Opponent surrounds explain diversity of contextual phenomena across visual modalities

David A. Mély[1,2,3] & Drew Linsley[1,2,3] & Thomas Serre[1,2,3]

[1]*Department of Cognitive, Linguistic & Psychological Sciences*
[2]*Brown Institute for Brain Science*
[3]*Brown University, Providence, RI 02912, USA.*

Running head: Opponent surrounds explain diversity of contextual phenomena

Corresponding Author: Thomas Serre, Department of Cognitive, Linguistic & Psychological Sciences, Brown University, 190 Thayer St, Providence, RI 02912, USA.

Abstract

Context is known to affect how a stimulus is perceived. A variety of illusions have been attributed to contextual processing — from orientation tilt effects to chromatic induction phenomena, but their neural underpinnings remain poorly understood. Here, we present a recurrent network model of classical and extra-classical receptive fields that is constrained by the anatomy and physiology of the visual cortex. A key feature of the model is the postulated existence of near- vs. far-extra-classical regions with complementary facilitatory and suppressive contributions to the classical receptive field. The model accounts for a variety of contextual illusions, reveals commonalities between seemingly disparate phenomena, and helps organize them into a novel taxonomy. It explains how center-surround interactions may shift from attraction to repulsion in tilt effects, and from contrast to assimilation in induction phenomena. The model further explains enhanced perceptual shifts generated by a class of patterned background stimuli that activate the two opponent extra-classical regions cooperatively. Overall, the ability of the model to account for the variety and complexity of contextual illusions provides computational evidence for a novel canonical circuit that is shared across visual modalities.

Keywords: extra-classical receptive field, visual cortex, illusion, induction, assimilation, tilt effect.

Spatial context has been known to affect perception since at least Aristotle (Eagleman, 2001). The past several decades of work in visual psychophysics have revealed a plethora of seemingly disparate contextual phenomena (Series et al., 2003) whereby subtle differences in experimental conditions yield a wide variety of effects (Figure 1). In the classical tilt illusion (O'Toole and Wenderoth, 1977; Goddard et al., 2008), the perceived orientation of a center stimulus tilts either towards or away from that of a surround stimulus, depending on their relative orientations. Many variants have been tested with a variety of stimulus parameters including spatial frequency, color, luminance, contrast differences between center and surround stimuli as well as their spatial and temporal separation (see Clifford, 2014; for a review).

Similar effects have been reported in the motion domain – for both direction and speed (Marshak and Sekuler, 1979; Murakami and Shimojo, 1993; 1996; Kim and Wilson, 1997). In color induction, both the spatial frequency and phase of the surround controls the direction of the perceived shift in hue of a center stimulus relative to that of the surround (Smith et al., 2001; Monnier and Shevell, 2003; Shevell and Monnier, 2005). In the disparity domain, a center stimulus appears closer or further away from an observer, depending on the relative depth and spacing between center and surround stimuli (Westheimer, 1986; Westheimer and Levi, 1987). While much is known about the psychological basis of these phenomena, our understanding of the underlying neural mechanisms remains, at best, fragmentary.

A widely held assumption is that such contextual phenomena are mediated in the cortex by extra-classical receptive field (eCRF) mechanisms (reviewed in Series et al., 2003; Angelucci and Shushruth, 2013): The presentation of a stimulus in the eCRF alone does not typically elicit any response

from the neuron but modulates its response to a stimulus presented in the classical receptive field (CRF). Such center-surround interactions have been reported across visual modalities including orientation and spatial frequency (DeAngelis et al., 1994), motion (Li et al., 1999; Jones et al., 2001), color (Schein and Desimone, 1990; Wachtler et al., 2003) and disparity (Bradley and Andersen, 1998).

Although several eCRF models have been developed to describe specific phenomena (reviewed in Series et al., 2003; Schwartz et al., 2007; Angelucci and Shushruth, 2013; see also Discussion), a unifying theory, which would integrate disparate aspects of contextual integration and, ultimately, link primate neurophysiology to human behavior, is still lacking. We have thus developed a large-scale recurrent network model of classical and extra-classical receptive fields that distinguishes itself from previous work – allowing us to simulate realistic cortical responses to a variety of full-field, real-world, contextual stimuli defined across visual modalities (we model orientation, color, motion, and binocular disparity). The model is constrained by anatomical data and shown in our experiments to be consistent with V1 neurophysiology. A key feature of the model is the postulated existence of near vs. far extra-classical eCRF regions with complementary contributions (facilitatory vs. suppressive) to the CRF response. Using an ideal neural observer, we show that the model is consistent with human behavioral responses for a variety of contextual phenomena – revealing commonalities between seemingly disparate phenomena and helping to establish a novel taxonomy of contextual illusions.

Results

The visual cortex is modeled as a dense, regular topographic grid of cortical (hyper)columns which tile the visual field (Figure 2A). Each hypercolumn contains a complete set of units with coinciding CRFs. Their tuning curves are idealized (see Materials and Methods), and centered at regular intervals (e.g., between 0^o and 180^o for orientation-tuned units). For simplicity, we do not take into account cortical magnification and assume a fixed sampling of the visual field at all eccentricities. The model takes into account connections both within and across hypercolumns in order to explain several CRF and eCRF properties. The resulting circuit motif is replicated for every hypercolumn.

Intra-columnar recurrent circuits

Recurrent connections (Figure 2A, red connections) within a column (i.e., originating from within the CRF) include both local excitatory and inhibitory connections. Inhibitory CRF contributions constitute one of the key mechanisms in an influential model of gain control (divisive normalization, reviewed in Carandini and Heeger, 2012). This model accounts for cross-orientation normalization phenomena (when a grating stimulus is masked by another one at any orientation, see Heeger, 1993; Carandini and Heeger, 1994) and was later extended to capture neural population responses, in order to account for the competitive interactions within a single hypercolumn (Busse et al., 2009; Sit et al., 2009). A recent optogenetic study demonstrated that the underlying circuits are recurrent rather than feedforward (Nassi et al., 2015). We have also confirmed that this form of recurrent intra-columnar inhibition is critical for the model to reproduce these types of competitive interactions within the CRF and for model units to exhibit a realistic contrast response

(see Supplementary Experiments; Figures S1-S5.)

Because this form of suppression does not seem to depend on the orientation of the afferent and target cells, it is often called "untuned" inhibition (Rust et al., 2006). In our model, we speculate that such untuned inhibitory recurrent connections within hypercolumns exist for all other visual domains (including color, motion as well as binocular disparity, see Discussion).

In addition to short-range inhibitory connections within hypercolumns, the model also incorporates short-range excitatory connections. In the cortex, such excitatory connections may drive neurons up to ten times more strongly than their feed-forward inputs (Douglas et al., 1995; Stepanyants et al., 2008). As suggested by Shushruth et al. (2012), we have found that recurrent excitation within the CRF is essential to account for some of the more complex aspects of surround suppression (see Supplementary Experiments, Figures S3 and S5), by placing the column in a regime dominated by recurrent as opposed to feed-forward inputs. However, experimental data on the selectivity of these recurrent excitatory connections are scarce. Here, we assume that the corresponding local excitatory connections within a hypercolumn are only weakly tuned, as perfectly untuned excitation would effectively "flatten out" population response curves.

Inter-columnar recurrent circuits

Expanding the optimal stimulus of a cortical neuron immediately beyond its CRF (also commonly referred to as the "minimum response field" or mRF) may facilitate its response (Bringuier et al., 1999; Sengpiel et al., 1997; Sceniak et al., 1999; Angelucci et al., 2002a;b; Briggs and Usrey,

2011). The area which covers the CRF and its immediate eCRF is sometimes referred to as "peak spatial summation area". It is considered distinct from the CRF because a direct stimulation of this eCRF region in isolation does not elicit any action potential. Since this region is located immediately beyond the CRF, we deem it the *near eCRF* (or near surround; green annulus and connections in Figure 2A).

A potential neural substrate for the near eCRF includes the short-range, tuned excitatory networks (Lee et al., 2016) which span a spatial extent consistent with that of the eCRF facilitation (Angelucci et al., 2002a;b) and amplify co-occurring local inputs at similar orientations (Sengpiel, 1997; Sceniak et al., 1999; Angelucci et al., 2002a;b; Briggs and Usrey, 2011). In the model, we assume that all excitatory connections from other hypercolumns centered in the near surround are tuned, irrespective of the visual modality (i.e., the stimulus with the preferred orientation, or direction of motion, etc. in the CRF is also most effective in the near eCRF). Also note that our definition of the near eCRF is purely anatomical and might thus differ from that of others (e.g., Angelucci and Shushruth, 2013), whose definition is functional in nature.

Expanding the optimal stimulus beyond the near eCRF results in neural suppression (first reported by Hubel and Wiesel, 1968; as hypercomplex tuning). Critically, the presentation of the suppressing stimulus in the eCRF alone does not elicit any activity from the recorded cell (see Angelucci and Shushruth, 2013; for review). The tuned nature of these suppressive mechanisms is well documented across visual modalities: from orientation (Hubel and Wiesel, 1968; DeAngelis et al., 1994; Weliky et al., 1995; Petrov et al., 2005; Ozeki et al., 2009) to color (Schein and Desimone, 1990; Wachtler et al., 2003), spatial frequency (DeAngelis et al., 1994), temporal frequency (Li

et al., 1999; Jones et al., 2001), motion direction and speed (Allman et al., 1985) as well as binocular disparity (Bradley and Andersen, 1998).

Thus, we also define an *inhibitory far eCRF* (or far surround; blue annulus and connections in Figure 2A), located immediately beyond the excitatory near eCRF. In our model, a hypercolumn receives tuned inhibition from hypercolumns centered in its far surround.

To summarize, contributions from the eCRF as a whole arise from surround regions with opposite polarities. We do not assume any gap between the CRF and the near eCRF, nor between the near eCRF and the far eCRF. The model assumes that the near excitatory and far inhibitory eCRFs do not overlap spatially based on partial anatomical evidence (Angelucci et al., 2002b). In practice, we found that allowing these two eCRF regions to overlap did not affect the model's ability to fit experimental data (see Supplementary Experiments).

The first key assumption of the model is that, unlike local recurrent interactions within a hypercolumn, interactions across hypercolumns are "tuned" as only units that share the same preferred stimulus are directly connected. The second key assumption of the model is an asymmetry between excitation and inhibition: In the model, excitation only depends on pre-synaptic activity and is purely additive. Inhibition, on the other hand, from either the CRF or eCRF, depends on both pre- and post-synaptic activity, and ultimately results in a combination of subtractive and divisive effects (Carandini and Heeger, 2012). Similar forms of inhibition have been used in previous recurrent network models to achieve divisive normalization (Grossberg and Todorović, 1988). In practice, this means that, given a fixed amount of pre-synaptic inhibition, weakly active units receive less effective inhibition

than more active ones. In contrast, any given amount of pre-synaptic excitation results in the same amount of effective post-synaptic excitation.

Neural field model

Upon the presentation of a stimulus, recurrent interactions between units yield complex model dynamics. In particular, population responses at any location are modulated first by their immediate (near and far) eCRFs, then by responses across the visual field as transient activity propagates through the network, until all unit responses settle into a steady-state. Such short and long-range interactions are modeled using coupled differential equations and the steady-state solution of the resulting neural field model is computed using numerical integration methods (see Materials and Methods).

Next, we describe experiments conducted *in silico* to compare model responses to published psychophysics data. Psychophysics studies typically record perceptual judgments related to a center stimulus under varying surround conditions. To approximate these judgments, we use an ideal neural observer which maps center population responses to a sensory value. Note that in most cases, several columns may be located within the center stimulus; while any of these columns would be suitable for readout by the ideal observer, we selected the center-most column for simplicity (unless specified otherwise). Surround modulation thus gets translated into measurable perceptual changes in the center that can then be compared to human behavioral data.

We have organized these experiments into three broad categories, each reflecting a key computational

mechanism and highlighting commonalities across visual modalities. As we will show, these experiments allow a clear picture to emerge: The diversity of observed contextual phenomena may result from a balance between two opposing "forces" that arise from complementary excitatory-inhibitory eCRF mechanisms. Figure 2 shows examples of CRF and eCRF population responses recorded from the model together with representative transformations they undergo as a result of these two forces (see Discussion for more details).

All model parameters (Table S1) governing the dynamics and relative strengths of the interactions between the CRF and the eCRF sub-regions were initially adjusted for the model to reproduce a host of V1 neurophysiology data (see Supplementary Experiments; Figures S1-S5) including a comparison with data from Busse et al. (2009) and Trott and Born (2015). They were held fixed for all subsequent comparisons with psychophysics data. After scaling the stimulus, the only model parameter that was optimized for individual modalities was the tuning bandwidth of individual model units.

Competitive activation of the near vs. far surrounds

Our comparison between experimental and model data starts with a set of three experiments that span the orientation, motion and color domains. All experiments involve simple center-surround stimuli, in which the surround stimulus is expected to jointly activate both the excitatory and inhibitory components of the eCRFs. Thus, these experiments should reveal a fundamental aspect of the model: the outcome of a competition between the near facilitatory and the far suppressive

eCRFs when they are simultaneously activated by a surround stimulus.

The orientation tilt occurs when the perceived orientation of a center stimulus is biased either towards (Figure 3A) or away (Figure 3B) from the orientation of a surround stimulus, also called the inducing stimulus or inducer (O'Toole and Wenderoth, 1977; Goddard et al., 2008). Figure 3C shows representative psychophysics data (digitally extracted from Figure 4 in O'Toole and Wenderoth, 1977; only data averaged across subjects are available from that study). These data are characterized by two regimes: a *repulsive* regime (i.e., the perceived center orientation shifts away from the surround orientation, corresponding to positive ordinates) when the surround orientation is similar to that of the center and an *attractive* regime (i.e., the perceived center orientation shifts towards the surround orientation, corresponding to negative ordinates) when the surround orientation is different enough from that of the center.

The model successfully reproduces this balance between attraction and repulsion (Figure 3C; a similarly good fit was also obtained using broadband oriented textures as done in Goddard et al., 2008; data not shown). The key mechanism which enables the emergence of these two regimes is the postulated asymmetry between facilitatory and suppressive interactions originating from the near and far eCRFs, respectively. The net inhibition in the model, unlike excitation which is only dependent on pre-synaptic activity, increases monotonically with the level of post-synaptic activity of a target unit. We have confirmed this hypothesis via selective lesioning of the model key components (see Figure S14).

As a result, when neural population responses in the CRF and eCRF overlap significantly (as

when center and surround orientations are similar), inhibition predominates and center population responses get comparatively more suppressed at orientations close to that of the surround. The center of mass of center population response curves shifts away from the surround orientation, biasing the neural decoding accordingly (Figure 3A). The surround thus acts as a repellent in this regime. In contrast, when neural population responses in the CRF and eCRF are far more offset (as when the surround orientation is near orthogonal to that of the center), excitation from the near eCRF predominates, and increases the activity of center units selective for the surround orientation. This results in a force that pushes the center population response towards the surround orientation. This, in turn, biases the decoding of the center orientation in the direction of the surround orientation (Figure 3B). The surround thus acts as an attractor in this regime.

Beyond the orientation domain, tilt effects have also been reported for the perception of motion direction. Figure 3D shows representative psychophysics data (digitally extracted from Figure 3 "periphery" condition in Kim and Wilson, 1997). Unlike in the orientation domain, however, perceptual shifts are always repulsive (the perceived motion direction of the center grating tilts away from that of the surround grating; both gratings have the same contrast and speed). This phenomenon can also be induced using coherently moving random dots (Marshak and Sekuler, 1979); the effect seems to peak for similar center-surround differences in motion direction (between 40^o and 60^o) for either kind of stimuli.

We found both a qualitatively and quantitatively good fit between the model and psychophysics data as shown in Figure 3D. In the model, the disappearance of the attractive regime is accounted for by a broadening of the tuning curves (compared to orientation; see Supplementary Materials

and Methods). Interestingly, this seems consistent with neurophysiology data from the primary visual cortex (Ringach et al., 2002; Albright et al., 1984), which suggest that tuning for motion direction tends to be broader than for orientation.

In our next experiment, we show that the model is also able to account for tilt effects in the hue domain, more widely known as color induction. The model reproduces the known shifts in human judgment obtained when a center hue is surrounded by an isoluminant background of a different hue (digitally extracted from Figure 2 in Klauke and Wachtler, 2015; averaged across multiple combinations of center-surround hues sampled uniformly and independently as done in the original experiment).

As with motion induction, only a repulsive tilt effect is observed with hue. The model's ability to account for these data is evident from Figure 4B, which confirms the hypothesis by Klauke and Wachtler (2015) that color induction is in fact just another tilt effect (i.e., a "hue tilt effect"). Furthermore, the same mechanisms that are responsible for the tilt effect in the orientation and motion domains, namely the balance between facilitatory and suppressive forces originating from the eCRF, are also at play in color induction (Figure 4A). However, opponent color coding yields populations from the center and the surround with high overlap, which explains the absence of an attractive regime for this phenomenon.

Exclusive activation of the near vs. far surrounds

Previous experiments involved stimuli that reflected the outcome of a competition between the near and far eCRF, which were activated jointly. Here, instead, we consider experiments that are based on surround stimuli that activate the near or the far eCRFs separately.

In classical depth induction experiments (Westheimer, 1986; Westheimer and Levi, 1987), human observers are presented binocularly with a center test stimulus (e.g., a thin bar) flanked by two surround stimuli (e.g., parallel thin bars or small squares). The disparities of the flanker stimuli are adjusted so that they appear in the same depth plane, either slightly in front of or behind the center stimulus. The planar separation between the center and flanker stimuli (i.e., their distance in the fronto-parallel plane) is varied systematically. Examples where the flankers appear behind the center stimulus for a shorter and a larger separation are shown in Figure 5A-B).

Results from the original study (data digitally extracted from Figure 1, upper panels in Westheimer and Levi, 1987) are shown in Figure 5C. When the flankers are close enough to the center stimulus, they seem to attract it in depth (corresponding to a negative shift in perceived disparity for very small flankers/center separations). That is, the center stimulus appears closer to (further away from) the observer when the flankers are in front of (behind) the center stimulus. Instead, when the flankers are moved far enough laterally, they start to repel the center stimulus (corresponding to a positive shift in perceived disparity for larger flankers/center separations).

The observed shifts in depth found in the model (Figure 5C) matches qualitatively with human psychophysics data: Flanker stimuli located close enough to the test stimulus activate solely the

near eCRF, resulting in a purely facilitatory net eCRF influence. As with the aforementioned tilt effects, net facilitatory eCRF contributions yield attraction of the center towards the surround. Conversely, flankers that are far enough from the test stimulus activate solely the far eCRF. This results in a net suppressive eCRF influence, which translates into repulsion of the center away from the surround.

A classical stimulus used in motion direction induction (Murakami and Shimojo, 1993; 1996) is a center-surround stimulus consisting of randomly moving dots with some coherence in the surround but no coherence in the center. The presentation of coherently moving dots in the surround elicits the illusory perception of coherent motion in the center – either in the same or in the opposite direction to that of the surround (depending on the experimental condition). In (Murakami and Shimojo, 1993; 1996), the diameter of the center and surround was fixed to w and $2w$, respectively, and the parameter w was varied systematically. This allowed the dimension of the overall stimulus to vary while the relative size of the center and surround regions were maintained.

Figure 6C shows psychophysical data (digitally extracted from Figure 6 and 7 in Murakami and Shimojo, 1996). For small stimulus sizes, the induced center movement is in the same direction as that of the surround. This corresponds to an attractive regime measured as a negative shift in the point of subjective equality (PSE). For larger sizes, the center induced movement reverses direction. This corresponds to a repulsive regime measured as a positive shift in PSE.

We found these results to be consistent with the model (Figure 6C). With a small enough stimulus, the coherently-moving surround dots activate the model near eCRF exclusively (Figure 6A). This

leads to a perceptual shift in the direction of the surround, consistent with the analogous case discussed in depth induction. At the population level, facilitation from the near eCRF tends to cause a sharpening in the population response around the surround stimulus value in an otherwise flat population response (as all motion directions are present in the center stimulus). As a result, the center stimulus looks "more like" the surround stimulus. As the stimulus size increases, the coherently-moving surround dots start to activate an increasingly large proportion of the far surround (Figure 6B), which yields the opposite (repulsive) effect. At the population level, suppression from the far eCRF causes a small notch around the surround stimulus value and the center stimulus appears to look "less like" the surround.

Cooperative activation of the near and far surrounds

Thus far, we have seen that a variety of contextual phenomena can be explained as resulting from a balance between two opposing forces: an attractive force derived from facilitatory mechanisms originating from the near eCRF vs. a repulsive force derived from suppressive mechanisms originating from the far eCRF. This competition can be tipped from attraction to repulsion by increasing the relative contribution of suppressive mechanisms originating from the far eCRF (relative to facilitatory mechanisms from the near eCRF) either by increasing the spatial extent of the stimulus (so as to activate an increasingly large proportion of the far eCRF) or by increasing the similarity between the center and surround stimulus (so as to increase the overlap between center and near surround population responses). However, we reasoned that if a surround stimulus takes on distinct and appropriate values in the near and far eCRFs (which we deem the near and far values),

attraction towards the near value could go in the same direction as repulsion from the far value. Thus, the joint activation of the two eCRF sub-regions would cooperate rather than compete, resulting in an even larger perceptual shift compared to what would be achieved by presenting either the near or the far stimulus values alone.

The implication for color perception would be that assimilation, the attraction of the perceived center hue towards a neighboring inducing hue (i.e., the near hue), could be amplified by adding an appropriate outer hue (i.e., the far hue). This idea seems consistent with an "enhanced color shift" illusion discovered by Monnier and Shevell (2003), for which we provide a novel explanation. In the classical color assimilation illusion, a colored test ring (e.g., orange) is presented within a narrow uniform surround (e.g., purple or lime), which then attracts the test ring towards its own hue. This effect was found to be greatly amplified when patterned rings (e.g., alternating, thin rings of purple and lime) at an appropriate spatial frequency and phase were used in place of the uniform colored surround (Figure 1C). Such enhancement has also been documented with achromatic stimuli (Anstis, 2006) and in brightness perception (White, 1979; Anstis, 2006).

Our model provides a simple explanation: As we have established, attraction (i.e., assimilation) towards say purple is caused by the activation of the near surround by a purple stimulus, with respect to a center region coinciding with the test ring. For the appropriate spatial frequency (Figure 7A), the additional lime-colored stimulus activates the far surround, leading to repulsion (i.e., contrast) *away* from lime, thus amplifying the perceptual shift towards purple as purple and lime are roughly perceptual opposites. By reversing the phase of the color grating (Figure 7B), the colors stimulating the near and far eCRFs switch, leading to the same effect in the opposite

direction.

The original psychophysics data (digitally extracted from the "6 min test" curves of Figure 5 from Shevell and Monnier, 2005) and model data are shown in Figure 7C. The model explains the existence of an optimal spatial frequency value, which maximizes the magnitude of the illusion. The spatial frequency of the stimulus controls the strength of the illusion because it determines how cleanly each of the inducing colors (e.g., lime and purple) activate the near and far eCRFs respectively for a CRF centered on the test ring. The model also postdicts that reversing the phases of the color grating leads to an effect with the same amplitude but opposite direction.

Critically, our explanation only depends on the appropriate hues falling within the near and far eCRFs regions; thus, we predict that the periodicity of the inducing stimulus per se is not important, as long as both regions are correctly stimulated. We show this with our own versions of the illusion in Figures S6, suggesting that the illusion is just as strong, if not stronger, when the outer rings are replaced with a single uniform region that activates the far surround optimally (which is not the case for the original stimulus by Monnier and Shevell, 2003).

Discussion

We have described a computational neuroscience model of recurrent cortical circuits to account for classical (CRF) and extra-classical receptive field (eCRF) effects. The model was constrained by anatomical data and shown in our experiments to be consistent with V1 neurophysiology. In particular, the model unifies several electrophysiology phenomena such as (cross-orientation) normalization within the CRF (Busse et al., 2009) and modulation by the eCRF (including feature-selective

suppression, see Trott and Born, 2015) into a computational neuroscience model of contextual integration.

The model further provides computational evidence for the existence of two eCRF mechanisms with complementary contributions to the CRF (a facilitatory near vs. suppressive far eCRF). In addition, the model predicts that an asymmetry between excitation and inhibition in the eCRF is needed: In our implementation, excitation depends on pre-synaptic activity only, whereas inhibition depends on both pre- and post-synaptic activities. Another model prediction is that short-range connections within a hypercolumn are weakly tuned or untuned, whereas long-range connections across hypercolumns are tuned. We ran a systematic "lesioning" study on the model, whereby each of the hypothesized mechanisms was removed individually while all remaining parameters were optimized to fit behavioral data across all experiments (see Supplementary Experiments; Figures S7-S11.)

Although our analysis revealed that a model which includes all assumed mechanisms performs best, we also found that some of the assumptions could be relaxed. Most importantly, a spatial segregation between the near excitatory and far inhibitory eCRF does not appear necessary and the model was found to be robust to significant overlap between these two regions (Figure S8). More generally, the model was robust to a range of parameter values (Figure S9-S11) even when relaxing the strict one-to-one mapping for the "tuned" connections from the eCRF onto the CRF (Figure S10). At the same time, while this study has focused on explaining behavioral data for an average observer, the model's variations associated with changes in individual parameter values may help explain inter-subject variations observed experimentally (Figure S13).

The model distinguishes itself from previous work in succeeding to account for an array of disparate contextual phenomena spanning experimental conditions. Previous computational models have focused on explaining one or a few eCRF phenomena with an emphasis on surround suppression phenomena (see Series et al., 2003; Angelucci and Shushruth, 2013; for reviews): Phenomenological models of center-surround processing (Sceniak et al., 2001; Cavanaugh et al., 2002) and other normative models of visual coding (Coen-Cagli et al., 2012; Zhu and Rozell, 2013) have been shown to provide a good fit to single-unit contrast and size tuning responses. Recurrent network models have provided a mechanistic account for some of these phenomena (see Angelucci and Shushruth, 2013; for review) and have even led to testable predictions for single-unit electrophysiology (e.g., Rubin et al., 2015). But, none of these models have been systematically compared to a broad and diverse set of psychophysical experiments.

Furthermore, our model suggests that several contextual phenomena result from not one, but two opposing forces that yield systematic distortions on center population responses: repulsion from the far suppressive eCRF vs. attraction towards the near facilitatory eCRF (see Figure 2B for representative population response dynamics). By revealing commonalities between seemingly disparate perceptual phenomena, the model has helped us establish a novel taxonomy of visual illusions: We have found that the way in which individual stimuli activate these near and far eCRFs (competitively, exclusively or cooperatively; organized by columns in Figure 8) affects the qualitative behavior of the model.

A novel taxonomy of contextual phenomena

Contextual stimuli that yield competitive activation of the near vs. far eCRFs were found for a set of tilt illusions including orientation (O'Toole and Wenderoth, 1977; Goddard et al., 2008), motion (Kim and Wilson, 1997) and hue (Klauke and Wachtler, 2015; also known as color induction). In these stimuli, the surround spatially overlaps with both the near facilitatory and far suppressive eCRFs – activating them both competitively. Because of the asymmetry between excitation and inhibition in the model, repulsion from the surround stimulus prevails when the center and surround population responses overlap, i.e., when the center and surround stimuli are perceptually similar. Conversely, attraction towards the surround stimulus prevails when such overlap is minimal, such as when the center and surround stimuli are perceptually dissimilar.

Previous authors (Klauke and Wachtler, 2015; Goddard et al., 2008; Kim and Wilson, 1997; Clifford, 2014) have suggested that surround inhibition may be key to explaining the repulsive regime in tilt effects (see Supplementary Discussion for a more in-depth discussion). The proposed mechanisms, which include shifts in neural tuning curves (Klauke and Wachtler, 2015), varying inhibition strength depending on the relative center-surround orientation (Goddard et al., 2008), or recurrent center-surround interactions (Kim and Wilson, 1997) are all consistent with the proposed mechanistic model. In addition, the present study offers a plausible computational explanation for not only the existence of a repulsive regime but also an attractive one for certain classes of stimuli, in agreement with a host of experimental data (O'Toole and Wenderoth, 1977; Goddard et al., 2008; Kim and Wilson, 1997; Westheimer and Levi, 1987).

Another model postdiction is the absence of such attractive regime for contextual stimuli that yield broad-band population responses (arising because of broad neural tuning for the perceptual domain or because the stimulus is inherently ambiguous as in textures with little coherent orientation). For such stimuli, the overlap between center and surround population responses remains large even for maximally dissimilar center and surround stimuli, and the only discernible contextual effect is governed by the repulsive regime. Interestingly, the model achieves its quantitative fit for motion induction experiments via a broadening of neural tuning curves for motion direction compared to orientation, which is consistent with V1 electrophysiology data (Ringach et al., 2002; Albright et al., 1984) (see also Supplementary Discussion and Figure S12 for a more in-depth discussion).

Stimuli that activate exclusively the near or the far eCRF have been used in classical induction experiments in the domain of depth (Westheimer and Levi, 1987) and motion (Murakami and Shimojo, 1996). In the model, consistent with the proposal by Murakami and Shimojo (1993), re-scaling a stimulus display (or similarly, varying the relative spacing between center and surround stimuli) yields a reversal from attraction to repulsion. A surround stimulus close to the center or presented at a small scale tends to predominantly activate the facilitatory near eCRF, yielding attraction towards the surround. A surround stimulus farther from the center or presented at a larger scale tends to activate the suppressive far eCRF to a greater extent, yielding repulsion away from the surround.

For the last set of illusions called enhanced color shifts (Shevell and Monnier, 2005), the contextual (or surround) stimulus took on "opposite" optimal values in the near and the far eCRFs. As a result, shifts induced by either region of the eCRF tended to cooperate rather than compete with

one another. This resulted in a perceptual shift greater than a purely attractive effect involving only the near eCRF or a purely repulsive effect involving only the far eCRF. In addition, the spatial antagonism of the model eCRF captures the existence of an optimal spatial frequency (and phase) such that a cycle of the surround stimulus coincides maximally with the near and far eCRFs. More generally, the model confirms the consensus that assimilation predominates at higher spatial frequencies and finer scales whereas contrast emerges at lower spatial frequencies and coarser scales (Murakami and Shimojo, 1993; 1996; Monnier and Shevell, 2003; Shevell and Monnier, 2005; White, 1979; 1981; Anstis, 2006).

Shevell and Monnier (2005) have previously modeled enhanced color shifts through an S-cone color opponent model (see Supplementary Discussion). As in our model, such center-surround spatial antagonism results in the existence of an optimal spatial frequency. By design, their model, however, predicts the existence of enhanced perceptual shifts for S-cone stimuli only. One the other hand, our model predicts that such enhanced color shifts should persist for surround stimuli that do not activate S cones. We have created such stimuli (Figure S6) for the reader to judge for themselves but careful psychophysical work using a properly calibrated monitor will be needed to test this model prediction.

We have found a further subdivision of the above taxonomy (rows in Figure 8) based on a more detailed characterization of the center stimulus and, in particular, whether it is ambiguous (e.g., incoherently moving random dot or achromatic stimuli) or not (e.g., high-contrast gratings and bars, highly coherent moving random dot or saturated chromatic stimuli). Unambiguous center-surround stimuli yield peaked, narrow population responses (simulation results in Figure 2B-C) across the

visual field. The effect of the surround on a peaked center population response is to shift its center of mass, biasing the associated decoded value accordingly (see Supplementary Discussion for a discussion of the evidence of such shifts in neurophysiology studies). The shift is either towards (attraction) or away (repulsion) from the peak of the surround population response depending on whether the net effect of the eCRF is facilitatory (Figure 2B) or suppressive (Figure 2C). Ambiguous center stimuli yield broad-band (or even flat) center population responses (Figure 2D-E). These can be distorted by a peaked surround population in two ways: a bump centered at the surround stimulus value when tuned facilitation from the eCRF prevails (Figure 2D) or a notch at the surround value when tuned suppression does (Figure 2E).

Table S2 shows how the literature fits in the proposed taxonomy. Note that some table entries are missing for certain visual modalities, which suggests more contextual phenomena remain to be found (e.g., cooperative shifts in orientation, which would result in an "enhanced orientation tilt"). Overall, the present study thus provides a vivid example of how computational models may help re-interpret results as well as summarize and integrate disparate phenomena.

Open questions

The neural tuning curves considered in this work (orientation, disparity, motion direction, color opponent) can be found in relatively low-level areas of the visual cortex, such as V1, V2 or MT. Thus, the consistency between model and behavioral data is all the more remarkable as many of the illusions studied here are likely to also involve higher-level visual processes which are known

to affect perception including perceptual organization and grouping (e.g., Manassi et al., 2016), attention and other top-down feedback (Gilbert and Li, 2013) including surface-based and other filling-in processes (Grossberg and Todorović, 1988). The model's ability to account for contextual interactions may be limited to the relatively simple stimuli such as the bars and gratings tested here. We expect the model to fail to account for human data for more complex contextual stimuli defined by objects or shapes (e.g., Manassi et al., 2016). At the very least, a more complete model would likely require multiple stages of processing as well as mechanisms of filling-in and contour extraction (Grossberg and Todorović, 1988). Similarly, considering tuning curves found in higher-level areas, such as tuning to hue observed in V4/PIT neurons (as opposed to color-opponent V1 neurons considered here, see Conway et al., 2007) could also improve the fit with experimental data (though hue tuning remains controversial, see Mollon, 2009; Conway, 2009).

More generally, the present model leaves open any role for attention. Indeed, recent work has shown that attention seems to be shifting both the CRF and eCRF independently towards the attended location (Anton-Erxleben et al., 2009). It is likely that attention (not accounted for in the present model) may have played a role in shaping the pattern of observed behavioral results. In our simulations, the CRF size was scaled to the center of the stimuli – a role that could possibly be endued to attention (Carandini, 2012). Indeed, one of the main mechanisms in the present model – that of complementary excitatory and inhibitory surround mechanisms – is a key mechanism in one of the leading models of spatial attention (Tsotsos et al., 2001). In this model, an annular region of inhibition creates a negative attentional field surrounding the region of perceptual facilitation centered on the attended target. In addition, modeling work has also suggested that top-down

influences may "gate" the effective contextual interactions mediated by long-range horizontal connections (Setić and Domijan, 2008).

Anatomical data to constrain the patterns of recurrent connectivity (both within and across hypercolumns) in the model are scarce. The near and far eCRFs as modeled are likely to constitute, at best, coarse approximations for more complex patterns of anatomical connections. In particular, both the spatial extent and the relative strength of the near and far eCRFs relative to that of the CRF were held constant across experiments. Given that the experiments considered throughout spanned a range of visual stimuli across modalities and sizes, it is likely that these phenomena recruit neural populations in different cortical areas and visual eccentricities. It is also likely that variations in experimental factors lead to differences in how the center and surround capture attention. We thus expect improvements in the model's quantitative fit by considering additional parameters to control the spatial extent and the relative strength of the near and far eCRFs (e.g., as done in Goddard et al., 2008).

We have also left open the question of whether the connectivity in the near and far eCRFs would draw on slow intra-areal lateral connections or fast intra-areal feedback connections (Angelucci et al., 2002a;b; Shushruth and Ichida, 2009). Such lack of refinement, in addition to a lack of realistic modeling of excitatory and inhibitory synapses and their relative timing (Vinck et al., 2013), negatively impacts our ability to make predictions about the precise time course of the contextual effects modeled. We also expect that a resolution on the question of feedback vs. lateral connectivity will be needed to account for some of the electrophysiology phenomena we left aside in the present study including the known contrast dependence of the eCRF size (see Angelucci and

Shushruth, 2013; for review) or cross-orientation enhancements (Levitt and Lund, 1997; Sillito et al., 1995). Another hypothesis of the model that has yet to be confirmed is the existence of cortical columns for all visual domains beyond orientation (see Sincich and Horton, 2005; for review). At present, the existence of cortical columns for color (Dow, 2002), motion (DeAngelis et al., 1999) and binocular disparity (DeAngelis and Newsome, 1999) is only partially supported by neurophysiology evidence.

We have assumed for simplicity that the near eCRF is circular (i.e., isotropic with respect to the topography of the visual field). There is, however, evidence for anisotropies in the pattern of horizontal connections between cortical columns (as orientation-tuned cells tend to be more often connected when they share the same selectivity and their CRFs are aligned along an axis parallel to their preferred orientation, see Bosking et al., 1997). There is also more direct evidence for anisotropies in the shape of the eCRF (i.e., various elongations over a wide range of orientations and widths, see Tanaka and Ohzawa, 2009). The function of these anisotropies has been attributed to the computation of higher-order features (including contrast- or texture-defined boundaries) as well as contour integration and pop-out (Stemmler et al., 1995; Hess et al., 2003; Tanaka and Ohzawa, 2009). Future work should test whether these phenomena can be accounted for with a model extension that incorporates such eCRF anisotropies.

More generally our study did not address the role of the perceptual biases and the altered discriminability that arise because of surround mechanisms. It has been suggested that surround mechanisms could constitute one of the primary mechanisms for predictive coding and Bayesian inference type of computations (see Schwartz et al., 2007; for review). We speculate that the computational

mechanisms revealed by the contextual illusions studied here play a key role in shaping invariant population codes for object constancy. We have obtained preliminary results suggesting that tuned suppression from the far eCRF may improve the accurate decoding of surface reflectances across changes in illumination (i.e., color constancy; see Mély & Serre, abstract presented at the 2015 Vision Science Society meeting), by helping to discount undesirable variations in center population responses caused by changes in the light source. (This is reminiscent of a color constancy algorithm by Land and McCann (1971) known as the Retinex.) This raises the intriguing possibility that at least some of the mechanisms unraveled here may support other forms of perceptual constancy beyond color. Further work will be needed to quantify how object transformations such as changes in illumination or depth affect neural population responses tuned to orientation or binocular disparity and what computational mechanisms are needed to help discount these nuisances. Nonetheless, the ability of the model to account for the variety and complexity of contextual illusions provides computational evidence for a novel canonical cortical circuit shared across visual modalities.

Materials and Methods

Additional methods may be found in Supplemental Materials and Methods.

Model connectivity

A column centered at location (x,y) contains a complete set of N units with CRFs centered at (x,y) and tuning values covering the full range $\theta_{k=1...N}$ (e.g., orientation tuning curves are regularly centered at values $\theta_k \in [0, 180^o]$). Tuning curves are idealized – either bell-shaped for

disparity (Cumming and Parker, 1997), motion direction (Albright et al., 1984) and orientation (Ringach et al., 1997) or monotonic for color opponency (Johnson et al., 2001).

Each unit (x,y,k) receives excitation Q_k^{xy}, assumed to be weakly tuned and originating from within the same hypercolumn:

$$Q_k^{xy} = \sum_{j=1\ldots N} w_{jk} X_j^{xy} \text{ s.t. } w_{jk} \sim \mathcal{N}(\theta_k, \varsigma), \tag{1}$$

where w_{jk} corresponds to excitatory weights between units k and j (with selectivity θ_k and θ_j, respectively) and X_j^{xy} to input activity at location (x,y,j). We assume these weights to be normally distributed, centered at a target unit tuning preference θ_k with standard deviation ς. Some tuning (albeit weak) is necessary in order to prevent intra-columnar excitation from flattening the population responses to well-defined stimuli. In the color domain, we consider color-opponent model units with monotonic tuning curves. Instead of drawing weights from a normal distribution, which only makes sense for bell-shaped tuning curves, we set $w_{kk} = (\varsigma\sqrt{2\pi})^{-1}$ and $w_{jk} = \text{const.}$ (when $j \neq k$; under the constraint that the weights sum up to $r1$).

Each unit (x,y,k) also receives some inhibition U^{xy}, assumed to be untuned and originating from within the same hypercolumn:

$$U^{xy} = \frac{1}{N} \sum_{j=1\ldots N} Y_j^{xy}, \tag{2}$$

where Y_j^{xy} is the output activity of unit j at location (x,y). Unlike the excitation which is linear,

inhibition is non-linear, of the shunting kind (Grossberg and Todorović, 1988), and acts on the output of the pre-synaptic units (Equation 5). Combined with broad tuning, this allows populations of cells with coinciding CRFs to be significantly driven by a common input. Such mechanism was used in the model by Shushruth et al. (2012) and found experimentally to be critical to reproduce nonlinear neural effects such as stimulus-matched surround suppression (Trott and Born, 2015). Because the local inhibition is untuned, its strength is independent of a unit selectivity θ_k, and we drop the subscript k for simplicity.

Furthermore, unit (x, y, k) also receives tuned excitation P_k^{xy} from other units with the same selectivity θ_k that are located within its near eCRF $\mathbb{N}^{xy}$, defined relatively to position (x, y):

$$P_k^{xy} = \frac{1}{|\mathbb{N}^{xy}|} \sum_{u,v \in \mathbb{N}^{xy}} X_k^{u,v} \tag{3}$$

Similarly, unit (x, y, k) also receives tuned inhibition T_k^{xy} from other units with the same selectivity θ_k that are located within its far eCRF $\mathbb{F}^{xy}$, defined relatively to position (x, y):

$$T_k^{xy} = \frac{1}{|\mathbb{F}^{xy}|} \sum_{u,v \in \mathbb{F}^{xy}} Y_k^{u,v} \tag{4}$$

As in Equation 2, inhibition is non-linear and acts on the output $Y_k^{u,v}$ of unit k at location (u, v).

Neural field model

Neural field dynamics obey the following equations:

$$\begin{aligned}\eta\,\partial_t X_k^{xy} + \varepsilon^2 X_k^{xy} &= \left[\xi L_k^{xy} - (\alpha X_k^{xy} + \mu)U^{xy} - (\beta X_k^{xy} + \nu)T_k^{xy}\right]_+ \\ \tau\partial_t Y_k^{xy} + \sigma^2 Y_k^{xy} &= \left[\gamma P_k^{xy} + \delta Q_k^{xy}\right]_+\end{aligned} \tag{5}$$

where the feed-forward input L_k^{xy} drives every unit (x, y, k) across the visual field; each is represented by its recurrent input X_k^{xy} and output Y_k^{xy}. The parameters $\alpha, \beta, \delta, \gamma, \mu$ and ξ can be interpreted as synaptic weights (see Table S1 for values used) which control the amount of intra- and inter-columnar excitation and inhibition (Equations 1–4). The steady-state solution is computed using numerical integration (with convergence typically taking ~ 50 iterations). Population responses at the steady-state $\bar{Y}_k^{xy}$ are a very nonlinear function of the model input L_k^{xy}.

For each unit, the steady-state input and output are given by $\bar{X}_k^{x,y}$ and $\bar{Y}_k^{x,y}$, resp. Due to the rectifying non-linearity in the dynamics (Equation 5), at steady-state, $\bar{X}_k^{x,y}$ and $\bar{Y}_k^{x,y}$ can either be equal to zero, or to the values below:

$$\begin{aligned}\bar{X}_k^{x,y} &= \frac{\xi L_k^{x,y} - \mu\bar{U}^{x,y} - \nu\bar{T}_k^{x,y}}{\varepsilon^2 + \alpha\bar{U}^{x,y} + \beta\bar{T}_k^{x,y}} \\ \bar{Y}_k^{x,y} &= \frac{\gamma\bar{P}_k^{x,y} + \delta\bar{Q}_k^{x,y}}{\sigma^2}, \text{ with:}\end{aligned} \tag{6}$$

$$
\begin{aligned}
\bar{U}^{x,y} &= \frac{1}{N} \sum_{j=1\ldots N} \bar{Y}_j^{x,y} \\
\bar{T}_k^{x,y} &= \frac{1}{|\mathbb{F}^{x,y}|} \sum_{u,v \in \mathbb{F}^{x,y}} \bar{Y}_k^{u,v} \\
\bar{P}_k^{x,y} &= \frac{1}{|\mathbb{N}^{x,y}|} \sum_{u,v \in \mathbb{N}^{x,y}} \bar{X}_k^{u,v} \\
\bar{Q}_k^{x,y} &= \sum_{j=1\ldots N} w_{j,k} \bar{X}_j^{x,y}
\end{aligned}
\tag{7}
$$

Tuning curves

The model constitutes an example of tuning curve population model (Schwartz et al., 2007; Rust et al., 2006). We considered two kinds of tuning curves: bell-shaped (orientation, motion direction, binocular disparity) and monotonic, non-saturating tuning curves (color). All tuning curves were normalized, i.e., the maximum unit activity was set to be equal to 1. For non-angular variables (e.g., disparity), bell-shaped tuning curves were parametrized as Gaussian functions:

$$
f(\theta \mid \theta_k, \sigma) = \exp(-\frac{(\theta - \theta_k)^2}{2\sigma^2}), \tag{8}
$$

with preferred stimulus value θ_k and tuning bandwidth σ. When the variable was circular (e.g., orientation, motion direction), we modeled the tuning curve as a von Mises function instead:

$$
f(\theta \mid \theta_k, \sigma) = \exp(\frac{\cos((\theta - \theta_k)\frac{2\pi}{I}) - 1}{2\sigma^2}) \tag{9}
$$

where I indicates the length of domain of the tuning curve (e.g., π for orientation vs. 2π for direction). We generally sampled on the order of 30 tuning curve centers regularly spaced in the domain of the considered visual modality. We found that the number of tuning curve centers considered did not impact our results as long as it was large enough.

Monotonic, non-saturating tuning curves for color were derived by converting stimuli to idealized cone responses first, which were then mapped to opponent color channels similarly to Zhang et al. (2012). These included red-on/green-off (R^+G^-), green-on/red-off (G^+R^-), blue-on/yellow-off (B^+Y^-), and yellow-on/blue-off (Y^+B^-), alongside with a pair of luminance-sensitive channels, selective for lighter (Wh^+Bl^-) and darker (Bl^+Wh^-) stimuli.

Model parameters

All circuit parameters were held constant in all comparisons with psychophysics data. They were determined a priori in order to reproduce key neurophysiology data (see Supplementary Experiments) and were held constant for all visual modalities except color because of a qualitative difference in tuning curve (see Equation 1). In all subsequent experiments, only two variables were allowed to vary: the stimulus scale and the tuning bandwidth for model units.

The stimuli used in psychophysics studies varied greatly – recruiting neural populations subtending a wide range of CRF (and eCRF) sizes and eccentricities, possibly spanning different visual areas. Rather than adjusting the size of the model CRFs and eCRFs for individual experiments, which would have required structural changes to the model, we instead varied the stimulus scale. Because

the connectivity between model hypercolumns was held fixed, this is somewhat akin to varying the magnification factor in the model. Critically, this yielded broad estimates for CRF (and eCRF) sizes within a biologically realistic range (from a fraction of a degree of visual angle to a couple of degrees). The width of the idealized tuning curves which is common to all model units was optimized separately for each experiment (see Supplementary Materials and Methods for details). We have confirmed that our key model predictions were robust over a range of these parameter values.

Ideal neural observer model

We used an ideal neural observer model to map model population responses to decoded sensory variables, which can then be compared to behavioral judgments collected experimentally. We used a population vector model (Georgopoulos et al., 1986), in which each unit votes for its preferred sensory value in proportion to its activity (normalized by the summed activities of all units within the same column). This model is not appropriate for color because of the tuning along opponent color pairs rather than a hue angle. Instead, we used cross-validated ridge regression to decode the sine and cosine of hue.

Acknowledgments

This work was supported by DARPA young faculty award [grant number N66001-14-1-4037] and NSF early career award [grant number IIS-1252951] to TS.

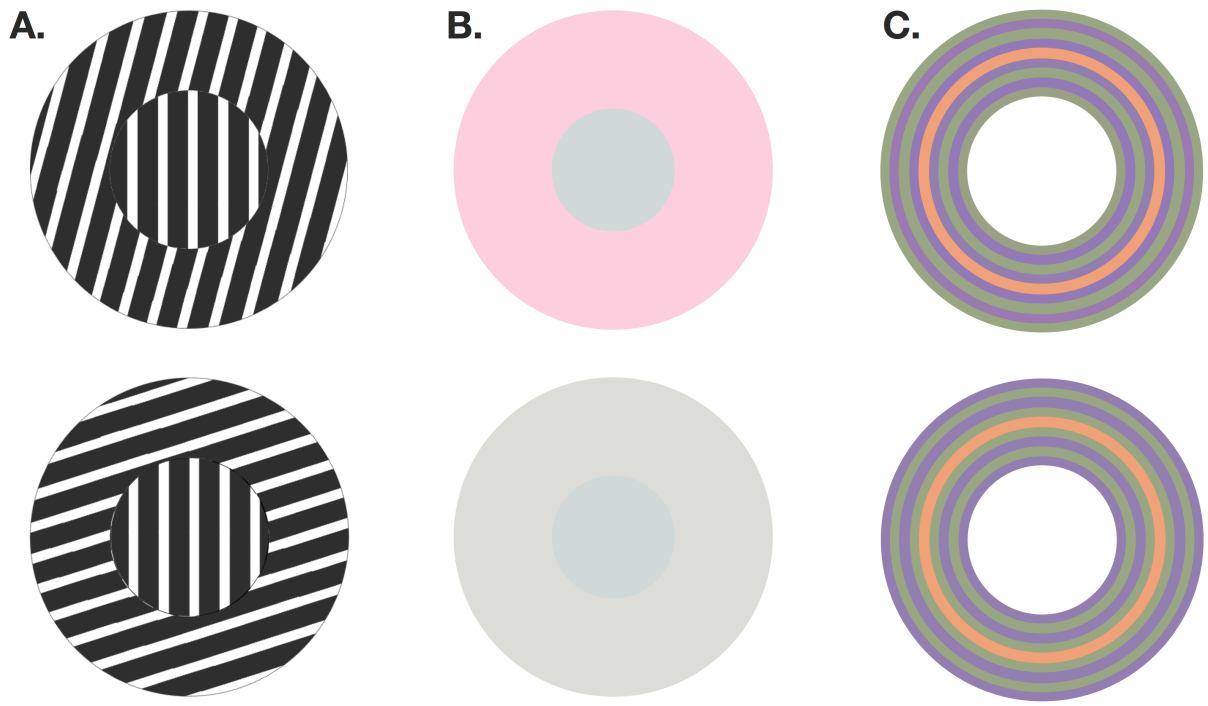

Figure 1. ***Representative contextual phenomena explained by the model.*** **A. Orientation tilt:** The perceived center (or test) orientation appears tilted from its true physical orientation, away from the surround (or contextual) orientation when center and surround stimuli are similar (top) and towards the surround orientation when they are dissimilar (bottom). **B. Color induction:** A central gray stimulus appears greener when embedded in a pink surround (top) compared to a neutral gray surround (bottom). **C. Enhanced color shifts:** The test stimulus is a central, orange ring, embedded in a surround stimulus composed of alternating purple and lime rings. The test ring looks vividly more pink when the adjacent color is purple, followed by lime (top), and looks more yellow when lime is the adjacent color, followed by purple (bottom).

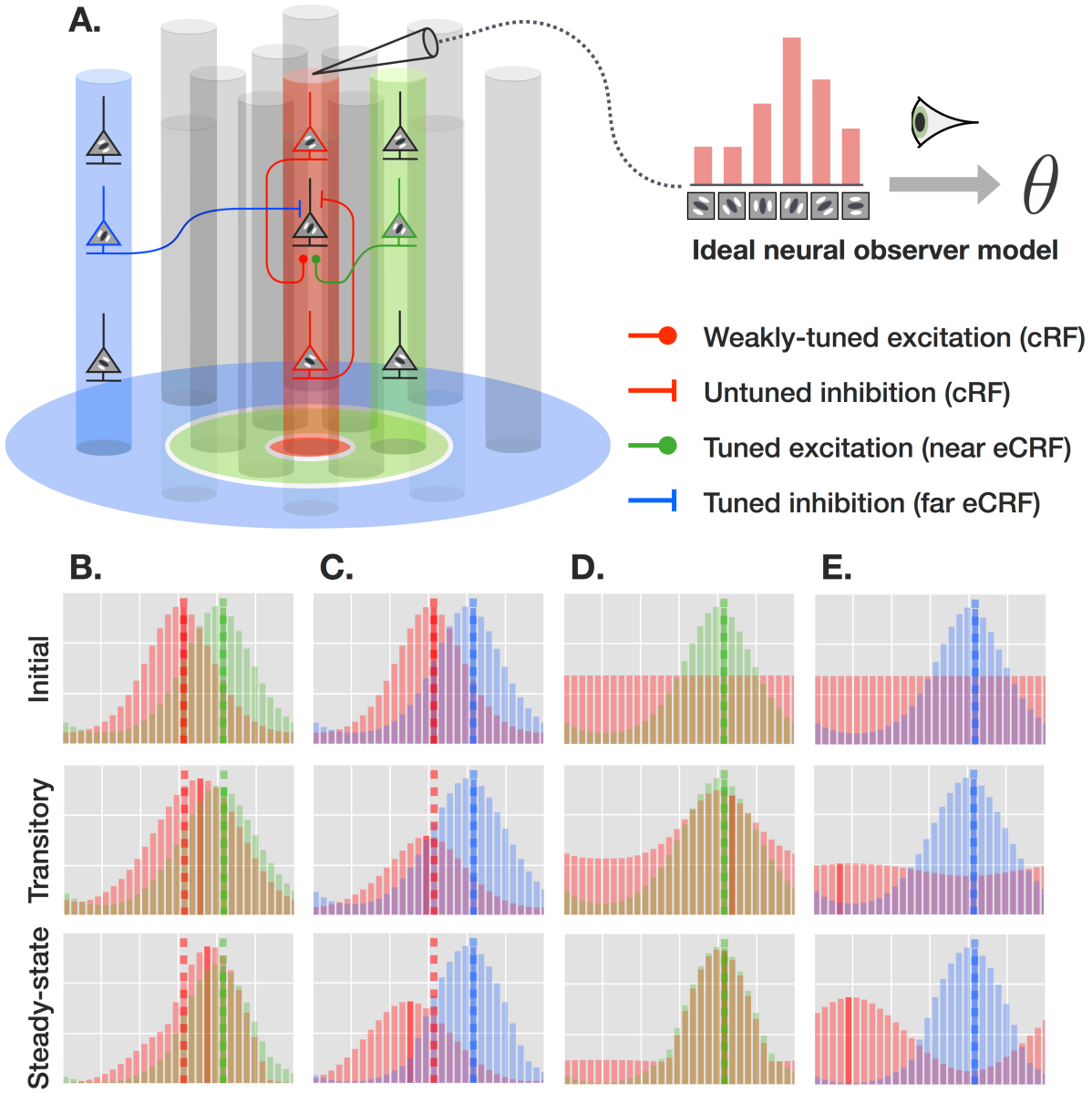
A.
θ
Ideal neural observer model
Weakly-tuned excitation (cRF)
Untuned inhibition (cRF)
Tuned excitation (near eCRF)
Tuned inhibition (far eCRF)
B.
C.
D.
E.
Initial
Transitory
Steady-state

Figure 2. ***Recurrent network model of center-surround interactions.*** **A. Connectivity:** The model implements excitatory and inhibitory connections both short-range (within hypercolumns) and long-range (between hypercolumns). The regions shown in red, green and blue correspond to the CRF, near eCRF and far eCRF, respectively (defined for the reference column in red). Model inhibitory connections are such that the net inhibition onto a target unit is a function of not just the pre-synaptic activity, but also the post-synaptic activity (see text for details). Color conventions for the CRF and the near and far eCRFs are used consistently throughout the paper. **B–E. Representative model dynamics:** Example population responses (32 direction-tuned model units) following the presentation of a contextual stimulus corresponding to the initial, transitory and steady state (rows). Population responses correspond to locations in the CRF, near eCRF and far eCRF. Highlighted bars represent directions decoded from the corresponding populations (undefined for flat responses); dashed lines represent initial decoded values at stimulus onset. Each column corresponds to a representative transformation undergone by the center population under the proposed taxonomy of contextual phenomena derived from the model: **B.** attractive shift, **C.** repulsive shift, **D.** bump, **E.** notch. (see also Discussion and Figure 8). Abscissas span the range $[-180^o, 180^o]$ and ordinates are normalized independently for readability.

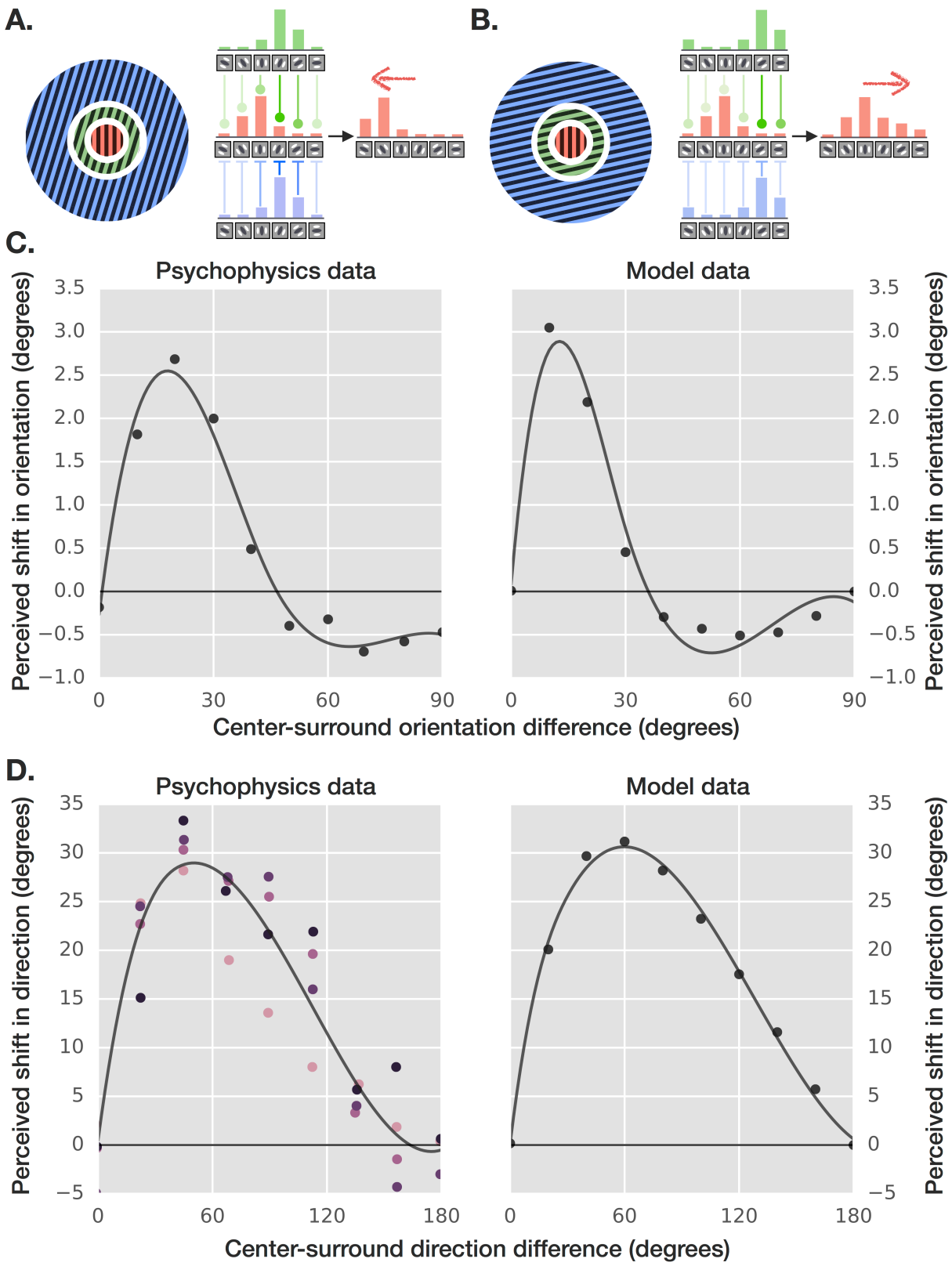
A.
B.
C.
Psychophysics data
Model data
Perceived shift in orientation (degrees)
Center-surround orientation difference (degrees)
D.
Psychophysics data
Model data
Perceived shift in direction (degrees)
Center-surround direction difference (degrees)

Figure 3. ***Tilt effects.*** Competitive activation of the near vs. far eCRFs explains the shift in tilt from one direction to the other. **A. Repulsion:** For similar center-surround orientations, tuned inhibition from the far eCRF outweighs excitation from the near eCRF, which yields a net repulsive force on the center population responses (away from that the surround orientation). **B. Attraction:** For dissimilar center-surround orientations, tuned excitation from the near eCRF prevails, which yields a net attractive force on the center population responses (towards the surround orientation). Note that gaps between the CRF and the near and far eCRFs were added for improved readability only and are not present in the actual model. **C. Orientation tilt: Psychophysics vs. model data.** Psychophysics data were digitally extracted from Figure 4 in (O'Toole and Wenderoth, 1977) and fitted with splines. The model explains the characteristic shift from perceptual repulsion (positive ordinates) to attraction (negative ordinates). **D. Motion tilt: Psychophysics vs. model data.** Psychophysics data were digitally extracted from Figure 3 ("periphery" condition) in (Kim and Wilson, 1997) and fitted with splines. Different colors correspond to different subjects. Both psychophysics and model data exhibit a similar dependency on the direction difference between center and surround, as well as a lack of an attractive regime.

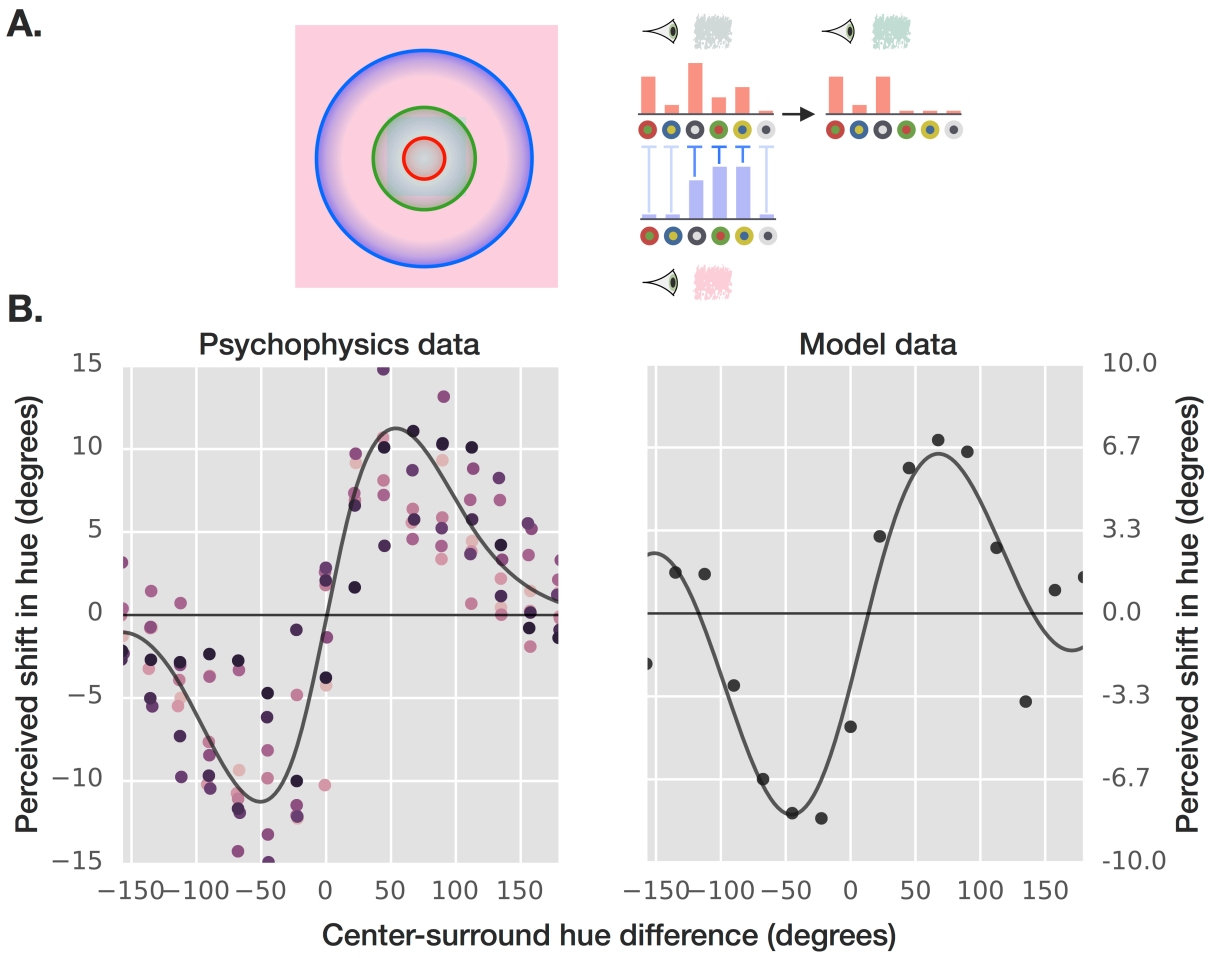

Figure 4. ***Color induction (or hue tilt effect).*** This experiment generalizes the tilt effect to opponent population codes. **A. Repulsion:** As with the classical tilt effect, the key model mechanism behind perceptual repulsion is the tuned inhibition from the far eCRF. In this example, the pink surround suppresses "red" center neurons, therefore reducing the "redness" of the gray center patch yielding a shift in the perceived center hue towards green. The same explanation also applies to chromatic center stimuli. Colored patches shown next to the eyes correspond to the color decoded under the ideal observer. **B. Psychophysics vs. model data:** Psychophysics data were digitally extracted from Figure 2 in (Klauke and Wachtler, 2015) and fitted with splines (averaged across eight surround hues). Both model and behavioral data exhibit a characteristic two-lobed shape peaking around $\pm 50^o$. Please note the difference in ordinate scale between the psychophysics and model data.

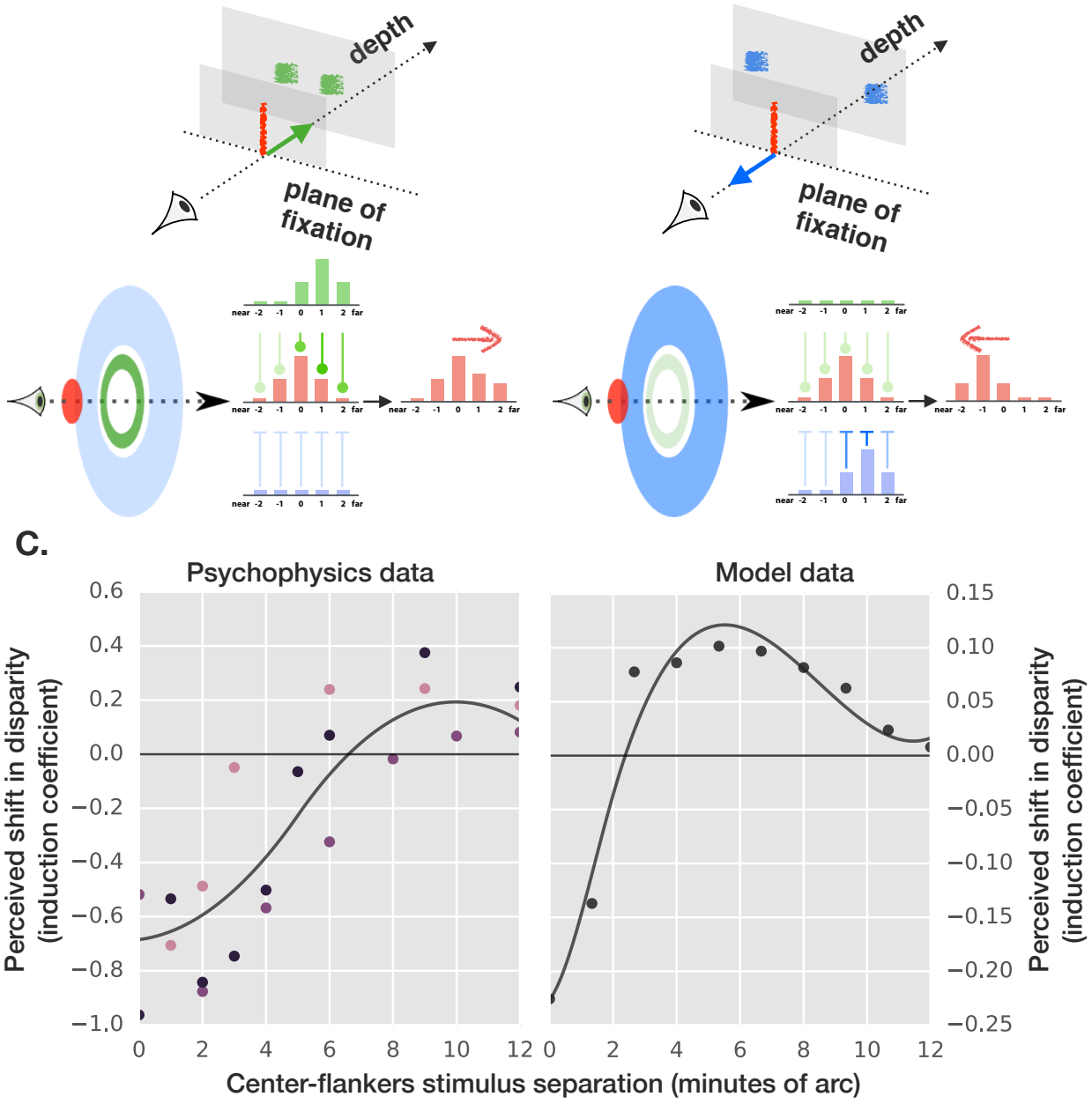
depth
plane of fixation
depth
plane of fixation
near -2 -1 0 1 2 far
C.
Psychophysics data
Model data
Perceived shift in disparity (induction coefficient)
Perceived shift in disparity (induction coefficient)
0.6
0.4
0.2
0.0
−0.2
−0.4
−0.6
−0.8
−1.0
0.15
0.10
0.05
0.00
−0.05
−0.10
−0.15
−0.20
−0.25
0 2 4 6 8 10 12
0 2 4 6 8 10 12
Center-flankers stimulus separation (minutes of arc)

Figure 5. ***Depth induction.*** The exclusive activation of either the near or far eCRFs by flankers (as their separation vary) explains the existence of the shift from assimilation to contrast. The perceived depth of a binocular center stimulus (at zero disparity) is affected by binocular flankers located on either sides, presented at either crossed or uncrossed disparities. **A. Attraction:** For short separations, flankers activate the near eCRF, which yields a net attractive force on the center population responses (towards the surround disparity) corresponding to a negative perceived shift. **B. Repulsion:** For larger separations, flankers activate the far eCRF, which yields a net repulsive force on the center population responses (away from the surround disparity) corresponding to a positive perceived shift. **C. Psychophysics vs. model data:** Psychophysics data were digitally extracted from Figure 1 (upper panels) in (Westheimer and Levi, 1987) and fitted with splines. Both behavioral and model data capture the balance between stronger attraction towards the flankers at small separations, and weaker repulsion at larger separations. Note that the agreement between the model and human data is only qualitative as the perceived shifts in disparity are on different scales (the model underestimates the strength of the attractive regime in this illusion).

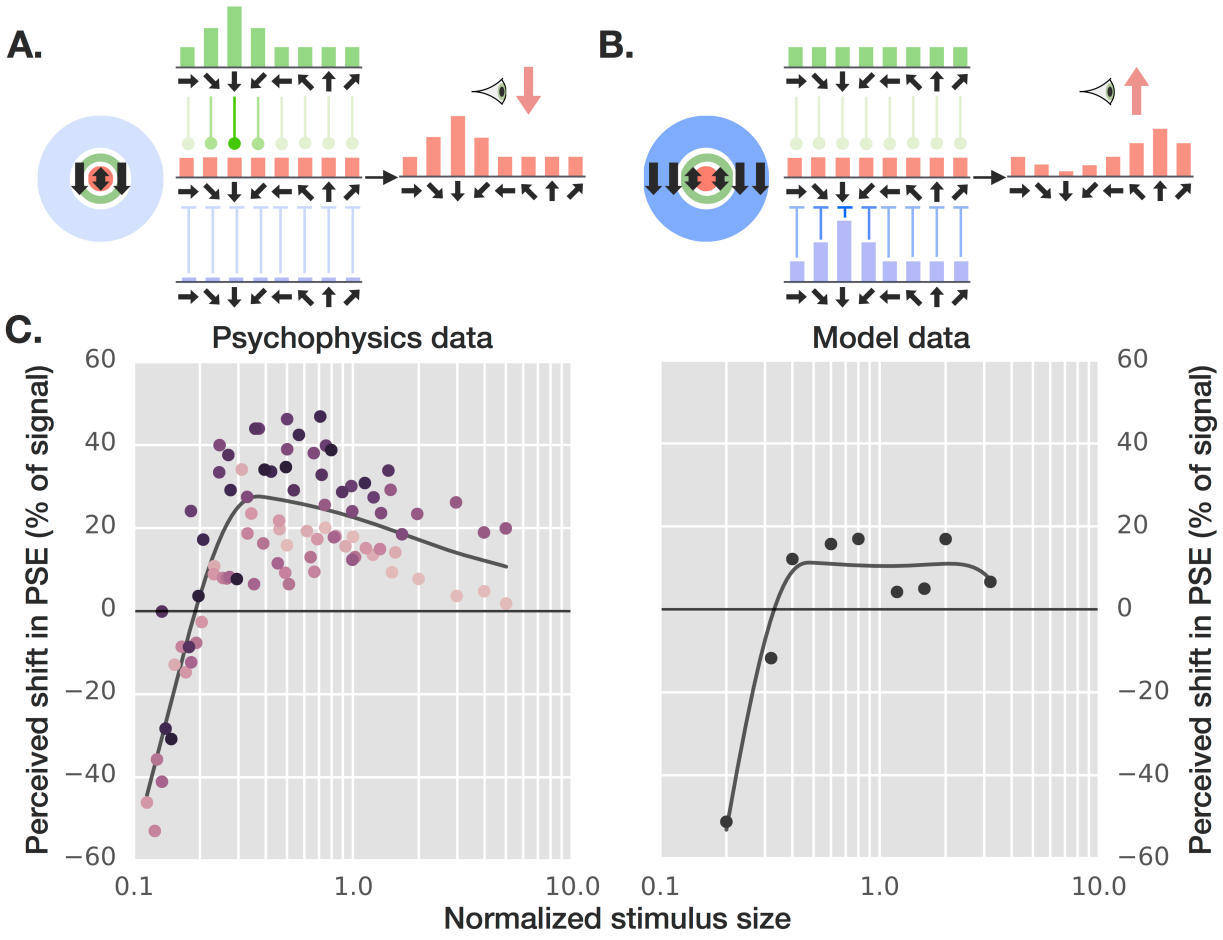

Figure 6. ***Motion induction.*** The increasing proportion of the far eCRF activated by larger stimuli explains the shift from assimilation to contrast. **A. Attraction:** When the overall stimulus is small enough, the coherently moving surround dots activate the near eCRF exclusively, leading to motion assimilation (i.e., the center dots' direction appears the same as that of the surround dots). **B. Repulsion:** Beyond a critical size, activation of the far inhibitory eCRF prevails, which leads to the opposite motion contrast effect (i.e., the center dots' direction appears opposite to that of the surround dots). **C. Psychophysics vs. model data:** Psychophysics data were digitally extracted from Figure 5 and 6 (Murakami and Shimojo, 1996) and fitted with splines. Both exhibit stronger attraction (negative ordinates) for smaller stimulus sizes, and weaker repulsion (positive ordinates) for larger sizes. Shifts in the point of subjective equality (PSE) were used as a proxy for shifts in perceived motion direction.

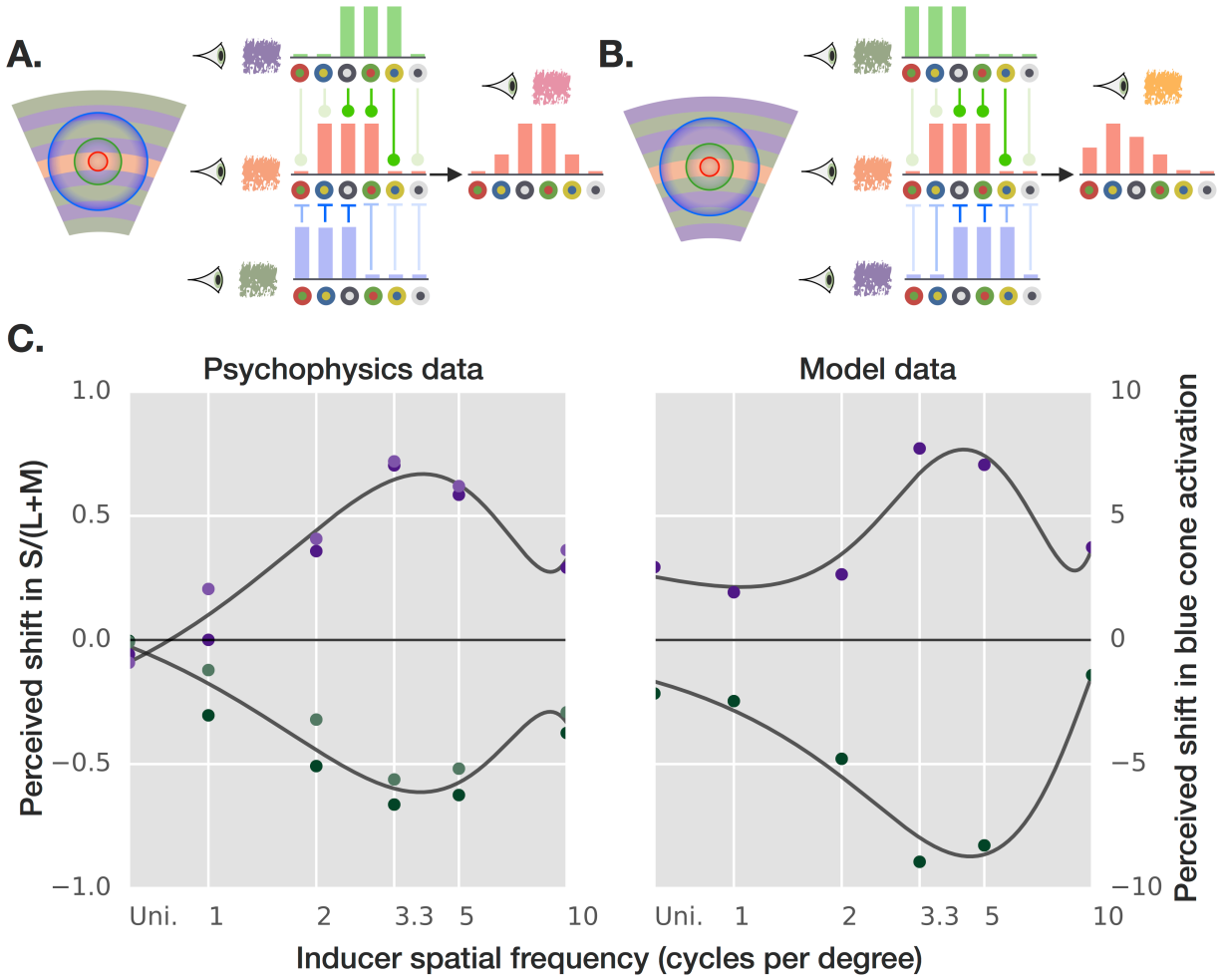

Figure 7. ***Enhanced color shifts.*** Cooperative activation of the near and far surrounds explains enhanced perceptual shifts. When distinct and "opposite" hues are used in a patterned surround (or inducer), the resulting shift in color perception of a test hue (here, orange) is amplified relative to a uniform surround of either hue. **A. Shift in one direction:** For the optimal spatial frequency, one surround hue (e.g., purple) overlaps optimally with the near eCRF and the other one (e.g., lime) with the far eCRF. For the right color combination (as here with purple and lime which are complementary colors), this results in cooperating perceptual forces: a shift towards purple / away from lime. The colored patches next to the eyes correspond to the color decoded under the ideal observer. **B. Shift in the other direction:** when purple and lime are switched. **C. Psychophysics vs. model data:** Psychophysics data were digitally extracted from Figure 5 (6 minutes test condition) in (Shevell and Monnier, 2005) and fitted with splines. Purple/green dots correspond to condition A/B. 'Uni.' stands for a uniform inducer composed of a single hue. For

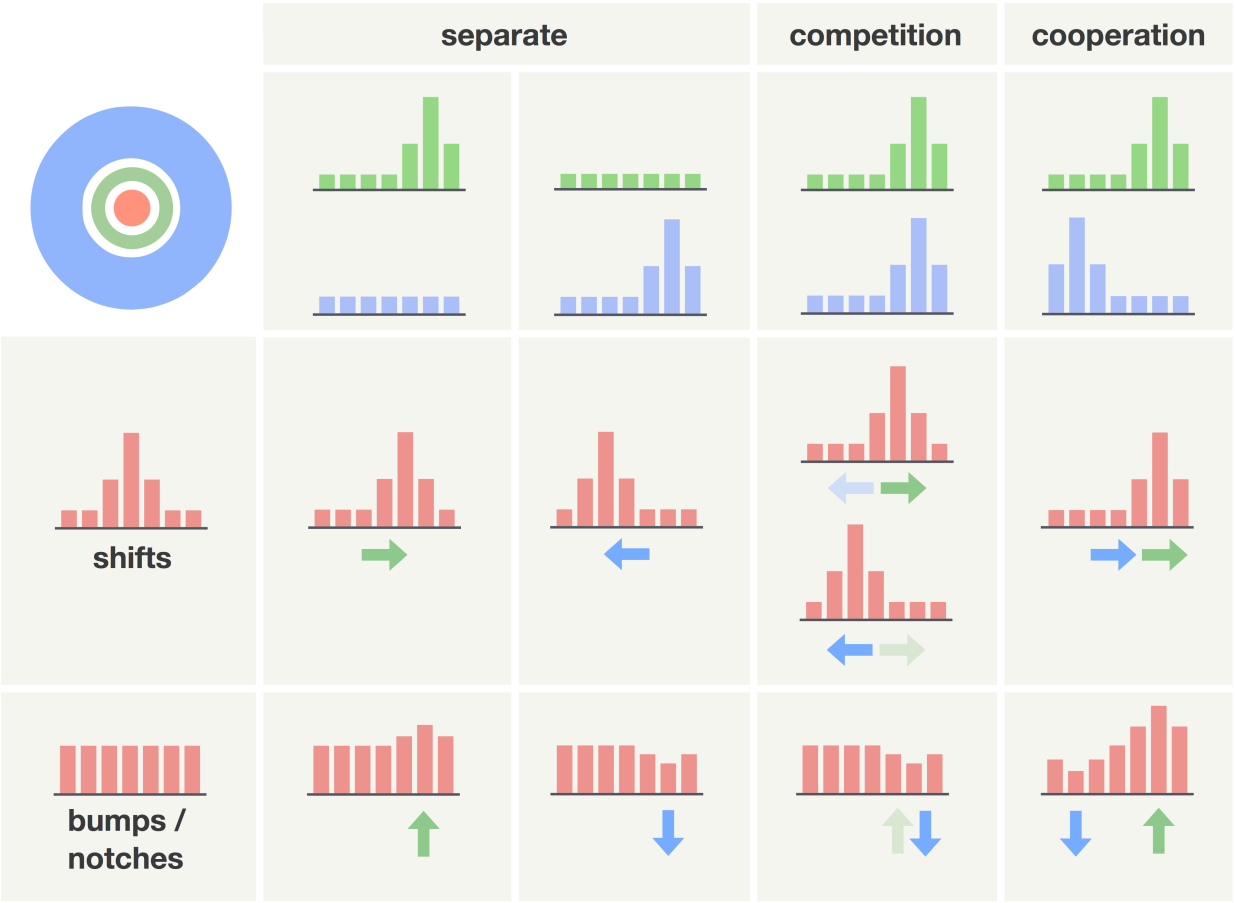

Figure 8. ***A new taxonomy of contextual phenomena.*** **Rows:** Contextual phenomena manifest themselves in the model either as (i) shifts with peaked center population response curves (unambiguous stimuli), or (ii) bumps/notches with broad/uniform center population response curves (ambiguous stimuli). **Columns:** Center-surround stimuli activate the near and far eCRFs in three typical ways: (i) either one separately, (ii) both competitively (i.e., near and far eCRFs each induce shifts that tend to stymie each other; green and blue arrows, resp.), and (iii) both cooperatively (i.e., the shifts induced by the near and far eCRFs are in the same direction, amplifying the perceptual shift). See Table S2 for a version of this table populated with representative psychophysics studies for each individual case.

Supplementary Online Material

Opponent surrounds explain diversity of contextual phenomena across visual modalities

David A. Mély[1,3] & Thomas Serre[1,2,3]

[1]*Department of Cognitive, Linguistic & Psychological Sciences*
[2]*Brown Institute for Brain Science*
[3]*Brown University, Providence, RI 02912, USA.*

The Supplementary Online Material includes supplementary methods, experiments and discussion with tables and figures.

1 Supplementary Methods

For all experiments, sensory variables could be read out from population responses from any column whose CRF overlapped with the center stimulus (in practice we used the column located exactly in the middle of the center stimulus). Stimuli were scaled such that the surround overlapped with both the near excitatory and the far inhibitory eCRF.

Orientation tilt

We modeled orientation-selective units with bell-shaped tuning curves which tile the visual field. The tuning bandwidth was 23^o. The shift in orientation perceived by the model was computed as the difference between the orientation decoded from the center of a stimulus, with and without presenting the surround stimulus.

Motion direction tilt

We modeled populations of motion direction selective units with bell-shape tuning. We used the same stimuli as in the original study by Kim and Wilson (1997), holding the direction of motion of the center stimulus fixed, and varying the direction of motion of the surround stimulus between 0^o and 180^o. The tuning bandwidth was set to 72^o. The shift in motion direction perceived by the model was computed as the difference between the direction decoded from the center of the stimulus, with and without presenting the surround stimulus.

Hue tilt

We considered population units based on the V1 opponent color channels described in (Zhang et al., 2012). Isoluminant center-surround stimuli with uniformly and independently sampled hues were used for both the center and surround as done in the original study by Klauke and Wachtler (2015). The shift in hue perceived by the model was computed as the difference between the hue decoded from the center of the stimulus, with and without presenting the surround stimulus.

Instead of using a population vector model as done for the other modalities, we used cross-validated

ridge regression to decode the sine and cosine of hue. For training and testing, we used a dataset of 144 colored patches uniformly sampled in the hue domain ($n = 36$), presented on an approximately isoluminant achromatic background, at two saturation levels (0.20, 0.50) and two value levels (0.60, 1.0). Varying those levels or adding more of them had no significant effect either on the results or the accuracy of the decoding. We shuffled all the stimuli and randomly excluded 20% of them from the training procedure for testing; accuracy on the resulting test set was always near perfect (well above $R^2 = 0.95$.)

Depth induction

We modeled this experiment using population of units tuned to binocular disparity. We decoded depth from population responses centered on the test stimulus as a function of the lateral separation between the center stimulus and the flankers. We used stimuli similar to those used in the original study by Westheimer and Levi (1987), where a center line, presented at zero disparity, was flanked by two smaller squares, presented at crossed or uncrossed disparities. The tuning bandwidth was 1.15 minutes of visual angle. To assess the strength of the illusion, we used the "induction coefficient" measure as done in (Westheimer, 1986; Westheimer and Levi, 1987): we first computed the difference between the decoded disparities for the center stimulus, with and without presenting the flanker stimuli. We then reported that difference as a percentage of the difference in disparity between center and flankers.

Motion direction induction

We modeled this experiment using the same stimuli and protocol as used in the original study by Murakami and Shimojo (1996). The tuning bandwidth was set to 57^o. We measured the perceptual shift through the Point of Subjective Equality (PSE), i.e., corresponding to the coherence of the center moving dots (or center coherence) at which the model had an equal probability of detecting either direction of motion. For each stimulus size, estimates of the PSE were derived from fitted model psychometric curves, each computed across 25 trials for 21 regularly spaced values between -100% and 100%. Similarly, we also computed the Point of Maximum Variance (PMV) across trials, i.e., the center coherence at which the population code for the center motion direction was maximally flat (ambiguous). Following the prescription of Ma et al. (2006), the variance of the population response curve was interpreted as a measure of uncertainty in the value decoded under the ideal neural observer model; i.e., narrow-band population response curves resulted in estimates of the sensory variable associated with high confidence levels, whereas wide-band or flat population response curves resulted in low-confidence estimates. The PMV reliably matched the PSE, so we averaged the PMV and PSE together and across trials to yield the final estimate of the PSE for each stimulus size.

Enhanced color induction

We modeled this experiment using the same stimuli as used in the original study (Shevell and Monnier, 2005). We used the b* axis of the CIE L*a*b* color space to measure perceptual shifts along the blue-yellow opponent axis, defined as the difference between the b* coordinate decoded

from the test stimulus and the true b* value of the test stimulus.

2 Supplementary Experiments

Cross-orientation normalization

Inhibition in the CRF has long been documented through cross-orientation normalization (Heeger, 1993; Carandini and Heeger, 1994), where the CRF response to a primary oriented grating may be suppressed by superimposing an orthogonal masking grating. A widely-known model of single-cell and population responses alike (divisive normalization, reviewed in Carandini and Heeger, 2012), has been successful at capturing the remarkable contrast dependency of cross-orientation normalization. When the two gratings are presented at approximately equal contrasts, the population response to the resulting plaid is best described as the average of the population responses to either gratings if they were presented alone (the summation regime). However, if the two gratings contrasts are disparate enough, the resulting population response looks like the response to the grating with the strongest contrast, as if the weaker-contrast grating was not presented (the winner-take-all regime).

Our model, which includes untuned inhibition within the CRF, captures the balance between the summation vs. the winner-take-all regimes and its dependence on both gratings' contrasts. We digitally extracted electrophysiology data from Figure 4A, 4E and 4G from (Busse et al., 2009) in Figure S2 and show the corresponding model data in Figure S3.

Feature-matched surround suppression

Center-surround stimuli are widely used in neurophysiology to probe the mechanisms underlying surround suppression (Hubel and Wiesel, 1968; DeAngelis et al., 1994; Weliky et al., 1995; Petrov et al., 2005; Ozeki et al., 2009). In many studies, a cell is driven by a center grating at the optimal orientation; presenting a surround grating with the same orientation (or extending the center stimulus beyond the CRF and into the eCRF) suppresses the CRF response.

However, since the center grating was chosen to be optimal for the recorded cell, it remains unclear whether the surround grating is maximally suppressive because it matches the cell's orientation preference, or because it matches the orientation of the center grating. Two studies in V1 (Trott and Born, 2015; Shushruth et al., 2012) support the latter hypothesis, i.e., surround suppression is strongest when the surround stimulus matches the center in orientation, irrespective of the cell's orientation preference. This excludes simpler models where a center cell is simply selectively targeted by tuned inhibitory connections from its surround (Schwabe et al., 2006).

Similarly to Shushruth et al. (2012), we have found that this phenomenon is best accounted for by weakly-tuned recurrent excitation within an (orientation) hypercolumn. Such excitation places the hypercolumn in a highly recurrent regime where a cell may be driven by a stimulus at a sub-optimal orientation. Then, the most suppressive surround stimulus matches the center orientation because it withdraws most of the cell's input.

Our model explains neurophysiology data (digitally extracted from Figure 1B of Trott and Born (2015), curve labelled as "surround tuning") for such feature-matched surround suppression (see

Figure S3). It also explains another experiment where the presentation of a surround grating selectively suppresses population responses to the identically-oriented component of a two-grating plaid presented in the CRF, in effect "cancelling" the contribution of the plaid component matched to the surround (see Figure S5).

The necessary mechanisms of the contextual model

Our model successfully recreated human perceptual and primate neurophysiological responses across a variety of contextual phenomena. This suggests that the model's full suite of mechanisms are sufficient for a general explanation of contextual processing, but are they all necessary? We investigated this with a large-scale lesion screening procedure that identified the minimal version of the model. This procedure involved lesioning a mechanism in the model by setting its weights to 0, and then optimizing the model's remaining free parameters for explaining contextual phenomena. We hereafter refer to this as the lesioning procedure. We used the lesioning procedure to separately measure the necessity of each of the model's critical components. In total, we tested: (1) the necessity of each CRF and eCRF mechanism, (2) the validity of our assumptions about their spatial configurations, (3) the importance of asymmetric facilitatory vs. suppressive contributions for explaining contextual phenomena, and (4) the sensitivity of the model to a range of stimulus tuning properties and relaxed constraints on its patterns of connectivity. We performed this analysis on all contextual phenomena explored in the main and supplemental text except for Figure 6, which was excluded because its computational complexity rendered it intractable for this procedure.

The lesioning procedure was performed separately on each component of the model. We also

applied the free parameter optimization routine to the full model without lesioning any of its components. This supported identification of what parameters most accurately explained each of the contextual phenomena. This procedure is equivalent to solving:

$$C_{lesion} = \underset{\forall j \in J}{\arg\max}\, f(\rho_j) \tag{10}$$

Where C_{lesion} is the selected model from the optimization procedure over $J = 1{,}000$ iterations for a specific model configuration (e.g., a model with a lesion to its far eCRF). This optimization is performed over hand-tuned model parameters α, β, μ, ν, γ, δ in the space of all possible combinations of these parameters across the perceptual phenomena discussed in the main text. ρ is a vector of correlation scores for a selected model configuration describing the quality of its simulations for each perceptual phenomenon. f is a monotonic function of ρ as follows:

$$f(\rho) = \mu(\rho_j)/\sigma(\rho_j) \tag{11}$$

Where μ calculates the mean correlation and σ calculates the standard devation of correlations across perceptual phenomena for a sampled model configuration. In effect, f penalizes models by σ for overfitting on a subset of perceptual phenomena, making it a better method than simply maximizing correlation for selecting model configurations that are generally successful at simulating perception. Pearson correlation is desirable for this optimization procedure because it does not penalize irrelevant differences between model data and perceptual data, such as shifts of the mean and rescaling.

Each set of parameters applied to a model configuration was selected by searching over an exponentially spaced grid around each of the hand-tuned model parameters listed above. We define this sampling procedure S, which was applied on every iteration to each of the hand-tuned model parameters p discussed in the main text:

$$S(p) = |p + unif[-1,1]^{unif[-2,2]}| \tag{12}$$

where $unif$ denotes uniform sampling in the specified range. The parameters that maximize C_{lesion} yield optimal performance (as measured by Pearson correlation) for a lesioned version of the model in recreating observers' responses across contextual phenomenon.

While we found all model mechanisms to be necessary, we did find that the full version of the model was relatively tolerant to parameter perturbations caused by this optimization procedure. See Figure S9 for histograms of these scores. Plotting performance of the full model across these parameters on any particular problem revealed that the majority of them yielded simulations that were qualitatively similar to the perceptual data. Variations in performance potentially provide insight into individual differences in perception of contextual phenomena. See Figure S13 for an example of the qualitative variability we observed for a representative phenomenon.

A combination of quantitative and qualitative evidence demonstrates that our model performs best when it contains the full suite of CRF and eCRF mechanisms. Lesioning either of the eCRF mechanisms (near excitation or far inhibition) diminishes the model's ability to explain observers' behavior for several phenomena (Figure S8). Lesioning the near eCRF tuned excitation degrades

the ability of the model to explain phenomena demonstrated in Figure 3C, Figure 5, Figure 7, and Figure S5. Lesioning the far eCRF tuned inhibition diminishes the model explanatory power of Figure 3C (see also Figure S14), Figure 4, Figure 7,and Figure S5. Indeed, the full model was significantly better at capturing these phenomena than either lesioned version, as measured with 2-tailed t-tests comparing the maximum correlations accrued by each model across each of the eight phenomena included in the lesion optimization (Full model vs. near eCRF excitatory lesion: $t(7) = 5.102$, $p = 0.001$; Full model vs. far eCRF inhibitory lesion: $t(7) = 3.911$, $p = 0.006$).

Lesions to either of the model's CRF mechanisms reveal that they were less important for explaining the high-level contextual phenomena discussed in the main manuscript than its eCRF mechanisms (Figure S7). A closer inspection of lower level phenomena, however, revealed their importance for explaining contrast-dependence tuning in V1. We measured responses from the full optimized model and versions of the model with lesioned CRF mechanisms to a stimulus that varied in size and contrast. Lesioning either the weakly tuned excitation or untuned inhibition CRF mechanisms qualitatively harmed the model's ability to discriminate between stimuli of different contrasts (Figure S1). For the model with lesioned weakly tuned excitation, this is immediately apparent: it is qualitatively worse than the full model at discriminating stimulus contrast until the size of the stimulus extends into the near eCRF. For the model with lesioned untuned inhibition, the opposite phenomenon is observed: contrast discrimination fails as soon as the stimulus extends into the inhibitory far eCRF.

Lesion optimization also highlighted the importance of our key assumption of asymmetry between excitation versus inhibition for explaining contextual phenomena. Excitation in the model is purely

additive and only depends on pre-synaptic activity. But inhibition depends on both pre- and post-synaptic activity and results in a combination of subtractive and divisive effects (Carandini and Heeger, 2012). A model with lesioned presynaptic shunting inhibition was significantly worse at explaining the contextual phenomena than the full model ($t(7) = 2.631$, $p = 0.034$; CRF α and far eCRF β, Figure S8).

This lesion-screening framework allowed us to measure the necessity of having spatially seperate versus overlapping CRF and near eCRF mechanisms. We created a version of the model in which the influence of CRF and near eCRF mechanisms were averaged together. This configuration yielded significantly worse than the full model at explaining the contextual phenomena ($t(7) = 2.908$, $p = 0.023$; Spatially overlapping CRF and near eCRF, Figure S8). The impact of having spatially distinct CRF and near eCRF mechanisms was most apparent on Figure 4A-B, Figure 5 and Figure 7.

We also investigated the necessity of seperate near and far eCRF regions for explaining these perceptual phenomena. We optimized a model with completely overlapping near and far eCRFs – extending from the proximal point near eCRF to the distal point of the far eCRF. The spatially seperate eCRFs of the full model yielded significantly better performance than this spatilly overlapping version at explaining contextual phenomena ($t(7) = 2.552$, $p = 0.038$; Spatially overlapping near eCRF and far eCRF, Figure S8).

Having found evidence that each of the contextual model's mechanisms are necessary to explain the full array of contextual phenomena, we explored its robustness to variations in the shape

of the model unit tuning curves. We did this over $1,000$ iterations by resampling the tuning curve properties for each contextual phenomenon with the sampler S. For disparity (Cumming and Parker, 1997), motion direction (Albright et al., 1984) and orientation (Ringach et al., 1997) this involved resampling tuning curve bandwidth with the sampler S. For color opponent tuning phenomena, for which we did not assume a bell-shape tuning curve, this involved resampling the response threshold from a uniform distribution in $[-0.5, 0.5]$. On every iteration of the procedure, we simulated each contextual phenomenon with the optimized full model after resampling its tuning properties. This approach revealed that, with the exception of Figure 4, the full optimized model was robust to a wide range of tuning properties. See Figure S11 for histograms of the model's ability to explain contextual phenomena as these parameters were varied.

The contextual model has strictly "tuned" connections from the eCRF onto the CRF, with a one-to-one mapping between computational units preferring the same stimulus features. We tested how important this constraint is to the model's performance by relaxing this strict one-to-one mapping into a weakly tuned eCRF-CRF mapping (Figure S10). eCRF units in the orientation, motion, and disparity domains had normally distributed connectivity, centered at a target unit tuning preference θ_k with standard deviation ς. Because we consider color-opponent eCRF units with monotonic tuning curves in the color domain, assumptions of normality are inappropriate. For these phenomena we instead set $w_{kk} = (\varsigma\sqrt{2\pi})^{-1}$ and $w_{jk} = \text{const.}$ (when $j \neq k$; under the constraint that the weights sum up to $r1$). In each case the lesion screening procedure was used to search the extent to which eCRF unit connectivity could be weakened without destroying the model's ability to simulate contextual phenomena. Model performance was recorded over $1,000$

iterations while the standard deviation ς of these connectivity schemes was randomized with the sampler S (see 12 for details on the sampling procedure). For each contextual phenomenon this procedure yielded many successful simulations, and at times outperformed strict tuning. We expect that additional work on incorporating more anatomically plausible connectivity into the full model will yield even better performance than we report here.

Taken together, our large-scale lesion screening procedure indicates that the mechanisms in our full model are not only sufficient for explaining contextual phenomena, but also necessary.

3 Supplementary Discussion

Neurophysiology evidence for shifts in single-cell tuning curves

A key prediction of the model is that shifts in population responses may underlie perceptual shifts in phenomena as in the two regimes of the tilt effect. Several electrophysiology studies have provided evidence for shifts in single-unit tuning curves that matched behavioral data on perceptual shifts. In the motion domain, Li et al. (1999) recorded from speed-tuned neurons in V1 and found that the presence of a slower (faster) moving stimulus in a neuron's eCRF shifts its preferred speed towards faster (slower) values. In the color domain, shifts in tuning curves that are consistent with color contrast where found both in V1 (Wachtler et al., 2003) and V4 (Kusunoki et al., 2006). In two of these studies (Li et al., 1999; Kusunoki et al., 2006), single-unit recordings were shown to be consistent with the animals' behavioral responses. In the disparity domain, Thomas et al. (2002) recorded from single cells in monkey V2 and found that presenting stimuli with different disparities in their CRF and eCRF resulted in tuning curve shifts towards the disparity presented

in the eCRF. This behavior is consistent with the computation of relative disparity. Interestingly, this yields a rather counter-intuitive model prediction for the associated perception: single-cell tuning curve shifts *towards* the eCRF disparity translate, at the population level, to a shift of the estimated disparity in the CRF *away* from that of the eCRF. That is, apparent attraction at the level of individual tuning curves effectively corresponds to repulsion at the perceptual level; this paradox emphasizes the necessity for a population-level analysis to fully appreciate the perceptual effect of extra-classical modulation.

Evolution of perceptual shifts as a function of population response curve bandwidths

In the orientation and motion direction tilt effects, our model suggests that repulsive shifts happen when the eCRF suppresses one side of the CRF population response ("pushing" it away from its own mode) and that the strength of such a push grows with the overlap between these populations (because of the inhibition increasing with post-synaptic activity). Consistently, Goddard et al. (2008) have found that in the orientation tilt effect, using stimuli embedded in noise, broadening the orientation power of the center stimulus (yielding broader-band center population response curves) results in larger repulsive shifts than with narrower-band stimuli (e.g., the gratings used in (O'Toole and Wenderoth, 1977)).

Figure S12 shows a prediction of the model regarding how the largest achievable repulsive shift should evolve as a function of the bandwidths of the center and the surround populations. For example, large effects are predicted when the center bandwidth is much larger than the surround bandwidth (up to a factor of 2), presumably because high overlap between population responses

from the CRF and the eCRF can then be achieved while keeping them offset from each other in the orientation domain (perfect alignment would result in equal inhibition on either side of the CRF population response curve, yielding no shift).

Conversely, our model predicts that the attractive regime of the tilt effect can only exist when CRF and eCRF populations do not overlap significantly. By extension, this implies that the attractive regime weakens, or even disappears altogether, when typical population bandwidths are so large that there is no "room" left in the tuning domain for the CRF and eCRF populations to exist without overlapping. This is consistent with the observation that the attractive regime in the motion direction tilt effect is much less pronounced than in the orientation tilt effect (Figure 3), as neurophysiology data from V1 suggests that direction-tuned cells have a higher tuning bandwidth than orientation-tuned cells (Ringach et al., 2002; Albright et al., 1984).

Surround inhibition explains perceptual repulsion

To explain color induction, Klauke and Wachtler (2015) have proposed a phenomenological model based on weakly-tuned surround suppression, which reduces the gain of individual bell-shaped, hue tuning curves in the center. This results in center population shifts that are qualitatively consistent with their behavioral data. Our model thus provides computational evidence for their population-level explanation, using a similar form of surround suppression with sharper tuning. Interestingly, we show that neural populations with explicit bell-shaped hue tuning are actually not necessary and perceptual shifts consistent with psychophysical data can be accounted for with neural populations tuned to cardinal color opponency. Additionally, our experiments involving

visual modalities with bell-shaped tuning curves (e.g., orientation, motion direction tilt effects) also confirmed their intuition that tuning bandwidth controls the maximal amplitude of the perceptual shift.

Goddard et al. (2008) also explains the orientation tilt effect with a computational model based on tuned surround inhibition. Their model also incorporates a divisive normalization term but lacks a realistic model of recurrent connections. In order to explain the emergence of an attractive regime at dissimilar orientations in the center and the surround, these authors modeled the surround orientation tuning curve as a "Mexican hat", that has a positive peak around the center orientation, and becomes negative for values further away. This is superficially similar to the balance between the influences of the near and far eCRFs in our model. However, such a model assumes that excitatory and inhibitory influences are evenly distributed across the surround, and thus would fail to explain the dependence of attraction vs. repulsion on stimulus size in our second set of illusions. Interestingly, the authors allow the relative shape of the surround tuning curve (i.e., the excitation-inhibition balance), as well as the overall strength of surround modulation, to vary across experiments to fit experimental data. This suggests that carefully allowing the analogues of these parameters in our model to vary across visual modalities may result in closer quantitative fits.

Enhanced color shift effect: beyond S-cone contrast

Monnier and Shevell (2003); Shevell and Monnier (2005) modeled enhanced color shifts through an antagonist CRF organization corresponds to an S-ON center (i.e., the neuron response increases

with short cone activation in its center) and S-OFF surround (i.e., the neuron response decreases with short cone activation in its surround). Their explanation is consistent with ours: at the optimal spatial frequency (which exists because spatial frequency controls the overlap of the different parts of the stimulus with the different sub-regions of the CRF or the eCRF), the two colors present in the contextual stimulus create cooperative shifts in S cone activation. However, such a model does not predict enhanced perceptual shifts using patterned contextual stimuli that have zero contrast in S-cone activation; the authors report little perceptual shifts along L and M cone activations, which could be due to the specific stimuli they used. We rendered a simple equivalent of their stimuli with contextual colors that should elicit little to no contrast in S-cone activations (Figure S6E-H). In other words, CRFs based on S-cone activations should respond uniformly across the surround stimulus of this modified illusion; yet, the shift seems just as vivid (see caption for details) as with the original stimuli (Shevell and Monnier, 2005). Interestingly, the model reported in (Shevell and Monnier, 2005) to best account for behavior after parameter fitting results in a surround up to 5 times wider than the center, and a center almost 2 times as responsive to visual stimulation as the surround. This seems to agree better with an eCRF model like ours, which has an inhibitory far eCRF that is (relatively) much wider than the OFF region of traditional center-surround CRFs.

	parameter description	**value**
η	Input time constant	6.00
ε	Input gain	0.50
ξ	Afferent strength	4.50
τ	Output time constant	6.00
σ	Output gain	0.50
α	Untuned suppression strength (divisive)	1.00
μ	Untuned suppression strength (subtractive)	1.00
β	Tuned suppression strength (divisive)	3.00
ν	Tuned suppression strength (subtractive)	0.30
γ	Tuned facilitation strength	1.00
δ	Untuned facilitation strength	1.00
ς	Standard deviation of tuned facilitation weights	0.15
$R[\mathbb{N}_{x,y}]$	Radius of near eCRF of (x,y) (in number of hypercolumns)	9
$R[\mathbb{F}_{x,y}]$	Radius of far eCRF of (x,y) (in number of hypercolumns)	29

Table S1. ***Model parameters.*** Hypercolumns are organized in a regularly-spaced, square grid. Radii for near and far eCRFs are then defined in number of hypercolumns on that grid.

	Separate	**Competition**	**Cooperation**
Shifts	Goddard et al. (2008) (o.), Westheimer (1986); Westheimer and Levi (1987) (d.)	Gibson and Radner (1937); O'Toole and Wenderoth (1977) (o.), Marshak and Sekuler (1979) (m.), Loomis and Nakayama (1973); Norman et al. (1996); Baker and Graf (2010) (s.)	Monnier and Shevell (2003); Shevell and Monnier (2005) (c.)
Bumps/Notches	Goddard et al. (2008) (o.), Murakami and Shimojo (1993; 1996) (m.)	Smith et al. (2001); Klauke and Wachtler (2015) (c.)	Anstis (2006) (b., c.)

Table S2. ***How existing psychophysics studies on contextual phenomena across modalities fit in the proposed taxonomy.*** See Figure 8in the main text for a companion figure. **Abbreviations:** b: brightness, c: color, d: depth, m: motion direction, o: orientation, s: motion speed.

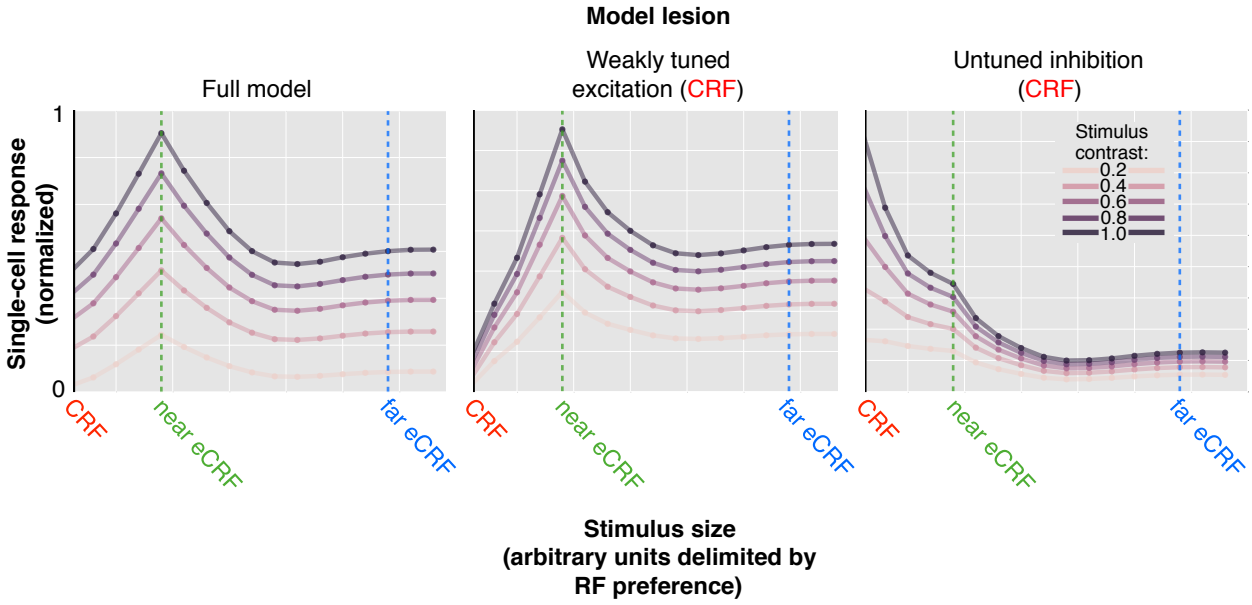

Figure S1. ***Both intra-columnar recurrent excitation and inhibition are necessary for explaining neural contrast responses.*** Model response is plotted as a function of the size (x-axis) and contrast (line color) of a stimulus. Vertical dashed lines show the spatial extent of the CRF (red) and near excitatory (green) and far inhibitory (blue) eCRFs. All remaining model parameters were optimized for the contextual phenomena in the main manuscript after selectively lesioning either of the two mechanisms. Only the full model exhibits expected tuning properties: the CRF response discriminates stimuli and reaches its peak response once stimuli expand into the excitatory near eCRF receptive field.

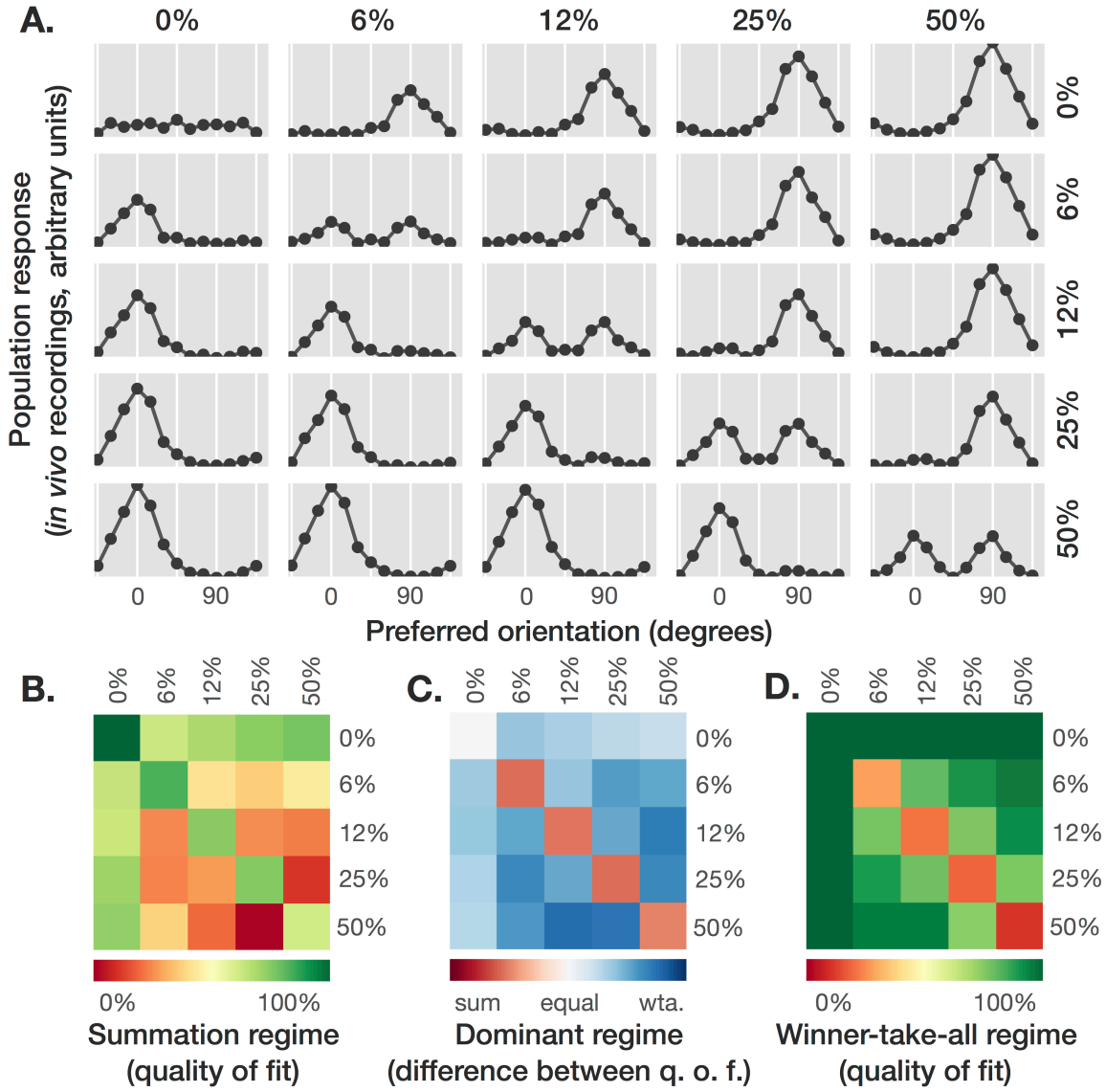
A.
0%
6%
12%
25%
50%
Population response
(in vivo recordings, arbitrary units)
0
90
Preferred orientation (degrees)
B.
Summation regime
(quality of fit)
0%
100%
C.
sum
equal
wta.
Dominant regime
(difference between q. o. f.)
D.
Winner-take-all regime
(quality of fit)

Figure S2. ***Balance between summation and winner-take-all regimes during the presentation of cross-oriented gratings.*** **Upper plots:** cat V1 electrophysiology data digitally extracted from Figure 4A in (Busse et al., 2009). Population recordings of orientation-tuned cells in response to the presentation of orthogonally-oriented gratings at variable contrast levels (rows indicate the contrast of the 0^o grating and columns indicate the contrast of the 90^o grating). When the contrasts of the two gratings are approximately equal, the population response is best described as the sum of the population responses to either grating presented in isolation. When they differ markedly, the population response is best described as the population response to the grating with the strongest contrast presented in isolation. This latter regime is also known as a winner-take-all regime as it seems that the response to the stronger stimulus predominates and suppresses the response to the weaker one. **Lower plots:** data reproduced from Figure 4E and Figure 4G in (Busse et al., 2009). Quality of fit of either a summation model or a winner-take-all model as a function of the contrasts of either component grating. As before, the relative strengths of either component grating controls the balance between summation and winner-take-all.

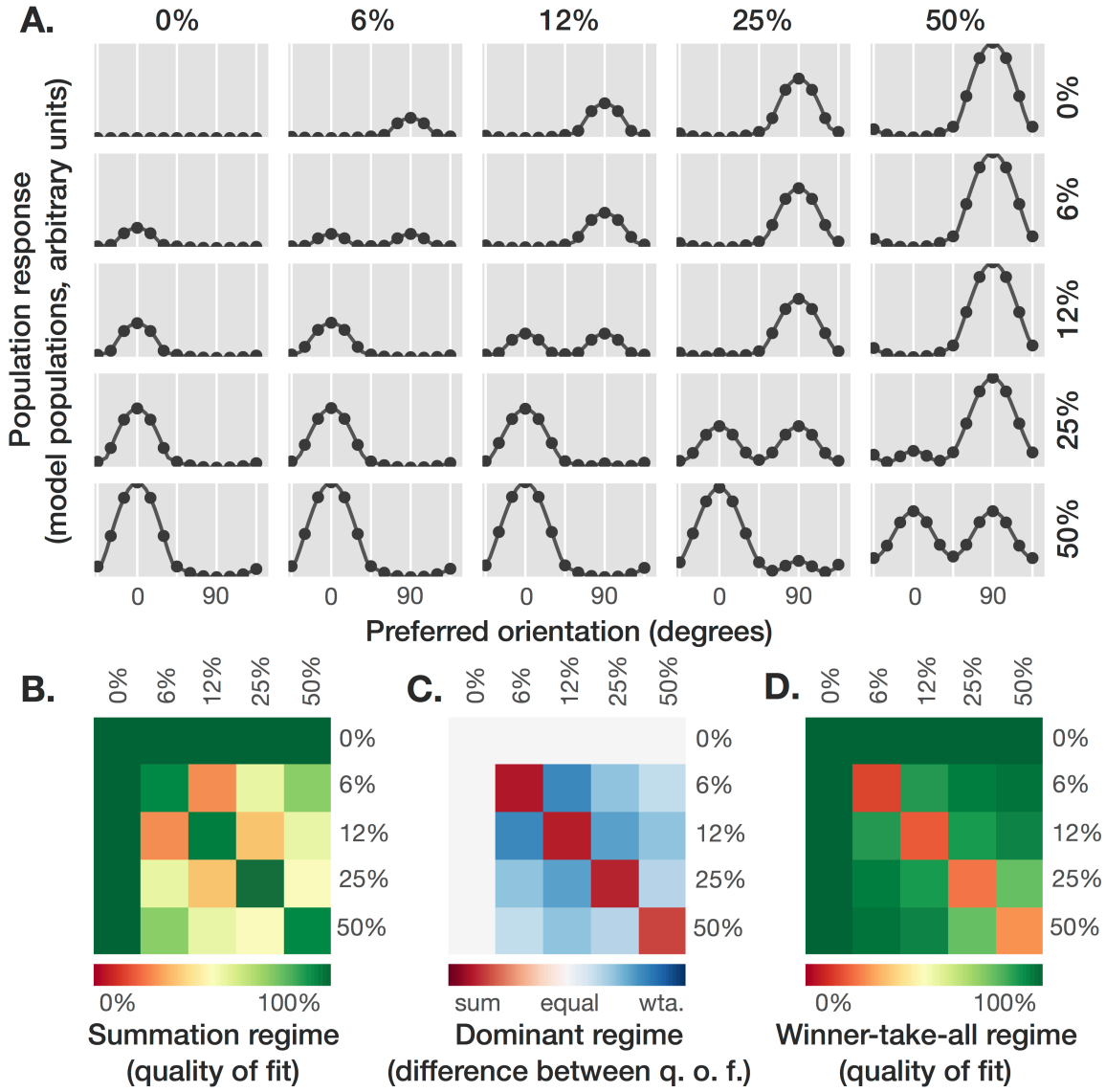

Figure S3. ***Balance between summation and winner-take-all regimes during the presentation of cross-oriented gratings in the model.*** Using populations of orientation-tuned cells, our model reproduces the neurophysiology data presented in Figure S2.

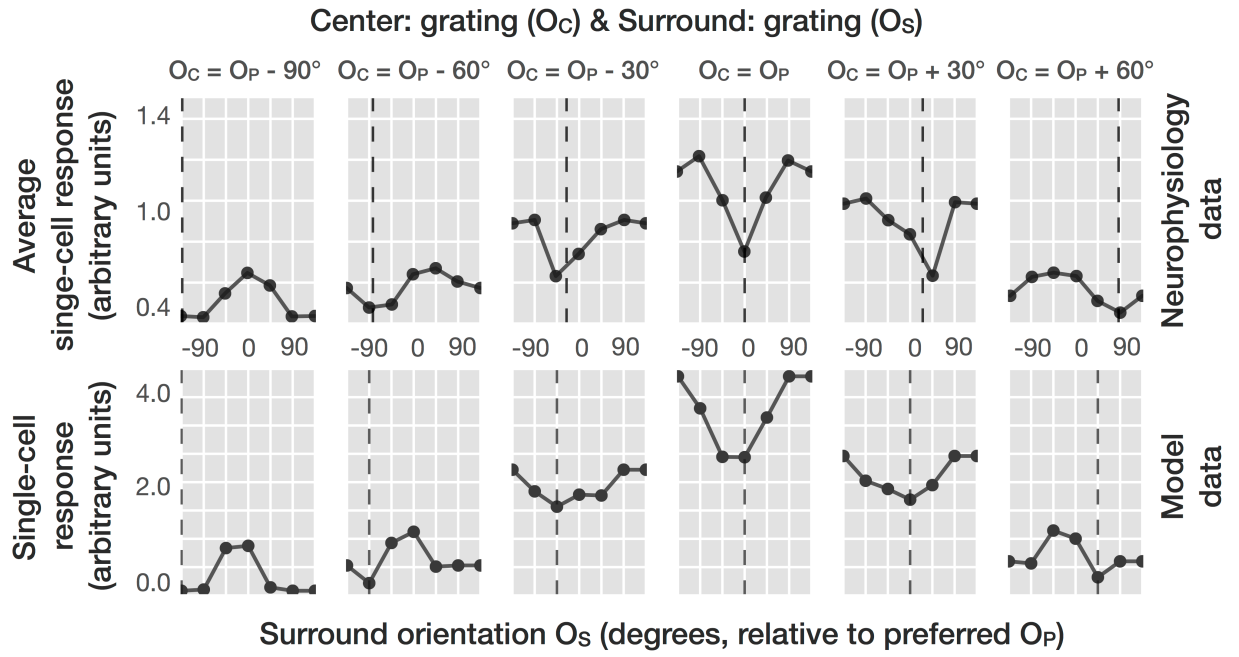

Figure S4. ***Feature-specific surround suppression: the most suppressive surround stimulus matches that of the center stimulus.*** **Top row:** monkey V1 electrophysiology data digitally extracted from Figure 1B (curve labelled as "surround tuning" in Trott and Born, 2015). Single-cell recordings from orientation-tuned cells (averaged and normalized across cells such that 0^o on the abscissas corresponds to O_P, the cell's preferred orientation) were made using center-surround gratings. Different center grating orientations O_C were presented in the CRF. For each value of O_C, the orientation of the surround grating O_S was systematically varied (abscissas). The key result is that the most suppressive surround orientation (vertical dashed black line) always matches the orientation of the center stimulus O_C, even for non-optimal center orientations. **Bottom row:** The model exhibits a similar feature-selective suppression as found experimentally.

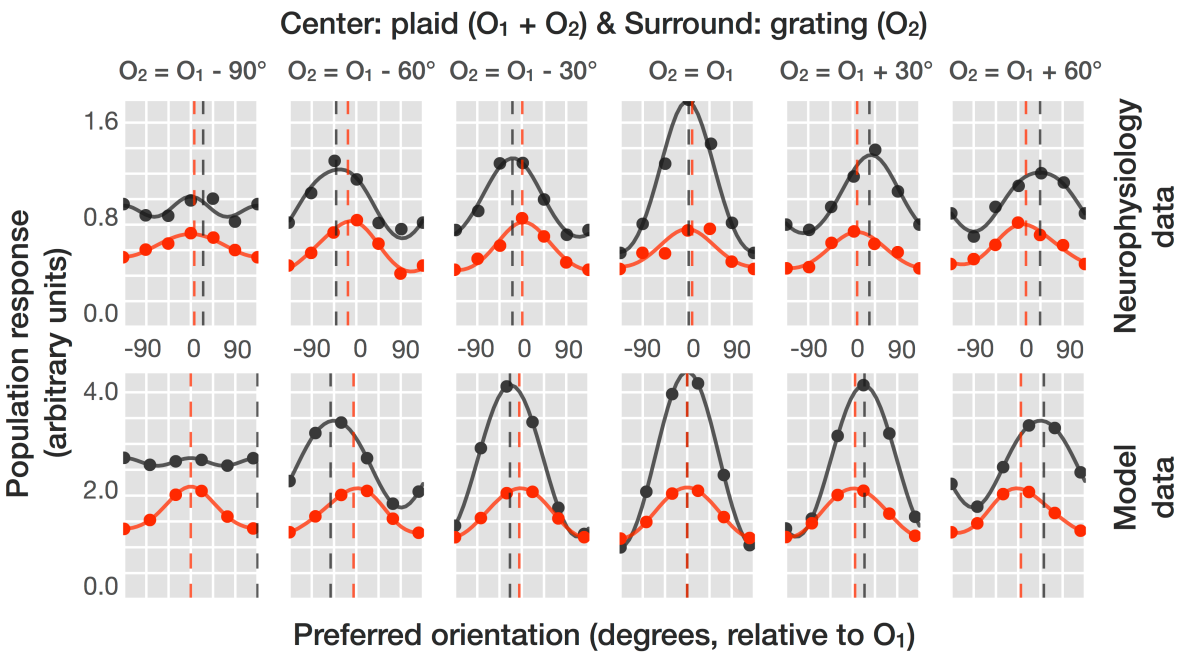
Center: plaid (O1 + O2) & Surround: grating (O2)
O2 = O1 - 90°
O2 = O1 - 60°
O2 = O1 - 30°
O2 = O1
O2 = O1 + 30°
O2 = O1 + 60°
Population response (arbitrary units)
1.6
0.8
0.0
-90
0
90
4.0
2.0
Neurophysiology data
Model data
Preferred orientation (degrees, relative to O1)

Figure S5. ***Feature-specific surround suppression: the surround stimulus suppresses the iso-oriented component of a center plaid.*** **Top row:** monkey V1 electrophysiology data reproduced from Figure 4 B from (Trott and Born, 2015). Population-level recordings were made from orientation-tuned cells. A plaid was first presented in the center (with components at orientations O_1 and O_2), alone. The resulting population response curves look like the averaged population responses to the presentation of either plaid component in isolation (black curves; black dashed lines placed at the orientation encoded by the corresponding populations; compare also to cross-orientation normalization, Figures S2 and S3). When a surround grating with the same orientation as one of the center plaid's components (e.g., O_2) was added to the surround, the corresponding plaid component (e.g., O_2) was selectively suppressed, resulting in a population response tuned to the other plaid component (e.g., O_1; red curves; red dashed lines corresponds to the orientation encoded by the corresponding populations). **Bottom row:** The model exhibits feature-selective suppression as well (the red dashed lines remain centered at O_1, the plaid component that does not match the surround orientation).

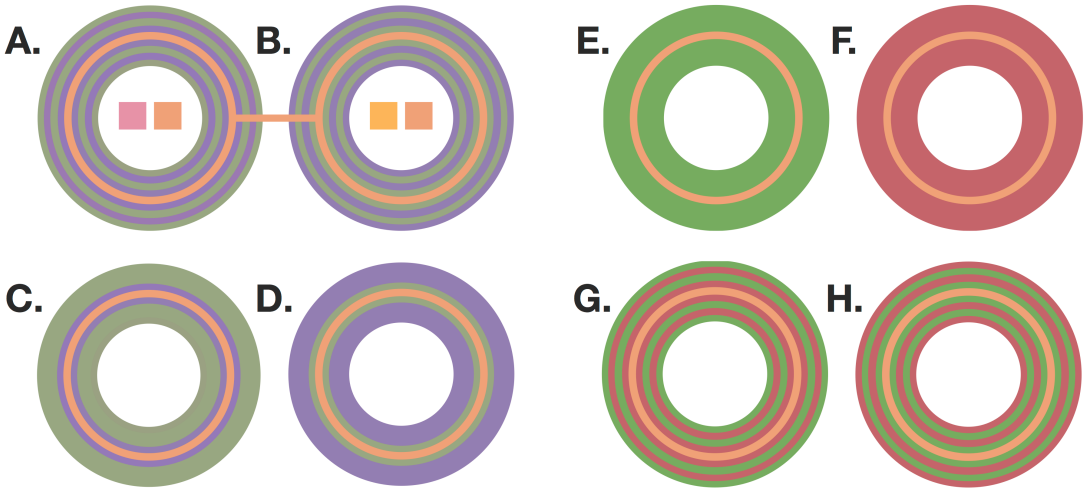

Figure S6. ***Enhanced color shifts: model readout and novel predictions.*** **A-B.** Stimuli were adapted from Figure 1 in (Monnier and Shevell, 2003). Though the test ring has the same color in either case (orange), patterned contextual stimuli of opposite phases result in color shifts in opposite directions. The squares in the middle of the rings represent the hue of the test ring decoded from our model (left square) and from a CRF-only model of opponent color processing for baseline comparison (Zhang et al., 2012; right square). The hue decoded from our model appears more consistent with human perception. **C-D.** Our model explains the illusion through the cooperative stimulation of the near and far eCRF by near-opposite hues (see main text and Figure 7), here purple and lime. As a result, we predict that the illusion should have the same intensity when the contextual patterned grating is replaced by two uniform regions with the right size in order to coincide with the near and far eCRFs (compare the appearance of the test ring in **C/D** to that in **A/B**. **E-H.** A previous explanation for this illusion involved a center-surround antagonism for S-cone activity (Monnier and Shevell, 2003; Shevell and Monnier, 2005). We rendered a new version of the illusion predicted by our model to be just as vivid as the original one, but whose patterned context lacks S-cone contrast (colors rendered approximately; appearances may differ on paper). Thus, whereas a S-cone antagonist model should see no difference in the appearance of the test ring in **E-H**, we predict a slight difference between **E-F**, and a very vivid difference

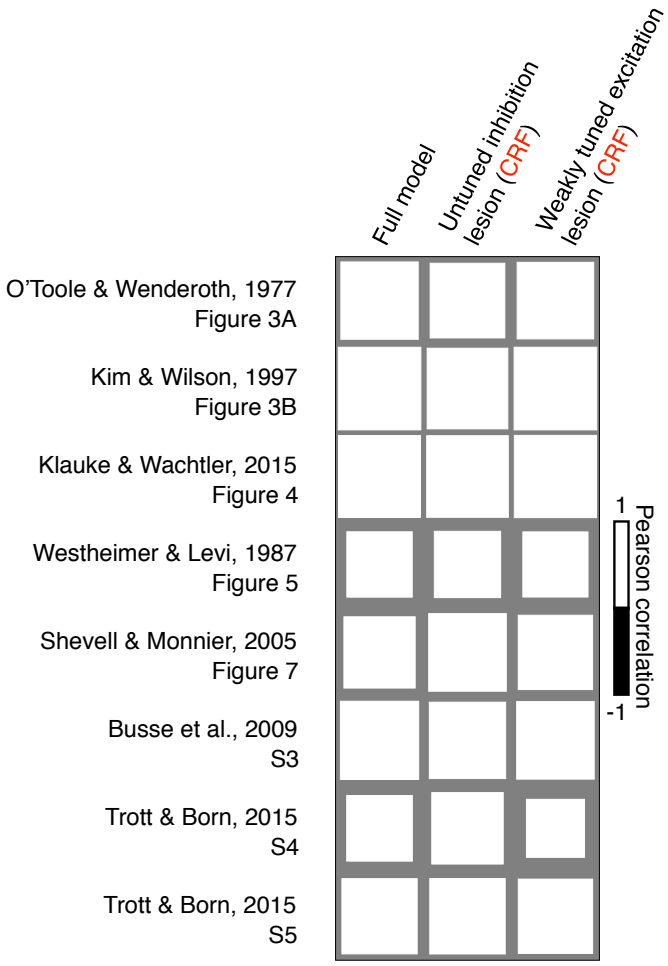
Full model
Untuned inhibition lesion (CRF)
Weakly tuned excitation lesion (CRF)
O'Toole & Wenderoth, 1977
Figure 3A
Kim & Wilson, 1997
Figure 3B
Klauke & Wachtler, 2015
Figure 4
Westheimer & Levi, 1987
Figure 5
Shevell & Monnier, 2005
Figure 7
Busse et al., 2009
S3
Trott & Born, 2015
S4
Trott & Born, 2015
S5
1
-1
Pearson correlation

Figure S7. ***Lesioning model CRF mechanisms mostly leaves its performance intact.*** We measured the model's ability to explain a variety of contextual phenomena (depicted along rows) when it was intact versus when one of its CRF circuits was lesioned (depicted along columns). Correlations between each experiment's behavioral data and model simulations are plotted with a Hinton diagram, where each square's size depicts correlation magnitude and color depicts the sign of the correlation. All correlations > 0. Model simulations follow 1000 iterations of parameter optimization. Figure 6 is omitted because the complexity of that simulation made the parameter search computationally intractable.

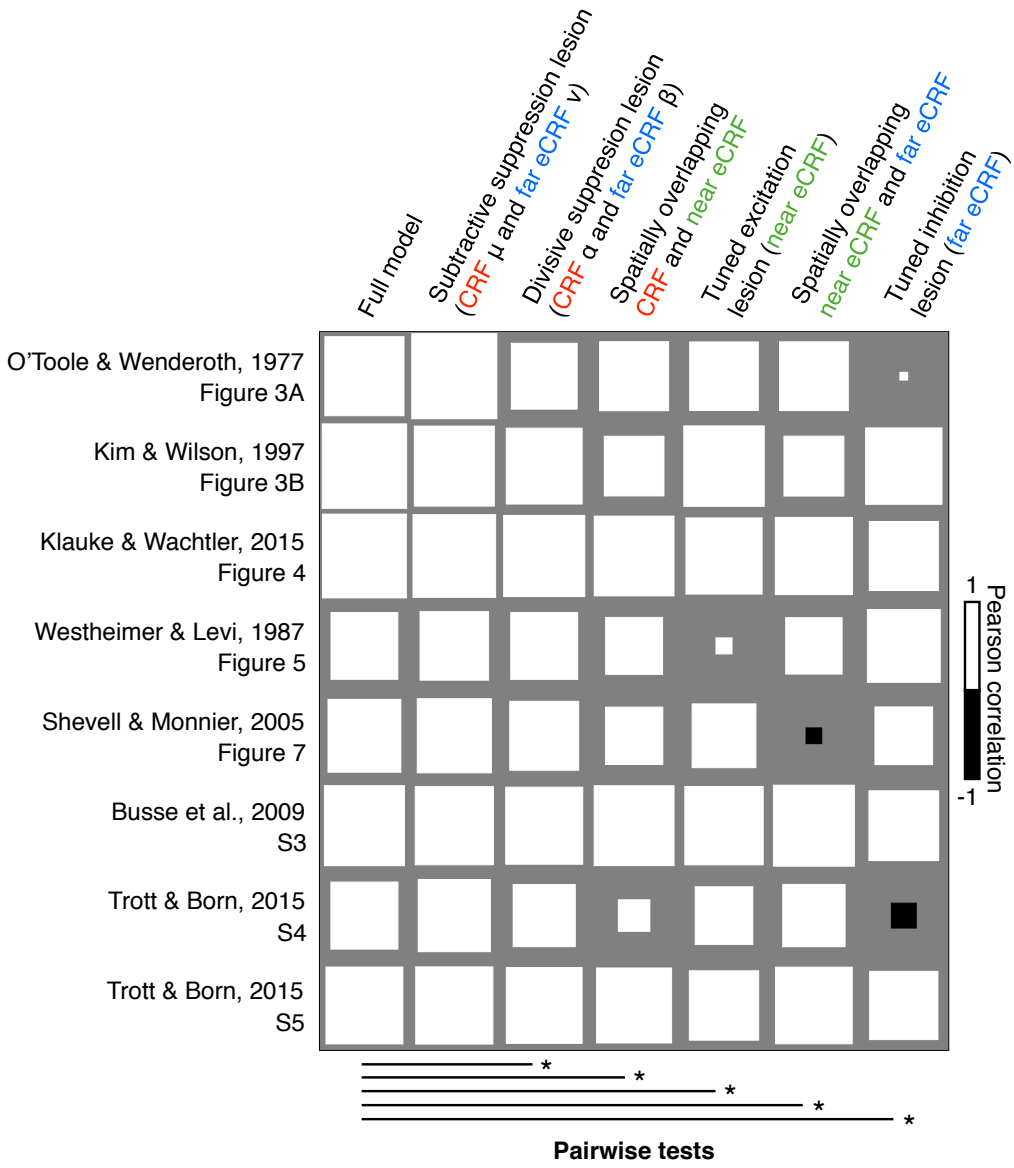
Full model
Subtractive suppression lesion (CRF μ and far eCRF ν)
Divisive suppresion lesion (CRF α and far eCRF β)
Spatially overlapping CRF and near eCRF
Tuned excitation lesion (near eCRF)
Spatially overlapping near eCRF and far eCRF
Tuned inhibition lesion (far eCRF)
O'Toole & Wenderoth, 1977 Figure 3A
Kim & Wilson, 1997 Figure 3B
Klauke & Wachtler, 2015 Figure 4
Westheimer & Levi, 1987 Figure 5
Shevell & Monnier, 2005 Figure 7
Busse et al., 2009 S3
Trott & Born, 2015 S4
Trott & Born, 2015 S5
1
Pearson correlation
-1
*
*
*
*
*
Pairwise tests

Figure S8. ***Model eCRF mechanisms are necessary for its ability to explain contextual phenomena.*** We measured the model's ability to explain a variety of contextual phenomena (depicted along the rows) when it was intact versus when one of its circuits was lesioned (depicted along the columns). Correlations between each experiment's behavioral data and model simulations are plotted with a Hinton diagram, where each square's size depicts correlation magnitude and color depicts the sign of the correlation. Model simulations follow 1000 iterations of parameter optimization. Figure 6 is omitted because the complexity of that simulation made the parameter search computationally intractable. Also note that we select the optimal model over all experiments which is not necessarily the optimal model for any given individual experiment. Lines and * at the bottom identify lesioned versions of the full model with correlations to perceptual data that are significantly worse than the full model. Comparisons are made with two-tailed *t*-tests. All tests shown are $p < 0.05$.

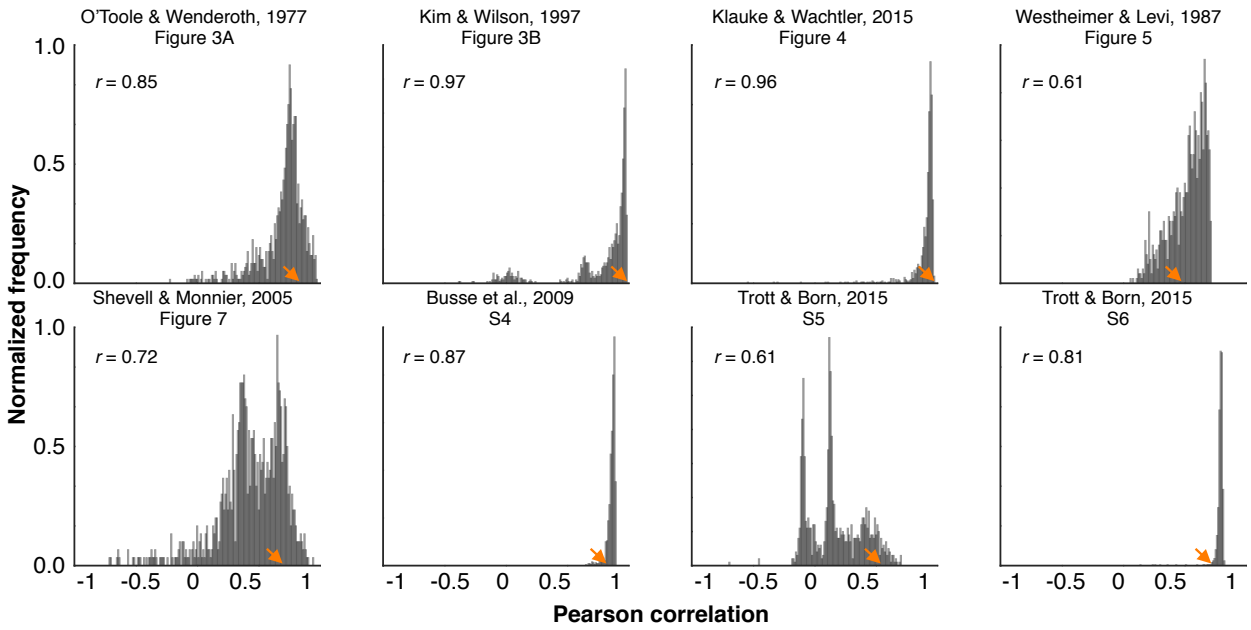

Figure S9. ***Distributions of fit derived from parameter optimization of the full model.*** Histograms depict accuracy of the full model in explaining each contextual phenomenon as its free parameters are randomly sampled with an exponential search grid over model parameters. Performance is measured with Pearson correlation. Correlations are derived for each of the 1000 iterations of the lesion-screening procedure. Correlations reported in the top left-hand corner of each panel denote performance of the optimized full model on each problem. These values are also highlighted with orange arrows. Figure 6 is not included because it is computationally intractable for the lesion-screening.

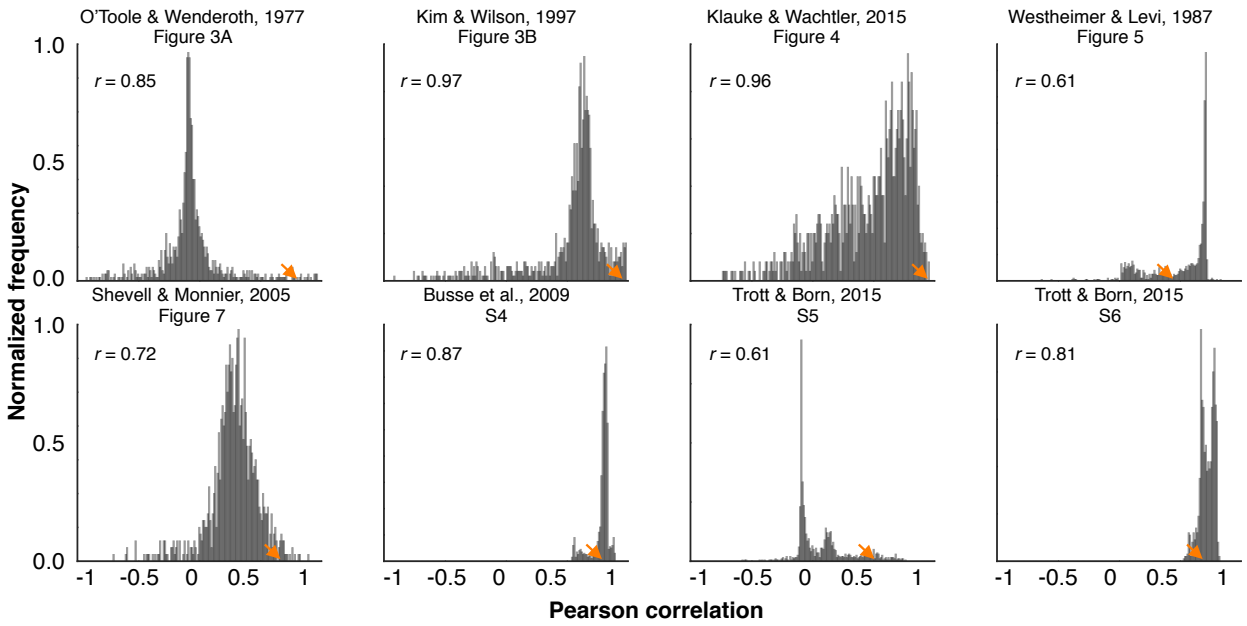
O'Toole & Wenderoth, 1977
Figure 3A
r = 0.85
Kim & Wilson, 1997
Figure 3B
r = 0.97
Klauke & Wachtler, 2015
Figure 4
r = 0.96
Westheimer & Levi, 1987
Figure 5
r = 0.61
Shevell & Monnier, 2005
Figure 7
r = 0.72
Busse et al., 2009
S4
r = 0.87
Trott & Born, 2015
S5
r = 0.61
Trott & Born, 2015
S6
r = 0.81
Normalized frequency
1.0
0.5
0.0
-1
-0.5
0
0.5
1
Pearson correlation

Figure S10. ***Distributions of fit for the optimized full model when its strict one-to-one CRF-eCRF connectivity is relaxed.*** Histograms depict the optimized full model's ability to explain contextual phenomena (measured by Pearson correlation coefficients between model predictions and perceptual data) as the one-to-one connectivity of its CRF-eCRF connections are relaxed into weakly-tuned, "many-to-many" CRF-eCRF connectivity. For orientation, motion, and disparity phenomena, each eCRF unit had normally distributed connectivity, which was centered at a target unit tuning preference θ_k with standard deviation ς. For the color domain we consider color-opponent model units with monotonic tuning curves and set $w_{kk} = (\varsigma\sqrt{2\pi})^{-1}$ and $w_{jk} = \text{const.}$ (when $j \neq k$; under the constraint that the weights sum up to $r1$). The lesion screening procedure was used to evaluate the model's ability to explain perception across a range of relaxed CRF-eCRF connectivity patterns. This involved randomizing ς over $1,000$ iterations, and then measuring the model's success in simulating the contextual phenomena with each pattern of connectivity. See 12 for details of the procedure for sampling ς. Pearson correlation values reported in the top left-hand corner of each panel denote performance of the optimized full model with strictly tuned connections on each problem. These values are also highlighted with orange arrows. Figure 6 is not included because it is computationally intractable for the lesion-screening.

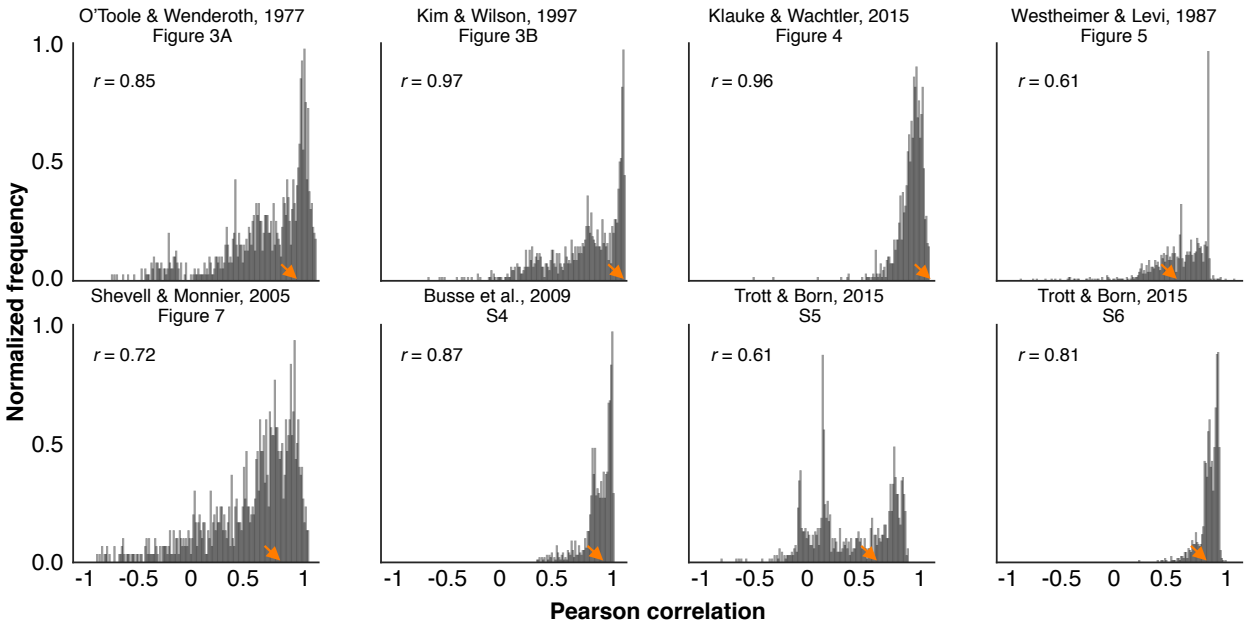
O'Toole & Wenderoth, 1977
Figure 3A
r = 0.85
Kim & Wilson, 1997
Figure 3B
r = 0.97
Klauke & Wachtler, 2015
Figure 4
r = 0.96
Westheimer & Levi, 1987
Figure 5
r = 0.61
Shevell & Monnier, 2005
Figure 7
r = 0.72
Busse et al., 2009
S4
r = 0.87
Trott & Born, 2015
S5
r = 0.61
Trott & Born, 2015
S6
r = 0.81
Normalized frequency
1.0
0.5
0.0
-1
-0.5
0
0.5
1
Pearson correlation

Figure S11. ***Distributions of fit for the optimized full model derived by varying the tuning bandwidth of the input units' tuning curves.*** Histograms depict the optimized full model's ability to explain perception (measured by Pearson correlations between model predictions and perceptual data) as the shapes of its tuning curves were randomized over $1,000$ iterations for each contextual phenomenon. We explored a range of parameters around the hand-tuned values reported in the manuscript (see Eq. 12 for sampling details). For disparity (Cumming and Parker, 1997), motion direction (Albright et al., 1984) and orientation (Ringach et al., 1997) this involved resampling tuning curve bandwidth. For color opponent tuning this involved resampling the response threshold since we did not assume a bell-shape tuning curve for the modality. See 12 for details of the sampling procedure. Note that for the color domain the sampling rangPearson correlation values reported in the top left-hand corner of each panel denote performance of the optimized model with tuning bandwidth optimized for each experiment. These values are also highlighted with orange arrows. Figure 6 is not included because it is computationally intractable for the lesion-screening.

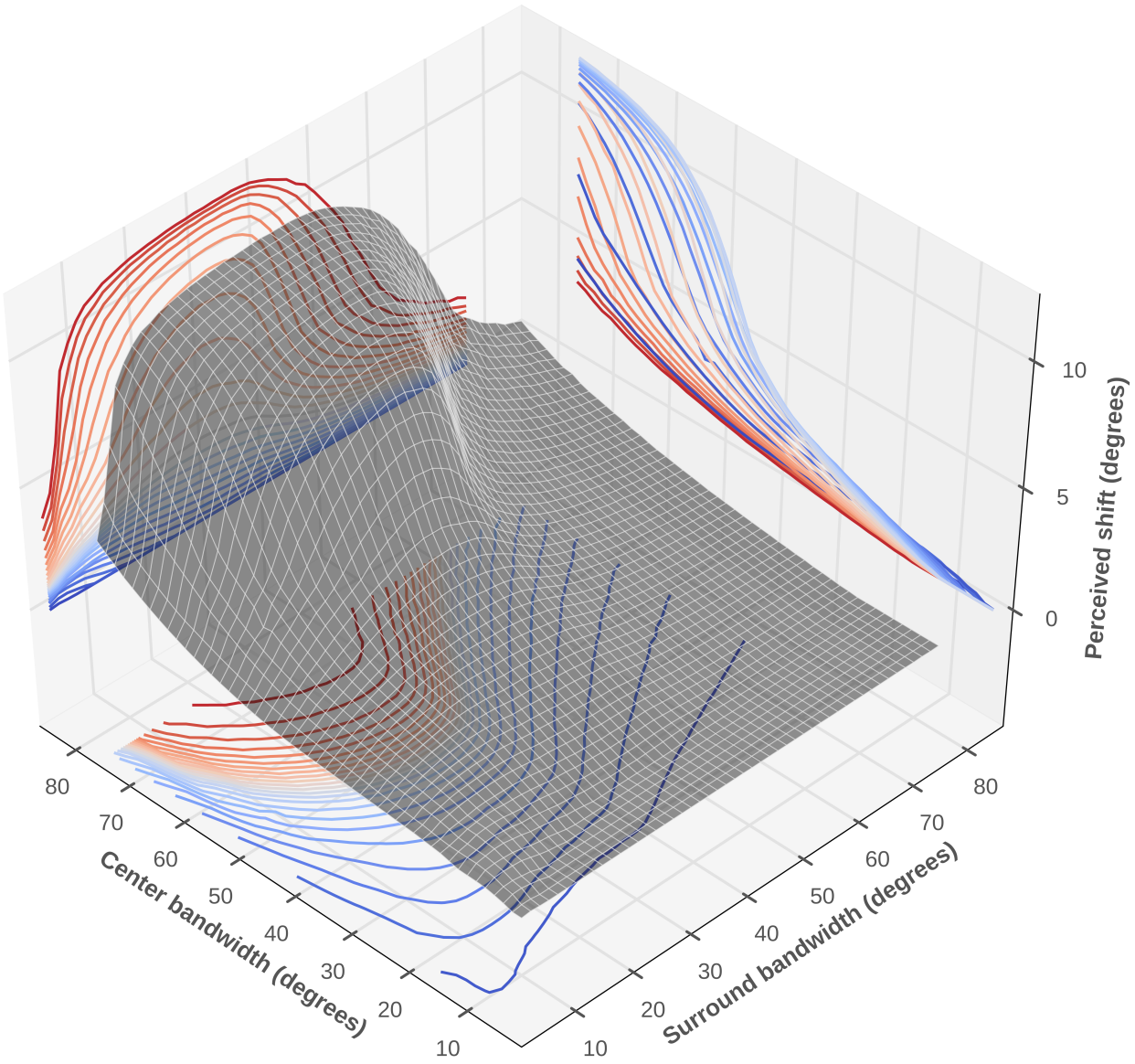
Perceived shift (degrees)
10
5
0
Center bandwidth (degrees)
80
70
60
50
40
30
20
10
Surround bandwidth (degrees)
10
20
30
40
50
60
70
80

Figure S12. ***Evolution of the repulsive shift as a function of tuning curve bandwidths.*** We examined how the maximal amplitude of the repulsive orientation tilt effect (Figure 3A) changes as center and surround stimulus bandwidths are adjusted. The maximal amplitude is defined at the largest repulsive shift that is achieved as the orientation of the surround is systematically varied relative to the center.

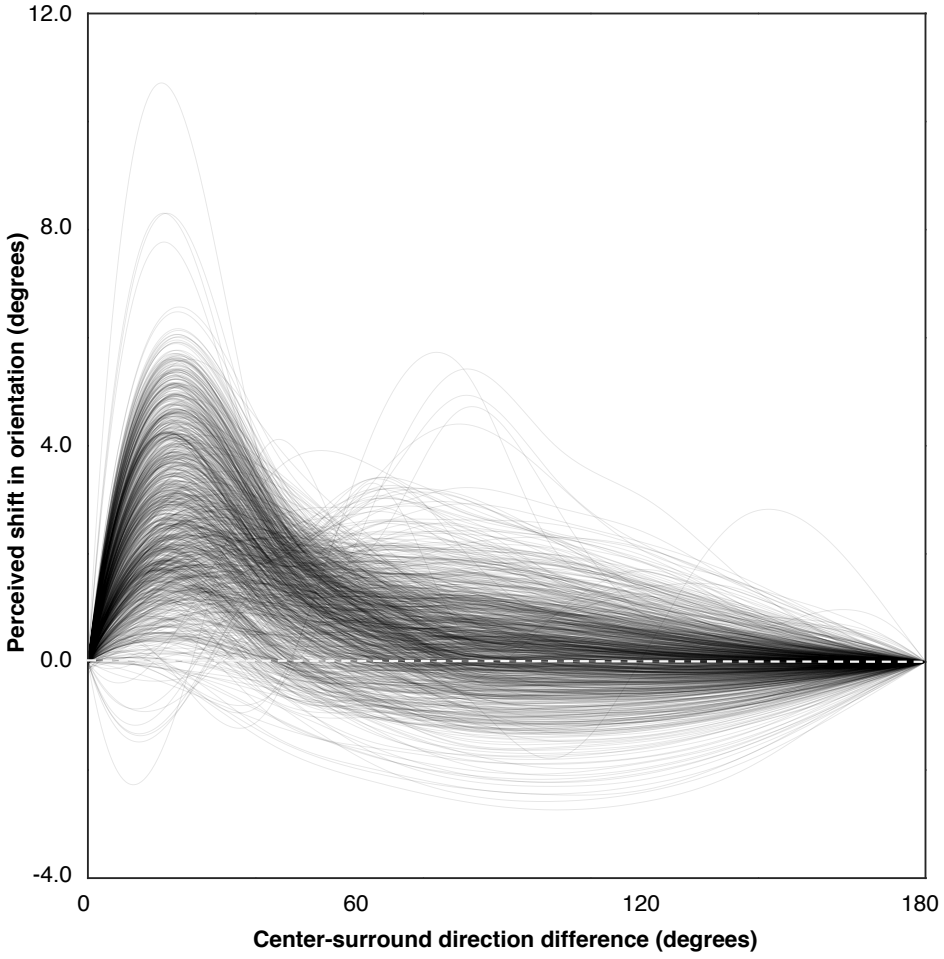

Figure S13. ***Model instantiations derived from the lesion-screening procedure may account for some of the inter-participant variability observed experimentally.*** Plot depicts full model behavior of the orientation tilt effect (Figure 3C) for every combination of parameters sampled (see Eq. 12 for sampling details).

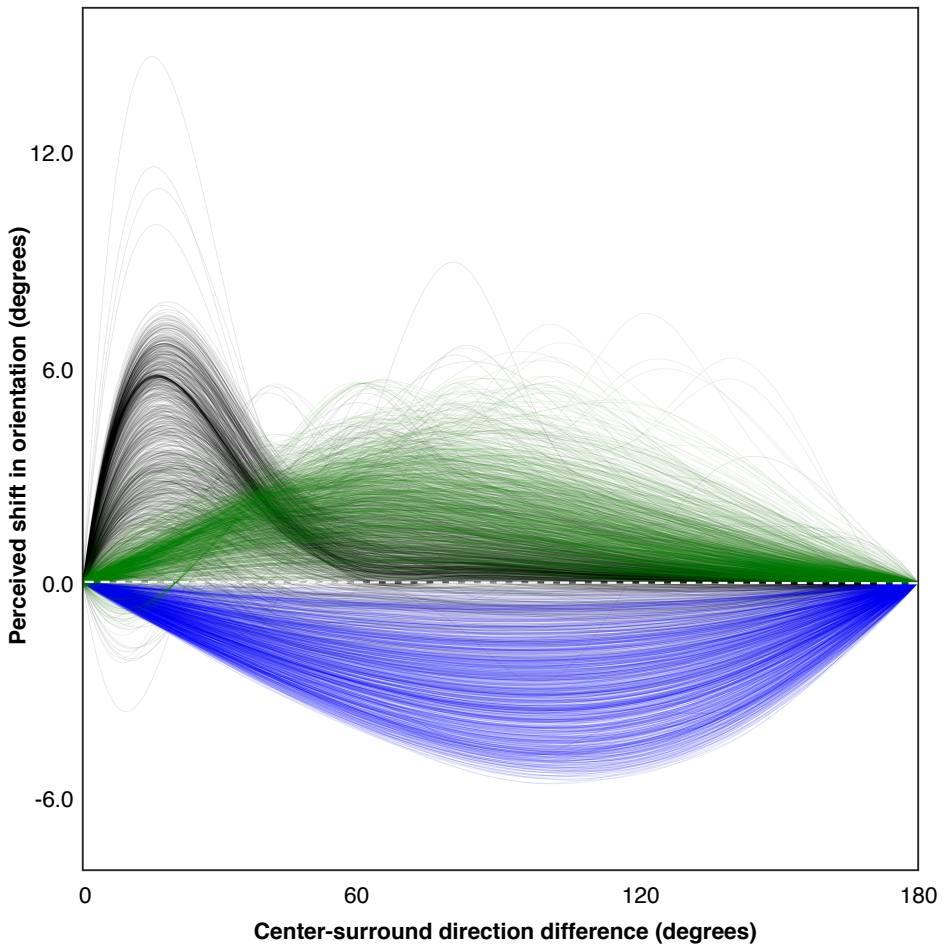

Figure S14. ***Model instantiations with lesioned eCRF or divisive normalization mechanisms are unable to explain the orientation tilt effect.*** Plot depicts simulations from models with lesions applied to either the facilitatory near eCRF (green), suppressive far eCRF (blue), or divisive normalization (black) on the orientation tilt effect (Figure 3C) for every sampled combination of parameters (see Eq. 12 for sampling details). Lesioning the facilitatory near eCRF destroys the attraction regime. Lesioning the suppressive far eCRF destroys the repulsion regime. Without divisive normalization the model is unable to balance these two regimes.

References

Albright, T. D., Desimone, R., and Gross, C. G. (1984). Columnar organization of directionally selective cells in visual area MT of the macaque. *Journal of Neurophysiology*, 51(1):16–31.

Allman, J., Miezin, F., and McGuinness, E. (1985). Direction- and velocity-specific responses from beyond the classical receptive field in the middle temporal visual area (MT). *Perception*, 14(2):105–26.

Angelucci, A., Levitt, J. B., and Lund, J. S. (2002a). Anatomical origins of the classical receptive field of single neurons in macaque visual cortical area V1. *Progress in Brain Research*, 136:373–88.

Angelucci, A., Levitt, J. B., Walton, E. J. S., Hupe, J.-M., Bullier, J., and Lund, J. S. (2002b). Circuits for local and global signal integration in primary visual cortex. *Journal of Neuroscience*, 22(19):8633–8646.

Angelucci, A. and Shushruth, S. (2013). Beyond the classical receptive field: surround modulation in primary visual cortex. In *The New Visual Neurosciences*, pages 425–444. MIT Press.

Anstis, S. (2006). White's effect in lightness, color, and motion. In *Seeing Spatial Form*, pages 91–100. Oxford University Press.

Anton-Erxleben, K., Stephan, V. M., and Treue, S. (2009). Attention reshapes center-surround receptive field structure in macaque cortical area MT. *Cereb. Cortex*, 19(10):2466–2478.

Baker, D. H. and Graf, E. W. (2010). Contextual effects in speed perception may occur at an early stage of processing. *Vision Research*, 50(2):193–201.

977 Bosking, W. H., Zhang, Y., Schofield, B., and Fitzpatrick, D. (1997). Orientation selectivity and the
978 arrangement of horizontal connections in tree shrew striate cortex. *The Journal of Neuroscience*,
979 17(6):2112–27.

980 Bradley, D. C. and Andersen, R. a. (1998). Center-surround antagonism based on disparity in
981 primate area MT. *Journal of Neuroscience*, 18(18):7552–7565.

982 Briggs, F. and Usrey, W. M. (2011). Distinct mechanisms for size tuning in primate visual cortex.
983 *The Journal of neuroscience*, 31(35):12644–12649.

984 Bringuier, V., Chavane, F., Glaeser L., and Frégnac, Y. (1999). Horizontal propagation of visual
985 activity in the synaptic integration field of area 17 neurons. *Science*, 283(5402):695–699.

986 Busse, L., Wade, A. R., and Carandini, M. (2009). Representation of concurrent stimuli by
987 population activity in visual cortex. *Neuron*, 64(6):931–42.

988 Carandini, M. (2012). From circuits to behavior: a bridge too far? *Nature neuroscience*,
989 15(4):507–9.

990 Carandini, M. and Heeger, D. (2012). Normalization as a canonical neural computation. *Nature*
991 *Reviews Neuroscience*, 13:51–62.

992 Carandini, M. and Heeger, D. J. (1994). Summation and division by neurons in primate visual
993 cortex. *Science*, 264:1333–1336.

994 Cavanaugh, J. R., Bair, W., and Movshon, J. A. (2002). Selectivity and spatial distribution

995 of signals from the receptive field surround in macaque V1 neurons. *J Neurophysiol*,
996 88(5):2547–2556.

997 Clifford, C. W. G. (2014). The tilt illusion: Phenomenology and functional implications. *Vision*
998 *Research*, 104:3–11.

999 Coen-Cagli, R., Dayan, P., and Schwartz, O. (2012). Cortical surround interactions and perceptual
1000 salience via natural scene statistics. *PLoS Computational Biology*, 8(3):e1002405.

1001 Conway, B. R. (2009). Color vision, cones, and color-coding in the cortex. *The Neuroscientist : a*
1002 *review journal bringing neurobiology, neurology and psychiatry*, 15(3):274–90.

1003 Conway, B. R., Moeller, S., and Tsao, D. Y. (2007). Specialized color modules in macaque
1004 extrastriate cortex. *Neuron*, 56(3):560–73.

1005 Cumming, B. G. and Parker, A. J. (1997). Responses of primary visual cortical neurons to
1006 binocular disparity without depth perception. *Nature*, 389(6648):280–3.

1007 DeAngelis, G. C., Freeman, R. D., and Ohzawa, I. (1994). Length and width tuning of neurons in
1008 the cat's primary visual cortex. *Journal of neurophysiology*, 71(1):347–74.

1009 DeAngelis, G. C., Ghose, G. M., Ohzawa, I., and Freeman, R. D. (1999). Functional
1010 Micro-Organization of Primary Visual Cortex: Receptive Field Analysis of Nearby Neurons.
1011 *J. Neurosci.*, 19(10):4046–4064.

1012 DeAngelis, G. C. and Newsome, W. T. (1999). Organization of disparity-selective neurons in
1013 macaque area MT. *J. Neurosci.*, 19(4):1398–1415.

1014 Douglas, R. J., Koch, C., Mahowald, M., Martin, K. A., and Suarez, H. H. (1995). Recurrent
1015 excitation in neocortical circuits. *Science*, 269(5226):981–985.

1016 Dow, B. M. (2002). Orientation and color columns in monkey visual cortex. *Cerebral cortex*,
1017 12(10):1005–1015.

1018 Eagleman, D. M. (2001). Visual illusions and neurobiology. *Nature Reviews Neuroscience*,
1019 2(12):920–926.

1020 Georgopoulos, A. P., Schwartz, A. B., and Kettner, R. E. (1986). Neuronal population coding of
1021 movement direction. *Science*, 233:1416–1419.

1022 Gibson, J. J. and Radner, M. (1937). Adaptation, after-effect and constrast in the perception of
1023 tilted lines. *Journal of Experimental Psychology*, 20:453–467.

1024 Gilbert, C. D. and Li, W. (2013). Top-down influences on visual processing. *Nature reviews.*
1025 *Neuroscience*, 14(5):350–63.

1026 Goddard, E., Clifford, C. W. G., and Solomon, S. G. (2008). Centre-surround effects on perceived
1027 orientation in complex images. *Vision Research*, 48(12):1374–1382.

1028 Grossberg, S. and Todorović, D. (1988). Neural dynamics of 1-D and 2-D brightness perception: a
1029 unified model of classical and recent phenomena. *Perception and psychophysics*, 43(3):241–277.

1030 Heeger, D. J. (1993). Modeling simple-cell direction selectivity with normalized, half-squared,
1031 linear operators. *Journal of Neurophysiology*, 70(5):1885–1898.

1032 Hess, R. F., Hayes, A., and Field, D. J. (2003). Contour integration and cortical processing. *Journal*
1033 *Of Physiology Paris*, 97(2-3):105–119.

1034 Hubel, D. and Wiesel, T. N. (1968). Receptive fields and functional architecture of monkey striate
1035 cortex. *The Journal of physiology*, 195(1):215–43.

1036 Johnson, E. N., Hawken, M. J., and Shapley, R. (2001). The spatial transformation of color in the
1037 primary visual cortex of the macaque monkey. *Nature neuroscience*, 4(4):409–16.

1038 Jones, H. E., Grieve, K. L., Wang, W., and Sillito, A. M. (2001). Surround suppression in primate
1039 V1. *Journal of neurophysiology*, 86:2011–2028.

1040 Kim, J. and Wilson, H. R. (1997). Motion integration over space: interaction of the center and
1041 surround motion. *Vision Research*, 37(8):991–1005.

1042 Klauke, S. and Wachtler, T. (2015). "Tilt" in color space: hue changes induced by chromatic
1043 surrounds. *Journal of Vision*, 15(13):17.

1044 Kusunoki, M., Moutoussis, K., and Zeki, S. (2006). Effect of background colors on the tuning of
1045 color-selective cells in monkey area V4. *Journal of neurophysiology*, 95(5):3047–59.

1046 Land, E. H. and McCann, J. J. (1971). Lightness and Retinex theory. *Journal of the Optical Society*
1047 *of America*, 61(1):1–11.

1048 Lee, W.-c. A., Bonin, V., Reed, M., Graham, B. J., and Hood, G. (2016). Anatomy and function of
1049 an excitatory network in the visual cortex. *Nature*, 532(7599):370–374.

1050 Levitt, J. B. and Lund, J. S. (1997). Contrast dependence of contextual effects in primate visual
1051 cortex. *Nature*, 387(6628):73–76.

1052 Li, C. Y., Lei, J. J., and Yao, H. S. (1999). Shift in speed selectivity of visual cortical neurons: a
1053 neural basis of perceived motion contrast. *Proceedings of the National Academy of Sciences of*
1054 *the United States of America*, 96(7):4052–4056.

1055 Loomis, J. M. and Nakayama, K. (1973). A velocity analogue of brightness contrast. *Perception*,
1056 2(4):425–428.

1057 Ma, W. J., Beck, J. M., Latham, P. E., and Pouget, A. (2006). Bayesian inference with probabilistic
1058 population codes. *Nature Neuroscience*, 9(11):1432–8.

1059 Manassi, M., Lonchampt, S., Clarke, A., and Herzog, M. H. (2016). What crowding can tell us
1060 about object representations. *J. Vis.*, 16(3):35.

1061 Marshak, W. and Sekuler, R. (1979). Mutual repulsion between moving visual targets. *Science*
1062 *(New York, N.Y.)*, 205(4413):1399–401.

1063 Mollon, J. D. (2009). A neural basis for unique hues? *Current Biology*, 19(11):441–442.

1064 Monnier, P. and Shevell, S. K. (2003). Large shifts in color appearance from patterned chromatic
1065 backgrounds. *Nature neuroscience*, 6(8):801–802.

1066 Murakami, I. and Shimojo, S. (1993). Motion capture changes to induced motion at higher
1067 luminance contrasts, smaller eccentricities, and larger inducer sizes. *Vision Research*,
1068 33(15):2091–2107.

Murakami, I. and Shimojo, S. (1996). Assimilation-type and contrast-type bias of motion induced by the surrond in a random-dot display: evidence for ceter-surround antagonism. *Vision Research*, 36(22):3629–3639.

Nassi, J., Avery, M., Cetin, A., Roe, A., and Reynolds, J. (2015). Optogenetic Activation of Normalization in Alert Macaque Visual Cortex. *Neuron*, 86(6):1504–1517.

Norman, H. F., Norman, J. F., Todd, J. T., and Lindsey, D. T. (1996). Spatial interactions in perceived speed. *Perception*, 25(7):815–830.

O'Toole, B. and Wenderoth, P. (1977). The tilt illusion: repulsion and attraction effects in the oblique meridian. *Vision research*, 17(3):367–374.

Ozeki, H., Finn, I. M., Schaffer, E. S., Miller, K. D., and Ferster, D. (2009). Inhibitory stabilization of the cortical network underlies visual surround suppression. *Neuron*, 62(4):578–92.

Petrov, Y., Carandini, M., and McKee, S. (2005). Two distinct mechanisms of suppression in human vision. *The Journal of neuroscience*, 25(38):8704–8707.

Ringach, D. L., Hawken, M. J., and Shapley, R. (1997). Dynamics of orientation tuning in macaque primary visual cortex. *Nature*, 387(6630):281–284.

Ringach, D. L., Shapley, R. M., and Hawken, M. J. (2002). Orientation selectivity in macaque V1: diversity and laminar dependence. *Journal of neuroscience*, 22(13):5639–5651.

Rubin, D. B., Hooser, S. D. V., and Miller, K. D. (2015). The stabilized supralinear network : A unifying circuit motif underlying multi-input integration in sensory cortex. *Neuron*, 85(1):1–51.

1088 Rust, N. C., Mante, V., Simoncelli, E. P., and Movshon, J. A. (2006). How MT cells analyze the
1089 motion of visual patterns. *Nature Neuroscience*, 9(11):1421–31.

1090 Sceniak, M. P., Hawken, M. J., and Shapley, R. (2001). Visual spatial characterization of macaque
1091 V1 neurons. *Journal of Neurophysiology*, 85(5):1873–1887.

1092 Sceniak, M. P., Ringach, D. L., Hawken, M. J., and Shapley, R. (1999). Contrast's effect on spatial
1093 summation by macaque V1 neurons. *Nature neuroscience*, 2(8):733–9.

1094 Schein, S. J. and Desimone, R. (1990). Spectral properties of V4 neurons in the macaque. *Journal*
1095 *of Neuroscience*, 10(10):3369–3389.

1096 Schwabe, L., Obermayer, K., Angelucci, A., and Bress (2006). The role of feedback in shaping
1097 the extra-classical receptive field of cortical neurons: A recurrent network model. *Journal of*
1098 *Neuroscience*, 26(36):9117–9129.

1099 Schwartz, O., Hsu, A., and Dayan, P. (2007). Space and time in visual context. *Nat. Rev. Neurosci.*,
1100 8:522–535.

1101 Sengpiel, F. (1997). Binocular rivalry: ambiguities resolved. *Current Biology*, 7:447–450.

1102 Sengpiel, F., Sen, A., and Blakemore, C. (1997). Characteristics of surround inhibition in cat area
1103 17. *Experimental Brain Research*, 116(2):216–228.

1104 Series, P., Lorenceau, J., and Frégnac, Y. (2003). The "silent" surround of V1 receptive fields:
1105 theory and experiments. *Journal of physiology-Paris*, 97:453–474.

Setić, M. and Domijan, D. (2008). Modeling the top-down influences on the lateral interactions in the visual cortex. *Brain Res.*, 1225:86–101.

Shevell, S. K. and Monnier, P. (2005). Color shifts from S-cone patterned backgrounds: Contrast sensitivity and spatial frequency selectivity. *Vision Research*, 45:1147–1154.

Shushruth, S. and Ichida, J. (2009). Comparison of spatial summation properties of neurons in macaque V1 and V2. *Journal of neurophysiology*, 1:2069–2083.

Shushruth, S., Mangapathy, P., Ichida, J. M., Bressloff, P. C., Schwabe, L., and Angelucci, A. (2012). Strong recurrent networks compute the orientation tuning of surround modulation in the primate primary visual cortex. *Journal of Neuroscience*, 32(1):308–321.

Sillito, A. M., Grieve, K. L., Jones, H. E., Cudeiro, J., and Davis, J. (1995). Visual cortical mechanisms detecting focal orientation discontinuities. *Nature*, 378:492–496.

Sincich, L. and Horton, J. (2005). The circuitry of V1 and V2: integration of color, form, and motion. *Annu. Rev. Neurosci.*, 28(1):303–326.

Sit, Y. F., Chen, Y., Geisler, W. S., Miikkulainen, R., and Seidemann, E. (2009). Complex Dynamics of V1 Population Responses Explained by a Simple Gain-Control Model. *Neuron*, 64(6):943–956.

Smith, V. C., Jin, P. Q., and Pokorny, J. (2001). The role of spatial frequency in color induction. *Vision Research*, 41(8):1007–1021.

Stemmler, M., Usher, M., and Niebur, E. (1995). Lateral interactions in primary visual cortex: a model bridging physiology and psychophysics. *Science*, 269(5232):1877–1880.

Stepanyants, A., Hirsch, J. A., Martinez, L. M., Kisvarday, Z. F., Ferecsko, A. S., and Chklovskii, D. B. (2008). Local potential connectivity in cat primary visual cortex. *Cerebral Cortex*, 18(1):13–28.

Tanaka, H. and Ohzawa, I. (2009). Surround suppression of V1 neurons mediates orientation-based representation of high-order visual features. *Journal of neurophysiology*, 101(3):1444–62.

Thomas, O. M., Cumming, B. G., and Parker, a. J. (2002). A specialization for relative disparity in V2. *Nature neuroscience*, 5(5):472–8.

Trott, A. R. and Born, R. T. (2015). Input-gain control produces feature-specific surround suppression. *Journal of neuroscience*, 35(12):4973–1982.

Tsotsos, J. K., Culhane, S., and Cutzu, F. (2001). Visual attention and cortical circuits. In Braun, J., Koch, C., and Davis, J. L., editors, *From Theoretical Foundations to a Hierarchical Circuit for Selective Attention*, pages 285–306. The MIT presss.

Vinck, M., Womelsdorf, T., and Fries, P. (2013). Gamma-Band synchronization and information transmission. In *Principles of Neural Coding*, pages 449–470. CRC Press.

Wachtler, T., Sejnowski, T. J., and Albright, T. D. (2003). Representation of color stimuli in awake macaque primary visual cortex. *Neuron*, 37(4):681–91.

Weliky, M., Kandler, K., Fitzpatrick, D., and Katz, L. C. (1995). Patterns of excitation and

1143 inhibition evoked by horizontal connections in visual cortex share a common relationship to
1144 orientation columns. *Neuron*, 15(3):541–552.

1145 Westheimer, G. (1986). Spatial interaction in the domain of disparity signals in human stereoscopic
1146 vision. *The Journal of physiology*, 370:619–29.

1147 Westheimer, G. and Levi, D. M. (1987). Depth attraction and repulsion of disparate foveal stimuli.
1148 *Vision Research*, 27(8):1361–1368.

1149 White, M. (1979). A new effect of pattern on perceived lightness. *Perception*, 8(4):413–416.

1150 White, M. (1981). The effect of the nature of the surround on the perceived lightness of grey bars
1151 within square-wave test gratings. *Perception*, 10(2):215–230.

1152 Zhang, J., Barhomi, Y., and Serre, T. (2012). A new biologically inspired color image descriptor.
1153 In *European Conference on Computer Vision*, pages 312–324.

1154 Zhu, M. and Rozell, C. J. (2013). Visual nonclassical receptive field effects emerge from Sparse
1155 coding in a dynamical system. *PLoS Computational Biology*, 9(8):1–15.