Time to wave good-bye to phase scrambling: Creating controlled scrambled images using diffeomorphic transformations

Bobby Stojanoski

Brain and Mind Institute, University of Western Ontario, London ON, Canada
Department of Psychology, University of Western Ontario, London ON, Canada

Rhodri Cusack

Brain and Mind Institute, University of Western Ontario, London ON, Canada
Department of Psychology, University of Western Ontario, London ON, Canada
Medical Biophysics, University of Western Ontario, London ON, Canada

To isolate the neural mechanisms associated with recognizing objects from those processing basic visual properties, control stimuli are required that contain the same perceptual properties as the objects but are unrecognizable. We demonstrate that conventional methods for generating control stimuli (phase scrambling, box scrambling, texture scrambling) yield poor controls because they dramatically distort the basic visual properties (e.g., spatial frequency, perceptual organization) to which even the earliest stages of visual processing are sensitive. We developed a new scrambling method, using a diffeomorphic transformation that preserves the basic perceptual properties of the image while removing meaning. We acquired perceptual ratings to determine the least amount of scrambling necessary to remove recognition. We hypothesized that our "diffeomorphic" images would produce neural activity at the earliest stages of the visual system that more closely matched activity in response to intact images relative to the other scrambling methods. To test this hypothesis, we used the HMAX computational model of object recognition and compared the simulated neural activity at the earliest stages of the visual system (layers S1, C1, and S2) between a set of 149 images scrambled using each distortion method to their intact version. We found that scrambled "diffeomorphed" images were indistinguishable to intact images in each layer of the model, but all of the other distortion methods yielded quite different patterns. Our results indicate that "diffeomorphed" images serve as more appropriate control stimuli in neuroimaging studies that aim to disentangle the representations of perceptual and semantic object properties.

Introduction

To create a rich perceptual representation of a real-world object, the visual system must integrate its many constituent features. In a hierarchical fashion, low-level perceptual features (e.g., orientation, spatial frequency) are extracted in early visual cortex (striate and extrastriate cortex; Hubel & Wiesel, 1968) while increasingly complex features (e.g., shapes) are processed in more anterior regions of the ventral visual stream (Felleman & Essen, 1991). At still higher stages of the visual cascade, these representations undergo an additional level of abstraction when an object is recognized (Peelen & Caramazza, 2012) and is assigned to a category (Ishai, Ungerleider, Martin, & Haxby, 2000; Tanaka, 1996).

To isolate a brain region's responsiveness to a particular visual feature, that feature must be manipulated in a way that is not confounded with others. To investigate the neural mechanisms underlying object recognition in anterior visual areas using neuroimaging, it is common to compare the brain's response in these regions to stimuli that can be recognized with control

Citation: Stojanoski, B., & Cusack, R. (2014). Time to wave good-bye to phase scrambling: Creating controlled scrambled images using diffeomorphic transformations. *Journal of Vision, 14*(12):6, 1–16, http://www.journalofvision.org/content/14/12/6, doi:10.1167/14.12.6.

doi: 10.1167/14.12.6 Received October 9, 2013; published October 9, 2014 ISSN 1534-7362

stimuli that cannot (Epstein & Kanwisher, 1998; Grill-Spector et al., 1998; Kanwisher, McDermott, & Chun, 1997; Kourtzi & Kanwisher, 2000; Lerner, Hendler, Ben-Bashat, Harel, & Malach, 2001). When using this experimental approach, stimuli should be designed carefully to manipulate how easily they are recognized while preserving their basic visual properties. This is pertinent when examining object recognition and semantic processing along the visual hierarchy up to the inferior temporal cortex. Traditionally, two methods have been used to create control stimuli: phase scrambling and box scrambling, but we believe these are inadequate for most applications. Our concern is that in the process of making images unrecognizable their basic visual features are changed, either by introducing artifacts (box scrambling; Vogels, 1999) or by removing crucial information (phase scrambling; Oppenheim & Lim, 1981; Thomson, 1999) to which even the early visual system is sensitive. When basic visual features are confounded with recognizability, it is not clear whether a difference in neural activity in anterior regions between control stimuli and their intact counterparts is a result of the manipulation of meaning or whether it is merely a result of different information being fed forward from early visual regions.

The aim of the current study is to (a) introduce novel control stimuli that we call "diffeomorphed" images that have a number of elegant properties and (b) investigate whether "diffeomorphed" images are superior at preserving fundamental features of the intact image while sufficiently distorting it to the point at which it is no longer recognizable. To investigate which of the scrambling methods introduced differences in basic visual features, we simulated neural activity at three stages of visual processing (S1, C1, C2) using the HMAX model (Riesenhuber & Poggio, 2002). These stages of the HMAX model are designed to simulate the processes that precede object recognition, but crucially, they do not have any stored representations that would give them the capacity for object recognition. Any differences between intact and scrambled stimuli at these stages of processing, therefore, must reflect poor control of basic visual features. The HMAX response at each level of processing evoked by intact images was compared to that evoked by images generated using our new scrambling method and images created using three methods: box scrambling, phase scrambling, and a method called texture scrambling designed by Portilla and Simoncelli (2000), which has been implemented in many studies on perceptual processing (Balas, Nakano, & Rosenholtz, 2009; Greene & Oliva, 2009; Rousselet, Pernet, Bennett, & Sekuler, 2008).

To preview our results, existing scrambling methods provide poor control over basic visual features, and diffeomorphed images evoked simulated visual activity that closely mirrored that evoked by intact objects and serve as better controls for any neuroimaging investigation in which object recognizability is manipulated.

Methods

Stimuli

A set of 149 images was drawn from the Hemera image database (Hemera Images: http://www.hemera.com/). Images were in color, 500 × 500 pixels in size, and positioned centrally on a white background (see Figure 1a for sample images). The images were divided into 13 categories: sporting equipment, shoes, electronics, kitchen supplies, office supplies, instruments, tools, clothing, faces, mammals, birds, bikes, and fruit. The categories could also be divided into subsets thought to be important in human semantic representation, such as living versus nonliving (Costanzo et al., 2013) or mobile versus stationary (Kriegeskorte et al., 2008). We compared our "diffeomorphed" images to three other image scrambling methods prevalent in the perception literature.

Diffeomorphic transformation

Diffeomorphic transformations are smooth, continuous, and invertible: Imagine printing the image on a rubber sheet and then distorting it without tearing. There is a one-to-one mapping between the source and target spaces (i.e., no duplication or removal of parts). As the transformation is continuous across space, if there were N islands of contiguous nonbackground pixels in the intact image, there would be N islands in the scrambled image. This is an important control, as figure–ground grouping through spatial proximity is known to be perceptually salient (Koffka, 1935), and there are brain regions that respond with the number of objects (Cusack, 2005; Cusack, Mitchell, & Duncan, 2009).

Our diffeomorphic transformation was created by repeatedly applying a flow field generated from a set of two-dimensional cosine components with random phase and amplitude. Each pixel had an equal chance of being expanded or contracted (i.e., mean Jacobian = 1), so on average, the diffeomorphic scrambled objects had the same number of nonbackground pixels as the intact objects. One iteration of the flow field was implemented through the following transformation:

$$I_{n+1}(x,y) = I_n\Big(f_0(x,y), f_1(x,y)\Big) \quad (1)$$

where I_n is the image on iteration n, x and y are the pixel coordinates, and n is the iteration number.

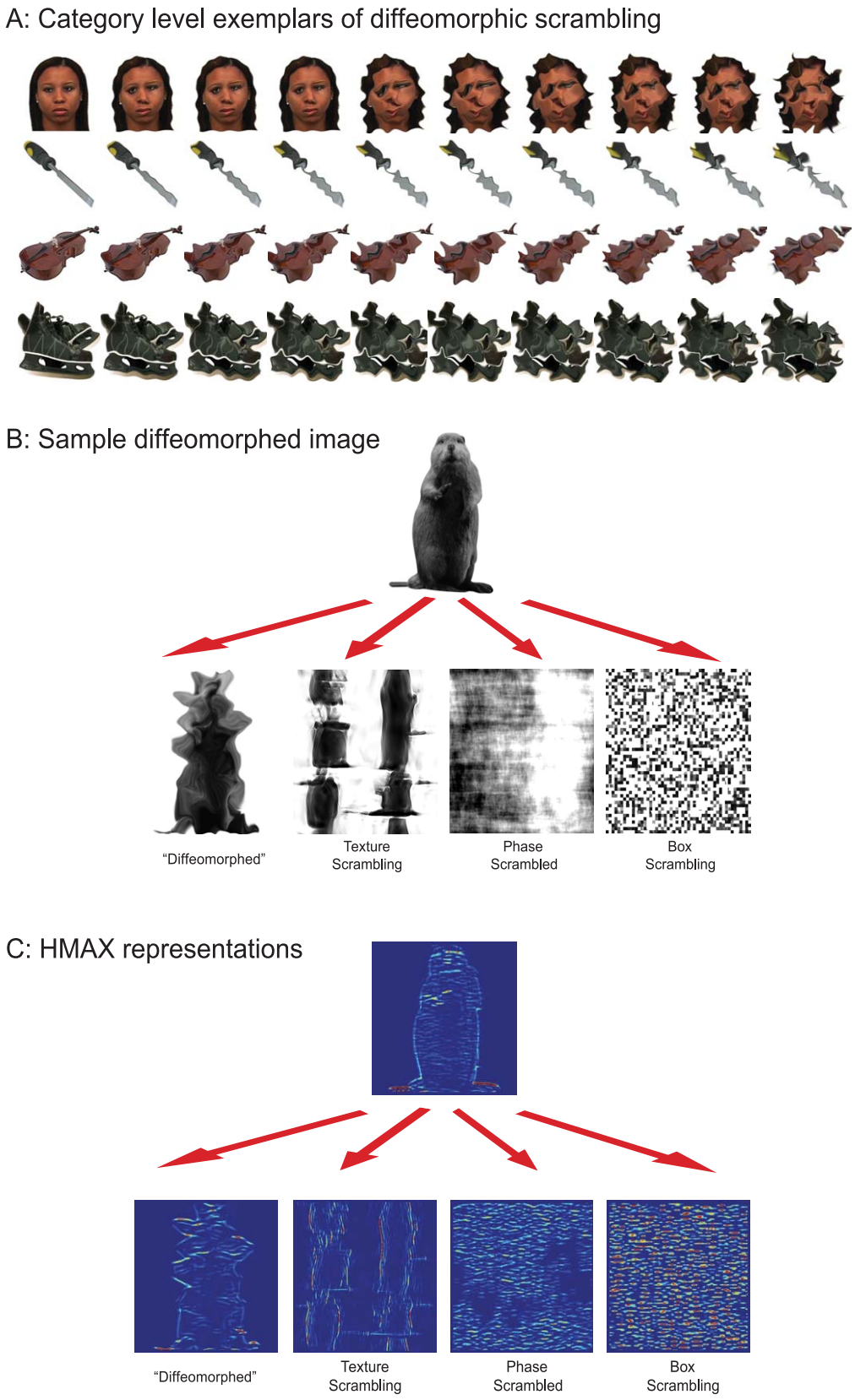

Figure 1. (A) Examples of diffeomorphed images from four object categories at different stages of scrambling. (B) A sample object (intact; top row) with the same object after it was scrambled using the four scrambling methods of interest ("diffeomorphing," texture, phase, and box). (C) The same object (and all scrambled versions) represented at a single scale in layer S1 of the HMAX model.

The flow fields were defined by

$$f_i(x,y) = \sum_{k=1}^{6}\sum_{l=1}^{6} a_i(k,l)\cos\left(\frac{2\pi kx}{N} + p_i(k,l)\right) \quad (2)$$

where $a_i(k,l)$ and $p_i(k,l)$ are the amplitude and phase of the cosine components, each selected from a random uniform distribution.

There were 20 steps. To reduce blurring through interpolation errors, the images were upsampled by a factor of two prior to warping. Linear interpolation was used, and the same warp fields were applied to each color plane (R, G, B). To remove the potential for artificial differences in the spatial frequency between diffeomorphic and intact images due to residual blurring, after the transformation, we matched the spatial power spectrum of the intact images to the diffeomorphic images (sample image in Figure 1a). In the current study, we created diffeomorphed images of isolated real-world objects; however, this scrambling method can be applied to any stimulus. The Matlab script for generating diffeomorphed images can be downloaded at http://www.cusacklab.org/?page_id=222.

Rating experiment: Identifying the amount of diffeomorphic warping necessary to remove meaning

The diffeomorphic transformation allows for straightforward manipulation of the amount images are scrambled. To identify the minimum amount of scrambling necessary for our diffeomorphed images to become unrecognizable, we acquired perceptual ratings for each image. We incrementally scrambled objects into 20 distinct levels (level one contained the least amount of scrambling and level 20 the most), creating 20 images for each object, corresponding to every level of scrambling. The flow fields f were normalized so that they had a root mean square of unity and then multiplied by a scaling factor chosen in pilot viewings (3.75) of the upsampled pixels between adjacent images.

Using Amazon's Mechanical Turk (a crowdsourcing resource), we asked "workers" to provide perceptual ratings for each object. Ethical approval was obtained from the Western Health Sciences Research Ethics Board. A total of 415 "workers" completed 15,600 trials and were paid approximately $2 per hour. Twenty sets of the stimuli were created so that each object was presented once at a single warping level within each set, but across all sets, every warping level was tested for every object. Each worker was permitted to respond to all objects, but they were limited to rating only one of the 20 sets that were assigned randomly to preclude biased ratings due to prior exposure to a different level of scrambling for the same object. Their task was to indicate how easily they could identify the object in the image by selecting one of the following perceptibility ratings: 1 (no, not at all), 2 (I'm not sure, but I can take a guess), 3 (I can see it fairly well), and 4 (I know exactly what it is). We also asked the workers to make a forced-choice report on the identity of an object. We concatenated all responses from each worker and computed an average perceptual rating for each category. We then selected the warping level that corresponded with a perceptibility rating of 1.5 separately for each category and applied that level of scrambling to all images within that category. These images, along with images created using the other forms of scrambling, were used in the HMAX simulations to compare with intact images.

Box scrambling

The first method we compared, box scrambling, is commonly used probably due to its simplicity. Like the children's puzzle, the image is divided by an invisible grid and the squares rearranged. To create sufficiently scrambled images, we divided the intact image into 1,250 independent "boxes" and randomly repositioned each box within the confines of the images (Figure 1b). Varying the number of boxes changes the resolution of the scrambling, which interacts with how easily the scrambled image is discerned. For instance, reducing the number of sections produces larger boxes that may leave enough information per box to help identify the object.

Phase scrambling

The second method we tested was phase scrambling. Images were scrambled by computing a two-dimensional fast Fourier transform (FFT), yielding a complex (magnitude and phase) representation. The phase values were then randomized by assigning a random value to each element taken from a uniform distribution across the range $(-\pi, \pi)$; an inverse FFT was applied to the resulting magnitude/phase maps to produce a scrambled version (Figure 1b). Variations to this method have been applied in previous studies (Koenig-Robert & Vanrullen, 2013; Malach et al., 1995) with new variations developed more recently, such as Dakin, Hess, Ledgeway, and Achtman's (2002) weighted mean phase algorithm (Ales, Farzin, Rossion, & Norcia, 2012). The underlying intention across all variations is that each scrambled image contains the same frequency spectrum as the intact image. Although this is true when calculated in the pixel space of the

image, it ceases to become true once the image enters the filters of the eye and the visual system, which are sensitive to the visual features created by smooth changes in phase with frequency that are typical in natural images and lost during phase scrambling. This will be quantified in the S1 simulation of the earliest cortical processing in the visual system.

Texture scrambling

We also tested a newer scrambling method developed by Portilla and Simoncelli (2000) using a texture synthesis algorithm. This approach extracts over 700 parameters per image using linear filters at a range of orientations and scales to resemble orientation and spatial frequency tuning of simple striate cortex cells (Figure 1b). Included in this set of parameters are pixel statistics (variance, skewness and kurtosis, and maximum and minimum intensity values) in an attempt to ensure texture-scrambled images maintain equal luminance. The set of parameters also includes correlations across filter pairs that have the same orientation and spatial frequency but different phase, giving rise to periodic structure, such as contours. A synthesis process, which starts from a sample of Gaussian white noise, is then conducted and is iteratively modified so that its statistics converge on the parameters extracted from the intact image.

Neural simulations: HMAX model of object perception

The HMAX standard model of object perception was designed to simulate neural activity at distinct stages along the visual hierarchy based on biologically feasible principles (Riesenhuber & Poggio, 2002). For a more detailed description of their model along with the Matlab code, refer to http://maxlab.neuro.georgetown.edu/hmax/#standard. Briefly, in the model, perception of complex objects is driven solely by feed-forward processes, whereby input to each stage comes from the maximized neural output at the immediately preceding stage. We ran 149 intact images, along with diffeomorphic transformed, phase scrambled, box scrambled, and texture scrambled versions of each image, through the HMAX model and used the simulated output at layers S1, C1, and C2 to determine how similar each of the scrambled images were in their basic visual features relative to the intact versions (see Figure 1c for a sample image from each scrambling method from a single scale in layer S1).

Layer S1 is an approximation of classic simple cells found in V1 that take the form of Gabor functions. To match the tuning properties of V1 neurons, units in layer S1 were set to be most responsive to orientation, spatial location, and spatial frequency. The receptive field of each of the cells was simulated by Gabor filters in a pyramidal structure ranging from seven to 37 pixels in steps of two pixels at four orientations (0°, 45°, 90°, and 135°), producing 64 different S1 receptive field types (16 scales × four orientations). The next layer, C1, corresponds to complex cortical cells in the striate cortex. Units in this layer are generated using a MAX pooling function, selecting the strongest outputs of neighboring units in layer S1 that have similar orientation preference, spatial frequency tuning and receptive field locations (eight scales × four orientations). As a consequence, these units have larger receptive fields, which show more tolerance to size, shape, and spatial position. The last stage in the model, C2, is computed by taking the global maximum over all scales and positions from the S2 layer; S2 units are inputs from C1 units with similar tuning properties that are pooled over local neighborhoods that respond most strongly to specific prototype images. This results in an S1 layer that is selective to more complex features, reflecting neurons in the V2 or V4 extrastriate cortex. The pooled inputs, using the MAX operation, drive the C2 layer (256 units) that is invariant to size (pooling all filter sizes) and position (pooling over scales), reflecting the properties of a neuron in visual area V4 or posterior inferior temporal (pIT) cortex (for a detailed overview, refer to Serre, Wolf, Bileschi, Riesenhuber, & Poggio, 2007). Outputs at each of the three layers of interest (S1, C1, and C2) were concatenated across all scales and orientations to get an overall representation for each image and scrambling method.

Evaluation of HMAX results

To see how well matched the image sets were in their basic visual features, we examined the simulated neural activity produced by the HMAX model at each layer (S1, C1, and C2) in response to the intact and four sets of scrambled images. We inspected the mean activity across all simulated neurons produced by each object and the distribution (reduced bins at layers S1 and C1 to smooth spikes in activity solely due to misaligned patterns of neural activity across the different scales) of mean activity across objects. Within each layer, we conducted paired t tests to see if there was a difference between each possible pair of image sets (intact vs. diffeomorphed and so on).

Even if the mean activity does not discriminate between image sets, perhaps the pattern or distribution of activity within a layer does? To investigate this, we trained a linear discriminant classifier (using Matlab's classify function) to discriminate between intact and scrambled objects from either the vector of activations

or the histogram of activity across all simulated neurons. We used a leave-one-object-out cross-validation approach. That is, for each comparison, the output from all images but one were combined to form the training set, and the output from the remaining image was used as the test set (ensuring the training and test sets are independent). The classifier used the activity patterns in the training set to predict to which of the two classes (intact or scrambled) in the training set the activity in the test set belonged. The total number of correct classifications across train/test folds was calculated for each of the scrambling methods, and the inverse binomial distribution function was used to test whether prediction was above chance (50%).

The output from layer S1 was intractably large, and so we randomly sampled a subset of simulated neurons from layer S1 to match the size to the output from layer C1 before running the classifications. To ensure the results were not an artifact of a particular random sample, we repeated the group level and classification analyses for 10 iterations. It was highly stable across random samples, and we report the averaged statistical output across all iterations. To ensure estimability of the covariance matrix for the linear discriminant analysis, prior to classification of layers S1 and C1, we conducted dimension reduction using principal components analysis, selecting sufficient top components to explain 99% of the variance.

Potential of modified phase- and box-scrambling procedures

The procedure described thus far compared diffeomorphic warping, calibrated to be the minimal amount to impact recognizability, with the conventional phase- and box-scrambling procedures in which there is no calibration and substantial warping to ensure recognizability is impacted. In a final analysis, we examined whether, as an alternative to the diffeomorphic procedure, there is the potential for the conventional methods to be used in a modified way by reducing the amount of warping (thus reducing the visual distortions they introduce). To do this, we parametrically varied the degree of warping and examined its effect on simulated neural output and recognizability. We hypothesized that for diffeomorphic scrambling the magnitude of neural activity would be relatively invariant to the degree of warping whereas introducing even small amounts of phase and box scrambling would produce large differences in neural activity even before any impact on recognizability.

We randomly selected an image from each category (13 in total) and produced 20 images at different levels of phase, box, and diffeomorphic scrambling (no method was available to parametrically vary texture scrambling). Images were incrementally and increasingly distorted starting from level 1 (no scrambling and easily identifiable) to level 20 (maximum scrambling and, in our judgment, unrecognizable). For box scrambling, we manipulated the degree of warping by varying the number of swapped "boxes" from 0 (intact) to 1,250 (fully scrambled). We modified phase scrambling by varying the proportion of the phase values that were randomized (0% is an intact image, and 100% is fully phase scrambled). To maximize sensitivity to changes in neural signal with diffeomorphic warping, we used an even wider range of distortion than in the previous analyses (4× larger). Neural changes were quantified in the highest layer (C2) of the HMAX model, the closest putative input to recognition processes. To quantify change, we computed the mean absolute percentage deviation in simulated neural activity at each of the 20 levels of scrambling (across image categories) from the neural output in response relative to the intact versions.

Results

Perceptual ratings

We found that recognition ratings of scrambled objects differed depending on the category to which they belonged (Figure 2a). That is, categories such as *faces* and *bikes* were most impervious to scrambling and required a relatively higher degree of scrambling before objects were deemed unrecognizable. Conversely, *instruments* and *kitchen supplies* were most susceptible to scrambling. To estimate the warping level required to get a clarity rating of 1.5, we interpolated using nonlinear least-squares regression (Matlab's nlinfit function). When a category had no warping level with a clarity rating below 1.5, we extrapolated using this fit. In Figure 2b, we plot the resulting levels of scrambling for each object across category used in the subsequent modeling. Accuracy of participants' semantic responses in the forced-choice task with a clarity rating of 1.5 was estimated to be 25%, computed by interpolating between a rating of one (2.27%) and a rating of two (47.76%). This shows a strong (although not total) manipulation in the degree of meaning. Furthermore, it is likely that some of this residual recognition is a result of a cognitive level of deduction, exploiting cues, such as color, that may not reflect object recognition processes in the ventral visual stream.

Mean activity across neurons within each layer

Output from all three layers of the HMAX model indicated substantial visual differences between the intact images and those from previous scrambling

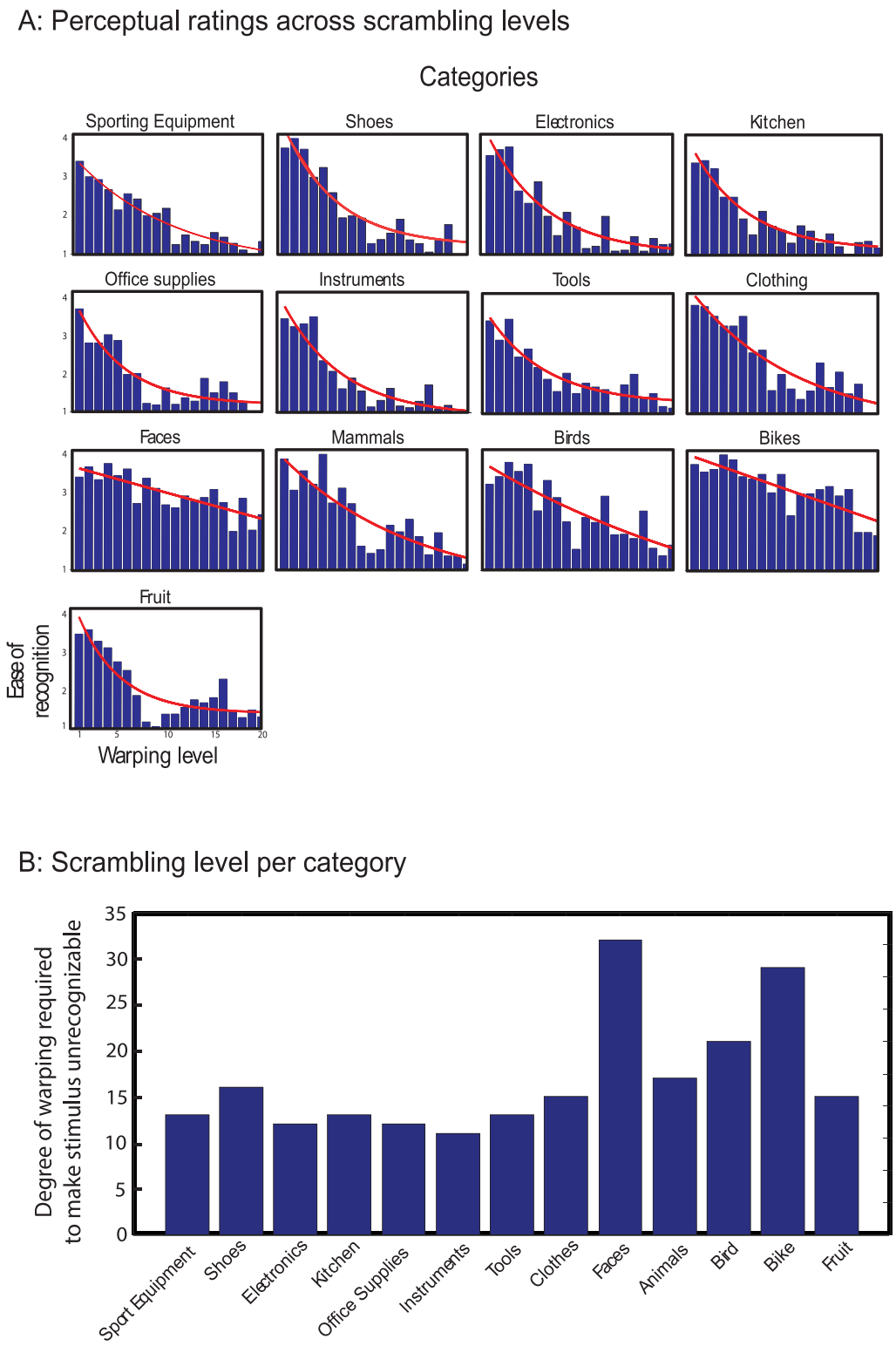

Figure 2. (A) Averaged perceptual ratings ($n = 415$ participants) collapsed across objects in each category. Participants indicated, on a scale from 1 to 4, how easily they could recognize objects presented at one of 20 diffeomorphic scrambling levels, shown on the x-axis. (B) For each category, the amount of diffeomorphic scrambling necessary to obtain a rating of 0.5.

methods. In contrast, the diffeomorphic images closely resembled intact images in their visual features (Figure 3). In layer S1, the results of paired t tests showed no significant difference in mean neural activity between intact and diffeomorphed images, $t(1,296) = 0.69$, $p = 0.49$. However, relative to the other scrambling methods (phase, box, and texture scrambling), diffeomorphed images produced significantly less activity,

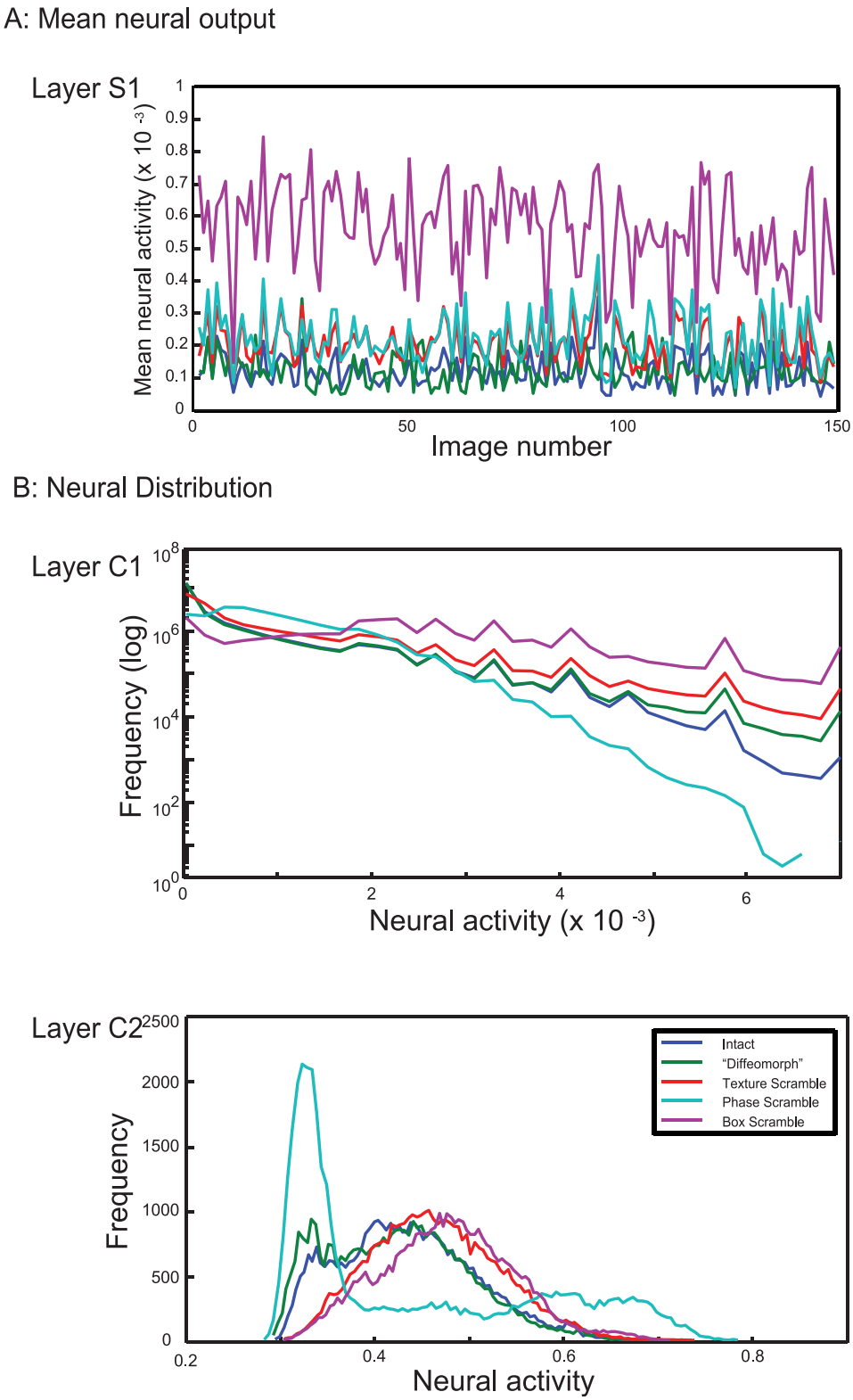

Figure 3. (A) Mean simulated neural activity at layer S1 of the HMAX model for all 149 objects presented as intact and scrambled using each of the scrambling methods. (B) Distribution of neural activity averaged across all of the 149 objects and simulated neurons from layer C1 and (C) layer C2.

$t(1,296) > 11.97$, $p < 0.0001$ (Figures 3 and 4, top panels). Further analysis revealed that the mean activity associated with texture scrambling was significantly different than that produced by phase scrambling, $t(1,296) = 2.18$, $p = 0.03$; however, this difference did not reach significance after applying Bonferroni correction, and box scrambling produced activity that was most different from all other image scrambling methods, $t(1,296) > 26.69$, $p < 0.0001$.

We found a similar pattern of results at layer C1 with equivalent levels of simulated neural activity for both diffeomorphed and intact images, $t(1,296) = 0.73$, $p = 0.46$, which were each significantly different from all other scrambling methods, $t(1,296) > 13.15$, $p < 0.0001$ (Figures 3 and 4, middle panels). At this level, the neural activity in response to texture scrambling and phase scrambling was not significantly different, $t(1,296) = 1.69$, $p = 0.093$, and box scrambling produced neural activity that was significantly higher relative to all other scrambling methods, $t(1,296) > 28.23$, $p < 0.0001$.

The results were slightly different at layer C2 (thought to be analogous to V4 and pIT). Although diffeomorphic and intact images produced similar levels of simulated neural activity, $t(1,296) = -1.38$, $p = 0.17$, so did three additional comparisons: phase scrambling and intact images, $t(1,296) = -0.502$, $p = 0.616$; phase scrambling and diffeomorphic images, $t(1,296) = 1.33$; $p = 0.18$; and texture scrambling and phase scrambling, after a Bonferroni correction, $t(1,296) = -1.98$, $p = 0.049$. All other comparisons with the intact images were significantly different, $t(1,296) > 4.28$, $p < 0.00003$. We also found neural activity for texture-scrambled and box-scrambled images were not significantly different from each other, $t(1,296) = 1.78$, $p = 0.08$, but were significantly different from the other scrambling categories, $t(1,296) > 2.98$, $p < 0.0032$ (Figures 3 and 4, bottom panels).

Pattern and distribution of activity within layers

We also tested whether the pattern or the distribution of activity within a layer was indicative of which set an image came from. Could a classifier be trained to use the neural outputs of the HMAX model (raw neural activity or the distribution of that activity) to identify whether a novel image was intact or scrambled? Classification accuracies are reported in Figure 5. At layer S1 (Figure 5, top row), classification based on the distribution of neural output across 10 permutations was at chance (determined by inverse binomial distribution functions, $p < 0.05$) when discriminating between diffeomorphed and intact images (mean: 48.42%; *SD*: 0.32%); all other comparisons between intact images and images produced by the different scrambling methods were significantly above chance (mean: >91.85%). For the raw pattern of activity, at layers S1, classification accuracy between intact and diffeomorphed images was at chance (mean: 50.03%; *SD*: 4.65%) as was classification accuracy between texture-scrambled and phase-scrambled images (mean: 49.93; *SD*: 2.69%). All other pairwise classifications were significantly above chance (mean: >62.65%). At layer C1 (Figure 5, second row), classification based on the raw neural output (50.67%) and distribution of activity (52.34%) was at chance when discriminating between diffeomorphed and intact images. The classifier significantly differentiated intact images from all other categories with a mean performance greater than 85.57%. In addition, the neural activity at layer C1 between texture scrambled and box scrambled was indistinguishable (47.65%).

At layer C2, we found, using the distribution of neural activity as the feature of interest, the classifier performed at chance when distinguishing between intact and diffeomorphed images (51%). However, the classifier also performed at chance when discriminating texture-scrambled and box-scrambled images (50.34%). The classifier significantly discriminated images for all other pairs of scrambling methods with an accuracy level of at least 54.7%. A slightly different pattern of results emerged for the distribution of neural activity. The classifier could successfully distinguish intact from diffeomorphed images (60.69%), phase-scrambled images (61.15%), and box-scrambled images (68.5%) but performed at chance when discriminating intact and texture scrambled (54.06%). Accuracies for all comparisons at each of the three layers are presented in Figure 5 (third row).

Potential of modified phase- and box-scrambling procedures

We calculated the neural output at layer C2 across a range of 20 parametrically varied levels of diffeomorphed, phase, and box scrambling to relate neural activity with image recognizability. To do this, the percentage deviation in neural activity relative to the intact version of was averaged across 13 images (one from each category) at each scrambling level. We found that the percentage deviation in neural activity remains relatively constant (ranging from 4.05% to 7.67%) whereas even very little box and phase scrambling produces drastic changes in neural activity, ranging from 8.55% to 28.31% for box scrambling and 7.29% to 39.34% for phase scrambling (Figure 6) but has little impact on recognizability.

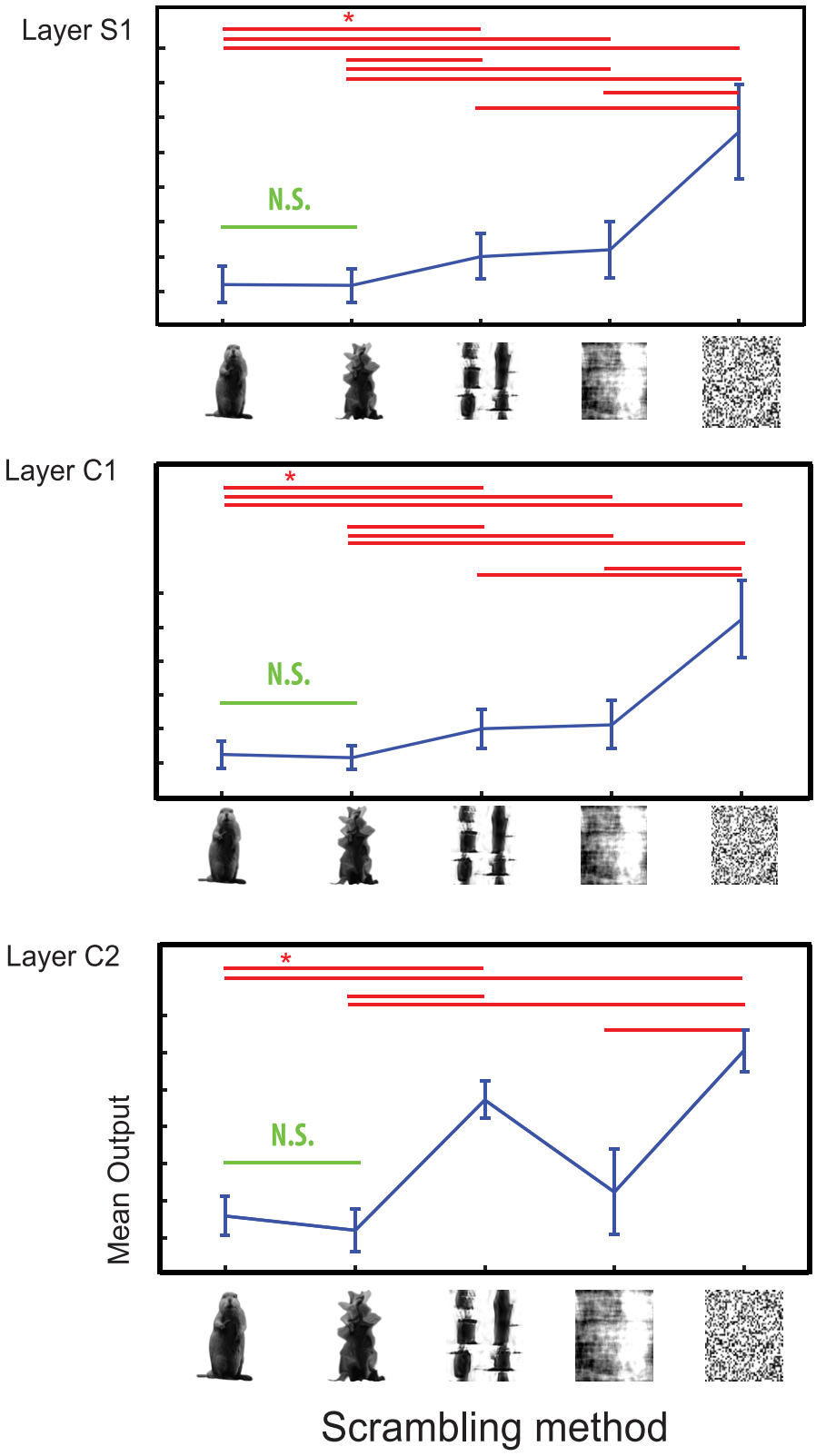

Figure 4. To determine whether "diffeomorphed" images are more visually similar to intact images relative to the other scrambling methods, we compared the simulated visual neural response they evoke. Paired *t* tests revealed that across each layer (S1, C1, and C2) of the HMAX model the neural activity in response to "diffeomorphed" images did not differ from the neural activity in response to intact images.

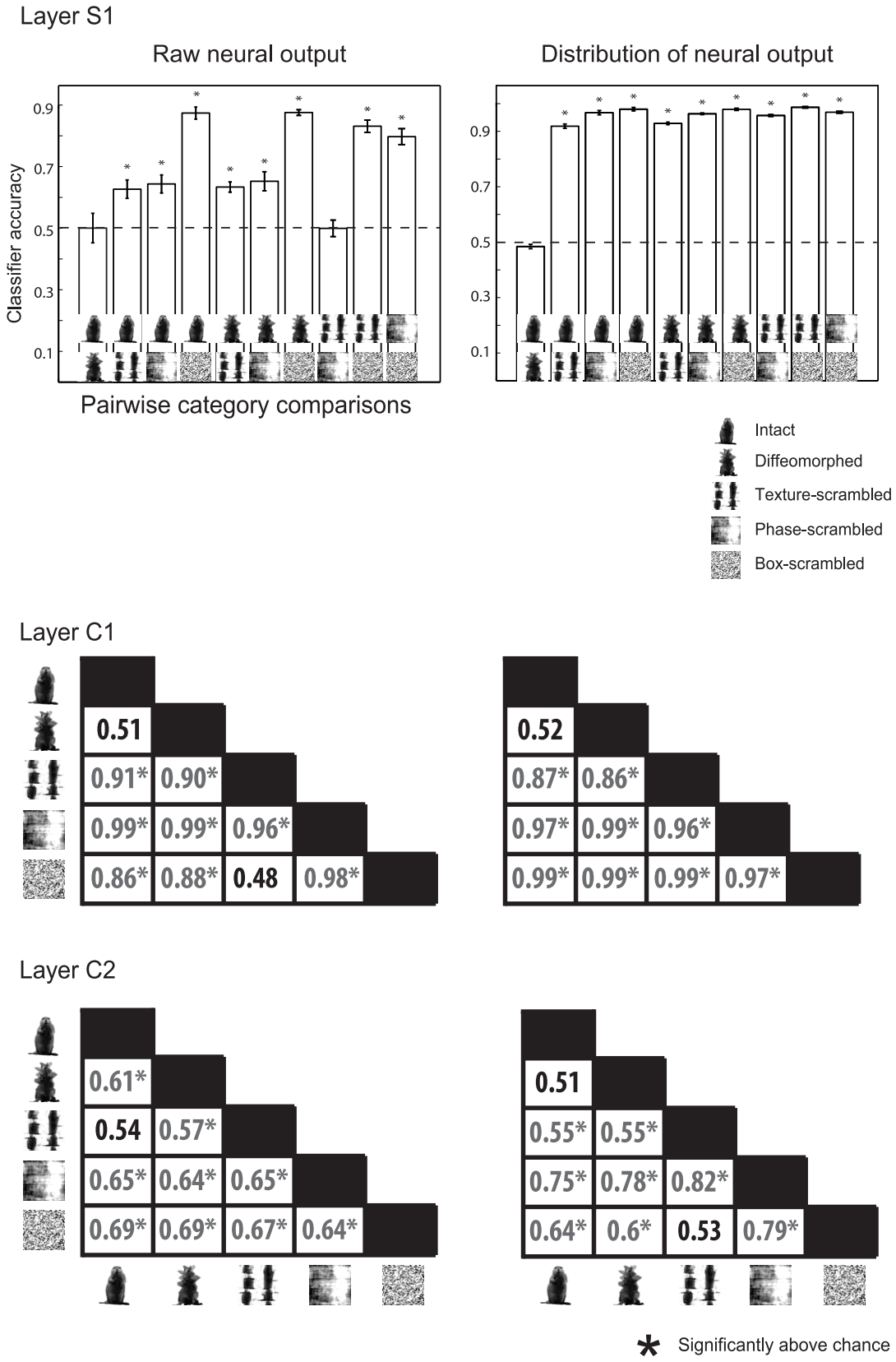

Figure 5. A linear discriminant classifier was used to discriminate between intact and scrambled objects using both the pattern of neural activity across neurons (left column) and the distribution of this activity (right column) at layers S1 (top row), C1 (second row), and C2 (third row) of the HMAX model. For layer S1, the bar graph represents mean classifier accuracy (with *SEM*) across 10 permutations. For layers C1 and C2, each cell contains the classifier accuracy. Each comparison across the three layers was marked where the classifier performed significantly above chance, determined by inverse binomial distribution functions; $p < 0.05$.

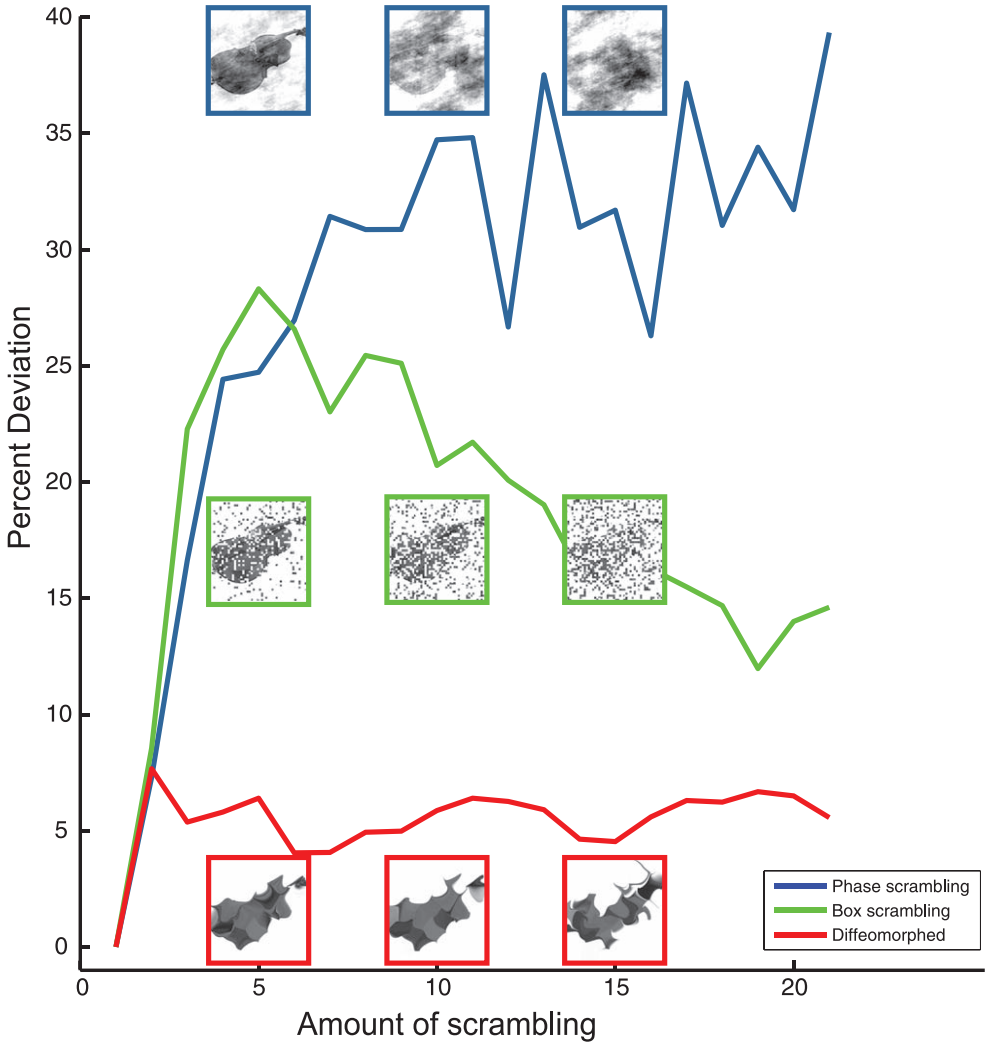

Figure 6. The absolute percentage deviation in neural activity relative to the intact images was computed for 20 parametrically spaced levels of phase, box, and diffeomorphic scrambling. A representative image is presented at three arbitrary levels of scrambling (5, 10, 15) to highlight that at very little phase and box scrambling (easily identifiable) there was a drastic difference in neural activity with little effect on recognizability whereas for diffeomorphic scrambling the neural output remained similar even for extreme levels of distortion.

Discussion

Using the HMAX model (Riesenhuber & Poggio, 2002), we showed that the earliest stages of the visual system do not respond to control stimuli generated by phase, box, or texture scrambling in the same way as they do to intact images. We found differences for most measures within each layer of the model (S1, C1, C2); the magnitude of the differences was substantial with average neural activity around two to seven times higher for conventionally scrambled stimuli relative to intact stimuli. This indicates that these image sets differ in their basic visual properties and will obstruct our ability to isolate object recognition processes. We conclude conventional scrambling methods make poor controls in experiments that intend to manipulate semantic content.

In contrast, at each layer of the HMAX model, we found that the mean neural signal was the same for the intact and diffeomorphed image sets. Furthermore, a linear discriminant classifier was generally unable to differentiate between intact and diffeomorphed images based on either their pattern or distribution of neural activity in each layer. Moreover, our results are not restricted to the specific levels of scrambling assigned to each scrambling method. Across 20 equally spaced levels of distortion, we found the mean percentage deviation in neural activity relative to intact images in layer C2 remained relatively constant for diffeomorphed images even when we increased the amount of scrambling by 400%. In contrast, the percentage deviation rose sharply for clearly recognizable phase-

and box-scrambled images with very little distortion. These results indicate two things: (a) There was little in the visual content of diffeomorphed images that differentiates them from intact images and (b) the basic visual properties are better preserved in diffeomorphed images than even slight amounts of phase and box scrambling in which the content of the image is easily identifiable.

A number of factors contribute to the differences in neural processing for stimuli generated using the three conventional scrambling methods. During box scrambling, images are divided into an equal number of segments (sometimes as small as individual pixels) with each box randomly shuffled to a new location in the image. It was assumed that because they were created with unaltered segments of the original image, they would be visually matched. Unfortunately, repositioning pixels to different locations in the image produces artificial edges at the borders between discontinuous segments. The result is images that no longer retain their original Gestalt and, even more concerning, contain spatial frequency artifacts contingent on the scrambling resolution. These changes produce different patterns of neural activity (Vogels, 1999) relative to the intact version of the image. Even with attempts to ameliorate the effects of edges by using spatial vignetting (convolving the edges with a linear ramp of 25-pixel width), Rainer, Augath, Trinath, and Logothetis (2002) found that area V1 showed a linear relationship between activity and the amount of box scrambling; the more scrambled an image, the higher the activity (up to the second highest level of scrambling when activity dropped precipitously). In extrastriate cortex (V1, V2, V3A, V4), activity for highly box-scrambled and intact images were similar until the highest level of scrambling at which activity dropped, much like in V1. These findings are consistent with Singh, Smith, and Greenlee (2000), who found that blood oxygenation level dependency increased in response to grating that increased from low to medium spatial frequencies.

The phase-scrambling method controls better for the spatial frequency content of the image. Intact images are decomposed into their constituent spatial frequencies using a Fourier transform. The phase values are then randomized, and the emerging scrambled versions are reconstructed using an inverse Fourier transform with the scrambled images containing the same power spectrum as the corresponding intact versions. However, the visual system (or the HMAX model) is sensitive to image features that result from the smoothly changing, continuous phase variations with frequency that is typical in natural images (Oppenheim & Lim, 1981; Thomson, 1999). Artifacts therefore result from changes to image properties as a result of randomized phase spectra, an inherent byproduct of this method. Thomson (1999) has demonstrated that intact images contain higher-order statistical properties that are absent in the phase-scrambled images primarily driven by distortions in the local phase coherence. In fact, phase spectra have been shown to contain perceptually more important information than the power spectra (Oppenheim & Lim, 1981; Thomson, 1999). That is, local phase coherence is responsible for vital information, such as localized features, including lines, edges, and contours (Morrone & Burr, 1988). The loss of these structural properties and the importance of phase coherence can have significant perceptual consequences; the visual system is sensitive to harmonic phase relationships even at the earliest processing stages, such as V1 (Wang & Simoncelli, 2004), which is reflected in an increase in perceptual sensitivity to detecting distortions in phase-scrambled images (Bex, 2010; Kingdom, Field, & Olmos, 2007).

There have been attempts to improve the phase-scrambling method and remove some of its limitations. For instance, the approach proposed by Dakin et al. (2002) improves second- (contrast) and fourth-order (kurtosis) statistics and avoids overrepresentation of certain phases but leads to nonuniform phase angle steps (Ales et al., 2012). Ales et al. (2012) modified it to produce images with coherent phase but at the cost of randomized amplitude. Future iterations of this method might lead to further improvements, but we believe that spatial domain techniques such as diffeomorphic warping are more readily suited to scrambling without altering distinct visual features.

Of the three test scrambling methods, it was the neural activity associated with texture-scrambled images that most closely resembled that of intact images. However, even at the earliest stages (S1, C1), differences emerged that became more pronounced at the latest stage in the model (C2). Texture scrambling was originally designed to synthesize visual textures with homogeneous and consistently repeating elements that are ideal for modeling higher-order statistics (Portilla & Simoncelli, 2000) and was only later applied to natural scenes. This scrambling method was never intended for isolated objects, which contain different properties from textures, and scenes that may partially explain the subtle differences in neural output. More critically, however, like the other methods, texture scrambling distorts the Gestalt of the image while creating irregularly shaped closed contours. The large differences at C2 may occur because of the grouping discrepancies between the intact and texture-scrambled images. Distorting the grouping properties of objects also produces perceptual consequences at early processing stages. Given the sensitivity of visual area V1 to spatial frequencies and varying sizes of the image set, the spatial frequency content of larger objects (lower spatial frequency) will be differentially altered relative

to smaller images (higher spatial frequency) by long continuous contours (Rust & Dicarlo, 2010).

The diffeomorphic transformation did not change basic visual properties like the other scrambling methods. Diffeomorphic transformations are smooth, continuous, and invertible, so the topology (with no folding) was preserved, and the process could be reversed to re-recreate the original intact image. Moreover, the range of spatial frequencies was restricted using a discrete cosine basis, ensuring that high spatial frequency artifacts were not introduced in the scrambled image. The final result is that the early visual system (as modeled by Riesenhuber & Poggio, 2002) processed diffeomorphed images in much the same way as intact images.

This method provides a first step toward overcoming the limitations inherent to the conventional scrambling methods. Although diffeomorphic images are a significant improvement in the design of appropriate control stimuli, generating fully controlled stimuli is limited by our knowledge of the visual system. Preserving all basic visual features of each image would require a complete understanding of the tuning properties of all neurons along the visual pathway in addition to how they are influenced by connections to other neurons and behavioral objectives (e.g., top-down effects). To date, this has proven to be a considerable challenge. Attempts to find critical features driving neural activity at each processing stage often results in the need to generate idiosyncratic stimuli tailored for each neuron that make comparisons along the visual hierarchy untenable (Kobatake & Tanaka, 1994). In fact, scrambled images can be used to directly examine the properties of the visual system (Murray, 2011). For instance, Freeman, Ziemba, Heeger, Simoncelli, and Movshon (2013) generated a model for creating synthesized images that served as visual metamers (perceptually indistinguishable from the intact version) to outline the receptive field sizes and sensitivity at different eccentricities of visual area V2, which predicts visual degradation in the periphery associated with crowding.

It should be noted that conventional image-scrambling methods may be useful in contexts that focus on specific object properties, such as demarcating cortical regions sensitive to the presence of edges (Kovesi, 2003). Or, for example, one may be interested in brain regions that process shape (e.g., lateral occipital complex), in which case it might be helpful to contrast stimuli with a defined shape with those that do not have one. Diffeomorphic stimuli have been designed for a particular scientific question in which visual properties are not the focus of interest.

A complementary approach to examine the representations along the visual pathway is to keep the stimuli constant but change the task requirements. As outlined by Schyns, Gosselin, and Smith (2009), diagnostic features of images and reverse correlation can be used to link brain activity to functional cognitive states. We believe using diffeomorphed images might complement this approach well; the gradual warping of diffeomorphed images allows for the selection of certain diagnostic features (e.g., shape or semantics), which can be used to differentiate confounded neural activity due to correlated visual properties of images within categories (Rousselet, Pernet, Caldara, & Schyns, 2011).

The category-dependent perceptual ratings offer another instance of how diffeomorphed images can help further our understanding of object perception. Some categories were more affected by scrambling than others, suggesting that the visual system relies on differing sets of features to categorize objects. Banno and Saiki (2011) found that humans use higher-order statistics when detecting animals in scenes, suggesting certain higher-order image properties may be more telling of object structure in some categories over others. Statistically regular features are also important for recognizing objects (Gerhard, Wichmann, & Bethge, 2013). Faces and bikes contain highly regular properties that occur in almost every exemplar (i.e., eyes above a nose above a mouth), and a category like fruit contains highly discernible properties, but there is little consistency across exemplars. Diffeomorphed images can be used to help outline these properties of object perception, which can then be used to guide the selection of better-matched control stimuli. In turn, extracting the mechanisms governing perception at earlier stages of perception can then be used to design better-matched control stimuli to help explicate the mechanisms at the highest perceptual stages.

In conclusion, we demonstrate that the simulated neural signal in response to diffeomorphed images more closely resembled the neural signal associated with intact images relative to the other scrambling methods. Moreover, the advantage of diffeomorphed images over phase and box scrambling persists across many levels of scrambling. This similarity is consistent across distinct stages of the visual hierarchy (modeled by the HMAX model). Because processing at the earliest stages is held constant, differences in neural activity at anterior visual areas cannot be influenced by the properties of the image or the nature of the information feeding in from posterior areas. We suggest that diffeomorphed images serve as better control stimuli and should be used in neuroimaging studies that aim to disentangle early from later visual processing in order to more rigorously examine the neural mechanisms underlying perception, attention, and memory of the real world in later stages of visual processing.

Keywords: image scrambling, diffeomorphic transformation, control stimuli, object perception

Acknowledgments

This research was funded by the Natural Sciences and Engineering Research Council of Canada. We are thankful to the Centre for Excellence Research Chairs.

Commercial relationships: none.
Corresponding author: Bobby Stojanoski.
Email: bobby.stojanoski@gmail.com.
Address: Brain and Mind Institute, University of Western Ontario, London ON, Canada.

References

Ales, J. M., Farzin, F., Rossion, B., & Norcia, A. M. (2012). An objective method for measuring face detection thresholds using the sweep steady-state visual evoked response. *Journal of Vision, 12*(10): 18, 1–18, http://www.journalofvision.org/content/12/10/18, doi:10.1167/12.10.18. [PubMed] [Article]

Balas, B., Nakano, L., & Rosenholtz, R. (2009). A summary-statistic representation in peripheral vision explains visual crowding. *Journal of Vision, 9*(12):13, 1–18, http://www.journalofvision.org/content/9.12.13, doi:10.1167/9.12.13. [PubMed] [Article]

Banno, H., & Saiki, J. (2011). Higher-order image statistics is a cue for animal detection. *Journal of Vision, 11*(11):837, http://www.journalofvision.org/content/11/11/837, doi:10.1167/11.11.837 [Abstract]

Bex, P. J. (2010). Sensitivity to spatial distortion in natural scenes. *Journal of Vision, 10*(2):23, 1–15, http://www.journalofvision.org/content/10/2/23, doi:10.1167/10.2.23. [PubMed] [Article]

Costanzo, M. E., McArdle, J. J., Swett, B., Nechaev, V., Kemeny, S., Xu, J., & Braun, A. R. (2013). Spatial and temporal features of superordinate semantic processing studied with fMRI and EEG. *Frontiers in Human Neuroscience, 7,* 293.

Cusack, R. (2005). The intraparietal sulcus and perceptual organization. *Journal of Cognitive Neuroscience, 17*(4), 641–651.

Cusack, R., Mitchell, D. J., & Duncan, J. (2009). Discrete object representation, attention switching, and task difficulty in the parietal lobe. *Journal of Cognitive Neuroscience, 22*(1), 32–47.

Dakin, S., Hess, R., Ledgeway, T., & Achtman, R. (2002). What causes non-monotonic tuning of fMRI response to noisy images? *Current Biology, 12*(14), R476–R477.

Epstein, R., & Kanwisher, N. (1998). A cortical representation of the local visual environment. *Nature, 392*(6676), 598–601.

Felleman, D. J., & Essen, D. C. V. (1991). Distributed hierarchical processing in the primate cerebral cortex. *Cerebral Cortex, 1*(1), 1–47.

Freeman, J., Ziemba, C. M., Heeger, D. J., Simoncelli, E. P., & Movshon, J. A. (2013). A functional and perceptual signature of the second visual area in primates. *Nature Neuroscience, 16*(7), 974–981.

Gerhard, H. E., Wichmann, F. A., & Bethge, M. (2013). How sensitive is the human visual system to the local statistics of natural images? *PLoS Computational Biology, 9*(1), e1002873.

Greene, M. R., & Oliva, A. (2009). The briefest of glances: The time course of natural scene understanding. *Psychological Science, 20*(4), 464–472.

Grill-Spector, K., Kushnir, T., Hendler, T., Edelman, S., Itzchak, Y., & Malach, R. (1998). A sequence of object-processing stages revealed by fMRI in the human occipital lobe. *Human Brain Mapping, 6*(4), 316–328.

Hubel, D. H., & Wiesel, T. N. (1968). Receptive fields and functional architecture of monkey striate cortex. *The Journal of Physiology, 195*(1), 215–243.

Ishai, A., Ungerleider, L. G., Martin, A., & Haxby, J. V. (2000). The representation of objects in the human occipital and temporal cortex. *Journal of Cognitive Neuroscience, 12*(supplement 2), 35–51.

Kanwisher, N., McDermott, J., & Chun, M. M. (1997). The fusiform face area: A module in human extrastriate cortex specialized for face perception. *The Journal of Neuroscience, 17*(11), 4302–4311.

Kingdom, F. A. A., Field, D. J., & Olmos, A. (2007). Does spatial invariance result from insensitivity to change? *Journal of Vision, 7*(14):11, 1–13, http://www.journalofvision.org/content/7/14/11, doi:10.1167/7.14.11. [PubMed] [Article]

Kobatake, E., & Tanaka, K. (1994). Neuronal selectivities to complex object features in the ventral visual pathway of the macaque cerebral cortex. *Journal of Neurophysiology, 71*(3), 856–867.

Koenig-Robert, R., & Vanrullen, R. (2013). SWIFT: A novel method to track the neural correlates of recognition. *NeuroImage, 81C,* 273–282.

Koffka, K. (1935). *Principles of Gestalt psychology*. London: Lund Humphries.

Kourtzi, Z., & Kanwisher, N. (2000). Cortical regions involved in perceiving object shape. *The Journal of Neuroscience, 20*(9), 3310–3318.

Kovesi, P. (2003). Phase congruency detects corners and edges. In C. Sun, H. Talbot, S. Ourselin, & T. Adriaansen (Eds.), *The Australian pattern recogni-*

tion society conference: DICTA (pp. 309–318). Sydney, Australia: CSIRO Publishing.

Kriegeskorte, N., Mur, M., Ruff, D. A., Kiani, R., Bodurka, J., Esteky, H., & Bandettini, P. A. (2008). Matching categorical object representations in inferior temporal cortex of man and monkey. *Neuron, 60*(6), 1126–1141.

Lerner, Y., Hendler, T., Ben-Bashat, D., Harel, M., & Malach, R. (2001). A hierarchical axis of object processing stages in the human visual cortex. *Cerebral Cortex (New York, N.Y.: 1991), 11*(4), 287–297.

Malach, R., Reppas, J. B., Benson, R. R., Kwong, K. K., Jiang, H., Kennedy, W. A., & Tootell, R. B. (1995). Object-related activity revealed by functional magnetic resonance imaging in human occipital cortex. *Proceedings of the National Academy of Sciences, USA, 92*(18), 8135–8139.

Morrone, M. C., & Burr, D. C. (1988). Feature detection in human vision: a phase-dependent energy model. *Proceedings of the Royal Society of London. Series B, Containing Papers of a Biological Character. Royal Society (Great Britain), 235*(1280), 221–245.

Murray, R. F. (2011). Classification images: A review. *Journal of Vision, 11*(5):2, 1–25, http://www.journalofvision.org/content/11/5/2, doi:10.1167/11.5.2. [PubMed] [Article]

Oppenheim, A. V., & Lim, J. S. (1981). The importance of phase in signals. *Proceedings of the IEEE, 69*(5), 529–541.

Peelen, M. V., & Caramazza, A. (2012). Conceptual object representations in human anterior temporal cortex. *The Journal of Neuroscience, 32*(45), 15728–15736.

Portilla, J., & Simoncelli, E. P. (2000). A parametric texture model based on joint statistics of complex wavelet coefficients. *International Journal of Computer Vision*, *40*(1), 49–71.

Rainer, G., Augath, M., Trinath, T., & Logothetis, N. K. (2002). The effect of image scrambling on visual cortical BOLD activity in the anesthetized monkey. *NeuroImage, 16*(3), 607–616.

Riesenhuber, M., & Poggio, T. (2002). Neural mechanisms of object recognition. *Current Opinion in Neurobiology, 12*(2), 162–168.

Rousselet, G. A., Pernet, C. R., Bennett, P. J., & Sekuler, A. B. (2008). Parametric study of EEG sensitivity to phase noise during face processing. *BMC Neuroscience, 9,* 98–120.

Rousselet, G. A., Pernet, C. R., Caldara, R., & Schyns, P. G. (2011). Visual object categorization in the brain: What can we really learn from ERP peaks? *Frontiers in Human Neuroscience, 3*, 156.

Rust, N. C., & Dicarlo, J. J. (2010). Selectivity and tolerance ("invariance") both increase as visual information propagates from cortical area V4 to IT. *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience, 30*(39), 12978–12995.

Schyns, P. G., Gosselin, F., & Smith, M. L. (2009). Information processing algorithms in the brain. *Trends in Cognitive Sciences, 13*(1), 20–26.

Serre, T., Wolf, L., Bileschi, S., Riesenhuber, M., & Poggio, T. (2007). Robust object recognition with cortex-like mechanisms. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 29*(3), 411–426.

Singh, K. D., Smith, A. T., & Greenlee, M. W. (2000). Spatiotemporal frequency and direction sensitivities of human visual areas measured using fMRI. *NeuroImage, 12*(5), 550–564.

Tanaka, K. (1996). Inferotemporal cortex and object vision. *Annual Review of Neuroscience, 19*(1), 109–139.

Thomson, M. G. (1999). Visual coding and the phase structure of natural scenes. *Network (Bristol, England), 10*(2), 123–132.

Vogels, R. (1999). Effect of image scrambling on inferior temporal cortical responses. *Neuroreport, 10*(9), 1811–1816.

Wang, Z., & Simoncelli, E. P. (2004). Local phase coherence and the perception of blur. In S. Thrun, L. Saul, & B. Schölkopf (Eds.), *Adv. neural information processing systems* (pp. 1435–1442). Cambridge, MA: MIT Press.