nature neuroscience

Flexible gating of contextual influences in natural vision

Ruben Coen-Cagli[1,4], Adam Kohn[1–3,5] & Odelia Schwartz[1,3–5]

Identical sensory inputs can be perceived as markedly different when embedded in distinct contexts. Neural responses to simple stimuli are also modulated by context, but the contribution of this modulation to the processing of natural sensory input is unclear. We measured surround suppression, a quintessential contextual influence, in macaque primary visual cortex with natural images. We found that suppression strength varied substantially for different images. This variability was not well explained by existing descriptions of surround suppression, but it was predicted by Bayesian inference about statistical dependencies in images. In this framework, surround suppression was flexible: it was recruited when the image was inferred to contain redundancies and substantially reduced in strength otherwise. Thus, our results reveal a gating of a basic, widespread cortical computation by inference about the statistics of natural input.

The contextual modulation of perceptual and neural responses is ubiquitous in cognition and sensory processing[1–3]: how a stimulus is perceived or valued depends strongly on which other stimuli are currently present or have recently been encountered. Contextual effects underlie phenomena as varied as visual saliency[4,5] and illusions[1,6], adaptation aftereffects[2], multisensory integration[7], and value encoding of multiple alternatives[8].

An example of contextual modulation is the influence of a visual stimulus placed outside a neuron's spatial receptive field (RF)—in the 'surround'—on the response evoked by a stimulus in the RF. This type of spatial contextual effect is thought to be central to the integration and segregation of visual information[9,10], allowing objects to be identified and the relationships between them to be understood. The surround is also widely invoked to explain perceptual salience[4,5].

Spatial contextual effects have been documented at many stages of the visual system using simple stimuli such as bars and gratings[11]. This work has revealed that a stimulus in the surround is often suppressive, with the strength of its influence depending on its contrast[12–14], precise location[15,16] and relation to the stimulus in the RF[16,17]. The recruitment of the surround has also been shown to strongly influence responses to natural images[18,19], but whether this modulation can be explained with knowledge derived from experiments with simple stimuli is unclear[19,20]. In the RF itself, models derived from simpler stimuli have a limited ability to predict responses to natural images, constituting a recognized barrier in the study of visual processing[21].

We found that existing descriptive models derived from measurements with simple stimuli have a limited ability to explain the surround suppression recruited by natural images in primary visual cortex (V1) of macaque monkeys. To explain this suppression, we considered how surround suppression should behave to allow cortical neurons to encode natural images optimally. In this approach, one first considers the statistical properties of natural images, then hypothesizes about the computations involved in representing them optimally, and finally uses these hypotheses to develop predictions about neuronal response properties[22–28].

Our previous modeling work using this approach revealed that surround suppression should be engaged when an image provides statistically redundant (or homogeneous) signals to the RF and its surround, but should be strongly reduced otherwise, a form of gating[27]. Consistent with this prediction, we found that surround suppression in V1 neurons was stronger for images inferred to be homogeneous. Including this gating in existing descriptive models considerably improved their ability to predict responses to scenes. Our results show that a basic cortical computation, surround suppression, can be modulated by inference about the statistical structure of visual inputs.

RESULTS

We recorded from 207 V1 neurons in three anesthetized, paralyzed macaque monkeys. Neuronal spatial receptive fields were located 2–3 degrees from the fovea, in the lower visual field.

Variability of surround modulation with natural images

We measured responses of V1 neurons to brief presentations (100 ms) of 270 static natural images, as well as various static grating stimuli. Natural images were windowed to fit largely in the RF (1 degree) or up to 6.7-fold larger (**Fig. 1a**). The example neuron in **Figure 1a** illustrates the effects of the surround: the neuron fired vigorously to presentations of many of the smaller images and usually less when these were embedded in a larger image.

We quantified the modulation caused by enlarging the image with a common metric[14,29], namely the ratio between the responses to each paired large and small image (modulation ratio, MR; Online Methods). MR values smaller than 1 correspond to surround suppression; values larger than 1 to facilitation. The geometric mean of the MR for natural images that evoked a measurable response was 0.71 (68%

[1]D.P. Purpura Department of Neuroscience, Albert Einstein College of Medicine, Bronx, New York, USA. [2]Department of Ophthalmology and Visual Sciences, Albert Einstein College of Medicine, Bronx, New York, USA. [3]Department of Systems and Computational Biology, Albert Einstein College of Medicine, Bronx, New York, USA. [4]Present addresses: Department of Basic Neuroscience, University of Geneva, Geneva, Switzerland (R.C.-C.), Department of Computer Science, University of Miami, Miami, Florida, USA (O.S.). [5]These authors contributed equally to this work. Correspondence should be addressed to R.C.-C. (ruben.coencagli@gmail.com).

Received 3 July; accepted 7 September; published online 5 October 2015; doi:10.1038/nn.4128

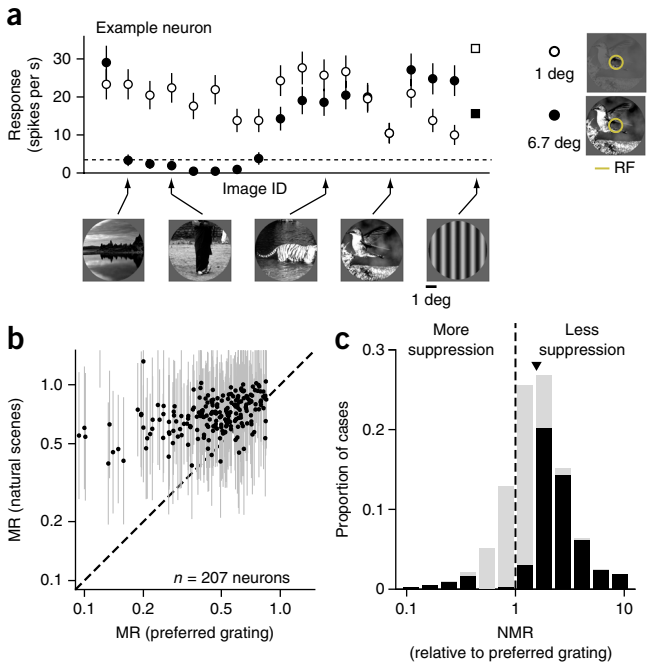

Figure 1 Variability of surround modulation with natural images. (**a**) Firing rate of a V1 neuron in response to static natural images (circles), and gratings at the preferred orientation and spatial frequency (squares), windowed to 1 degree of visual space (open symbols) or 6.7 degrees (filled). The dashed line denotes the spontaneous rate. (**b**) Each point represents the MR measured with gratings at the preferred orientation and spatial frequency (abscissa) and the average MR across all natural images (ordinate), for each neuron. (**c**) NMR across all images and neurons (N = 38,591 cases). Arrowhead indicates geometric mean and black bars indicate cases with NMR significantly different from 1 ($P < 0.05$). Error bars indicate 68% c.i. in **a** and s.d. in **b**.

confidence interval (c.i.) [0.70, 0.73], n = 207 neurons), indicating clear suppression on average. Across neurons, the mean MR for images was related to that measured with a preferred grating (Spearman's $\rho = 0.42$, permutation test $P < 0.0001$; **Fig. 1b**). Measurements with gratings were therefore informative as to whether a neuron would show strong or weak suppression, on average, across different images.

Although the mean suppression for images was related to that for a preferred grating, we found that the suppression was often much stronger for some images than others in each neuron. For instance, in the example neuron, enlarging some images nearly eliminated all spiking activity, but it had little influence on responses to other images (**Fig. 1a**). To quantify the variability in suppression across images, we computed a normalized MR (NMR) by dividing the MR for each image by the MR for that neuron measured with the preferred grating. The value of NMR thus provides a measure of relative suppression strength for each image, eliminating the confounding influence of variations in mean suppression across neurons. The NMR revealed strong variations in the suppression strength across images, with values varying from tenfold weaker than that provided by gratings (NMR = 10) to tenfold stronger (NMR = 0.1). The geometric mean NMR was 1.48 (c.i. [1.47, 1.49]; **Fig. 1c**), indicating that suppression for images was typically weaker than for the preferred grating. Furthermore, when the NMR value was significantly different from 1 (**Fig. 1c**), suppression was always weaker than with a preferred grating (91.5% of 17,200 cases).

Variability in surround suppression eludes existing models

We sought to understand the variability in suppression strength recruited by natural images. Previous work with simple stimuli has shown that surround suppression involves divisive normalization[13,30], a canonical computation thought to underlie several forms of contextual modulation[3]. In the standard descriptive model of surround suppression, the response to a stimulus in the RF is divisively suppressed by activity in a 'normalization pool' consisting of neurons whose RFs are driven by the surround stimulus (**Fig. 2a**). Depending on the drive it provides to the surround pool, an image might recruit either weak or strong suppression, perhaps explaining the variability that we observed.

We tested the ability of the standard normalization model to account for the observed suppression in two ways. First, we fit the model to neuronal responses to a subset of small and large images and evaluated the cross-validated performance at predicting responses to large images (Online Methods). Second, we quantified the surround drive provided by each image and tested whether images providing stronger drive recruited greater suppression. For these analyses, we focused on the subset of neurons with measurable average surround suppression (n = 126, Online Methods) to avoid trivial cases in which good performance could be achieved by assuming no surround modulation for any image. All reported results were similar when we included all neurons (**Supplementary Table 1**).

To evaluate the standard normalization model, we first needed to define the filters representing the RF and surround of each neuron. The drive to the RF was computed using a quadrature pair of linear filters. The surround was defined using eight pairs of filters, spaced equally around the RF (**Fig. 2a**). The properties of these filters were identical to those in the RF, except their orientation preference was chosen separately for each image, to match the dominant orientation content in the RF. This choice was motivated by previous work showing that the orientation tuning of the surround changes with the stimulus presented to the RF[13,17]. Finally, the model included a normalization pool inside the RF, comprising four pairs of filters that spanned a full range of orientations[31]. The specific parameters of the filters for each cell—the orientation preference and bandwidth, spatial frequency tuning, and size—were chosen to maximize fit quality to all images (Online Methods).

We found that the standard normalization model accounted for only about half of the explainable variance in the responses to large images, on average. The mean prediction quality was 0.53 (c.i. [0.46, 0.58]), where a value of 1 indicates an oracle model and 0 indicates a model that always predicts the cell's mean response across all images. The limited performance of the standard model appeared to arise largely from predicting strongly suppressed responses for images that provided little suppression, as illustrated for the example neuron (**Fig. 2b**). To test whether the failure of the standard model could be alleviated by defining the surround in a different manner, we considered several other choices of surround filters, as well as variations in the normalization model itself. The average performance of the standard normalization model remained poor (**Supplementary Fig. 1**).

We next tested the standard model by exploring the relationship between suppression strength and surround drive (defined as the magnitude of the output of the surround filters; Online Methods). For each cell, we compared the MR for the half of images providing the strongest surround drive with the half providing weakest drive, and found only slightly stronger suppression with greater drive (mean MR of 0.53 for strong drive versus 0.62 for weak drive, $P < 0.0001$; **Fig. 3**). Even comparing images providing particularly strong or weak

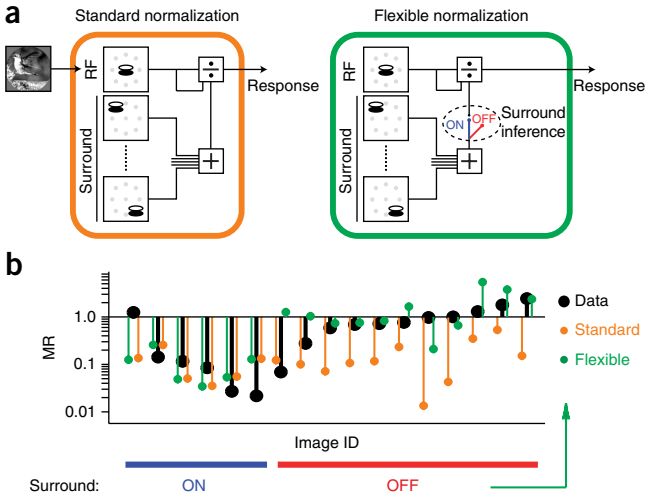

Figure 2 Standard and flexible normalization models of surround suppression. (**a**) Left, schematic of the standard normalization model. Visual input is first passed through linear filters representing the RF (top left) and its surround (bottom left). Gray symbols denote the location of the center of each filter. The output of the RF filters is divided by the filters representing the RF and surround. Right, the flexible normalization model is identical to the standard normalization except that the surround can be turned on and off, on an image-by-image basis, depending on an inference about image homogeneity. (**b**) Black symbols indicate MR for each pair of responses shown in **Figure 1a**; orange and green symbols indicate MR derived from the standard and flexible models, respectively, fit to the firing rates. In the flexible model, facilitation results when the surround stimulus provides additional drive to the RF, but surround suppression is inferred off.

drive (one s.d. above versus below the mean), we did not observe substantially more suppression with stronger drive (mean MR of 0.58 for both cases, $P = 0.73$).

We conclude that the surround drive provided by an image does not predict accurately the suppression it recruits. As a result, the standard normalization model provides a poor prediction of the suppression provided by images.

A principled account of surround suppression

Rather than attempting to improve the standard normalization model by searching for additional descriptive components, we instead sought to apply recent progress in understanding how contextual effects should function, given the statistics of natural images and the computational goal of representing those images efficiently. Previous work has successfully explained basic response properties of neurons in the early visual system using this approach[22–25,32,33]. Despite this success, this approach has not been used to overcome the limited ability of models derived from simple stimuli to predict responses to images[21,34].

Our starting point was the observation that natural images often produce a characteristic nonlinear dependence in the output of filters representing the RF and surround[24] (called a bowtie dependency because of its shape; **Fig. 4a**). This dependency is a result of global image properties, such as the contrast or orientation of a texture, which provide common modulation to neighboring filters[24]. This form of dependency can be effectively reduced by dividing the output of the RF by that of the surround, thereby resulting in a more efficient representation[24]. In this framework, spatial contextual effects take the form of divisive normalization because precisely this interaction is needed to remove the empirically observed dependencies between filter outputs.

The outputs of the RF and surround filters are dependent when they fall on homogeneous regions of an image (for example, on a single object). When the filters are driven by different objects, their outputs tend to be statistically independent (**Fig. 4a**). In such cases, no interaction between the RF and surround is needed or desired, as this would introduce a relationship between the two. Consideration of an optimal representation of images therefore suggests that surround suppression should be engaged when the image patch falling on the RF and surround is homogeneous, and absent when it is heterogeneous[27]. Previous work has shown that this principle can explain contextual effects in displays of simple stimuli[4].

Given a specific input to the RF and surround, how is one to decide if it is homogeneous or heterogeneous? If one were able to calculate the dependence between the RF and surround filters' outputs to a single input image, one could assess homogeneity and modulate suppression strength accordingly. Unfortunately, the dependence cannot be defined using a single set of filters' responses, much like correlation between variables (a simpler form of dependency than the bowtie) cannot be defined with a single sample. Instead, one needs to infer whether that single observation is more likely to have arisen from a homogeneous or heterogeneous image. Optimal inference relies on combining the likelihood of the evidence (how likely the pattern of filters outputs arose from a homogenous or heterogeneous image) with prior knowledge about how often natural images are homogeneous, according to Bayes' rule[27]. An analytical expression for the resultant inferred probability is provided in the Online Methods.

The inference about homogeneity amounts to a gating parameter on the surround influence. When the input is inferred to be homogeneous, surround suppression is fully active and takes the form of divisive normalization. When the input is inferred to be heterogeneous, the surround is muted, even if it is strongly driven by the image (**Fig. 4b**). When an input contains some evidence of being homogeneous and some evidence of being heterogeneous, surround suppression is engaged in proportion to the probability that the image is homogeneous, between fully active and muted.

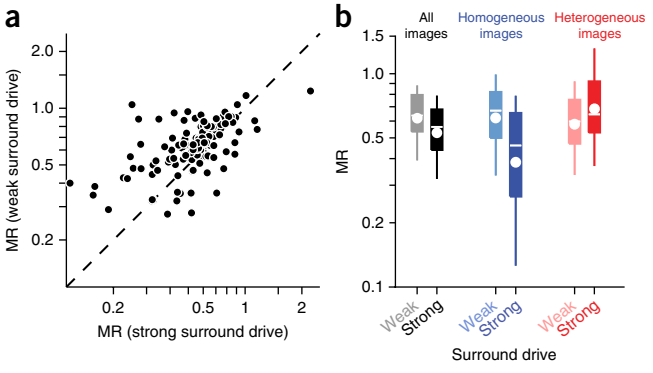

Figure 3 Drive to the surround does not explain surround suppression strength. (**a**) Each point represents, for each neuron, the average MR for images that provide the surround with below average (ordinate) versus above average (abscissa) drive. MR is only weakly modulated by surround drive, defined as the root mean square of the surround filters outputs. (**b**) Left pair, MR across all neurons, for images with weak versus strong surround drive as in (**a**). Boxes denote the 25–75th percentile, whiskers denote the 10–90th percentile, the white line indicates the median and the white circle indicates the geometric mean. Middle and right pairs, data presented as on the left, but including in the analysis only homogeneous (blue, light blue) or heterogeneous (red, pink) images.

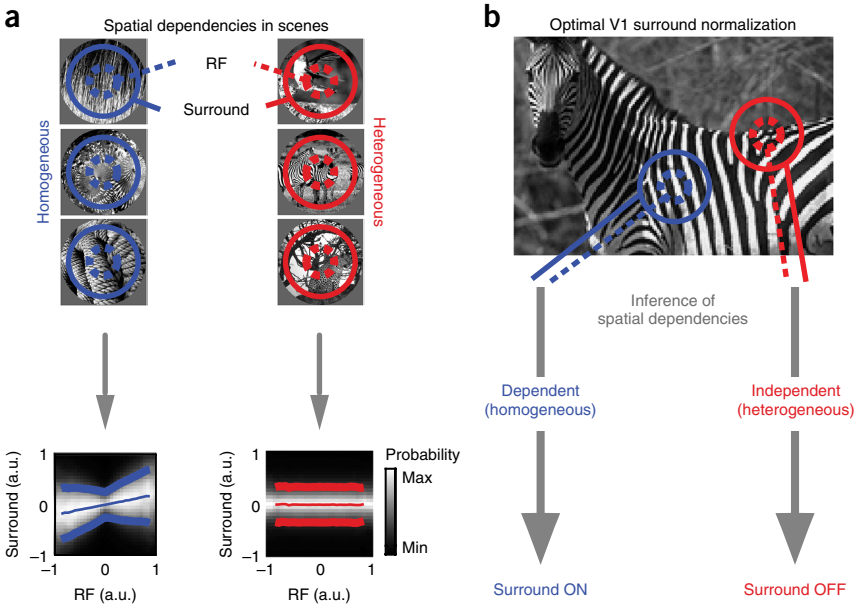

Figure 4 Surround divisive normalization is optimal only for statistically homogeneous stimuli. (**a**) Left, neighboring locations in homogeneous images contain redundant information. This produces a bowtie-shaped dependency in the outputs of filters representing the RF and surround. The dependency is illustrated in the conditional histogram (bottom): each column represents the histogram of surround filters' outputs (position on the ordinate) given a particular output of the RF filters (abscissa). Lighter shades of gray indicate larger occurrence probability. Thin and thick lines represent conditional mean ± conditional s.d., respectively. Right, neighboring locations in heterogeneous regions are independent, as evidenced by the absence of any structure in the conditional histogram (bottom). (**b**) Surround normalization reduces redundancy between RF and surround. Optimality requires that the normalization is turned off when the stimulus is heterogeneous. Schematic is adapted from ref. 27.

Flexible surround suppression

We used the inference about image homogeneity to define a flexible normalization model and asked whether it could improve our ability to predict neuronal responses to large natural images. The structure of the flexible model was identical to the standard model and we chose its parameters in the same manner (Online Methods). However, the flexible model also contained an additional gating variable that modulated the influence of the surround by the inference of homogeneity. Because we found that images were usually inferred to be homogeneous with a probability either near 1 or near 0 (**Supplementary Fig. 2**), we used a binary gating variable that muted the surround when the probability was less than 0.5 and left it unaffected otherwise (**Fig. 2a**).

Notably, the value of the gating variable was not fit to the neuronal data to improve predictions. Rather, its value was determined by learning the dependencies between the RF and surround filters representing each neuron, using an entirely independent ensemble of natural images. We then used these learned dependencies to infer whether each experimental image was homogeneous or heterogeneous, without regard to how the cell responded.

As illustrated for the example neuron, the flexible model captured the weak suppression observed for some images (**Fig. 2b**), as these were typically inferred to be heterogeneous. Across neurons, the flexible model improved prediction quality by 32% (0.53 for standard, 0.69 for flexible, $P < 0.0001$; **Supplementary Fig. 3**).

Although this improvement is substantial, it underestimates the true difference with the standard normalization model. In many cases, the subset of images that happened to drive a neuron was nearly entirely inferred to be either homogeneous or heterogeneous (**Fig. 5a**). For homogeneous images, the two models make identical predictions; thus, for neurons responding mostly to homogeneous images, we expect little difference in performance. Similarly, for neurons responding mostly to heterogeneous images, the surround can simply always be turned off in the standard model, minimizing any difference with the flexible model. The largest difference between the models occurs when a neuron is driven by a balanced ensemble of homogeneous and heterogeneous images. We therefore sorted the recorded neurons on the basis of the proportion of effective images that were inferred homogeneous. Consistent with our reasoning, the flexible model was only slightly better than the standard model (**Fig. 5b**) for those neurons driven primarily by homogeneous or heterogeneous images. In neurons driven with more balanced image ensembles, the improvement was marked. For the tertiles of neurons with mostly heterogeneous or homogeneous images the improvement was 20% ($P < 0.0001$) and 11% ($P = 0.03$), respectively; for the balanced tertile, there was a 100% improvement (0.27 versus 0.54, $P < 0.0001$). Notably, the proportion of homogenous and heterogeneous images was roughly balanced in a large library of natural images (**Supplementary Fig. 2**). Thus, the flexible model

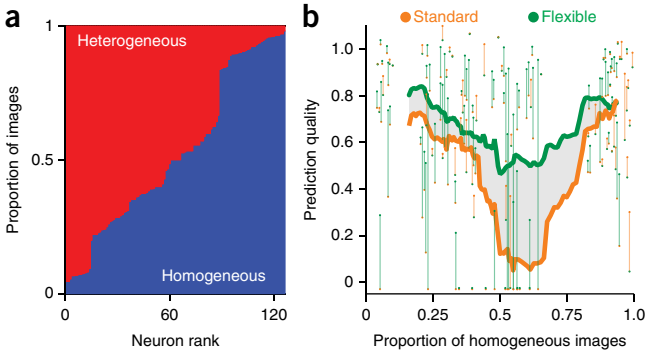

Figure 5 Standard and flexible normalization differ most for balanced image ensembles. (**a**) Proportion of images that were inferred homogeneous (blue) versus heterogeneous (red, stacked bars) for each neuron. (**b**) Cross-validated prediction quality for the flexible (green) and standard (orange) models, as a function of the proportion of effective images that were inferred homogeneous. Each dot denotes the average prediction quality for an individual neuron. Each vertical line connects the prediction quality for the two models for each neuron and is shaded green when the flexible model performed better and orange when it did not. Thick lines indicate a running average of the performance across neurons, including 30 data points, for each model. The shaded area denotes the difference in average prediction quality between flexible and standard models. The models performed similarly in cells driven primarily by homogeneous or heterogeneous images, but the flexible model performed much better in cells driven by balanced image ensembles.

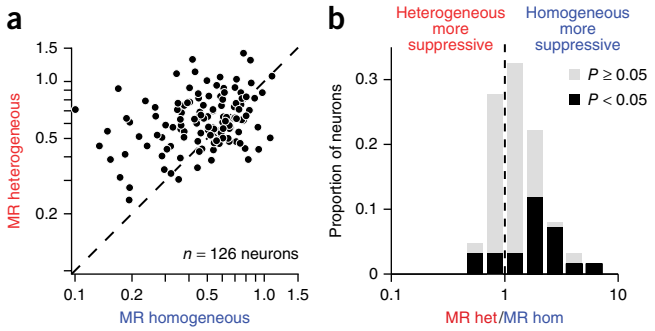

Figure 6 Surround suppression strength depends on image homogeneity. (**a**) MR for heterogeneous versus homogeneous images. Each symbol represents the average MR of a neuron, for each image class. (**b**) The ratio between MRs for the two image categories. Values larger than 1 correspond to neurons suppressed more by homogeneous than heterogeneous images. Black bars indicate neurons with a ratio significantly different from 1.

would strongly outperform the standard model across richer image ensembles than is feasible to show in an experiment.

In summary, modifying the standard normalization model by including a gating variable, whose value reflects an inference of image homogeneity, provided a marked improvement in the ability to predict responses to natural images. Notably, this improvement occurred without the addition of any parameters fitted to the neuronal data.

Surround suppression is stronger for homogeneous images

The better performance of the flexible normalization model suggests that the inference about image homogeneity captures an important source of variability in surround suppression. To test this further, we compared surround suppression for images inferred as homogeneous and heterogeneous in each neuron. We found that homogeneous images were significantly more suppressive than heterogeneous ones (mean MR = 0.47 versus 0.62, $P < 0.0001$; **Fig. 6a**). The difference in suppression strength between image classes was statistically significant in 40 of 126 neurons—a fraction that reflected primarily the number of images to which the cell responded. Among those neurons, suppression was stronger for homogeneous images in 80% of cases (**Fig. 6b**).

The difference in suppression between homogeneous and heterogeneous images was significantly larger than that between images providing strong compared with weak surround drive (32% for homogeneity versus 17% difference for drive shown in **Fig. 3a**, $P = 0.009$). However, for homogeneous images, when the surround is engaged, suppression was 63% stronger for images providing strong drive than those providing weak drive (MR of 0.38 versus 0.62, $P < 0.0001$; **Fig. 3b**). For heterogeneous images, surround drive had little influence on suppression strength (MR of 0.68 versus 0.58, $P = 0.002$), consistent with the muting of the surround. Thus, the inference about homogeneity distinguishes between images providing strong versus weak suppression better than simply measuring surround drive, the focus of the standard normalization framework. But surround drive is important for determining suppression strength in those images inferred to be homogeneous.

Finally, we asked whether the gating of the surround was binary or continuous. As mentioned previously, the inference about homogeneity may be any continuous probability value in our framework. In practice, it often produced probability values close to 0 or 1 for our images, so we used a binary classification and gating of the surround. To test whether gating was in fact continuous, we synthetized images with intermediate probability values. Suppression of V1 neurons increased gradually as the probability of homogeneity increased (**Supplementary Fig. 2**), indicating that the gating of suppression was continuous, consistent with our framework.

Comparing surround suppression for images, across neurons

We have thus far compared the suppression in each neuron for homogeneous and heterogeneous images. Notably, the same image may appear homogeneous to some neurons and heterogeneous to others. This is because each neuron only detects some of the structure in the image patch falling in its RF and surround, depending on its filter properties. For instance, the image shown in **Figure 7a** contains homogeneous structure in the RF and surround at a low spatial frequency, but appears to be heterogeneous at a finer scale. **Figure 7b**

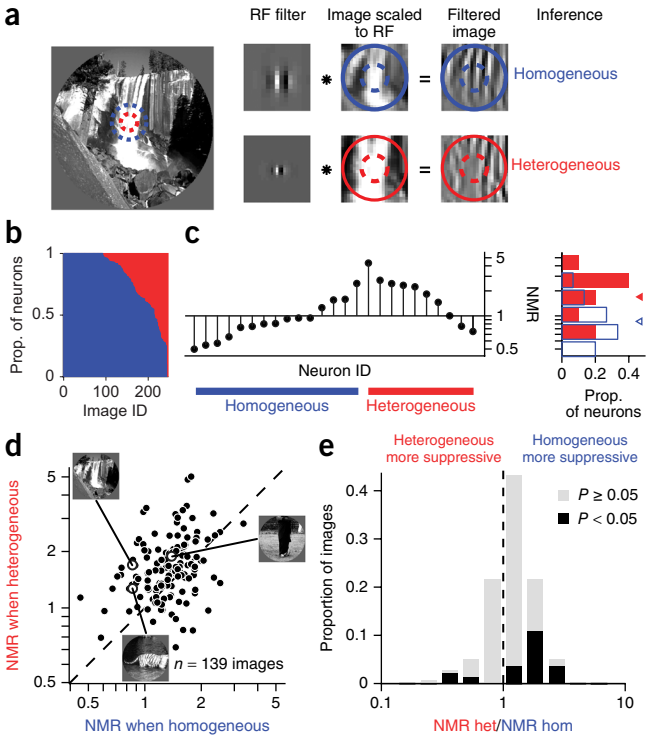

Figure 7 Homogeneity depends on neuronal tuning. (**a**) The same image (left) contains homogeneous structure for some neurons (RF position and size denoted with blue dashed circle, top) and heterogeneous for others (red dashed circle, bottom). First column in the boxes shows example filters representing two different neurons, the second column shows image patch scaled and centered to fit the RF and the third column shows the result of convolution between image and filter, indicating the image components visible to that filter. (**b**) Proportion of neurons for which a given image was inferred homogeneous (blue) versus heterogeneous (red, stacked bars). Many images could be classified as either type, depending on the neuron's tuning. (**c**) Left, stem plot of NMR for the example image in **a** across different neurons. Blue and red lines at the bottom denote neurons with surround inferred on and off, respectively. Right, histogram of NMR values for each class. Triangles denote geometric mean. (**d**) Each symbol represents the NMR for an image, averaged separately when it was classified as heterogeneous (ordinate) or homogeneous (abscissa). Only images classified in both ways by at least 5% of neurons were included. (**e**) Distribution of the ratio between NMR in the two conditions. Images that were more suppressive when classified as homogeneous have values larger than 1. Black bars indicate images with ratios significantly different from 1.

shows the proportion of neurons for which each image was inferred to be homogeneous or heterogeneous. For some images, the inference was always either homogeneous or heterogeneous, mainly because only a few neurons responded to them. However, for many images the inference differed across neurons.

Our framework predicts that an image should provide stronger suppression when it is inferred to be homogeneous than when that same image is inferred to be heterogeneous. For the example image (**Fig. 7a**), this was the case: normalized suppression was significantly stronger in neurons for which the image was inferred to be homogeneous (**Fig. 7c**) than in those neurons for which the inference was heterogeneous (average NMR of 0.86, c.i. [0.76, 0.98] versus 1.69, c.i. [1.39, 2.01]). For the full set of images, suppression was also stronger when an image was inferred to be homogeneous rather than heterogeneous (average NMR = 1.32 versus 1.58, $P < 0.0001$; **Fig. 7d**). The difference in suppression was significant for 30 of 139 individual images, and suppression was stronger when the inference was homogeneous in 83% of the cases (25 of 30; **Fig. 7e**).

In summary, homogeneous images provide stronger suppression than heterogeneous ones. The inference of homogeneity depends on both the image and the neuron's properties. Thus, its role is evident not only in comparing suppression between image classes in a given cell, but, notably, is also apparent in comparing the suppression caused by a given image across cells.

What determines image homogeneity?

Homogeneity is a generalized measure of similarity between RF and surround signals. An inference of homogeneity involves determining whether the outputs of the RF and surround filters for a given image are typical of those observed in homogeneous natural images (equation (2), Online Methods). The dependencies between the RF and surround outputs across different spatial positions and orientations are summarized by a set of templates (technically, the eigenvectors of a covariance matrix) learned from the ensemble of natural images. The better the match between an image and these templates, the more likely the image is to be homogeneous. The matching of RF and surround outputs to the templates has a strong intensity dependence, with a higher probability for homogeneity when the filter outputs are larger. We found that the full inference in our framework can be approximated by a heuristic in which the strength of the surround is gated by the product of RF and surround filter outputs divided by their sum (**Supplementary Modeling** and **Supplementary Fig. 4**).

The characterization of inference as a measure of similarity of signals in the RF and surround is reminiscent of previous descriptions based on neuronal responses to simple stimuli[9,10,16,17]. For instance, a large, uniform grating is a quintessential example of a homogeneous image[27], consistent with uniform gratings being more suppressive than most natural images (**Fig. 1b**,**c**). Conversely, a compound grating with orthogonal orientations in the RF and surround is a clear example of a heterogeneous image[27] and produces weak or no suppression[16,17,30]. However, the inference about homogeneity also captures effects that elude previous descriptions, in addition to its better ability to account for suppression with natural images. For instance, a large patch of white noise appears as a uniform texture, yet it is heterogeneous because random image structure inside the RF is independent from the random structure in the surround[27]. Consistent with this prediction, white noise images elicited much weaker suppression than uniform gratings (**Supplementary Fig. 5**). Similarly, when high-order structure in natural images was destroyed by randomizing their Fourier phase, images became heterogeneous and a release from suppression was observed (**Supplementary Fig. 5**).

DISCUSSION

Understanding cortical responses to natural images is a well-recognized bottleneck in the study of visual processing, as models based on measurements with simple stimuli have struggled to predict responses to images[21]. Our model predicts V1 responses to natural images substantially better than previous models. This improved performance involves a gating of a widespread computation in cortex—divisive normalization from the surround—by an inference about image homogeneity. Our framework not only predicts this flexible form of normalization, but also explains why it exists: namely, that surround suppression is only needed to encode images efficiently when the RF and surround are statistically dependent. Thus, our approach shows the advantage of studying and explaining cortical responses through an understanding of image statistics, rather than by attempting to derive more complex descriptive models.

We stress that our results in no way argue against the importance of normalization as a canonical computation. On the contrary, in our framework, normalization is critical for encoding homogeneous images. The shortcoming of the standard normalization model of the surround is that it assumes suppression is only determined by the drive to the normalization pool.

We considered many previously proposed constellations of surround filters, but none could account for the measured responses (**Supplementary Fig. 1**). Although there might be a complex arrangement of surround filters that would improve the performance of the standard model, the combinatorial explosion of surround configurations makes searching for this constellation impractical. Furthermore, defining such a complex descriptive model amounts to a data fitting exercise—one is left with no understanding as to why visual cortex uses such an arrangement to encode images. We instead found that understanding how natural image statistics should be encoded readily leads to a simple generalization of the standard model, allowing the influence of the surround to be minimized when it is driven by different image components from the RF. Although our data do not directly demonstrate that V1 performs the Bayesian optimal computation, they suggest that suppression is gated in a way consistent with reducing redundancy in the outputs of the RF and surround filters.

How could the inference in our model be implemented in cortex? Such a sophisticated computation is likely to be a network-level operation, possibly occurring in higher cortical areas in which neurons are sensitive to higher order image structure. Regardless of the cortical locus, previous theoretical work has shown that populations of neurons can represent and compute with probabilities using simple, realistic neural operations: inference might boil down to simple comparisons of synaptic inputs[35]. Another possibility is that cortical circuits do not perform the full inference, but rather an approximation, perhaps using the simple heuristic—in which the sum of RF and surround outputs is compared to their product—that we derived.

How could the inference of homogeneity then be used to gate the efficacy of the surround? Surround suppression is thought to rely primarily on lateral and feedback connectivity[11]. Gating surround suppression therefore requires adjusting the gain of lateral and feedback inputs. Previous work has revealed network mechanisms[20,36] that could achieve such a gain adjustment. In particular, previous studies have argued that inhibition functions to stabilize the recurrent excitation that is recruited by strong stimulus drive[37,38]. In these inhibition-stabilized networks, providing additional external excitatory drive can have the counter-intuitive effect of reducing responsivity in the network. If only homogeneous images recruit additional excitation from higher cortex, one would see the consequences of inhibitory stabilization (that is, weaker responses) only for these

images. Under this scenario, homogeneous, but not heterogeneous, images should result in a withdrawal of both excitation and inhibition, a distinguishing feature of an inhibition-stabilized network.

An alternative possibility is that distinct classes of inhibitory neurons might mediate surround modulation[19,39,40] and its gating[41]. Recently, interneurons expressing vasoactive intestinal peptide (VIP) have been shown to target specific subtypes of interneurons[41–43], particularly somatostatin-expressing cells that may have a central role in surround suppression[39,40]. If VIP neurons were driven by feedback signals encoding homogeneity, they could silence surround suppression. Under this hypothesis, the activity of VIP interneurons should correlate with the inferred probability of image homogeneity. Furthermore, inactivation of these interneurons should block the gating of surround suppression for heterogeneous images, yielding suppression similar to that predicted by the standard model. Recent characterization of distinct subtypes of inhibitory interneurons has also made clear that some have little surround suppression[19,39], perhaps explaining why we and others have found neurons that show little suppression even with gratings, which are highly homogeneous stimuli (**Fig. 1b**). The absence of suppression in these neurons may be necessary for providing signals that allow the efficient encoding of images by other subpopulations of neurons (for example, pyramidal cells which project downstream).

Although our framework clearly extends the standard normalization model, it did not capture all of the explainable variance in cortical responses. Model performance might be improved by relaxing the assumption that the surround normalization pool consists of a fixed set of filters for each neuron. This could be accomplished by considering models in which the inference not only gates the surround, but also determines which components of the normalization pool are included, on an image-by-image basis[44]. An alternative way to improve performance would be to consider other computational goals. Our framework aims to reduce the dependencies between filter outputs in the RF and surround. Other goals might also be important, such as representing relevant information in a format that could be easily read-out by downstream neurons[45] or dynamically adjusting neuronal tuning to changing perceptual demands[46]. Thus, although the principle of efficient coding that underlies our framework is consistent with the behavior of V1 neurons, we do not suggest that achieving an efficient representation is their sole goal.

It is important to note that heterogeneous (or more broadly, non-stationary) inputs are not rare or peculiar to early visual processing. They occur frequently in natural sensory processing, as signals are often generated by different physical sources: the light reflected off occluding objects, the smell of different foods on a table or the voices of different speakers in a room. Our approach should therefore be highly relevant to the study of other types of sensory processing, as well as to other forms of sensory contextual modulation (for example, adaptation). More generally, the inference about image homogeneity is representative of a general problem known as causal inference, which involves assessing the plausibility of different interpretations about the causal relation between objects or events in the world (which are not directly measurable) and the sensory input they provide. Causal inference is pervasive not only in sensory processing[47], but also in higher cognitive function[48], suggesting that our framework is relevant to higher cognitive processes, particularly those already linked to divisive normalization[8,49].

METHODS

Methods and any associated references are available in the online version of the paper.

Note: Any Supplementary Information and Source Data files are available in the online version of the paper.

ACKNOWLEDGMENTS

We thank P. Dayan, C.A. Henry and A. Huk for comments on an earlier version of this manuscript, members of the Kohn laboratory for help performing recordings, and S. Barthelme for discussion on estimating model performance. This work was supported by a US National Institutes of Health grant to O.S. and A.K. (CRCNS EY021371), an Irma T. Hirchl Career Scientist Award (A.K.), a Sloan Research Fellowship (O.S.) and by Research to Prevent Blindness.

AUTHOR CONTRIBUTIONS

R.C.-C., A.K. and O.S. designed the study. R.C.-C. collected and analyzed the data. R.C.-C., A.K. and O.S. wrote the manuscript.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

1. Schwartz, O., Hsu, A. & Dayan, P. Space and time in visual context. *Nat. Rev. Neurosci.* **8**, 522–535 (2007).
2. Kohn, A. Visual adaptation: physiology, mechanisms and functional benefits. *J. Neurophysiol.* **97**, 3155–3164 (2007).
3. Carandini, M. & Heeger, D.J. Normalization as a canonical neural computation. *Nat. Rev. Neurosci.* **13**, 51–62 (2012).
4. Li, Z. Contextual influences in V1 as a basis for pop out and asymmetry in visual search. *Proc. Natl. Acad. Sci. USA* **96**, 10530–10535 (1999).
5. Itti, L. & Koch, C. Computational modelling of visual attention. *Nat. Rev. Neurosci.* **2**, 194–203 (2001).
6. Clifford, C.W.G., Wenderoth, P. & Spehar, B. A functional angle on some after-effects in cortical vision. *Proc. Biol. Sci.* **267**, 1705–1710 (2000).
7. Ohshiro, T., Angelaki, D.E. & DeAngelis, G.C. A normalization model of multisensory integration. *Nat. Neurosci.* **14**, 775–782 (2011).
8. Louie, K., Khaw, M.W. & Glimcher, P.W. Normalization is a general neural mechanism for context-dependent decision making. *Proc. Natl. Acad. Sci. USA* **110**, 6139–6144 (2013).
9. Gilbert, C.D. & Wiesel, T.N. The influence of contextual stimuli on the orientation selectivity of cells in primary visual cortex of the cat. *Vision Res.* **30**, 1689–1701 (1990).
10. Knierim, J.J. & van Essen, D.C. Neuronal responses to static texture patterns in area V1 of the alert macaque monkey. *J. Neurophysiol.* **67**, 961–980 (1992).
11. Angelucci, A. & Bressloff, P.C. Contribution of feedforward, lateral and feedback connections to the classical receptive field center and extra-classical receptive field surround of primate V1 neurons. *Prog. Brain Res.* **154**, 93–120 (2006).
12. Sceniak, M.P., Ringach, D.L., Hawken, M.J. & Shapley, R. Contrast's effect on spatial summation by macaque V1 neurons. *Nat. Neurosci.* **2**, 733–739 (1999).
13. Cavanaugh, J.R., Bair, W. & Movshon, J.A. Nature and interaction of signals from the receptive field center and surround in macaque V1 neurons. *J. Neurophysiol.* **88**, 2530–2546 (2002).
14. Levitt, J.B. & Lund, J.S. Contrast dependence of contextual effects in primate visual cortex. *Nature* **387**, 73–76 (1997).
15. Walker, G.A., Ohzawa, I. & Freeman, R.D. Asymmetric suppression outside the classical receptive field of the visual cortex. *J. Neurosci.* **19**, 10536–10553 (1999).
16. Cavanaugh, J.R., Bair, W. & Movshon, J.A. Selectivity and spatial distribution of signals from the receptive field surround in macaque V1 neurons. *J. Neurophysiol.* **88**, 2547–2556 (2002).
17. Sillito, A.M., Grieve, K.L., Jones, H.E., Cudeiro, J. & Davis, J. Visual cortical mechanisms detecting focal orientation discontinuities. *Nature* **378**, 492–496 (1995).
18. Vinje, W.E. & Gallant, J.L. Sparse coding and decorrelation in primary visual cortex during natural vision. *Science* **287**, 1273–1276 (2000).
19. Haider, B. *et al.* Synaptic and network mechanisms of sparse and reliable visual cortical activity during nonclassical receptive field stimulation. *Neuron* **65**, 107–121 (2010).
20. Ozeki, H., Finn, I.M., Schaffer, E.S., Miller, K.D. & Ferster, D. Inhibitory stabilization of the cortical network underlies visual surround suppression. *Neuron* **62**, 578–592 (2009).
21. Carandini, M. *et al.* Do we know what the early visual system does? *J. Neurosci.* **25**, 10577–10597 (2005).
22. Olshausen, B.A. & Field, D.J. Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature* **381**, 607–609 (1996).
23. Bell, A.J. & Sejnowski, T.J. The "independent components" of natural scenes are edge filters. *Vision Res.* **37**, 3327–3338 (1997).
24. Schwartz, O. & Simoncelli, E.P. Natural signal statistics and sensory gain control. *Nat. Neurosci.* **4**, 819–825 (2001).
25. Karklin, Y. & Lewicki, M.S. Emergence of complex cell properties by learning to generalize in natural scenes. *Nature* **457**, 83–86 (2009).
26. Spratling, M.W. Predictive coding as a model of response properties in cortical area V1. *J. Neurosci.* **30**, 3531–3543 (2010).

27. Coen-Cagli, R., Dayan, P. & Schwartz, O. Cortical surround interactions and perceptual salience via natural scene statistics. *PLOS Comput. Biol.* **8**, e1002405 (2012).
28. Lochmann, T., Ernst, U.A. & Deneve, S. Perceptual inference predicts contextual modulations of sensory responses. *J. Neurosci.* **32**, 4179–4195 (2012).
29. Vinje, W.E. & Gallant, J.L. Natural Stimulation of the nonclassical receptive field increases information transmission efficiency in V1. *J. Neurosci.* **22**, 2904–2915 (2002).
30. Webb, B.S., Dhruv, N.T., Solomon, S.G., Tailby, C. & Lennie, P. Early and late mechanisms of surround suppression in striate cortex of macaque. *J. Neurosci.* **25**, 11666–11675 (2005).
31. Heeger, D.J. Normalization of cell responses in cat striate cortex. *Vis. Neurosci.* **9**, 181–197 (1992).
32. Barlow, H.B. Possible principles underlying the transformations of sensory messages. in *Sensory Communication* (ed. Rosenblith, W.A.) 217–234 (MIT Press, 1961).
33. Ruderman, D.L. & Bialek, W. Statistics of natural images: scaling in the woods. *Phys. Rev. Lett.* **73**, 814–817 (1994).
34. Felsen, G., Touryan, J., Han, F. & Dan, Y. Cortical sensitivity to visual features in natural scenes. *PLoS Biol.* **3**, e342 (2005).
35. Pouget, A., Beck, J.M., Ma, W.J. & Latham, P.E. Probabilistic brains: knowns and unknowns. *Nat. Neurosci.* **16**, 1170–1178 (2013).
36. Nauhaus, I., Busse, L., Carandini, M. & Ringach, D.L. Stimulus contrast modulates functional connectivity in visual cortex. *Nat. Neurosci.* **12**, 70–76 (2009).
37. Rubin, D.B., Van Hooser, S.D. & Miller, K.D. The stabilized supralinear network: a unifying circuit motif underlying multi-input integration in sensory cortex. *Neuron* **85**, 402–417 (2015).
38. Ahmadian, Y., Rubin, D.B. & Miller, K.D. Analysis of the stabilized supralinear network. *Neural Comput.* **25**, 1994–2037 (2013).
39. Adesnik, H., Bruns, W., Taniguchi, H., Huang, Z.J. & Scanziani, M. A neural circuit for spatial summation in visual cortex. *Nature* **490**, 226–231 (2012).
40. Nienborg, H. *et al.* Contrast dependence and differential contributions from somatostatin- and parvalbumin-expressing neurons to spatial integration in mouse V1. *J. Neurosci.* **33**, 11145–11154 (2013).
41. Pfeffer, C.K., Xue, M., He, M., Huang, Z.J. & Scanziani, M. Inhibition of inhibition in visual cortex: the logic of connections between molecularly distinct interneurons. *Nat. Neurosci.* **16**, 1068–1076 (2013).
42. Lee, S., Kruglikov, I., Huang, Z.J., Fishell, G. & Rudy, B. A disinhibitory circuit mediates motor integration in the somatosensory cortex. *Nat. Neurosci.* **16**, 1662–1670 (2013).
43. Pi, H.-J. *et al.* Cortical interneurons that specialize in disinhibitory control. *Nature* **503**, 521–524 (2013).
44. Schwartz, O., Sejnowski, T.J. & Dayan, P. Soft mixer assignment in a hierarchical generative model of natural scene statistics. *Neural Comput.* **18**, 2680–2718 (2006).
45. Yamins, D.L.K. *et al.* Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proc. Natl. Acad. Sci. USA* **111**, 8619–8624 (2014).
46. Gilbert, C.D. & Li, W. Top-down influences on visual processing. *Nat. Rev. Neurosci.* **14**, 350–363 (2013).
47. Kersten, D., Mamassian, P. & Yuille, A. Object perception as Bayesian inference. *Annu. Rev. Psychol.* **55**, 271–304 (2004).
48. Gershman, S.J. & Niv, Y. Learning latent structure: carving nature at its joints. *Curr. Opin. Neurobiol.* **20**, 251–256 (2010).
49. Green, C.S., Benson, C., Kersten, D. & Schrater, P. Alterations in choice behavior by manipulations of world model. *Proc. Natl. Acad. Sci. USA* **107**, 16401–16406 (2010).

ONLINE METHODS

Animal preparation and data collection. Data were collected from three adult male monkeys (*Macaca fascicularis*). Animal preparation and general methods were described previously[50]. In brief, anesthesia was induced with ketamine (10 mg per kg of body weight) and maintained during surgery with isoflurane (1.0–2.5% in 95% O_2). During recordings, anesthesia was maintained by sufentanil citrate (6–18 μg per kg per h, adjusted as needed for each animal). Vecuronium bromide (0.15 mg per kg per h) was used to suppress eye movements. The use of anesthesia allowed us to present a large ensemble of images while ensuring precise and reproducible retinal positioning across trials. All procedures were approved by the Albert Einstein College of Medicine at Yeshiva University and followed the guidelines in the United States Public Health Service Guide for the Care and Use of Laboratory Animals.

We recorded neuronal activity using arrays of 10 × 10 microelectrodes (400-μm spacing, 1-mm length) inserted in the opercular region of V1. Waveform segments that exceeded a threshold (a multiple of the RMS noise on each channel) were digitized (30 kHz) and sorted off-line. For all analysis we included signals from well-isolated single units as well as small multi-unit clusters, and refer to both as neurons. We obtained similar results when our analysis was based only on very well-isolated single units (**Supplementary Table 1**).

Visual stimuli and presentation. We displayed stimuli on a calibrated CRT monitor (1,024 × 768 pixels, 100-Hz frame rate, ~40 cd m^{-2} mean luminance) placed 110 cm from the animal, using custom software.

We measured the spatial RF of each neuron, using small gratings (0.5 degrees in diameter, four orientations, 250-ms presentation) presented at a range of positions. The RF center of each neuron was defined as the location of the peak of a two-dimensional Gaussian fit to the spatial activity map (across the population, median $R^2 = 0.59$).

We used natural images from several common databases: the Berkeley Segmentation Database, the Brodatz textures, and the Van Hateren database. We converted color images to gray scale values, and extracted 894 random patches of 320 × 320 pixels, for which the average pixel intensities were in the range 40–200 (where 255 was the maximal intensity that could be displayed) and RMS contrast was in the range 0.25–0.75. From these patches, we selected our experimental stimuli, with two objectives: to drive V1 cells sufficiently well that we could measure suppression, if it occurred, and to maximize the diversity of image types in the experimental ensemble. To select images that would provide robust drive to V1 cells, we computed the power spectrum of the central 1 degree (50 × 50 pixels) region of each image. We excluded images for which the centroid of the rotational average of the power spectrum (a measure of the average spatial frequency content) fell outside the range 0.5–10 cycles per degree, typically preferred by V1 neurons[50]. To ensure a diverse ensemble, we computed the orientation histogram of the central 1 degree using a steerable pyramid[51] with 16 orientations and 3 scales. We binned images into three categories, low, medium and high, based on the entropy of the orientation histogram, a measure of the 'orientedness' of the image. A lower entropy indicates that more of the image energy was confined in a single orientation band. We selected the 30 images closest to the middle of each bin. For images with low and intermediate entropy, we created four copies rotated in steps of 45 degrees. This provided a total of 270 images (30 without a dominant orientation in the center, and 2 × 4 × 30 with a dominant orientation in the center).

We adjusted the mean pixel value and RMS contrast of each image to values of 116 and 0.6, respectively. This was done to preclude any influence of different mean luminance or contrast across images on neuronal responses.

Each image was included in the experimental ensemble twice: once masked by a small annular window with diameter 50 pixels (1.04 degrees), and also by a larger annular window with inner diameter of 150 pixels (3.1 degrees of visual field, used in 1 animal) or 320 pixels (6.7 degrees, 2 animals). We used an annulus of 150 rather than 320 pixels in one animal to test whether model performance would be improved by activating a smaller portion of the surround. We observed no difference between the two data sets, so the results were pooled. To ensure that the experimental ensemble provided an accurate and unbiased view of the role of inferred homogeneity in determining suppression strength, we adopted a resampling procedure. For each cell, we resampled responses from those small images that drove the cell, to match the proportion of homogeneous and heterogeneous cases in small natural images for the filters representing that cell. Small images could appear homogeneous or heterogeneous because of the partial overlap between center and surround filters.

To measure the tuning properties of each neuron, we interleaved static grating stimuli with the natural images. We varied grating size (7 diameters logarithmically spaced between 0.33 and 6.25 degrees, presented at 4 orientations), orientation (16 values equally spaced between 0 and 180 degrees), and spatial frequency (8 values logarithmically spaced between 0.25 and 7.26 cycles per degree, presented at four orientations). Variations of size and orientation were presented at a spatial frequency of 1 cycle per degree. Variations of orientation and spatial frequency were presented at a size of 1 degree. All gratings were presented at four spatial phases, with mean luminance and RMS contrast matched to that of the natural images.

All stimuli were displayed in pseudo-random order for 100 ms each, followed by a 200-ms uniform gray screen. Each stimulus was presented 20 times. Stimuli were presented monocularly in a circular aperture surrounded by a gray field of average luminance.

Characterization of neuronal responses. We quantified neuronal activity by the mean firing rate (R) across repeated trials of the same image, measured in a window 50–150 ms following stimulus onset (to account for response latency). The spontaneous activity was measured during the last 50 ms of the blank screen following each image.

The strength of surround suppression, or modulation ratio (MR), was defined by a common metric[14,29], namely the ratio between the raw response to a large image (3.1 or 6.7 degrees) and its small counterpart (1.04 degree) $MR = R^{\text{large}} / R^{\text{small}}$. To obtain reliable estimates of MR, we considered for each neuron only those images with R^{small} at least 1 s.d. above the spontaneous rate.

To compare the suppression elicited by a given image across different neurons, we used a normalized MR (NMR) to account for differences in the overall suppression across neurons: $NMR = MR/MR^{\text{grating}}$, where MR^{grating} quantifies the suppression measured with the most preferred static grating in the stimulus ensemble. We normalized by MR^{grating} rather than by the average MR across natural images, because each neuron was typically driven by a different subset of images.

Inclusion criteria. The 207 neurons analyzed in **Figure 1** were selected because they met the following criteria: 1) The RF was centered on the stimulus, namely either the distance between RF center and stimulus center was less than 0.2 degree (two animals), or the maximal response to annular grating stimuli (1 degree inner diameter) was less than 15% of the maximal response across all stimuli (one animal), thus ensuring that surround stimuli did not provide substantial drive to the RF center. 2) The neuron responded above its spontaneous rate to at least 4 small natural images. The 126 neurons on which the majority of our analyses were based (**Figs. 2**, **3** and **5**–**7**), met the additional criterion that they showed evidence of surround modulation with natural images. This was defined by the responses being explained better by a model that included surround modulation (standard or flexible) than one that assumed none. Similar results held when we included all well-centered and responsive neurons ($n = 207$), but this diluted the difference between models because of the inclusion of neurons that showed no appreciable suppression (**Supplementary Table 1**).

Standard normalization model. Under the standard normalization model, the response R to an image is defined by the drive to the RF, divided by the drive to the normalization pool[3,13]:

$$R = \alpha \left(\frac{E_{RF}}{\sigma + \beta E_{\mathbf{k}} + \gamma E_{\mathbf{s}}} \right)^n \qquad (1)$$

The E_{RF} term measures image drive, or energy, to the RF, computed from the outputs of filters chosen individually for each neuron; $E_{\mathbf{k}}$ and $E_{\mathbf{s}}$ capture the drive to the normalization pool inside the RF and in the surround, respectively. β and γ determine the relative strength of these normalization pools, which may differ[13]. Note that γ captures the average surround suppression strength, which can vary widely across neurons[10,12–15] (**Fig. 1b**). Finally, the parameters α, σ and n capture the relationship between filter drive and spiking responses.

The drive to the RF, E_{RF}, is defined as follows. The RF of each cell was represented by a pair of linear filters (quadrature pair, of even and odd phase). The output

doi:10.1038/nn.4128

of each filter was defined by the dot product between the image and the filter. The drive was then defined as the square root of the sum of squares of the filters' outputs. We verified that the use of quadrature pairs (that is, modeling the RF drive as for complex cells) was justified for our data (**Supplementary Fig. 6**).

The normalization signal in the RF, $E_{\mathbf{k}}$, was computed from the output of four quadrature pairs of filters. These filters had the same size, position, and spatial frequency tuning as those of E_{RF}, but each pair had a different preferred orientation (0, 45, 90 and 135 degrees), yielding a normalization signal that was untuned for orientation[52]. The normalization signal from the surround $E_{\mathbf{s}}$ included eight pairs of filters, uniformly positioned around the RF center as on a clock face (**Fig. 2a**). The surround filters had the same spatial frequency preference and size as those in the RF. All surround filters had the same preferred orientation. This preferred orientation was changed from image-to-image, to match the dominant orientation in the image presented inside the RF, thus capturing the finding that the orientation tuning of surround suppression shifts with the orientation content within the RF[13,17]. The distance between the surround filters and the RF center was proportional to the RF size, allowing for a slight spatial overlap between the RF and surround and yielding a surround whose spatial extent was 2–3 times that of the RF[11–13].

Inference about homogeneity. The inference of image homogeneity was based on a Mixture of Gaussian Scale Mixtures model (MGSM), described previously[27,44] and summarized in **Supplementary Modeling**. This Bayesian model considers the statistical dependency between the outputs of the filters that define the RF and surround for a given neuron: **k**, which represents the outputs of the RF filters and the filters providing normalization within the RF; and **s**, which represents the outputs of the filters providing the normalization signals from the surround. The inclusion in **k** of both the RF filters and the within-RF normalization pool corresponds to the assumption that the outputs of spatially overlapping filters are always statistically dependent.

For a given input image, **k** and **s** are observed, but their dependence (homogeneous image) or independence (heterogeneous) must be inferred. The inferred probability that the image is homogeneous, given the observed filter responses, is[44]

$$p(\text{homogeneous}\,|\,\mathbf{k},\mathbf{s}) \propto p(\text{homogeneous}) \cdot \frac{\mathrm{B}\left(1-\frac{n}{2};\sqrt{(\mathbf{k},\mathbf{s})\mathbf{C}^{-1}(\mathbf{k},\mathbf{s})^{\mathrm{T}}}\right)}{\sqrt{(2\pi)^n \det(\mathbf{C})}\left(\sqrt{(\mathbf{k},\mathbf{s})\mathbf{C}^{-1}(\mathbf{k},\mathbf{s})^{\mathrm{T}}}\right)^{\frac{n-2}{2}}} \qquad (2)$$

where n is the total number of filters, B is the modified Bessel function, **C** is the covariance matrix between **k** and **s**, and p(homogenous) is the prior probability that **k** and **s** are dependent.

The covariance and the prior were learned by applying the filters representing each neuron to a separate database of natural images not used in the experiments. The goal of this optimization is for the model to learn which patterns in natural images distinguish homogeneous from heterogeneous images (the covariance), and how often they occur (the prior). Technically, this amounts to maximizing the likelihood of the training images under the MGSM. Further details are provided in **Supplementary Modeling**.

Flexible normalization model. The predicted response for the flexible normalization model was based on the standard model, equation (1), except that the surround term in the denominator, $E_{\mathbf{s}}$, was multiplied by a gating variable which took a value of 1 if the image was inferred homogeneous with a probability greater than 0.5, and 0 otherwise. Thus, when the image was inferred to be homogeneous, the flexible normalization model was identical to the standard model; when the image was inferred to be heterogeneous, there was no influence of the surround. The value of the gating variable was not fit to neural data. Therefore, the flexible and standard models used the same free parameters.

Model fitting. We first estimated the parameters of the normalization models (σ, n, α, β and γ) and the choice of filters for each neuron, by maximum-likelihood fitting to the raw responses to all images that drove the cell (as defined above). Likelihoods were computed assuming Gaussian variability of the spike counts, with a variance which scaled with the mean response[52]. Maximum likelihood fitting in this case amounts to minimizing the sum of squared errors between the predicted and measured responses, weighted by the inverse variances.

The maximization was performed numerically for the normalization model parameters, and by exhaustive search for the best filters. Specifically, we searched for the best filters from a set defined by a steerable pyramid[53]. The steerable pyramid is a popular method to construct oriented bandpass filters at multiple preferred orientations, spatial frequencies, and sizes[51]. We used an extension that allowed us to build simultaneously both filters in a quadrature pair[53]. The set contained filters with preferred orientations of 0, 45, 90, or 135 degrees, an orientation bandwidth (half width at full height) of 45, 22.5, or 11.25 degrees, a preferred spatial frequency (ω) of 1/12, 1/24, or 1/48 cycles/pixel, and one of four sizes (in pixels)

$$\left\{\frac{1}{\omega}+\frac{\omega}{12}m\right\}_{m=2,4,6,8}$$

In total, we thus considered 144 distinct filters. We constrained the choice of filter size to be within 25% of the measured RF size, defined as the peak of the neuronal area summation curve measured with gratings. Note that the relationship between the filters for the RF, the within-RF normalization pool, and the surround was fixed: the RF filters and those of the normalization pools of each neuron were the same, except that the orientation preference of the surround filters changed from image-to-image and the within-RF normalization pool always included all four orientation preferences, as described above.

Both standard and flexible models could account for the rare occurrence of response facilitation with the larger image, if that image provided additional drive to the RF but either little drive to the surround or was inferred heterogeneous.

Model comparison. We compared standard and flexible normalization models by cross-validation. We divided the measured responses into a training and test set, obtained by holding out the responses to one pair of large and small images. For the training set, we re-estimated the (σ, n, α, β and γ) parameters. We then quantified the prediction quality on the left-out data, using a normalized log-likelihood ratio[54]

$$\text{Prediction quality} = \frac{L - L_{\text{null}}}{L_{\text{oracle}} - L_{\text{null}}}$$

here L denotes the log-likelihood of the left-out data given the model parameters; L_{null} denotes the log-likelihood of a 'null' model that always predicts the mean response across all images of the training set; and L_{oracle} denotes the log-likelihood of an 'oracle' that is assumed to know the true mean response to each image, and whose error is due exclusively to the fact that the observed mean is estimated from a finite number of trials. Because of this uncertainty, the oracle's error can be marginally larger than the model's error on measured data, hence prediction quality can exceed a value of 1 (**Fig. 5b**). We repeated the procedure to compute prediction quality for all possible splits of the images into training and test sets, and then averaged the results across all held-out images (that is, leave-one-out cross-validation).

Note that we did not update our choice of filters for each subset of responses used for training in cross-validation, because that was computationally prohibitive. To be sure that prediction quality was not inflated by using filters chosen in part using responses to the test images, we used a second independent approach to choose filters for both the standard and flexible models in a subset of neurons: we measured the tuning for orientation, spatial frequency, and size, with static grating stimuli. We then chose the filters to match these measurements. All of our core results were quantitatively indistinguishable with this method (**Supplementary Fig. 7**).

Statistical analysis. Error bars represent bootstrap estimates of the 68% confidence intervals, obtained by sampling 10,000 times with replacement. Statistical significance was assessed for **Figure 1b** with a two-tailed permutation test, the standard choice for Spearman correlation. For all other analyses, we used non-parametric bootstrap t tests to relax the assumption of normality, using 25,000 bootstrap samples (unpaired, two-sided in **Figs. 6b** and **7e**; paired, one-sided for all other comparisons), except where noted. For the bootstrap paired t test, we created the bootstrap distribution of t statistics for the null hypothesis (mean difference equals zero) by subtracting the mean difference from each measured difference, sampling with replacement from this centered distribution, and calculating t statistics for the sampled differences. The P value was given by the proportion of cases in which the bootstrapped t statistics had a more extreme value than the t statistics for the actual data. The bootstrap unpaired t test was similar, except the distribution for the null hypothesis (identical means) was constructed by subtracting the mean of each data

set from the corresponding elements of that data set, combining the data sets, sampling with replacement to create two surrogate data sets, and computing the *t* statistics for the difference in the sampled surrogates. The average *MR* and *NMR* were geometric means, and statistical significance was evaluated for the logarithm of *MR* and *NMR*. No blinding or randomization was used. No statistical methods were used to pre-determine sample sizes but our sample sizes are similar to those of previous neuronal studies of surround suppression[12–17].

A **Supplementary Methods Checklist** is available.

50. Jia, X., Smith, M.A. & Kohn, A. Stimulus selectivity and spatial coherence of gamma components of the local field potential. *J. Neurosci.* **31**, 9390–9403 (2011).
51. Simoncelli, E.P., Freeman, W.T., Adelson, E.H. & Heeger, D.J. Shiftable multiscale transforms. *IEEE Trans. Inf. Theory* **38**, 587–607 (1992).
52. Carandini, M., Heeger, D.J. & Movshon, J.A. Linearity and normalization in simple cells of the macaque primary visual cortex. *J. Neurosci.* **17**, 8621–8644 (1997).
53. Portilla, J. & Simoncelli, E. A Parametric texture model based on joint statistics of complex wavelet coefficients. *Int. J. Comput. Vis.* **40**, 49–70 (2000).
54. Stocker, A.A. & Simoncelli, E.P. Noise characteristics and prior expectations in human visual speed perception. *Nat. Neurosci.* **9**, 578–585 (2006).