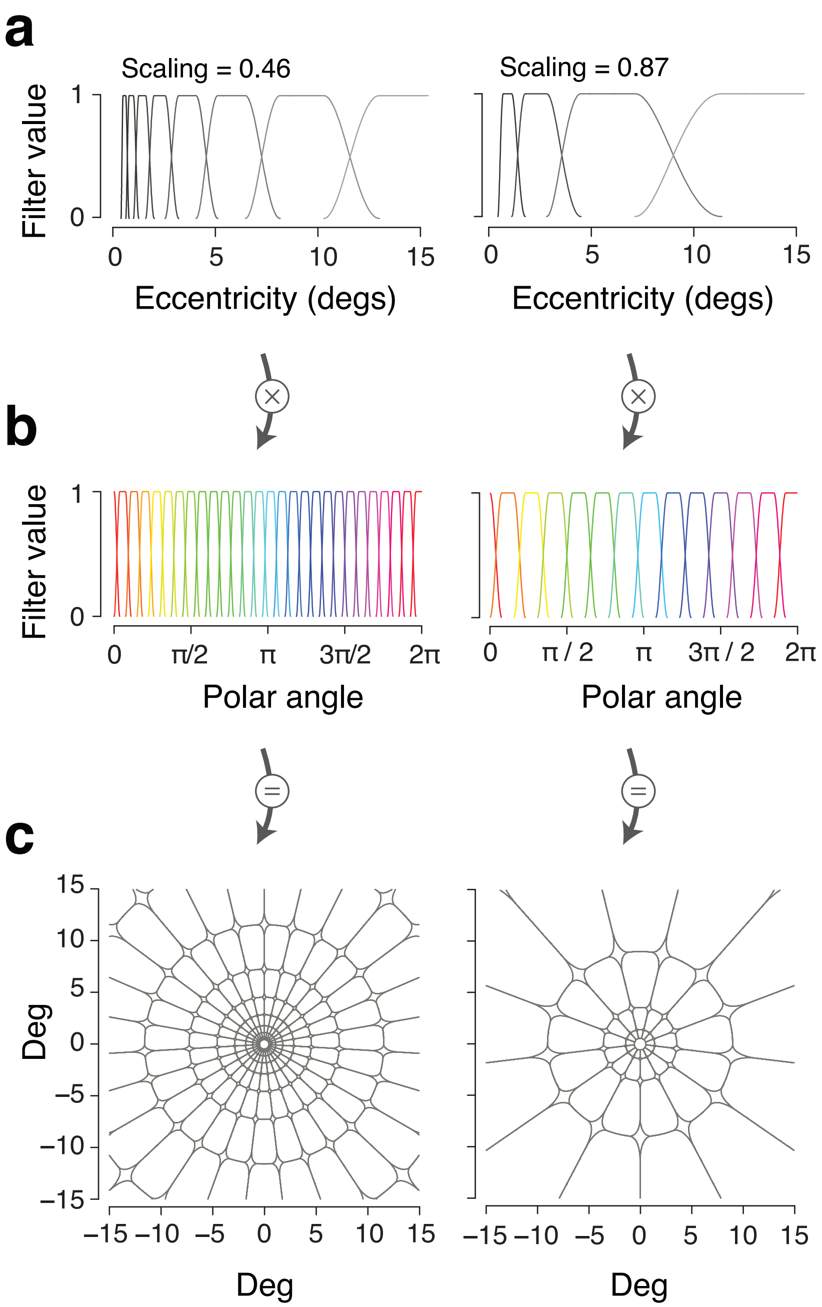
Supplementary Information

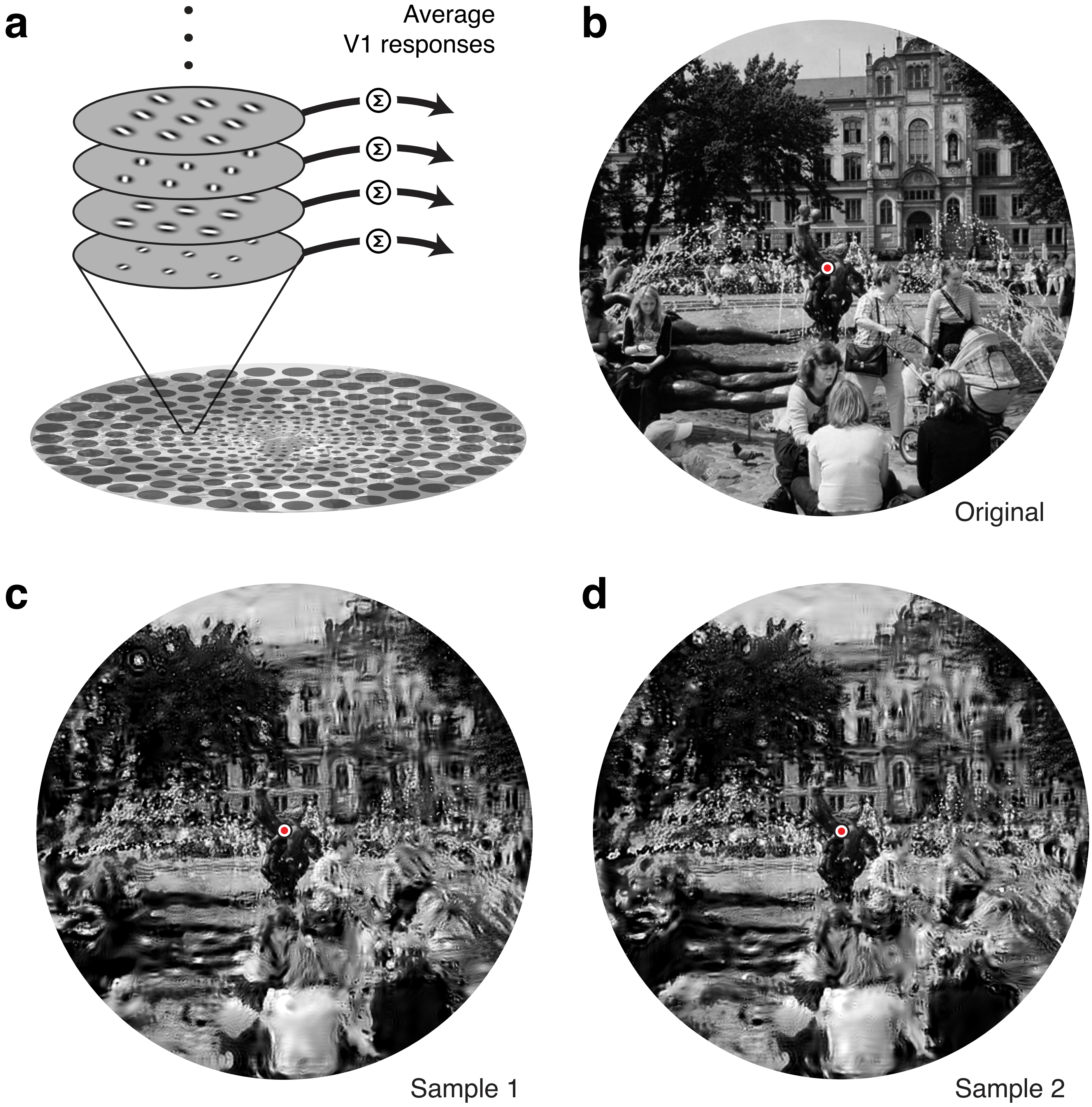
for Metamers of the Ventral Stream

Jeremy Freeman and Eero P. Simoncelli

Contents



Supplementary Figure 1. Construction of spatial pooling regions using filters that are separable in log eccentricity and polar angle. **(a)** Filters have a flat top and raised cosine transition regions, and are constructed to evenly tile log eccentricity, which yields filters that grow in size with eccentricity according to a fixed scaling (ratio of size to eccentricity). Filters are shown for two scalings, 0.46 and 0.87. Filters are constructed to have approximately 50% overlap. **(b)** Similarly constructed filters are spaced evenly to tile polar angle. Larger scalings yield broader polar angle filters, to ensure that the resulting two-dimensional spatial filters have a fixed ratio of width in eccentricity to width in polar angle (see Methods). **(c)** The two separable components are combined to obtain two-dimensional spatial filters. Contours indicate the full-width half-maximum of each filter. Actual filters have soft edges and overlap by approximately 50%, as shown for the separable components in (a) and (b). The full set of filters tile evenly, and sum to a constant.



Supplementary Figure 2. V1 model and stimuli. **(a)** In each spatial pooling region, the image is represented with a bank of model V1 complex cells, varying in their preferred orientation and spatial frequency. Model responses are averages of the squared filter responses over the pooling regions. The model captures local spectral energy, but not local correlations across orientations and scales. **(b)** An original photograph of the Brunnen der Lebensfreude fountain in Rostock, Germany (courtesy of Bruce Miner). **(c)** and **(d)** Image samples, randomly selected from the set of all images that generated V1 model responses identical to the original (panel b). The value of the scaling parameter (used to determine the pooling regions of the model) was selected to yield 75% correct performance in discriminating such synthetic images (Fig. 4). While fixating the center (red circle) the two images should appear nearly identical to the original and to each other.

Supplementary methods

1 Mid-ventral model responses

We provide mathematical details for computation of the second stage of model responses. The V1 responses are based on a complex steerable pyramid, in which the image is convolved with a bank of oriented bandpass filters, and their Hilbert transforms [13, 15]. The pyramid has a total of sixteen subbands (four orientations at four different scales), and a lowpass residual band. We write the nth V1 subband as $x_n(i,j)$, a two-dimensional array containing the complex-valued responses. We also use the more compact vector notation $\vec{x}_n$. The real part of this subband, which arises from convolution with the symmetric filter is denoted $s_n(i,j)$, and the magnitude of the subband (i.e., the square root of the sum of squared responses of symmetric and anti-symmetric filters) is $e_n(i,j)$.

The model responses are based on those developed in Portilla & Simoncelli (2000) for global texture modeling, but all averages are computed over localized pooling regions. A pooling region is defined by a weighting function whose values sum to one. In the following expressions, we consider a single pooling region, with weights denoted $w(i,j)$.

Simple autocorrelations are weighted spatial averages of products of the symmetric filter responses at nearby spatial locations:

$$A_w(n,k,l) = \sum \sqrt{w(i,j)}\,(s_n(i,j) - \mu_w(\vec{s}_n))\,\sqrt{w(i+k,j+l)}\,(s_n(i+k,j+l) - \mu_w(\vec{s}_n))\,, \tag{1}$$

where (k,l) specifies the spatial displacement (in horizontal and vertical directions), the summation is over (i,j), and $\mu_w(\vec{s}_n)$ is the weighted mean,

$$\mu_w(\vec{s}_n) = \sum w(i,j)s_n(i,j). \tag{2}$$

In our mid-ventral model, we include the autocorrelation for spatial displacements $(-3 \leq k \leq 3, -3 \leq l \leq 3)$. In the V1 model, we only include the central sample (i.e. $k = l = 0$), for which Eq. (1) reduces to a weighted variance.

Complex cell autocorrelations are analogous weighted averages of products of the magnitudes:

$$B_w(n,k,l) = \sum \sqrt{w(i,j)}\,(e_n(i,j) - \mu_w(\vec{e}_n))\,\sqrt{w(i+k,j+l)}\,(e_n(i+k,j+l) - \mu_w(\vec{e}_n))\,, \tag{3}$$

Again, we include displacements $(-3 \leq k \leq 3, -3 \leq l \leq 3)$ in the mid-ventral model. But complex cell autocorrelations are not included in the V1 model.

Cross-orientation and cross-scale correlations are computed between between magnitudes at different orientations and scales:

$$C_w(n,m) = \sum w(i,j)\,(e_n(i,j) - \mu_w(\vec{e}_n))\,(e_m(i,j) - \mu_w(\vec{e}_m))\,, \tag{4}$$

where indices (n, m) specify two subbands arising from filters at different orientations at the same scale, or at different orientations at adjacent scales. This yields 6 cross-orientation correlations at each scale, and 16 cross-scale correlations for each scale.

Finally, cross-scale correlations are also computed between the complex-valued responses at one scale, and the phase-doubled responses at a coarser scale:

$$S_w(n,m) = \sum w(i,j)\,(x_n(i,j) - \mu_w(\vec{x}_n)) \left(\frac{x_m^2(i,j)}{|x_m(i,j)|} - \mu_w \left(\frac{x_m^2(i,j)}{|x_m(i,j)|} \right) \right), \tag{5}$$

where indices (n, m) specify two subbands arising from filters at adjacent scales (with n corresponding to the finer scale).

Our model also includes weighted marginal statistics (mean, variance, skew, kurtosis), computed on the symmetric (real) filter responses. The weighted mean is given in Eq. (2). Higher-order weighted moments of order p are,

$$\mu_w^{(p)}(\vec{s}_n) = \sum w(i,j)\,(s_n(i,j) - \mu_w(\vec{s}_n))^p\,.$$

From this, the skew and kurtosis are:

$$\begin{aligned}\gamma_w(\vec{s}_n) &= \frac{\mu_w^{(3)}(\vec{s}_n)}{\left(\mu_w^{(2)}(\vec{s}_n)\right)^{3/2}}\\ \kappa_w(\vec{s}_n) &= \frac{\mu_w^{(4)}(\vec{s}_n)}{\left(\mu_w^{(2)}(\vec{s}_n)\right)^{2}}\end{aligned}$$

2 Observer model

We derive an observer model that can be fit to the data in our metamer experiments. Subjects viewed images that were matched for a set of model responses, within pooling regions that scaled in size with eccentricity according to scaling s. We varied this scaling across our experimental stimuli to obtain percent correct as a function of s. We assume the observer uses the same representation that is used to generate the images, with (unknown) critical scaling, s_0. Let $\vec{x}$ be a vector of values to be locally averaged (e.g., the array containing pairwise products of two orientation subbands). Let M be a matrix whose rows contain the weighting functions (with sizes scaling according to s), that are used to compute the local averages for the purposes of synthesis. Assume that $\vec{x}$ and $\vec{y}$ have been adjusted so that $M\vec{x} = M\vec{y}$. Define the projection matrix $P = M^\top(MM^\top)^{-1}M$, which projects vectors into the space spanned by M.

Now let R be the matrix that the observer uses to compute averages over regions scaling with s_0. We assume the discriminability of the two stimuli depends on the sum of squared differences between these averages. We can express the expected value of this quantity, taken over instantiations of $\vec{x}$

and $\vec{y}$ that match the same model measurements, as:

$$\begin{aligned} d^2 &= \mathbf{E}\left[||R\vec{x} - R\vec{y}||^2\right] \\ &= \mathbf{E}\left[||R(I-P)(\vec{x}-\vec{y}) + RP(\vec{x}-\vec{y})||^2\right] \\ &= \mathbf{E}\left[||R(I-P)(\vec{x}-\vec{y})||^2\right], \end{aligned}$$

where we use the fact that $M(\vec{x}-\vec{y}) = 0$, which implies (using the definition of P) that $RP(\vec{x}-\vec{y}) = 0$. Assuming that $\vec{x}$ and $\vec{y}$ are independent, and have the same covariance matrix, C, gives:

$$\begin{aligned} d^2 &= Tr\left(\mathbf{E}\left[R(I-P)(\vec{x}-\vec{y})(\vec{x}-\vec{y})^\top (I-P^\top)R^\top\right]\right) \\ &= Tr\left(R(I-P)2C(I-P^\top)R^\top\right) \\ &= Tr\left((R - RM^\top(MM^\top)^{-1}M)2C(R^\top - M^\top(MM^\top)^{-1}MR^\top)\right). \end{aligned}$$

Without loss of generality, we can assume that the variance along each dimension is equal, and thus that C is a multiple of the identity matrix. Under this assumption, and after some matrix algebra, we obtain

$$d^2 \propto \left(Tr(RR^\top) - Tr\left(R^\top RM^\top(MM^\top)^{-1}M\right)\right). \qquad (6)$$

This provides a generic closed-form expression for the overall error as a function of the measurement matrices M and R. Although the assumption that the components of $\vec{x}$ (and $\vec{y}$) are decorrolated may not hold for all of the parameters of our ventral model, we find through simulations that the resulting expression for discriminability still holds.

Finally, we wish to express this result in terms of the scaling parameters for the synthesis model and the observer. This is easily obtained from Eq. (6) if we assume that (i) M and R each compute local means within blocks of fixed sizes m and r, respectively, (ii) m is an integer multiple of r (iii) both m and r divide evenly into n, the length of $\vec{x}$. For matrices with this structure, we can express d^2 as a function of m:

$$d^2(m) \propto \begin{cases} \frac{n}{r^2}\left(1 - \frac{r}{m}\right) & m > r \\ 0 & m \leq r \end{cases} \qquad (7)$$

This expression has a natural continuous generalization to handle smoothly overlapping averages and non-integer ratios. The radial extent of our model pooling regions is proportional to the scaling s, so the average region size will be proportional to s^2, with a proportionality constant that depends on the shape of the region. Replacing m with s^2, and r with s_0^2, and absorbing the factor of n/r^2 into a single scale constant, gives the closed form approximation:

$$d^2(s) \approx \begin{cases} \alpha_0(1 - s_0^2/s^2) & s > s_0 \\ 0 & s \leq s_0 \end{cases} \qquad (8)$$

We empirically verified that this approximation holds for the smooth weighting functions used in our model implementation. The proportionality factor, α_0, is likely to differ for each measurement in the model. If we assume that the observer performs a weighted sum of the squared errors over the full set of measurements, then the overall error will be of the same form as that of Eq. (8). Notice that α_0 scales the magnitude of the squared difference, without affecting the point at which the curve exceeds 0 (i.e., when $s = s_0$). Thus, when fitting the data, the gain parameter captures variability in overall performance across observers and presentation conditions.

Finally, we use signal detection theory (Macmillan, 1977) to compute the probability of a correct response P_c in the ABX task as a function of the underlying difference d^2,

$$P_c = \Phi\left(d^2/\sqrt{2}\right)\Phi(d^2/2) + \Phi\left(-d^2/\sqrt{2}\right)\Phi(-d^2/2), \tag{9}$$

where Φ is the CDF of the Normal distribution. We use the MATLAB fminsearch routine to find the values of s_0 and the gain that maximize the likelihood of the data (proportion correct responses for each scaling) under this model, for each subject and condition. We use bootstrapping to obtain 95% confidence intervals for the parameter estimates: we resample the individual trials with replacement, and refit the resampled data to reestimate the parameters.

3 Physiological estimates of receptive field size

We performed a meta-analysis to estimate the relationship between physiologically measured receptive field size and eccentricity in non-human primates. Measurements of receptive field sizes are variable across different experiments because different labs use different stimuli and mapping procedures [14, 16, 4]. To compare our psychophysics to physiology, we considered a wide range of data sets: four in V2 [8, 9, 3, 1], five in V1 [8, 10, 4, 7, 2], and three in V4 [9, 12, 5]. Two of these data sets are from owl monkey [1, 2], one from capuchin [10], and the rest are from macaque.

For each visual area, we combined data across experiments and estimated variability by pooling the raw data (rather than the fits), matching sample sizes, and resampling multiple times to obtain a 95% confidence interval on the slopes. Specifically, we determined the minimum number of cells across the data sets, and on each iteration of a bootstrap, resampled that number with replacement from each data set, and reestimated the slope of size versus eccentricity from the pooled data. We fit the data with a two-parameter hinged line, with a constant minimum size over some small range of eccentricities, followed by a linear relationship with some slope. For consistency, we used this "hinged line" model to estimate all slopes, but we obtained similar results when using a linear fit through 0. We also considered a straight line with variable intercept and slope [6], but the hinged line fits the data well (error was comparable for the two fits) and is better matched to the parameterization of our model. Variability across data sets tended to be largest at far eccentricities, and given that our visual stimuli only extended to 12.25 deg, we restricted our analysis of the physiology data to this range. In some of the cited studies [8, 10, 9, 7, 3, 5], rectangular receptive field sizes were mapped using a minimum response field procedure. To convert these numbers to diameters of circular receptive fields, and partially compensate for the bias toward smaller values inherent in this mapping technique [16, 4], we took the average of the diameter associated with the corners and sides of the squares (i.e., we multiplied the reported diameters by $(1+\sqrt{2})/2$). Small modifications to any of these aspects of the data analysis did not qualitatively change the comparison between our psychophysics and the physiology.

4 Color

Applying our model to color images (e.g., Fig. 7) requires a minor modification. We initially tried applying the existing model directly to each color channel of the image (red, green, and blue), and then combining the channels into a synthesized color image. However, that procedure yields large color artifacts because the synthesis fails to respect pixel-domain correlations across the color channels. As a solution to this problem, we first use PCA to rotate the color space of the original image into a new three-dimensional space in which the pixels across color channels are maximally decorrelated [11]. We then apply the model as described, independently to each of the three PCA component channels. After each iteration of the synthesis, the PCA rotation is undone, the resulting three synthetic color channels are recombined, and the pixel statistics computed from the original color channels are imposed. The color space is then rotated again, and the procedure is repeated.

References

[1] J M Allman and J H Kaas. Representation of the visual field in striate and adjoining cortex of the owl monkey (aotus trivirgatus). *Brain Res*, 35(1):89–106, Dec 1971.

[2] J M Allman and J H Kaas. The organization of the second visual area (v ii) in the owl monkey: a second order transformation of the visual hemifield. *Brain Res*, 76(2):247–65, Aug 1974.

[3] A Burkhalter and D C Van Essen. Processing of color, form and disparity information in visual areas vp and v2 of ventral extrastriate cortex in the macaque monkey. *J Neurosci*, 6(8):2327–51, Aug 1986.

[4] J R Cavanaugh, W Bair, and J A Movshon. Nature and interaction of signals from the receptive field center and surround in macaque v1 neurons. *Journal of Neurophysiology*, 88(5):2530–46, Nov 2002.

[5] R Desimone and S J Schein. Visual properties of neurons in area v4 of the macaque: sensitivity to stimulus form. *Journal of Neurophysiology*, 57(3):835–68, Mar 1987.

[6] S O Dumoulin and B A Wandell. Population receptive field estimates in human visual cortex. *Neuroimage*, 39(2):647–60, Jan 2008.

[7] D C Van Essen, W T Newsome, and J H Maunsell. The visual field representation in striate cortex of the macaque monkey: asymmetries, anisotropies, and individual variability. *Vision Res*, 24(5):429–48, Jan 1984.

[8] R Gattass, C G Gross, and J H Sandell. Visual topography of v2 in the macaque. *J Comp Neurol*, 201(4):519–39, Oct 1981.

[9] R Gattass, A P Sousa, and C G Gross. Visuotopic organization and extent of v3 and v4 of the macaque. *J Neurosci*, 8(6):1831–45, Jun 1988.

[10] R Gattass, A P Sousa, and M G Rosa. Visual topography of v1 in the cebus monkey. *J Comp Neurol*, 259(4):529–48, May 1987.

[11] D Heeger and J Bergen. Pyramid-based texture analysis/synthesis. *Image Processing, 1995. Proceedings., International Conference on*, 3:648 – 651 vol.3, 1995.

[12] W M Maguire and J S Baizer. Visuotopic organization of the prelunate gyrus in rhesus monkey. *J Neurosci*, 4(7):1690–704, Jul 1984.

[13] J Portilla and EP Simoncelli. A parametric texture model based on joint statistics of complex wavelet coefficients. *International Journal of Computer Vision*, 40(1):49–70, 2000.

[14] S Shushruth, J M Ichida, B Levitt, and A Angelucci. Comparison of spatial summation properties of neurons in macaque v1 and v2. *Journal of Neurophysiology*, 102(4):2069–2083, Oct 2009.

[15] EP Simoncelli, WT Freeman, EH Adelson, and DJ Heeger. Shiftable multiscale transforms. *Information Theory, IEEE Transactions on*, 38(2):587 – 607, 1992.

[16] G A Walker, I Ohzawa, and R D Freeman. Suppression outside the classical cortical receptive field. *Vis Neurosci*, 17(3):369–79, Jan 2000.