However, crowding strength either increases (left panel) or stays constant (right panel), while target cueing always increases. All conditions are taken from (Manassi et al., 2012; Manassi et al., 2013; Manassi et al., 2015; Manassi et al., 2016).

Grouping effects and target cueing

Rosenholtz et al. (2019) argued that the results in (Manassi et al., 2012; Manassi et al., 2013; Manassi et al., 2015; Manassi et al., 2016) do not necessarily imply the existence of grouping processes in crowding. Instead, it was proposed that target cueing plays a crucial role. Different stimulus configurations may cue the location of the target in different ways, thus reducing target location uncertainty, leading to differences in crowding strength. Importantly, this explanation is entirely based on post-perceptual decision-making mechanisms. This is not a viable explanation for four main reasons.

First, cueing does not explain the results of Manassi et al., 2012; Manassi et al., 2013; Manassi et al., 2015; Manassi et al., 2016). In these experiments, some flanker conditions strongly cue the target location but still produce strong crowding. In each comparison in Figure 8, the vernier target location is more cued by the flankers on the right side than on the left side. According to the cueing argument, crowding should be weaker on the right side compared to the left side. However, the human data show the exact opposite trend. For example, on the first line of the left panel in Figure 8, in the condition on the right (6S1D), the target location is clearly cued by the central diamond. There is no ambiguity at all about where the target is: it is inside the central diamond. In the condition on the left (7S), the line of squares casts more doubts on the location of the target. Nevertheless, crowding is 7.5 times larger on the right than on the left (Manassi et al., 2013).

Second, in (Manassi et al., 2012; Manassi et al., 2013; Manassi et al., 2015), two vertical lines were placed above and below the vernier target as “pointers,” in order to clearly cue the target location in all conditions. As reported in (Manassi et al., 2012; Manassi et al., 2013; Manassi et al., 2015), the aim was to minimize the target location uncertainty. Rosenholtz et al.

(2019) argued that these pointers may instead increase crowding by creating multiple offsets among vernier, flankers and pointers lines (see figure 17 in Rosenholtz et al., 2019). However, in (Manassi et al., 2012; Manassi et al., 2013; Manassi et al., 2015), the pointers are actually quite far from the vernier, making this offset confusion argument unlikely (see Supplementary Information Figure SG). Moreover, we measured the performance of the TTM model with all conditions, with or without pointers. The model did not show any significant increase in crowding strength with the pointers (Supplementary Information Figure SH).

Third, the effects measured in (Manassi et al., 2012; Manassi et al., 2013; Manassi et al., 2015; Manassi et al., 2016) correspond to changes in threshold elevation up to 10 times the unflanked threshold. The strength of cueing effects in the literature has been consistently reported as small, with an average of 10% to 20% of difference in performance (Nazir, 1992; Scolari, Kohnen, Barton, & Awh, 2007; Wilkinson et al., 1997; Yeshurun & Rashal, 2010). Thus, cueing does not seem even remotely sufficient to be considered as a viable explanation for global effects in crowding.

Fourth, a large part of these grouping effects in visual crowding were also found in foveal vision (Malania et al., 2007; Sayim, Westheimer, & Herzog, 2008; Sayim, Westheimer, & Herzog, 2010; Waugh & Formankiewicz, 2020), where uncertainty is greatly reduced. Rosenholtz et al. (2019) argued that evidence for grouping effects in foveal vision casts doubts on whether these results are due to crowding. However, old and recent literature has shown evidence for crowding in foveal vision (Coates, Chin, & Chung, 2013; Coates, Levi, Touch, & Sabesan, 2018; Danilova & Bondarko, 2007; Flom, Heath, & Takahashi, 1963; Lev, Yehezkel, & Polat, 2014; Lev & Polat, 2015; Sayim, Greenwood, & Cavanagh, 2014; Siderov, Waugh, & Bedell, 2013; Westheimer & Hauske, 1975; but see Levi, Hariharan et al., 2002; Levi, Klein, & Hariharan, 2002), as well as grouping processes acting in foveal (Banks & White, 1984; Bock, Monk, & Hulme, 1993; Tannazzo, Kurylo, & Bukhari, 2014) and peripheral vision (Banks & Prinzmetal, 1976; Banks & White, 1984; Livne & Sagi, 2007; Tannazzo et al., 2014; Wolford & Chambers, 1983). In other words, showing evidence for grouping effects in foveal vision does not invalidate any claim about grouping effects in crowding, but instead strengthens them.

To sum up, post-perceptual cueing cannot account for the effects measured in (Manassi et al., 2012; Manassi et al., 2013; Manassi et al., 2015; Manassi et al., 2016). These effects must hence be yielded by more complex interactions than what was previously thought to happen in visual crowding, such as contextual grouping (Malania et al., 2007; Manassi et al., 2012; Saarela, Sayim, Westheimer, & Herzog, 2009).

TTM and face crowding

In the previous section, we showed that the TTM cannot explain the grouping effects found in (Manassi et al., 2012; Manassi et al., 2013; Manassi et al., 2015; Manassi et al., 2016) and that these effects cannot be explained by post-perceptual cueing. In this section, we tested the TTM with holistic face perception. Faces are considered as an invaluable tool to probe high-level visual processing, as they are analyzed holistically rather than as a set of separate features (Sergent, 1984). Mooney faces (Mooney, 1957), in particular, are the gold standard stimulus to test for holistic processing. Mooney faces (Figure 9) are two-tone shadow images that are readily perceived as faces despite the lack of bottom-up processes that can segment or parse the image into features like an eye or mouth (Cavanagh, 1991; Fan, Wang, Shao, Zhang, & He, 2020; Grützner, Uhlhaas, Genc, Kohler, Singer, & Wibral, 2010). That is, to see the mouth, eye, nose, eye separation, or other features, one must first recognize the stimulus as a face. This kind of holistic processing is necessary to recognize Mooney faces, and it has been operationalized in the literature by the inversion effect (McKone, 2004; Taubert, Apthorp, Aagten-Murphy, & Alais, 2011): upright faces are recognized more easily than inverted ones (Farah, Tanaka, & Drain, 1995; Kanwisher, Tong, & Nakayama, 1998; Latinus & Taylor, 2005; Rossion, 2008; Sergent, 1984; Yin, 1969). The inversion effect is especially strong for Mooney faces (Canas-Bajo & Whitney, 2020; McKone, 2004; Schwiedrzik, Melloni, & Schurger, 2018). Here, we tested the TTM with Mooney faces and found that it cannot predict two main results in holistic processing in crowding: (a) crowded object information is not lost at early stages

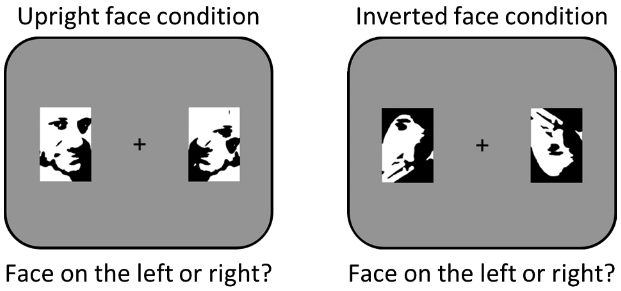

Figure 9. **Single face discrimination task.** Observers were asked to discriminate which of the two images was a face (left or right, 2AFC), by pressing the left or right arrow, while fixating the central cross. Across the experiment, the face could be either upright or inverted. In these examples, an upright face is presented on the left side (left panel), and an inverted face is presented on the right side (right panel). Mooney faces reprinted from Schwiedrzik et al. (2018). Distributed under a CC-BY license.

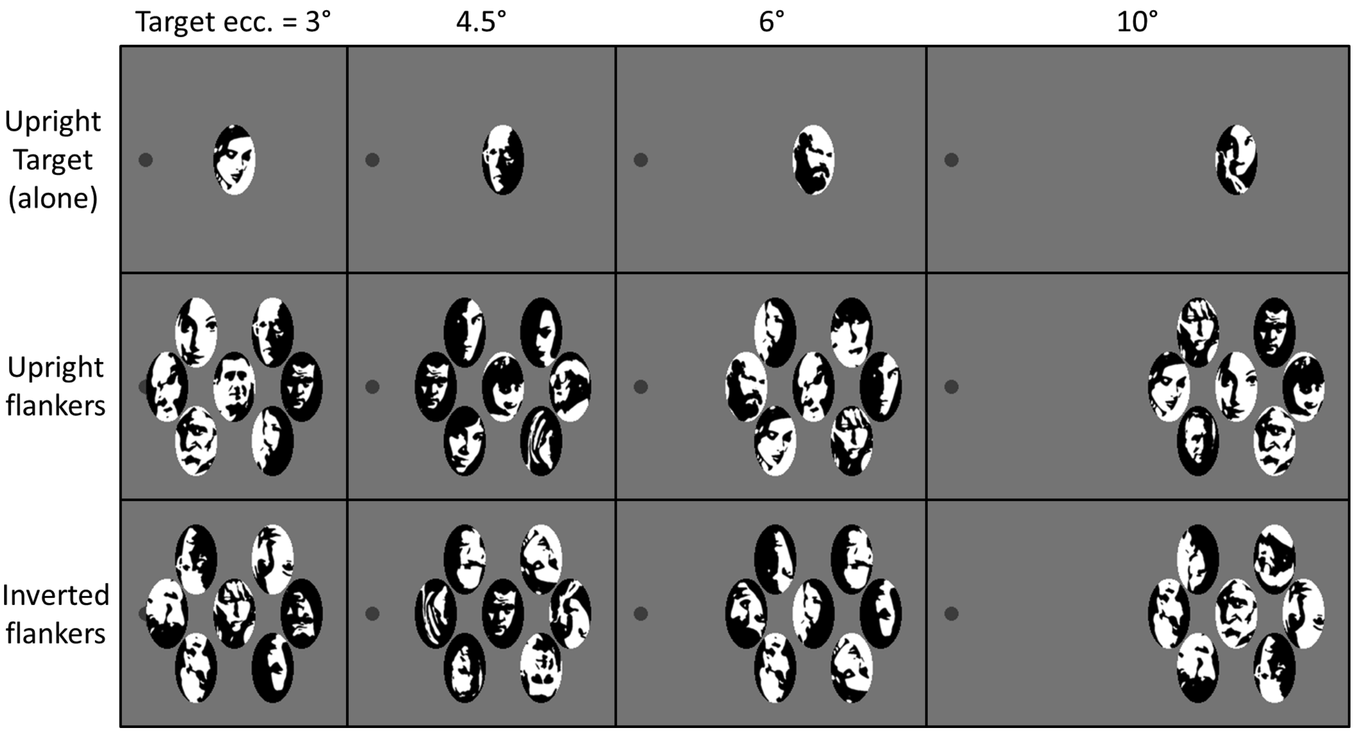

Figure 10. Examples of stimuli used in the face crowding task. There were three main conditions (upright target alone, target with upright flankers or target inverted flankers) presented at four different eccentricities. Mooney faces reprinted from Schwiedrzik et al. (2018). Distributed under a CC-BY license.

of visual processing (inversion effect in a single face discrimination task; Bayle, Schoendorff, Hénaff, & Krolak-Salmon, 2011; Boucart, Lenoble, Quettelart, Szaffarczyk, Despretz, & Thorpe, 2016; McKone, 2004) and (b) crowding occurs at high-level stages of visual processing between faces (crowding between holistic face representations; Farzin et al., 2009; Louie et al., 2007; Manassi & Whitney, 2018; Sun & Balas, 2015).

Methods

Single face discrimination task

We reproduced the single face discrimination task of Canas-Bajo & Whitney (2020). Observers were shown two images, one on each side of the visual field (see Figure 9). Both images subtended a visual angle of 6 degrees by 4.2 degrees and were presented at the same eccentricity on both sides (6 degrees, 10 degrees, 14 degrees, or 18 degrees). One image was always a Mooney face, whereas the other one was always a scrambled version of the same face. Mooney faces were taken from Schwiedrzik et al. (2018), with permission (freely available at https://doi.org/10.6084/m9.figshare.5783037). The face could either be upright or inverted. Observers' task was to discriminate which of the two images was a face by pressing the left or right arrow on a keyboard (2AFC), while fixating a cross in the center of the screen. The position on which the face appeared was randomized on each trial (either a face on the right and the corresponding scrambled face on the left or vice versa). There was no time constraint for giving a response, as unlimited viewing time has no effect on crowding (Wallace, Chiu, Nandy, & Tjan, 2013). The distance to the screen was 64 cm.

There were five different faces, for a total of 20 different stimuli per eccentricity (2 sides, 2 face orientations, and 5 different faces). Every stimulus was shown 10 times for a total of 200 trials per eccentricity. The experiment was run in blocks of fixed eccentricities. In each block, the stimuli were shown in a random left/right order. For each condition (upright versus inverted face) and eccentricity, we computed discrimination performance (error rate = 1-accuracy) and the corresponding standard error of the mean, computed over human observers.

In order to validate the TTM, we tested mongrel images with the same single face discrimination task as in Canas-Bajo and Whitney (2020). For each stimulus, 10 different mongrels were generated using the TTM. Face discrimination performance in mongrel images was quantified by performing the single face discrimination task in free-viewing conditions. The experiment was run by blocks of eccentricity, for a total of 200 mongrels shown per eccentricity. Seven observers (2 men and 5 women, 25.4 ± 1.2 years old) performed the task. For each condition (upright versus inverted face) and eccentricity, we computed discrimination performance (accuracy [%]) and the corresponding standard error of the mean computed across observers. Performance in the single face discrimination task was then compared to the mongrel validation task.

In addition, we measured the TTM performance for each condition with a template matching algorithm (Supplementary Information Figure SJ). As for the Vernier offset matching algorithm, a face target was searched in the mongrels by sliding target face templates over the image. The algorithm answered either left or right, as the side of the image on which the best matching score was obtained over all possible target face templates (see Equation 1 for the detailed computation). Accuracy was defined as the percentage of correct answers.

Gender face discrimination task

Mongrel images were generated, following experiment 6 from Farzin et al. (2009), which measured crowding induced by Mooney face flankers in a gender face discrimination task. Mooney faces were taken from Schwiedrzik et al. (2018), with permission (freely available at https://doi.org/10.6084/m9.figshare.5783037). The size of the faces was the same as in Farzin et al. (2009; i.e., 1.53 degrees by 2.48 degrees). In these stimuli, the target face, which was always presented upright, could either be alone or surrounded by six other randomly selected Mooney faces (Figure 10). Flankers could either be upright or inverted. There were three different flanking conditions (target alone, upright flankers, and inverted flankers) and four different target eccentricities (3 degrees, 4.5 degrees, 6 degrees, and 10 degrees). Compared to the original experiment, we had an additional eccentricity (4.5 degrees) in order to avoid floor and ceiling effects in the mongrel discrimination task. For each condition and eccentricity, 20 different Mooney faces were used as target (split equally between males and females), for a total of 240 original stimuli (20 faces × 3 flanking conditions x 4 eccentricities). Ten different mongrels were generated for each stimulus, for a total of 2400 unique samples shown to every participant. Seven observers (2 men and 5 women, 25.4 ± 1.2 years old) performed the task.

Crowding strength in the TTM was quantified by performing a gender discrimination task in free-viewing conditions. We presented the generated mongrel images and asked observers to indicate the gender of the target face (2AFC task). Mongrels were shown in a randomized order. Prior to the experiment, observers familiarized with the task as in the mongrel vernier offset discrimination task described above.

For each condition and eccentricity, we computed the discrimination performance (accuracy [%]) and the corresponding standard error of the mean computed across observers. Performance in the mongrel gender crowding discrimination task was then compared to the behavioral data of Farzin et al. (2009).

In addition to the behavioral experiment, we measured the gender discrimination performance with a template matching algorithm. The algorithm matched original target face templates to all mongrel images. As for the vernier offset matching algorithm, a face target was searched in the mongrels by sliding target face templates over the image (see Equation 1 for the detailed computation). For each mongrel, the algorithm outputted the gender of the target face template that had the best match. Accuracy was computed as the percentage of correct answers. The performance of the algorithm was also compared to the data of Farzin et al. (2009; Supplementary Information Figure SK).

Results

Single face discrimination task

The results of the single face discrimination task are plotted in terms of accuracy (Figure 11A). Data were analyzed using a linear mixed effect model, with eccentricity and face orientation as the two fixed effects and individual subjects as a random intercept. The two fixed effects showed no significant interaction ($\chi^2(1) = 0.062$, $p = 0.803$). The main effect of face orientation was significant ($\chi^2(1) = 30.99$, $p < 0.001$), but not the effect of eccentricity ($\chi^2(1) = 0.755$, $p = 0.385$). The difference in effect size between the full model and the reduced model, excluding the effect of eccentricity, was only 0.4% (full model: $\mathrm{r_m}^2 = 0.243$ and $\mathrm{r_c}^2 = 0.696$ and the reduced model: $\mathrm{r_m}^2 = 0.239$ and $\mathrm{r_c}^2 = 0.692$).

Observers were able to discriminate an upright/inverted face from a scrambled face at all tested eccentricities (see Figure 11A). Crucially, observers' accuracy was higher for upright than inverted faces (see Figure 11A, upright versus inverted), indicating a differential processing of inverted (low-level) and upright (holistic) faces, even at 18 degrees of eccentricity. The results suggest that face representations can survive any putative within-face low-level crowding, allowing holistic recognition of Mooney faces in the periphery.

Next, we tested whether the TTM could predict the inversion effect in individual Mooney faces (see Figure 11B). As before, we validated the mongrels with the single face discrimination task. Observers were shown the mongrels of the original stimuli and were asked to tell which mongrel image was a face (free unconstrained viewing; see Methods section for details). Data were analyzed using a linear mixed effect

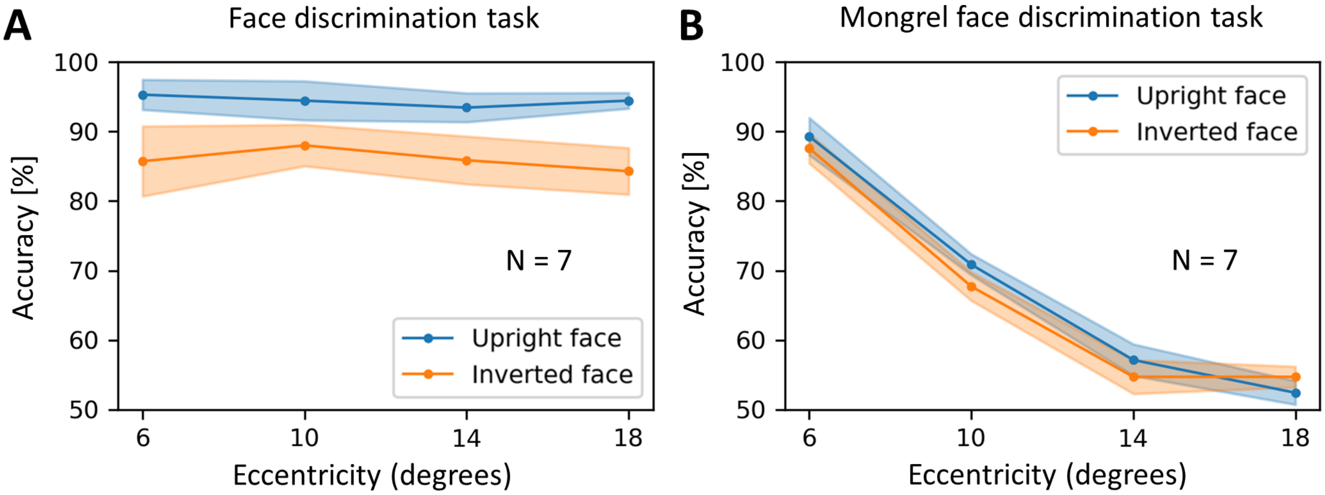

Figure 11. TTM and single Mooney face discrimination. (**A**) Face discrimination task. Observers were asked to discriminate an upright/inverted face from a scrambled face at all tested eccentricities. Accuracy remained on a constant high level for all eccentricities. Crucially, accuracy was higher for upright than for inverted faces. (**B**) Mongrel face discrimination task. Accuracy decreased with increasing eccentricity, contrary to the behavioral results. Using a linear mixed effect model, no significant difference between the upright and inverted face conditions was observed (i.e., no significant effect of face orientation on model performance). Shaded regions indicate the standard error of the mean.

model, with eccentricity and face orientation as the two fixed effects and individual subjects as a random intercept. The two fixed effects showed no significant interaction ($\chi^2(1) = 0.647$, $p = 0.421$). The main effect of eccentricity was significant ($\chi^2(1) = 88.779$, $p < 0.001$), but the effect of face orientation was not ($\chi^2(1) = 0.494$, $p = 0.482$). The difference in effect size between the full model, including both effects and the reduced model excluding the effect of face orientation, was only 0.2% (full model: $r_m^2 = 0.798$ and $r_c^2 = 0.802$ and the reduced model: $r_m^2 = 0.796$ and $r_c^2 = 0.800$).

These results show that the face discrimination performance in the TTM decreased with increasing eccentricity, contrary to the behavioral results (see Figure 11, A versus B). More importantly, there was no difference between the upright and inverted mongrel face conditions (see Figure 11B, orange versus blue). The lack of inversion effect shows that the high-dimensional pooling stage of the TTM does not preserve rich enough information to support holistic processing in a later post-perceptual stage, as suggested by Rosenholtz et al. (2019).

We ran another version of the mongrel validation task in which all mongrels generated with images comprising an inverted face were flipped upside-down. Hence, in this control task, observers were only shown upright mongrel faces, although they were processed either as upright or inverted faces in the TTM. This was done to isolate inversion effects in humans from inversion effects in the TTM as much as possible. The results were comparable (Supplementary Information Figure SI). Moreover, we also quantified the TTM performance using a template matching algorithm and obtained qualitatively similar results (Supplementary Information Figure SJ).

Taken together, the results show that holistic face recognition occurs also in peripheral vision, replicating and extending previous reports (Bayle et al., 2011; Boucart et al., 2016; Canas-Bajo & Whitney, 2020; McKone, 2004). Hence, crowded face-specific information is not lost at the early stages of visual processing but can be easily retrieved (see Figure 11A). The TTM cannot explain this class of results. The TTM causes an irretrievable loss of face-specific information: discrimination performance drops with eccentricity and the inversion effect is eliminated (see Figure 11B).

Gender face discrimination task

In Farzin et al. (2009), observers were asked to discriminate the gender of an upright face presented in the periphery. Accuracy decreased with increasing eccentricity (see Figure 12A, black line). This decline in performance for isolated faces is an unsurprising consequence of the small size of the faces and the difficulty of the gender discrimination task. More importantly, when the same upright face was flanked by inverted or upright flankers, accuracy decreased, a standard hallmark of crowding. Crucially, upright flankers crowded more compared to inverted ones (blue line falls below orange line). This is an inversion effect in crowding: it shows that stimuli seen as faces crowd

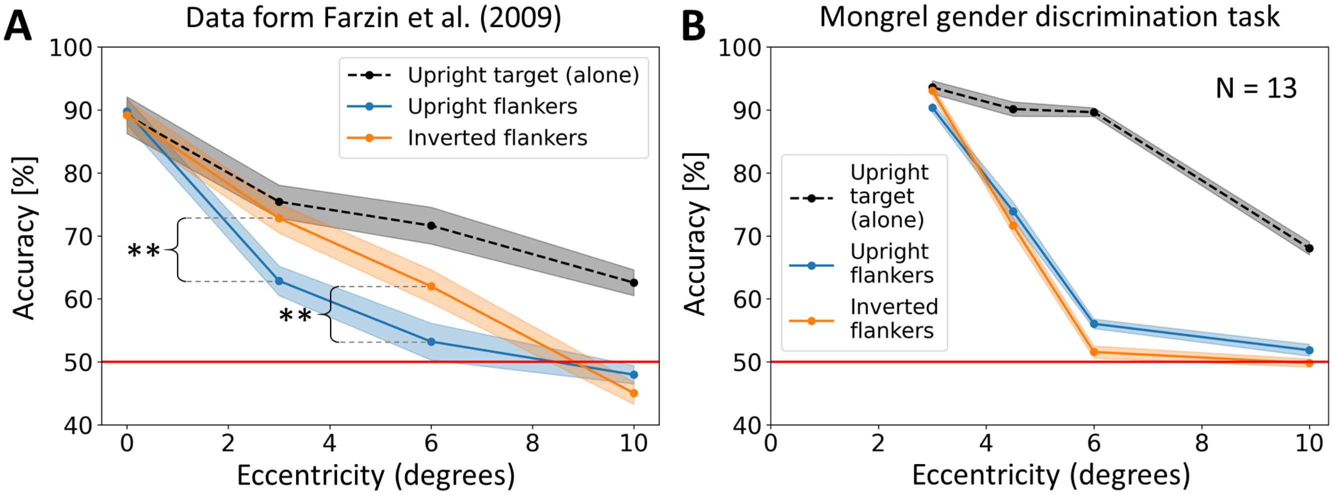

Figure 12. TTM and crowding of Mooney faces. (**A**) Face crowding task, data from Farzin et al. (2009). Target discrimination performance decreased when eccentricity increased. When the target face was flanked by inverted faces, crowding increased with increasing eccentricity (orange). When the target was flanked by upright faces, crowding increased even more with eccentricity (blue). Shaded regions indicate the standard error of the mean. Stars indicate a significant difference in crowding strength between the upright and inverted flanker face conditions (paired Student *t*-test, 2-tails). (**B**) Mongrel face crowding task. Accuracy decreased with eccentricity. When analyzing the results using a linear mixed effect model, no effect of flanker face orientation was exposed. Shaded regions indicate the standard error of the mean.

each other. When the same flanker stimuli are not seen as faces (i.e., are inverted), they do not crowd. Crowding is therefore gated by “similarity,” and the “similarity” must be at the level of holistic face representations. In the original publication (see Experiment 6 in Farzin et al., 2009), ANOVA resulted in a significant main effect of eccentricity and flanker orientation (paired-samples 2-tailed *t*-tests revealed that upright face flankers impaired performance more than inverted flankers at 3 degrees and 6 degrees of eccentricity). Here we tested whether the TTM makes a similar prediction.

We computed the TTM performance for this experiment in a mongrel gender discrimination task (see Methods section for details, gender face discrimination task). The results (see Figure 12B) were analyzed using a linear mixed effect model, with eccentricity and face orientation (upright versus inverted) as fixed effects and individual observers as a random intercept. The two fixed effects showed no significant interaction ($\chi^2(1) = 0.479$, $p = 0.489$). The main effect of eccentricity was significant ($\chi^2(1) = 121.11$, $p < 0.001$), but the effect of face orientation was not ($\chi^2(1) = 0.620$, $p = 0.431$). The difference in effect size between the full model, including both effects (eccentricity and face orientation) and the reduced model excluding the effect of face orientation, was only 0.2% (full model: $r_m^2 = 0.691$ and $r_c^2 = 0.691$ and the reduced model: $r_m^2 = 0.689$ and $r_c^2 = 0.689$).

As in Farzin et al. (2009; see Figure 12A), TTM performance decreased with eccentricity (see Figure 12B). However, unlike Farzin et al. (2009), the linear mixed effect model revealed no significant overall effect of flanker orientation, and no interaction between eccentricity and target orientation. Simply put, the TTM does not predict a systematic difference in crowding as a function of the flanker orientation. In addition, when TTM does predict a trending difference, it is often in a direction opposite that in the empirical data (blue-above-orange in Figure 12B compared to orange-above-blue in Figure 12A). We also quantified the TTM performance using a template matching algorithm and obtained qualitatively similar results (Supplementary Information Figure SK). These results show that the TTM can predict a general increase of crowding with eccentricity (i.e., low-level crowding) but it fails to predict face-selective or holistic effects in crowding.

Taken together, the results depicted in Figures 11 and 12 show that the TTM is not able to predict peripheral face recognition or the effects of high-level face processing in crowding. It fails to predict crowding of single faces (see Figure 11) and multiple faces (see Figure 12). In fact, target information in the TTM is irretrievably lost at a low-level pooling stage and crowding occurs only between low-level features (see Figure 7). In this light, the TTM may fail to explain a broad array of findings in the peripheral face recognition literature (Boucart et al., 2016; Farzin et al., 2009; Kovács, Knakker, Hermann, Kovács, & Vidnyánszky, 2017; Kreichman, Bonneh, & Gilaie-Dotan, 2020).

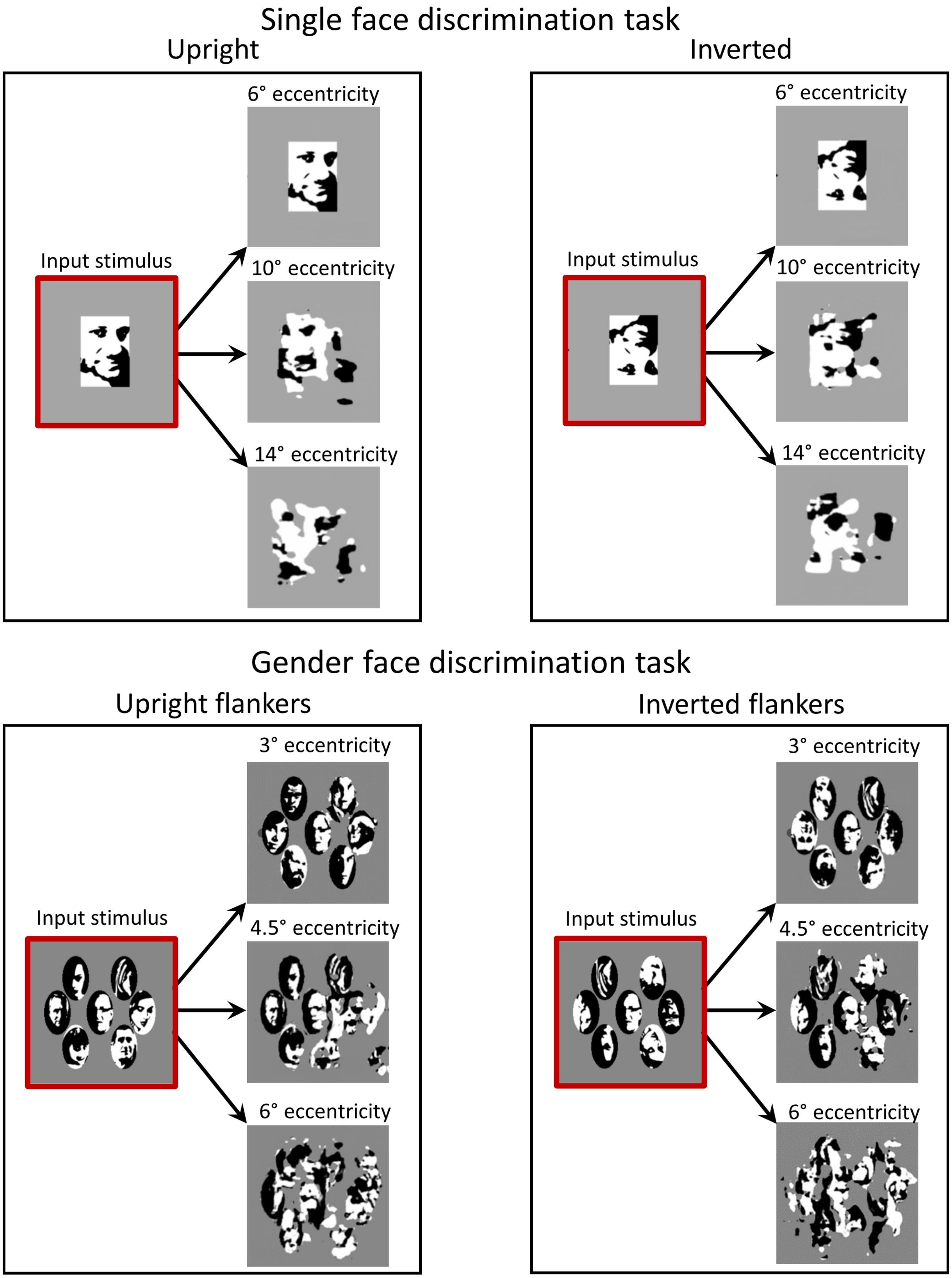

Figure 13. **TTM mongrel examples used in the single face and gender face discrimination tasks.** The stimuli (TTM input) are highlighted in red. To give a representative sample of the TTM outputs for each example, we show mongrels for different eccentricities. Note that we cropped the mongrels for ease of comparison. All mongrels can be found at https://github.com/albornet/TTM_Verniers_Faces_Mongrels. Mooney faces reprinted from Schwiedrzik et al. (2018). Distributed under a CC-BY license.

Discussion

Classic models describe crowding as a local and low-level phenomenon (Greenwood, Bex, & Dakin, 2009; Levi, Hariharan et al., 2002; Nandy & Tjan, 2012; Parkes et al., 2001; Van den Berg, Roerdink, & Cornelissen, 2010; Wilkinson et al., 1997). Recent studies, however, provided clear-cut psychophysical evidence that crowding is in fact more complex than previously thought, involving global interactions and occurring at multiple stages of visual processing (Farzin et al., 2009; Manassi et al., 2012; Manassi et al., 2013; Manassi et al., 2015; Manassi et al., 2016; Manassi & Whitney, 2018; Saarela et al., 2009; Saarela, et al., 2010; Whitney & Levi, 2011; for older studies about high-level effects in crowding, see also Banks et al., 1979; Banks & White, 1984; Egeth & Santee, 1981; Huckauf et al., 1999; Mason, 1982; Mewhort et al., 1982; Wolford & Chambers, 1983). More recently, it was shown that crowding is affected by emotional conditioning of the flankers (Pittino, Eberhardt, Kurz, & Huckauf, 2019) or by the high-level semantic information of visual scenes (Gong, Xuan, Smart, & Olzak, 2018). This large body of evidence for high-level effects in crowding suggest that current models of vision need to be radically updated. However, against this view of crowding, Rosenholtz et al. (2019) argued that (1) high-dimensional pooling is sufficient to explain the new results and (2) target cueing plays a crucial role in these effects. Here, we quantitatively tested these claims on a large array of experimental data and showed that (1) TTM fails to account for human crowding performance and (2) target cueing does not play a role.

Importantly, the current work is not about the TTM only. Instead, it asks the question whether a sophisticated pooling stage can preserve rich enough information about the stimulus to drive the global aspects of crowding in a post-perceptual stage. This argument has implications that go beyond a simple model controversy. For example, global configuration does not need to affect low-level information if Rosenholtz et al. (2019) is correct. In the following, we describe implications from our two sets of data on grouping effects and face recognition.

TTM and grouping effects

Using a mongrel offset discrimination task, we showed that the TTM did not reproduce any of the results of (Manassi et al., 2012; Manassi et al., 2013; Manassi et al., 2015; Manassi et al., 2016), in which: (1) increasing the number of flankers sometimes reduces crowding strength (see Figure 2); (2) adding a single element has a dramatic effect on crowding strength (see Figure 3; completion effect); (3) the overall configuration of the flankers determines crowding (see Figure 4); (4) high-level processing strongly affects low-level processing (see Figure 5), and (5) adding flankers beyond Bouma's window strongly modulates crowding strength (see Figure 6).

It was proposed that the best predictor of visual crowding is grouping between target and flankers: crowding increases when the target groups with the flankers, but decreases when the target ungroups and stands out from the flankers (Malania et al., 2007; Saarela et al., 2009; Saarela et al., 2010; Sayim et al., 2008, 2010). In line with this hypothesis, in Manassi et al. (2012) and in Saarela et al. (2009), subjective ratings on target-flankers grouping correlated with crowding strength. Furthermore, Doerig et al. (2019) showed that only models that included a grouping stage could explain these results (see also Doerig, Schmittwilken, Sayim, Manassi, & Herzog, 2020). In the TTM, crowding strength was never reduced, when additional flankers were added, regardless of flanker configuration (see Figures 2 to 6).

The only result that was reproduced by the TTM is the reduction in crowding strength when adding a straight-vernier mask at target location in the Completion experiment (see Figure 3, center and right, straight versus comp16). We attribute this reduction in crowding strength to a local effect of the mask. When the mask is added, the region around the target is summarized by different local statistics than when the mask is absent (higher spatial frequencies, locally). Hence, this region stands out from the rest of the image. It is thus better reconstructed by the TTM, yielding better performance. However, crowding in the TTM was still reduced in the control conditions (see Figure 3, center and right, comp16b and comp2), further supporting the notion that the mask induces a local effect only: when the configuration of the grating is broken by the presence of the long mask (comp16b) or by the absence of many flankers (comp2), crowding is still reduced. This is in contradiction to the human data, in which crowding is reduced by the global layout of the flankers. In addition, crowding strength with various numbers of same length flankers (see Supplementary Information Figure SA), was always weaker with than without the mask and always increased with more flankers, contrary to the human data.

Taken together, these results suggest that a pooling model, even a high-dimensional one, cannot account for the complexity of visual crowding. Comparing the performance of the TTM for all tested conditions to the corresponding human performance measured in Manassi et al., 2012; Manassi et al., 2013; Manassi et al., 2015; Manassi et al., 2016), we found no significant correlation (see Figure 7A). Moreover, we found that the TTM performance strongly correlates with the amount of flankers around the target (see Figure 7B), similar to a simple pooling model. Of course, this does not mean that the TTM is only measuring flanker pixel density, but rather that this factor is crucial in driving crowding strength and stimulus appearance in the TTM. Still, it seems that the TTM is blind to complex configurations and grouping cues. We propose that the main reason for this lies in the model architecture (i.e., feedforward pooling cannot explain high-level effects in crowding; Bornet, Doerig, Herzog, Francis, & Van der Burg, 2021; Doerig, Bornet, Choung, & Herzog, 2020; Doerig et al., 2019; Doerig, Schmittwilken et al., 2020; Choung, Bornet, Doerig, & Herzog, 2021).

There are several other reasons why the TTM failed. First, elements outside the pooling regions of the TTM can change crowding performance in humans but not in the TTM. Second, the strength of the TTM is the compression of information implemented by the computation of summary statistics, which may play a role for grouping. However, the TTM does not allow to change the scale of the pooling regions in function of the specificities of the stimuli. For this reason, the TTM filters out fine-grained information that is crucial for human performance. As expressed by Wallis, Funke, Ecker, Gatys, Wichmann, and Bethge (2017), "Based on our experiments we speculate that the concept of summary statistics cannot fully account for peripheral scene appearance. Pooling in fixed regions will either discard (long-range) structure that should be preserved or preserve (local) structure that could be discarded. Rather, we believe that the size of pooling regions needs to depend on image content." We think that the TTM summary statistics are important in crowding but need to adapt to the stimulus global configuration (including feedback processing) and not hard-wired.

Importantly, in contrast to what was proposed by Rosenholtz et al. (2019), cueing cannot account for grouping effects in crowding. Cueing may be an explanation for some configurations, but overall, it is a poor predictor of crowding strength (see Figure 8). Moreover, cueing studies only report small effect sizes (Nazir, 1992; Scolari et al., 2007; Yeshurun & Rashal, 2010), far beneath the effect sizes measured in (Manassi et al., 2012; Manassi et al., 2013; Manassi et al., 2015; Manassi et al., 2016). Hence, grouping effects in crowding are not post-perceptual (e.g., caused by differences in target visibility or target cueing). They are purely perceptual and are caused by complex target-flanker interactions occurring along the visual processing hierarchy.

Rosenholtz et al. (2019) argued that, because effects of contextual grouping were also found in foveal vision (Saarela & Herzog, 2008; Sayim et al., 2010; Sayim, Westheimer, & Herzog, 2011; Sayim, Manassi, & Herzog, 2014; Waugh & Formankiewicz, 2020), they may not be due to genuine crowding. However, literature showed that crowding can occur in foveal (Coates et al., 2013; Coates et al., 2018; Danilova & Bondarko, 2007; Flom et al., 1963; Lev et al., 2014; Lev & Polat, 2015; Sayim, Greenwood et al., 2014; Siderov et al., 2013; Westheimer & Hauske, 1975) and peripheral vision (Levi, 2008; Pelli, 2008). Importantly, the stimuli in foveal experiments were the same as in peripheral crowding and so were the results. In any case, the TTM needs either to explain the peripheral effects, independent of where or not there is foveal crowding, or to convincingly explain why not.

TTM and face crowding

In another set of experiments (see Figures 9 and 11), we focused on single face recognition in peripheral vision. Using a single Mooney face discrimination task, we showed that holistic face recognition occurs in peripheral vision (i.e., a better recognition performance for upright than for inverted faces; see Figure 11A, upright versus inverted), reproducing the results found in Canas-Bajo & Whitney (2020) and in line with old and recent literature (Farah et al., 1995; Rossion, 2008; Sergent, 1984; Yin, 1969). The advantage in recognizing upright Mooney faces speaks for a differential processing involved between inverted (low-level) and upright (holistic) faces. These results cannot be explained by models of crowding based on simple pooling. According to this class of models, the two-tone black and white blobs constituting a Mooney face should crowd themselves in peripheral vision (e.g., see Figure 11B), thus becoming more unrecognizable when increasing in eccentricity (Martelli, Majaj, & Pelli, 2005). Instead, our results show that the representation of these object parts nevertheless survives crowding (see also Manassi & Whitney, 2018), allowing holistic recognition of Mooney faces.

Using a mongrel Mooney face discrimination task, we showed that the low-level visual information that would allow to discriminate a face from a non-face object is irretrievably lost in the pooling stage of the TTM. Despite the high dimensionality of the pooling in the TTM, at increasing eccentricities the features that compose the faces crowd each other in the model and cannot be used for further processing in the mongrel face discrimination task (see Figure 11B). This is in contradiction with the results of the single face discrimination task we performed (see Figure 11A; Canas-Bajo & Whitney (2020), and with recent evidence that stimulus information on several levels of visual processing can survive crowding and influence subsequent perceptual judgments (Faivre & Kouider, 2011a; Faivre & Kouider, 2011b), including face-level information (Kouider, Berthet, & Faivre, 2011).

Next, we focused on holistic face crowding (as found in Experiment 6 of Farzin et al., 2009; see Figure 12A), in which upright flanker faces yielded more crowding than inverted ones in a gender face discrimination task. This inversion effect showed that crowding can occur selectively between high-level holistic representations conveyed by Mooney faces.

Rosenholtz et al. (2019) suggested that the TTM could predict these results without requiring high-level feature interactions. Instead, holistic effects might be driven, in a post-perceptual stage, by the rich information that survives high-dimensional pooling in the TTM.

We tested this hypothesis in practice. Using a mongrel gender crowding discrimination task (see Figure 10), we showed that the TTM did not reproduce holistic face crowding (see Figure 12B). Although crowding occurred in the TTM when face flankers were added, there was no effect of flanker face orientation on the TTM performance. In other words, the high-dimensional pooling stage of the TTM did not preserve enough information to drive holistic processing in a post-perceptual stage. This result gives more support to the hypothesis that crowding happens selectively between high-level representations and cannot arise from low-level accounts, even using a high-dimensional pooling stage.

It was recently argued that the face crowding results in Farzin et al. (2009) may be due to differences in flankers reportability (Reuther & Chakravarthi, 2019; Rosenholtz et al., 2019). When target and flankers belong to the same category (upright faces as target and flankers), crowding may arise in part from reporting the flankers' gender instead of the target one (substitution errors). However, when target and flankers belong to different categories (upright face as target and inverted faces as flankers), substitution errors are less likely to occur because flankers cannot be inadvertently reported. Hence, the decrease in crowding strength may be ascribed to the lack of substitution errors. As in the target cueing argument (see Figure 8), this explanation assumes that target location uncertainty (and substitution errors, as a consequence) plays a crucial role in crowding, driving the entire difference in crowding strength between upright and inverted face flankers. However, this argument assumes that, prior to target-flanker substitution, upright/inverted faces are processed differently, thus implying some kind of holistic face processing, just as Farzin et al. (2009) suggested. Indeed, if participants can avoid inadvertently reporting the gender of an inverted flanker face if it is swapped for the target due to location uncertainty, it means that this face needs to be identified as an inverted face. This requires holistic processing, especially for Mooney faces (which cannot be identified using low-level cues). Moreover, the results we obtain in the gender discrimination task (see Figure 12B) suggest that this is not what happens in the TTM.

Model assessment method

It could be argued that the TTM may reproduce high-level effects in crowding using a different set of model parameters. For example, some of the TTM failures could result from ceiling effects. Here, we used only parameters in the range preconized by the code documentation of the TTM (fovea radius with any value between 16 and 32 pixels). We originally used a fovea radius value of 32 pixels, which is what was used in Rosenholtz et al. (2019). However, for many of the tested conditions (especially with few flankers), the mongrels were almost untouched, which would have led to 100% accuracy, merely invalidating the TTM (i.e., no crowding). In addition, it may have obscured complex model behaviors, because of ceiling effects. For this reason, we decreased the fovea radius parameter from 32 to 16 pixels to increase crowding in all conditions (the main results reported in the current work). Still, for most stimuli that included large flanker configurations at large eccentricities (Shapes and Patterns experiments; see Figures 5 and 6, as well as Supplementary Information Figure SB), performance was at chance level and hence, high-level effects might have gone unnoticed. For all these stimuli, we ran a follow-up experiment in which we kept the fovea radius parameter as 32 pixels to make the task easier. This did not improve the model predictions, as measured by the template matching algorithm (see Supplementary Information Figure SC).

Moreover, it may be argued that assessing the TTM performance using behavioral mongrel discrimination tasks can introduce biases coming, for example, from different strategies used by human observers. First, it should be noted that the method we used is the same as in Rosenholtz et al. (2019) and their previous work (Balas et al., 2009; Keshvari & Rosenholtz, 2016; Rosenholtz, Huang, & Ehinger, 2012; Rosenholtz, 2011; Zhang, Huang, Yigit-Elliott, & Rosenholtz, 2015). Nevertheless, to control for unwanted human biases, we also quantified performance using a template matching algorithm (see Methods section for details). This did not change the results qualitatively (see Figures 2 to 6, as well as Supplementary Information Figures SA, SB, SC, SD, SE, SJ, and SK). The measured performances were similar to what was measured in the behavioral tasks, and none of the high-level effect of crowding were reproduced.

We want to point out that the template matching algorithms do not aim at reproducing human behavior results. They are an alternative (and more objective) way to measure TTM behavior, and to probe the information present in the model after high-dimensional pooling. Ultimately, the goal is to understand human perception, hence the main results of the present work are the ones that come from the human mongrel discrimination tasks. Nevertheless, we still tried to give the TTM an extra chance to reproduce uncrowding or holistic effects that could have been obscured by individual differences or biases of the participants during the mongrel discrimination tasks.

Model improvements

The TTM could account for a variety of perceptual properties of human vision, such as visual search (Alexander, Schmidt, & Zelinsky, 2014; Chang & Rosenholtz, 2016; Rosenholtz, 2011; Rosenholtz, Huang, Raj et al., 2012), gist perception and change blindness (Rosenholtz, 2014; Rosenholtz, et al., 2016; Ehinger & Rosenholtz, 2016), or visual metamers (Freeman & Simoncelli, 2011). Moreover, simply by using a rich set of image statistics, the TTM can explain many properties of visual crowding, such as substitution effects, its relationship to feature binding, or the selectivity of illusory feature conjunction (Keshvari & Rosenholtz, 2016). Finally, other models like the TTM that are based on image statistics can explain results from a large range of stimuli and tasks (Heeger & Bergen, 1995; Malik & Perona, 1990; Portilla & Simoncelli, 2000; Zhang et al., 2015; Ziemba & Simoncelli, 2021). Hence, our results should not be taken as a complete invalidation of the TTM or related image-statistic based models. Rather, they suggest that, to fully capture human behavior, models of crowding and of vision in general need to incorporate more specific mechanisms that account for complex visual processing. Our results provide evidence that high-level effects cannot emerge even from the most sophisticated and high-dimensional pooling models, such as the TTM.

How could these models be improved? First, to explain the complex effects in (Manassi et al., 2012; Manassi et al., 2013; Manassi et al., 2015; Manassi et al., 2016), we propose to add a recurrent grouping and segmentation stage to existing models of crowding. In such models, the high-level configuration of the stimulus affects lower-level target acuity, so that crowding interference only occurs within perceptual groups. Recent work confirmed that recurrent grouping and segmentation processes are a promising addition to capture global aspects of crowding (Bornet, Kaiser, Kroner, Falotico, Ambrosano, Cantero, Herzog, & Francis, 2019; Bornet et al., 2021; Doerig, Bornet et al., 2020; Doerig et al., 2019; Doerig, Schmittwilken et al., 2020; Francis, Manassi, & Herzog, 2017; Wallis, Funke, Ecker, Gatys, Wichmann, & Bethge, 2019). Along the same lines, it was shown that perceptual grouping is crucial to understand contextual effects in naturalistic scenes (Herrera-Esposito, Coen-Cagli, Gomez-Sena, 2021). Again, summary-statistics models (Balas et al., 2009; Freeman & Simoncelli, 2011; Parkes et al., 2001; Rosenholtz, 2016; Rosenholtz et al., 2019) could not predict this body of results (Herrera-Esposito et al., 2021).

Second, to explain why crowding happens at multiple levels, such as in holistic crowding between faces (Farzin et al., 2009; Manassi & Whitney, 2018; Whitney & Levi, 2011), we propose to include high-level statistics in high-dimensional pooling models, such as the TTM. Depending on the stimulus, interaction might occur at different levels of the visual processing hierarchy.

Alternatively, Chaney, Fischer, and Whitney (2014) proposed the Hierarchical Sparse Selection (HSS) model. In this model, fine-grained information is preserved by the feature integration process occurring in the visual cortex because of the high density of neurons paving the visual field (note that this is slightly different to the high-dimensional pooling stage of the TTM, in which fine-grained information is preserved because of the large number of pooled features). Crowding happens in the HSS model because, for the sake of efficient visual perception, the neurons that are selected to decode the target features are sampled sparsely.

We would like to point out that one of the advantages of the TTM is that it can easily be tested on various paradigms. The model provides a direct visualization of its output, which is not the case for most proposed models of vision. Importantly, the TTM does not need to be adapted or re-trained for new stimuli. This contrasts with, for example, the capsules network of Doerig, Schmittwilken et al. (2020), which needs to re-learn how to group stimuli for any new paradigm. This strong point of the TTM is also why it is easier to falsify it, as for example in the present work.

Finally, we cannot rule out that in the future, more complex or more flexible statistics may be used in the TTM to show that the model can exhibit uncrowding or holistic processing. For example, deep neural networks trained on natural images may be used as a source of complex summary statistics relevant to human perception (Ziemba & Simoncelli, 2021). However, we have reasons to believe that this will not be the case. Indeed, pooling is by nature ill-suited for this task, because adding more flankers always increases interference with the target representation. We do not see how this hurdle can be overcome. For example, feedforward convolutional neural networks, who are explicitly optimized for image recognition in a pooling framework, are biased towards local features (Baker, Lu, Erlikhman, & Kellman, 2018; Geirhos, Rubisch, Michaelis, Bethge, Wichmann, & Brendel, 2018; Wallis et al., 2019) and do not exhibit uncrowding (Doerig, Bornet et al., 2020; Doerig et al., 2019; Doerig, Schmittwilken et al., 2020), even when they are trained to ignore local features (Doerig, Bornet et al., 2020).

In conclusion, our results provide evidence that high-level effects cannot emerge even from the most sophisticated and high-dimensional pooling models, such as the TTM. Moreover, target cueing is not a viable explanation for these effects. Hence, crowding remains a complex, global and

multilevel perceptual phenomenon, as well as a precious and versatile probe to understand what may be missing from current models of human vision.

Keywords: crowding, grouping, face recognition, holistic processing, texture tiling model

Acknowledgments

The authors thank Ruth Rosenholtz for her detailed comments on this manuscript and for sharing the code of the TTM. We thank both reviewers for their insightful comments.

A.B. was supported by the European Union's Horizon 2020 Framework Programme for Research and Innovation under the Specific Grant Agreements No. 785907 (Human Brain Project SGA2) and No. 945539 (Human Brain Project SGA3). O.H.C. was supported by the Swiss National Science Foundation (SNF) 320030_176153 "Basics of visual processing: from elements to figures." A.D. was supported by the Swiss National Science Foundation grants No. 176153 "Basics of visual processing: from elements to figures" and No. 191718 "Towards machines that see like us: human eye movements for robust deep recurrent neural networks." D.W. was supported by the National Institutes of Health grant R01 CA236793.

Author contributions: All the authors contributed to design the study. A.B., O.H.C., and M.M. conducted the experiments. A.B. and O.H.C. analyzed the data. A.B. and M.M. wrote the manuscript. O.H.C., A.D., D.W., and M.H.H. provided feedback on earlier drafts of the manuscript.

Consent for publication: All participants gave permission for the publication of their data.

Availability of data and material: All relevant data are available from the authors at alban.bornet@epfl.ch.

Commercial relationships: none.
Corresponding author: Alban Bornet.
Email: alban.bornet@epfl.ch.
Address: Laboratory of Psychophysics, Brain Mind Institute, École Polytechnique Fédérale de Lausanne, Switzerland.

References

Alexander, R. G., Schmidt, J., & Zelinsky, G. J. (2014). Are summary statistics enough ? Evidence for the importance of shape in guiding visual search. *Visual Cognition, 22*(3–4), 595–609.

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

Baker, N., Lu, H., Erlikhman, G., & Kellman, P. J. (2018). Deep convolutional networks do not classify based on global object shape. *PLoS Computational Biology, 14*(12), e1006613.

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

Banks, W. P., Larson, D. W., & Prinzmetal, W. (1979). Asymmetry of visual interference. *Perception & Psychophysics, 25*(6), 447–456.

Banks, W. P., & Prinzmetal, W. (1976). Configurational effects in visual information processing. *Perception & Psychophysics, 19*(4), 361–367.

Banks, W. P., & White, H. (1984). Lateral interference and perceptual grouping in visual detection. *Perception & Psychophysics, 36*(3), 285–295.

Bayle, D. J., Schoendorff, B., Hénaff, M.-A., & Krolak-Salmon, P. (2011). Emotional facial expression detection in the peripheral visual field. *PLoS One, 6*(6), e21584.

Bock, J. M., Monk, A. F., & Hulme, C. (1993). Perceptual grouping in visual word recognition. *Memory & Cognition, 21*(1), 81–88.

Bornet, A., Doerig, A., Herzog, M. H., Francis, G., & Van der Burg, E. (2021). Shrinking Bouma's window : How to model crowding in dense displays. *PLoS Computational Biology, 17*(7), e1009187.

Bornet, A., Kaiser, J., Kroner, A., Falotico, E., Ambrosano, A., Cantero, K., Herzog, M. H., . . . Francis, G. (2019). Running large-scale simulations on the Neurorobotics Platform to understand vision-the case of visual crowding. *Frontiers in Neurorobotics, 13*, 33.

Boucart, M., Lenoble, Q., Quettelart, J., Szaffarczyk, S., Despretz, P., & Thorpe, S. J. (2016). Finding faces, animals, and vehicles in far peripheral vision. *Journal of Vision, 16*(2), 10.

Bouma, H. (1970). Interaction effects in parafoveal letter recognition. *Nature, 226*(5241), 177–178.

Canas-Bajo, T., & Whitney, D. (2020). Stimulus-specific individual differences in holistic perception of Mooney faces. *Frontiers in Psychology, 11*, 585921.

Cavanagh, P. (1991). What's up in top-down processing. *Representations of vision: Trends and tacit assumptions in vision research,* 295–304.

Chaney, W., Fischer, J., & Whitney, D. (2014). The hierarchical sparse selection model of visual crowding. *Frontiers in Integrative Neuroscience, 8*, 73.

Chang, H., & Rosenholtz, R. (2016). Search performance is better predicted by tileability than presence of a unique basic feature. *Journal of Vision, 16*(10), 13.

Choung, O. H., Bornet, A., Doerig, A., & Herzog, M. H. (2021). Dissecting (un)crowding. *Journal of Vision, 21*(10):10, 1–20.

Chung, S. T., Levi, D. M., & Legge, G. E. (2001). Spatial-frequency and contrast properties of crowding. *Vision Research, 41*(14), 1833–1850.

Coates, D. R., Chin, J. M., & Chung, S. T. (2013). Factors affecting crowded acuity : Eccentricity and contrast. *Optometry and Vision Science: Official Publication of the American Academy of Optometry, 90*(7), 628–638.

Coates, D. R., Levi, D. M., Touch, P., & Sabesan, R. (2018). Foveal crowding resolved. *Scientific Reports, 8*(1), 1–12.

Danilova, M. V., & Bondarko, V. M. (2007). Foveal contour interactions and crowding effects at the resolution limit of the visual system. *Journal of Vision, 7*(2), 25.

Doerig, A., Bornet, A., Choung, O. H., & Herzog, M. H. (2020). Crowding reveals fundamental differences in local vs. Global processing in humans and machines. *Vision Research, 167*, 39–45.

Doerig, A., Bornet, A., Rosenholtz, R., Francis, G., Clarke, A. M., & Herzog, M. H. (2019). Beyond Bouma's window : How to explain global aspects of crowding? *PLoS Computational Biology, 15*(5), e1006580.

Doerig, A., Schmittwilken, L., Sayim, B., Manassi, M., & Herzog, M. H. (2020). Capsule networks as recurrent models of grouping and segmentation. *PLoS Computational Biology, 16*(7), e1008017.

Egeth, H. E., & Santee, J. L. (1981). Conceptual and perceptual components of interletter inhibition. *Journal of Experimental Psychology: Human Perception and Performance, 7*(3), 506.

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

Faivre, N., & Kouider, S. (2011a). Increased sensory evidence reverses nonconscious priming during crowding. *Journal of Vision, 11*(13), 16.

Faivre, N., & Kouider, S. (2011b). Multi-feature objects elicit nonconscious priming despite crowding. *Journal of Vision, 11*(3), 2.

Fan, X., Wang, F., Shao, H., Zhang, P., & He, S. (2020). The bottom-up and top-down processing of faces in the human occipitotemporal cortex. *ELife, 9*, e48764.

Farah, M. J., Tanaka, J. W., & Drain, H. M. (1995). What causes the face inversion effect? *Journal of Experimental Psychology: Human perception and Performance, 21*(3), 628.

Farzin, F., Rivera, S. M., & Whitney, D. (2009). Holistic crowding of Mooney faces. *Journal of Vision, 9*(6), 18.

Flom, M. C., Heath, G. G., & Takahashi, E. (1963). Contour interaction and visual resolution : Contralateral effects. *Science, 142*(3594), 979–980.

Francis, G., Manassi, M., & Herzog, M. H. (2017). Neural dynamics of grouping and segmentation explain properties of visual crowding. *Psychological Review, 124*(4), 483.

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

Geirhos, R., Rubisch, P., Michaelis, C., Bethge, M., Wichmann, F. A., & Brendel, W. (2018). ImageNet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness. *arXiv preprint arXiv:1811.12231*.

Gong, M., Xuan, Y., Smart, L. J., & Olzak, L. A. (2018). The extraction of natural scene gist in visual crowding. *Scientific Reports, 8*(1), 1–13.

Greenwood, J. A., Bex, P. J., & Dakin, S. C. (2009). Positional averaging explains crowding with letter-like stimuli. *Proceedings of the National Academy of Sciences, 106*(31), 13130–13135.

Grützner, C., Uhlhaas, P. J., Genc, E., Kohler, A., Singer, W., & Wibral, M. (2010). Neuroelectromagnetic correlates of perceptual closure processes. *Journal of Neuroscience, 30*(24), 8342–8352.

Harrison, W. J., Retell, J. D., Remington, R. W., & Mattingley, J. B. (2013). Visual crowding at a distance during predictive remapping. *Current Biology, 23*(9), 793–798.

Heeger, D. J., & Bergen, J. R. (1995). Pyramid-based texture analysis/synthesis. *Proceedings of the 22nd annual conference on Computer graphics and interactive techniques,* 229–238.

Herrera-Esposito, D., Coen-Cagli, R., & Gomez-Sena, L. (2021). Flexible contextual modulation of naturalistic texture perception in peripheral vision. *Journal of Vision, 21*(1), 1.

Herzog, M. H., & Manassi, M. (2015). Uncorking the bottleneck of crowding : A fresh look at object recognition. *Current Opinion in Behavioral Sciences, 1*, 86–93.

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

Herzog, M. H., Sayim, B., Manassi, M., & Chicherov, V. (2016). What crowds in crowding? *Journal of Vision, 16*(11), 25.

Huckauf, A., Heller, D., & Nazir, T. A. (1999). Lateral masking : Limitations of the feature interaction account. *Perception & Psychophysics, 61*(1), 177–189.

Kanwisher, N., Tong, F., & Nakayama, K. (1998). The effect of face inversion on the human fusiform face area. *Cognition, 68*(1), B1–B11.

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

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.

Kouider, S., Berthet, V., & Faivre, N. (2011). Preference is biased by crowded facial expressions. *Psychological Science, 22*(2), 184–189.

Kovács, P., Knakker, B., Hermann, P., Kovács, G., & Vidnyánszky, Z. (2017). Face inversion reveals holistic processing of peripheral faces. *Cortex, 97*, 81–95.

Kreichman, O., Bonneh, Y. S., & Gilaie-Dotan, S. (2020). Investigating face and house discrimination at foveal to parafoveal locations reveals category-specific characteristics. *Scientific Reports, 10*(1), 1–15.

Latinus, M., & Taylor, M. J. (2005). Holistic processing of faces : Learning effects with Mooney faces. *Journal of Cognitive Neuroscience, 17*(8), 1316–1327.

Lev, M., & Polat, U. (2015). Space and time in masking and crowding. *Journal of Vision, 15*(13), 10.

Lev, M., Yehezkel, O., & Polat, U. (2014). Uncovering foveal crowding? *Scientific Reports, 4*, 4067.

Levi, D. M. (2008). Crowding—An essential bottleneck for object recognition : A mini-review. *Vision Research, 48*(5), 635–654.

Levi, D. M., Hariharan, S., & Klein, S. A. (2002). Suppressive and facilitatory spatial interactions in peripheral vision : Peripheral crowding is neither size invariant nor simple contrast masking. *Journal of Vision, 2*(2), 3.

Levi, D. M., Klein, S. A., & Hariharan, S. (2002). Suppressive and facilitatory spatial interactions in foveal vision : Foveal crowding is simple contrast masking. *Journal of Vision, 2*(2), 2.

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

Livne, T., & Sagi, D. (2007). Configuration influence on crowding. *Journal of Vision, 7*(2), 4.

Livne, T., & Sagi, D. (2010). How do flankers' relations affect crowding? *Journal of Vision, 10*(3), 1.

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.

Ly, A., Verhagen, J., & Wagenmakers, E.-J. (2016). Harold Jeffreys's default Bayes factor hypothesis tests : Explanation, extension, and application in psychology. *Journal of Mathematical Psychology, 72*, 19–32.

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

Malik, J., & Perona, P. (1990). Preattentive texture discrimination with early vision mechanisms. *JOSA A, 7*(5), 923–932.

Manassi, M., Hermens, F., Francis, G., & Herzog, M. H. (2015). Release of crowding by pattern completion. *Journal of Vision, 15*(8), 16.

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

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

Manassi, M., Sayim, B., & Herzog, M. H. (2013). When crowding of crowding leads to uncrowding. *Journal of Vision, 13*(13), 10.

Manassi, M., & Whitney, D. (2018). Multi-level crowding and the paradox of object recognition in clutter. *Current Biology, 28*(3), R127–R133.

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.

Mason, M. (1982). Recognition time for letters and nonletters : Effects of serial position, array size, and processing order. *Journal of Experimental Psychology: Human Perception and Performance, 8*(5), 724.

McKone, E. (2004). Isolating the special component of face recognition: Peripheral identification and a Mooney face. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 30*(1), 181.

Mewhort, D. J. K., Marchetti, F. M., & Campbell, A. J. (1982). Blank characters in tachistoscopic recognition : Space has both a symbolic and a sensory role. *Canadian Journal of Psychology/Revue Canadienne de Psychologie, 36*(4), 559.

Mooney, C. M. (1957). Age in the development of closure ability in children. *Canadian Journal of Psychology/Revue Canadienne de Psychologie, 11*(4), 219.

Nandy, A. S., & Tjan, B. S. (2012). Saccade-confounded image statistics explain visual crowding. *Nature Neuroscience, 15*(3), 463–469.

Nazir, T. A. (1992). Effects of lateral masking and spatial precueing on gap-resolution in central and peripheral vision. *Vision Research, 32(4)*, 771–777.

Parkes, L., Lund, J., Angelucci, A., Solomon, J. A., & Morgan, M. (2001). Compulsory averaging of crowded orientation signals in human vision. *Nature Neuroscience, 4*(7), 739–744.

Pelli, D. G. (2008). Crowding : A cortical constraint on object recognition. *Current Opinion in Neurobiology, 18*(4), 445–451.

Pittino, F., Eberhardt, L. V., Kurz, A., & Huckauf, A. (2019). Crowding with Negatively Conditioned Flankers and Targets. *Advances in Cognitive Psychology, 15*(1), 1.

Portilla, J., & Simoncelli, E. P. (2000). A parametric texture model based on joint statistics of complex wavelet coefficients. *International Journal of Computer Vision, 40*(1), 49–70.

Reuther, J., & Chakravarthi, R. (2019). Response selection modulates crowding : A cautionary tale for invoking top-down explanations. *Attention, Perception, & Psychophysics, 82*, 1763–1778.

Rosenholtz, R. (2014). Texture perception. *Oxford Handbook of Perceptual Organization, 167*, 186.

Rosenholtz, R. (2016). Capabilities and limitations of peripheral vision. *Annual Review of Vision Science, 2*, 437–457.

Rosenholtz, R. (2011). What your visual system sees where you are not looking. *Human Vision and Electronic Imaging XVI, 7865*, 786510.

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.

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

Rosenholtz, R., Yu, D., & Keshvari, S. (2019). Challenges to pooling models of crowding : Implications for visual mechanisms. *Journal of Vision, 19*(7), 15.

Rossion, B. (2008). Picture-plane inversion leads to qualitative changes of face perception. *Acta Psychologica, 128*(2), 274–289.

Rouder, J. N., Speckman, P. L., Sun, D., Morey, R. D., & Iverson, G. (2009). Bayesian t tests for accepting and rejecting the null hypothesis. *Psychonomic Bulletin & Review, 16*(2), 225–237.

Saarela, T. P., & Herzog, M. H. (2008). Time-course and surround modulation of contrast masking in human vision. *Journal of Vision, 8*(3), 23.

Saarela, T. P., Sayim, B., Westheimer, G., & Herzog, M. H. (2009). Global stimulus configuration modulates crowding. *Journal of Vision, 9*(2), 5.

Saarela, T. P., Westheimer, G., & Herzog, M. H. (2010). The effect of spacing regularity on visual crowding. *Journal of Vision, 10*(10), 17.

Sayim, B., Greenwood, J. A., & Cavanagh, P. (2014). Foveal target repetitions reduce crowding. *Journal of Vision, 14*(6), 4.

Sayim, B., Manassi, M., & Herzog, M. (2014). How color, regularity, and good Gestalt determine backward masking. *Journal of Vision, 14*(7), 8.

Sayim, B., Westheimer, G., & Herzog, M. H. (2008). Figural grouping affects contextual modulation in low level vision. *Journal of Vision, 8*(6), 436.

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

Sayim, B., Westheimer, G., & Herzog, M. H. (2011). Quantifying target conspicuity in contextual modulation by visual search. *Journal of Vision, 11*(1), 6.

Schwiedrzik, C. M., Melloni, L., & Schurger, A. (2018). Mooney face stimuli for visual perception research. *PLoS One, 13*(7), e0200106.

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.

Sergent, J. (1984). An investigation into component and configural processes underlying face perception. *British Journal of Psychology, 75*(2), 221–242.

Siderov, J., Waugh, S. J., & Bedell, H. E. (2013). Foveal contour interaction for low contrast acuity targets. *Vision Research, 77*, 10–13.

Strasburger, H., Rentschler, I., & Jüttner, M. (2011). Peripheral vision and pattern recognition : A review. *Journal of Vision, 11*(5), 13.

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

Tannazzo, T., Kurylo, D. D., & Bukhari, F. (2014). Perceptual grouping across eccentricity. *Vision Research, 103*, 101–108.

Taubert, J., Apthorp, D., Aagten-Murphy, D., & Alais, D. (2011). The role of holistic processing in face perception : Evidence from the face inversion effect. *Vision Research, 51*(11), 1273–1278.

Toet, A., & Levi, D. M. (1992). The two-dimensional shape of spatial interaction zones in the parafovea. *Vision Research, 32*(7), 1349–1357.

Van den Berg, R., Roerdink, J. B., & Cornelissen, F. W. (2010). A neurophysiologically plausible population code model for feature integration explains visual crowding. *PLoS Computer Biology, 6*(1), e1000646.

Vickery, T. J., Shim, W. M., Chakravarthi, R., Jiang, Y. V., & Luedeman, R. (2009). Supercrowding : Weakly masking a target expands the range of crowding. *Journal of Vision, 9*(2), 12.

Wallace, J. M., Chiu, M. K., Nandy, A. S., & Tjan, B. S. (2013). Crowding during restricted and free viewing. *Vision Research, 84*, 50–59.

Wallis, T., Funke, C., Ecker, A., Gatys, L., Wichmann, F., & Bethge, M. (2017). Towards matching peripheral appearance for arbitrary natural images using deep features. *Journal of Vision, 17*(10), 786.

Wallis, T. S., Funke, C. M., Ecker, A. S., Gatys, L. A., Wichmann, F. A., & Bethge, M. (2019). Image content is more important than Bouma's Law for scene metamers. *ELife, 8*, e42512.

Waugh, S. J., & Formankiewicz, M. A. (2020). Grouping Effects on Foveal Spatial Interactions in Children. *Investigative Ophthalmology & Visual Science, 61*(5), 23.

Westheimer, G., & Hauske, G. (1975). Temporal and spatial interference with vernier acuity. *Vision Research, 15*(10), 1137–1141.

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.

Wilkinson, F., Wilson, H. R., & Ellemberg, D. (1997). Lateral interactions in peripherally viewed texture arrays. *JOSA A, 14*(9), 2057–2068.

Wolford, G., & Chambers, L. (1983). Lateral masking as a function of spacing. *Perception & Psychophysics, 33*(2), 129–138.

Xia, Y., Manassi, M., Nakayama, K., Zipser, K., & Whitney, D. (2020). Visual crowding in driving. *Journal of Vision, 20*(6), 1.

Yeshurun, Y., & Rashal, E. (2010). Precueing attention to the target location diminishes crowding and reduces the critical distance. *Journal of Vision, 10*(10), 16.

Yin, R. K. (1969). Looking at upside-down faces. *Journal of Experimental Psychology, 81*(1), 141.

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

Ziemba, C. M., & Simoncelli, E. P. (2021). Opposing effects of selectivity and invariance in peripheral vision. *Nature Communications, 12*(1), 4597.