scientific reports

Check for updates

OPEN

Orthogonal neural representations support perceptual judgments of natural stimuli

Ramanujan Srinath[1,3], Amy M. Ni[1,2,3], Claire Marucci[2], Marlene R. Cohen[1,4] & David H. Brainard[2,4]✉

In natural visually guided behavior, observers must separate relevant information from a barrage of irrelevant information. Many studies have investigated the neural underpinnings of this ability using artificial stimuli presented on blank backgrounds. Natural images, however, contain task-irrelevant background elements that might interfere with the perception of object features. Recent studies suggest that visual feature estimation can be modeled through the linear decoding of task-relevant information from visual cortex. So, if the representations of task-relevant and irrelevant features are not orthogonal in the neural population, then variation in the task-irrelevant features would impair task performance. We tested this hypothesis using human psychophysics and monkey neurophysiology combined with parametrically variable naturalistic stimuli. We demonstrate that (1) the neural representation of one feature (the position of an object) in visual area V4 is orthogonal to those of several background features, (2) the ability of human observers to precisely judge object position was largely unaffected by those background features, and (3) many features of the object and the background (and of objects from a separate stimulus set) are orthogonally represented in V4 neural population responses. Our observations are consistent with the hypothesis that orthogonal neural representations can support stable perception of object features despite the richness of natural visual scenes.

Keywords Neural representations, Decoding, Psychophysics, Population recordings, V4, Natural vision

A major function of the visual system is to infer properties of currently relevant stimuli without interference from the tremendous amount of task-irrelevant information that bombards our retinas. Many laboratory studies of the neural basis of this ability use, for good reasons, relatively simple stimuli[1–9]. An advantage of this approach is experimental control: one can parametrically vary stimuli and completely specify the input to the visual system. However, a downside of using such stimuli is that their simplicity prevents them from fully illuminating the neural algorithms by which the brain sorts through the large quantity of visual information characteristic of natural viewing (see simulations in Ruff et al.[10]).

In contrast to simple artificial stimuli, natural images can vary in many features, and these features are jointly encoded by the responses of populations of neurons in visual cortex[11–16]. It is well known that neurons in visual cortex are tuned for various features[4,17], so a single neuron's response may not allow unique identification of multiple image feature values. In general, tuning for simple features is thought to be independent, meaning that, for example, a neuron's tuning for orientation does not predict its tuning for spatial frequency[18–20]. In this case, a population of neurons, comprised of neurons that on their own confound multiple stimulus features, will carry sufficient information as a group to allow robust estimation of any one feature[10]. Whether neural populations encode different features of natural images, which themselves contain many more statistical dependencies than typical artificial stimuli, in a similarly independent way remains unknown.

Knowing how neural population responses to different image features covary for natural images is important because this has profound implications for how those features can guide behavior. A given feature can guide behavior in a way that is uncorrupted by other features only if its representation is independent from those other features. We can test for independence by viewing feature representations in a high-dimensional space in which each dimension represents the firing rate of one neuron in a population[21,22]. Because tuning curves are

[1]Department of Neurobiology and Neuroscience Institute, The University of Chicago, Chicago, IL 60637, USA. [2]Department of Psychology, University of Pennsylvania, Philadelphia, PA 19104, USA. [3]Ramanujan Srinath and Amy M. Ni have contributed equally to this work. [4]Marlene R. Cohen and David H. Brainard have contributed equally to this work. ✉email: brainard@psych.upenn.edu

smooth and continuous, systematically varying one scene feature, such as the position of a banana, traces out a continuous trajectory in the population space, which can typically be approximated by a line[23,24]. If just one feature varies, then the value of that feature can be read out by projecting the population response onto this line (aka linear decoding).

If tuning for two features is independent across neurons, then we expect that systematically varying the second feature would move the population response in an orthogonal direction[10]. Such orthogonality would mean that the two features can each be linearly decoded without interference from the other. However, if tuning for the two features were highly correlated (e.g., all neurons that prefer vertical orientations also prefer high spatial frequencies), varying the two features would trace out similar lines, and it would be difficult, if not impossible, to read out the two features independently. Intermediate cases are also possible, in which the representations for different features trace out somewhat separate trajectories that are neither close to co-linear nor orthogonal.

If a population of neurons contains the stimulus representation that mediates behavior, and the representations of multiple features are not orthogonal or close to it in this population, then we expect that behavioral measurements requiring judgments about these features to be error prone. Specifically, changing a task-irrelevant feature would perturb estimates of a task-relevant feature to the detriment of visual performance and lead to misjudgments of the task-relevant feature. Therefore, and following Hong et al.[25], we reasoned that variation in task-irrelevant features of a natural scene should not impair performance on a visual task involving judgments about these features if two conditions are met: visual information is read out of a neural population in a way that approximates a linear decoder, and the representations of relevant and irrelevant features are orthogonal in the relevant neural populations.

Here, we studied how the representation of the position of a foreground target object depends on variations in other scene features. The first experiment studies the effect of the position of background objects in the scene, using both neural population recordings in monkey and human psychophysics. The second uses neural recordings in monkeys and considers a broader range of scene variations. Because we are ultimately interested in how neural population responses support behavior in the natural environment, we employed naturalistic stimuli in these first two experiments. In a third experiment, we measured the link between neural responses and behavior using somewhat complex but not fully naturalistic stimuli that parametrically vary across many feature dimensions.

We leveraged the power of computer graphics to take parametric control of stimulus features in naturalistic stimuli, enabling us to vary many naturalistic stimulus dimensions and test the hypothesis that the observers' perceptual abilities to make fine perceptual distinctions will not be perturbed on a threshold-level judgment task by task-irrelevant variations in stimuli and backgrounds if the neural representations of task-relevant and irrelevant features are orthogonal. Using a combination of human psychophysics and monkey neurophysiology, we demonstrate that (1) the population representation of a target object's position in V4 is orthogonal to those of several background features, (2) the ability of human subjects to make precise perceptual judgments about object position was largely unaffected by task-irrelevant variation in those background features, and (3) many features of the object and the background (position, color, luminance, rotation, and depth) in these naturalistic images, and also of artificial stimuli in which many features are parametrically varied, are independently decodable from V4 population responses. We also examined monkey's behavior for estimating features of the artificial stimuli. Together, these observations support the idea that orthogonal neuronal representations enable stable perception of objects and features despite the tremendous irrelevant variation inherent in natural scenes.

Results

Central hypothesis: orthogonal representations enable observers to ignore irrelevant visual information

We tested the hypothesis that task-irrelevant information will not affect a perceptual judgment if the representations of the task-relevant and irrelevant features are orthogonal[25]. Figure 1a-b depict how we used computer graphics to parametrically vary different scene features, such as the position of a central object, the rotational position of objects in the background, and the depth position of objects in the background. Using this set of stimulus variations, consider the effect of varying the position of the object on a hypothetical neural population response as illustrated in the left panel of Fig. 1c. Each point in the plot represents the noisy population response to one presentation of an image, illustrating how varying object position against a fixed background can trace out a line in the high-dimensional neural population space. For this background, the position of the background could be read out by projecting the population response onto the line shown (labeled 'position axis for one background' in the figure). The middle panel of Fig. 1c shows how varying the background objects could affect this line in an orthogonal manner. In this case, the line tracing out the neural population response to the object at various positions is shifted in a direction orthogonal to the position axis shown in the left panel. Although a different line is swept out by varying object position against this second background, projecting onto the line for the first background continues to decode the object position accurately. If, on the other hand, changing the background causes a change in the position axis that is not orthogonal (right panel of Fig. 1c), projecting onto the line for the first background will not provide an accurate linear position readout. Thus, we test the hypothesis that changing an irrelevant background feature (e.g., the position of background objects) will not impact the perception of the task-relevant feature if the irrelevant background changes are orthogonal to the relevant ones (Fig. 1c middle).

Naturalistic stimuli with parameterizable properties

To test these predictions, we created naturalistic stimuli with many parameterizable, interpretable features (Fig. 1a-b). We parametrically varied the position of a central object (banana) and the rotation and depth of background objects (leaves and branches) set against a larger fixed contextual scene (rocks, moss-covered stumps, mountains, and skyline). The context was consistent across images to anchor the representations of

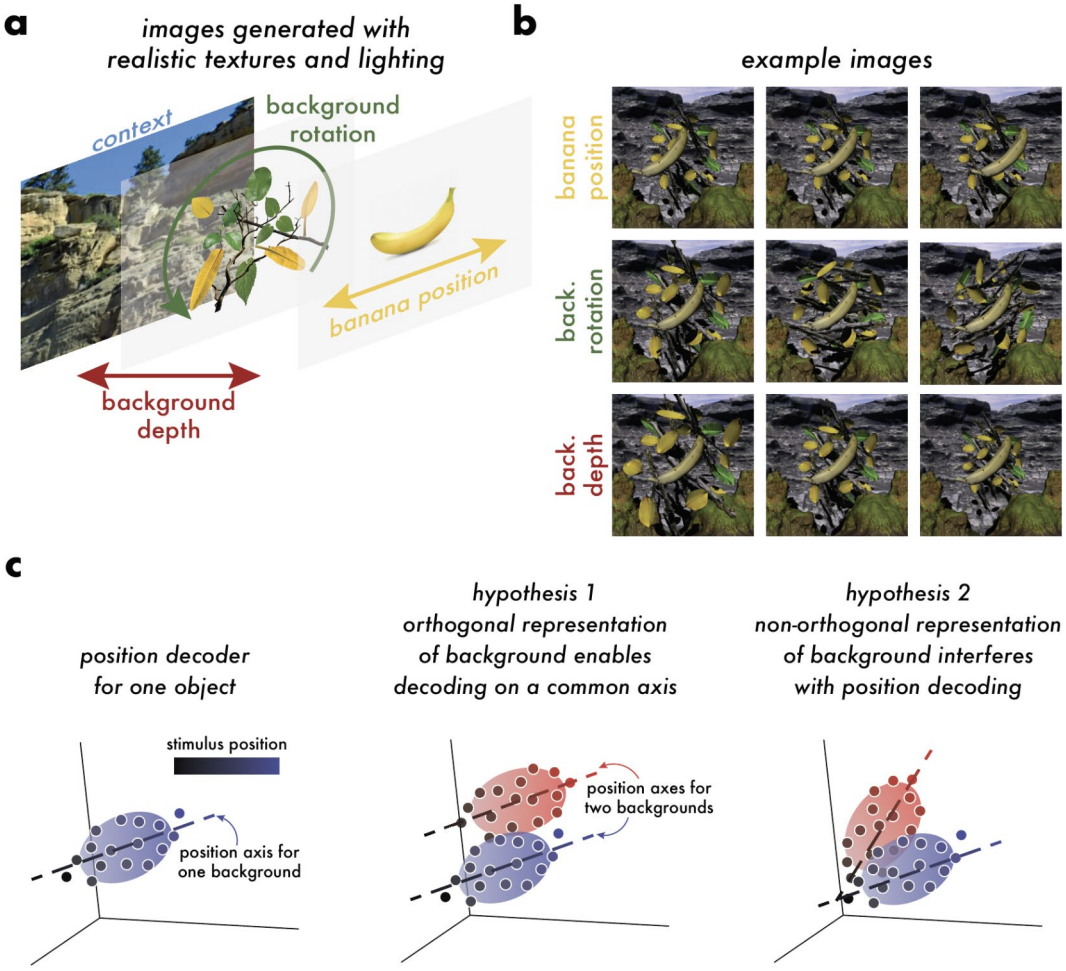

Fig. 1. Stimulus design and hypotheses about how neural representations enable generalizable decoding. (**a**) We generated photorealistic images in which we permuted the features of the central object (the banana) and background objects (sticks and leaves) using a Blender-based image generation pipeline that gave us control over central- and background-object properties (their position, size, pose, color, depth, luminance, etc.) The distant background (here referred to as the "context") is a static cue made of rock and grassy textures that were not varied. The monkey was rewarded for fixating a central point. Because the receptive fields of the recorded visual neurons were at several degrees of eccentricity (Supp. Fig. 1), the stimuli were placed within those receptive fields rather than centered over the fixation point. (**b**) Example images showing variations in three parameters – central-object position in the horizontal direction, a rotation of the background objects (leaves and branches), and the depth of the background objects. Five values of each of the three parameters were chosen for each monkey based on receptive field properties (see below), yielding an image set of $5 \times 5 \times 5 = 125$ images. The context was held constant across all images. (**c**) Hypothesized implications of the neural formatting of visual information on the ability to decode a visual feature. Consider the responses of a population of neurons in a high-dimensional space in which the response of each neuron is one dimension. The population responses to a series of stimuli that differ only in one parameter (e.g., the position of the central object) change smoothly in this space (left). Responses to a set of stimuli that differ in the same parameter but also have, for example, a difference in the background will trace out a different path in this space (e.g., the red points in the center and right panels). Relative to the first (blue) path, changing the same parameter on a different background could change the population response in a parallel way; more specifically, changing the background could move the population along an orthogonal dimension to the dimension encoding the parameter of interest (center). This scenario would enable linear decoding of the parameter of interest invariant to background changes. Alternatively, the direction that encodes the parameter of interest could depend on the background (right). Under the linear readout hypothesis, in this case, varying the background would impair the ability of a population of neurons to support psychophysical estimation of the parameter of interest.

object and background features, which means that the position of the central object could be judged relative to the edge of the monitor, the fixation point, the contextual elements, or a combination. We presented these stimuli within the joint receptive fields of recorded V4 neurons (Supp. Fig. 1) while each of the two monkeys fixated on a central point. In sum, we recorded V4 responses in 26 experimental sessions across two animals (85–94 visually

responsive multi-units per session). Most of the units in our measured population were sensitive to variations in both the position of the central object and the variations in background rotation and depth (Supp. Fig. 1d).

V4 neurons robustly encode stimulus position for each stimulus background

We first measured the extent to which V4 neurons encode object position by linearly decoding that position for each unique background stimulus. This decoding ability is a pre-requisite for meaningful tests of the orthogonality of the population representation when other image features are varied. Figure 2a shows that for each unique background configuration (rotation and depth), V4 neurons from a single session support good linear decoding of the position of the object (each of the 25 panels in Fig. 2a is for a specific background configuration; each gray point in the panels shows decoded object position for a single presentation; the mean and standard deviation for each position - open circles and error bars - summarize our ability to predict the position of the object from the activity of V4 neurons). The numbers at the upper left of each panel provide the correlation between the predicted and actual object stimulus position (mean performance = 0.698).

V4 representations of stimulus position and background features are approximately orthogonal

Because the first condition, that we could use a linear decoder to estimate the position of the object in each background (Fig. 2a), was met, we tested the second part of our orthogonality hypothesis. We tested this second part, that the neural representation of the position of the central object in V4 is approximately orthogonal to the representations of features of the background objects (depth and rotation of the leaves and branches) in our stimuli, in five ways.

First, we calculated how well we could use a linear decoder to decode object position with a single general decoder, across the various background conditions. This worked well. Indeed, across sessions, we could decode object position (Fig. 2b), background rotation (Supp. Fig. 2a), and background depth (Supp. Fig. 2b) accurately across variation in the values of the other two features. We trained these decoders with either all presentations in a leave-one-out fashion or matching presentation counts with the condition-specific decoders with qualitatively similar results (see Methods for the details of this comparison). The ability of a single decoder to read out banana position across all background variations is similar to that of the decoders that were optimized for each unique background stimulus and is high across all sessions for two monkeys (Fig. 2c). This result suggests that the population of V4 neurons we measured encodes all of the tested features well and near orthogonally.

Second, we found that on a trial-by-trial basis, errors in the decoded estimates of object position are not correlated with errors in decoding of background rotation and depth (Fig. 2d and Supp. Fig. 2e). This lack of correlation also suggests that the representations of object position are independent in V4 from representations of background rotation and depth.

Third, to probe the orthogonality of the representations in more detail, we also tested the performance of condition-specific decoders using responses obtained in other background conditions (cross-condition decoders; Supp. Fig. 3). In the case of perfect orthogonality, these decoders would perform as well on data from other background conditions as on held-out data from the condition they were trained on. We found that as the feature value of the background deviated further from the one the decoder was trained on, decoding accuracy did decrease slightly. This is a strong test of orthogonality, however, and the largest effect at the furthest difference in background feature value is a quite modest ~ 0.15. This suggests that the object position representations are largely, but not perfectly, orthogonal to the effect of background feature changes (see Discussion).

Fourth, we compared the weights assigned to each neuron when constructing each linear decoder. We found no significant correlations between the weights assigned while decoding any of the features (position, rotation, or depth; Supp. Fig. 4a). There was also no detectable relationship between each neuron's sensitivity to a feature and its decoding weight (Supp. Fig. 4d).

Finally, if the representations of stimulus position and background parameters are orthogonal, then the decoders optimized for each unique stimulus (Fig. 2a) should be mutually aligned in neural population space. Put another way, if the representation of stimulus position is robust to variation in the background, this representation should vary along the same direction across backgrounds. To probe this, we calculated the line in neural population space that best explains population responses to each stimulus position for each background. These are depicted in Fig. 3a for each unique background condition (plotted for the first two principal components of the population neural responses for visualization purposes only; each point is a trial, each bright point is the average population response to a particular object position, and gray to yellow point colors represent the five object position values). The angle in the population space between the decoders for each unique background and the decoder for the background configuration whose decoder is shown in the center of 3a marked with ∗ (chosen as a reference to define the origin of the angular measure) is indicated in degrees. The distribution of angles for this example session (Fig. 3b) and across all sessions in both animals (Fig. 3c) is skewed toward much smaller angles than expected by chance – gray distributions in Fig. 3c depicting a median of ~ 90° for decoders trained on shuffled (randomizing trial labels within background configuration) responses to each unique background condition (labeled "shuffle"; dark gray) and angles between random vectors in a response space with the same dimensionality as the neural decoding space (labeled "random"; light gray).

Together, these recording results support the idea that the neural representation of stimulus position is close to orthogonal to the representations of background rotation and depth in V4.

Human subjects discriminate stimulus position robustly with respect to background variation

Our central hypothesis predicts that when the representations of two stimulus features are orthogonal in the brain, varying one should not impact the ability of subjects to discriminate the other. We tested this hypothesis by measuring the ability of human observers to discriminate the position of the central object in our stimuli amid

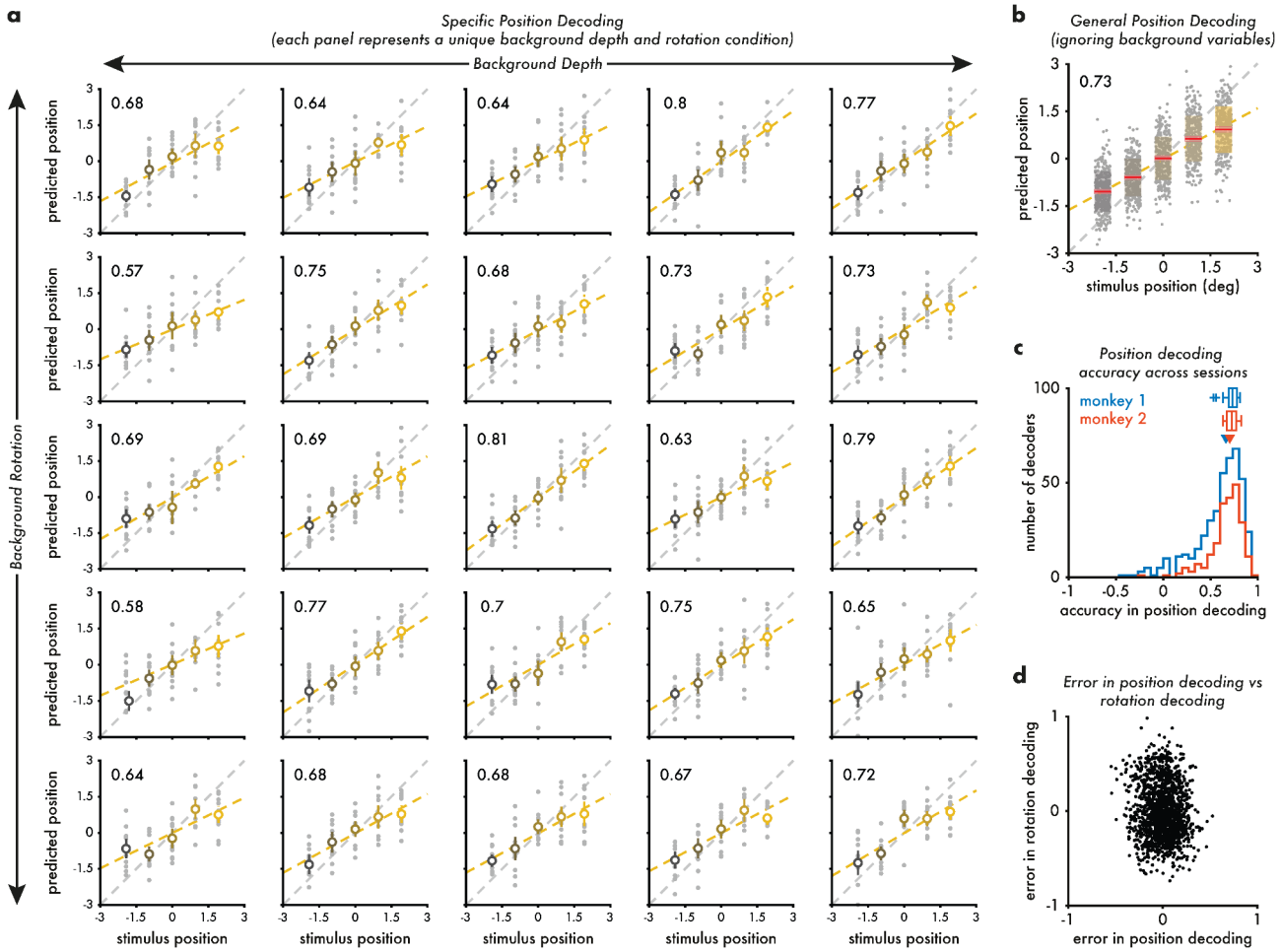

Fig. 2. Object position decoding from V4 population responses is consistent across background variations. (**a**) We can linearly decode object position for each background stimulus (for the example session shown here). Each panel represents a unique configuration of background rotation and depth, with rows representing variations in rotation and columns representing variations in depth. Each gray point shows the decoded position for a single image presentation in this session. These points depict the actual object position (x-axis, in visual degrees relative to the center of the image) and the decoded position (y-axis) using a separate, cross-validated linear decoder for each unique background. The open circles represent the trial-averaged predicted position (vertical length is the standard deviation). The number in the top-left is the correlation between the actual and decoded positions and the yellow dashed line is a linear fit (compared to null hypothesis that the correlation is 0 or 'constant model', the p-values of tests to reject this hypothesis range from 2.5×10^{-15} to 1.4×10^{-6} for this session). The dark gray to yellow gradient of the open circles is a redundant cue in the figure that also conveys stimulus position variation. The gray dashed line represents the identity. (**b**) Position decoding is largely consistent across background variations. This plot is in the same format as those in A. Here, a condition-general decoder that ignores variations in the background and incorporates all stimulus presentations is used ($r = 0.73$; vs. constant model, $p < 0.001$). We also computed the general decoder using the minimum number of presentations across all 25 specific decoders in A (60 trials) for 100 folds and found similar results ($r = 0.72$; vs. constant model, $p < 0.001$). Compare with Supp. Fig. 2a and b for background rotation and depth decoding. (**c**) Distribution of specific decoder accuracies (correlation) across all sessions for each monkey (each session contributed 25 values to the histogram). Blue and red arrows represent the median accuracy (0.662 for monkey 1, 0.703 for monkey 2). The box plots above the histograms summarize general decoder accuracy across sessions for each monkey. The central line in each box plot indicates the median (0.735 for monkey 1, 0.724 for monkey 2), box edges indicate 25 and 75 percentiles, whiskers indicate minimum and maximum values, and + symbols indicate outliers. We found similar results for the trial count matched general decoders (median 0.732 for monkey 1, 0.722 for monkey 2). Compare with Supp. Fig. 2c and d for background rotation and depth decoding. (**d**) Error in decoding object position (across background variations) for each trial compared with the error in decoding background rotation. The correlation between the two types of errors is very small ($r = -0.083$), although it is statistically significant ($p = 0.001$). See Supp. Figure 2e for comparison with error in trial-wise background depth decoding.

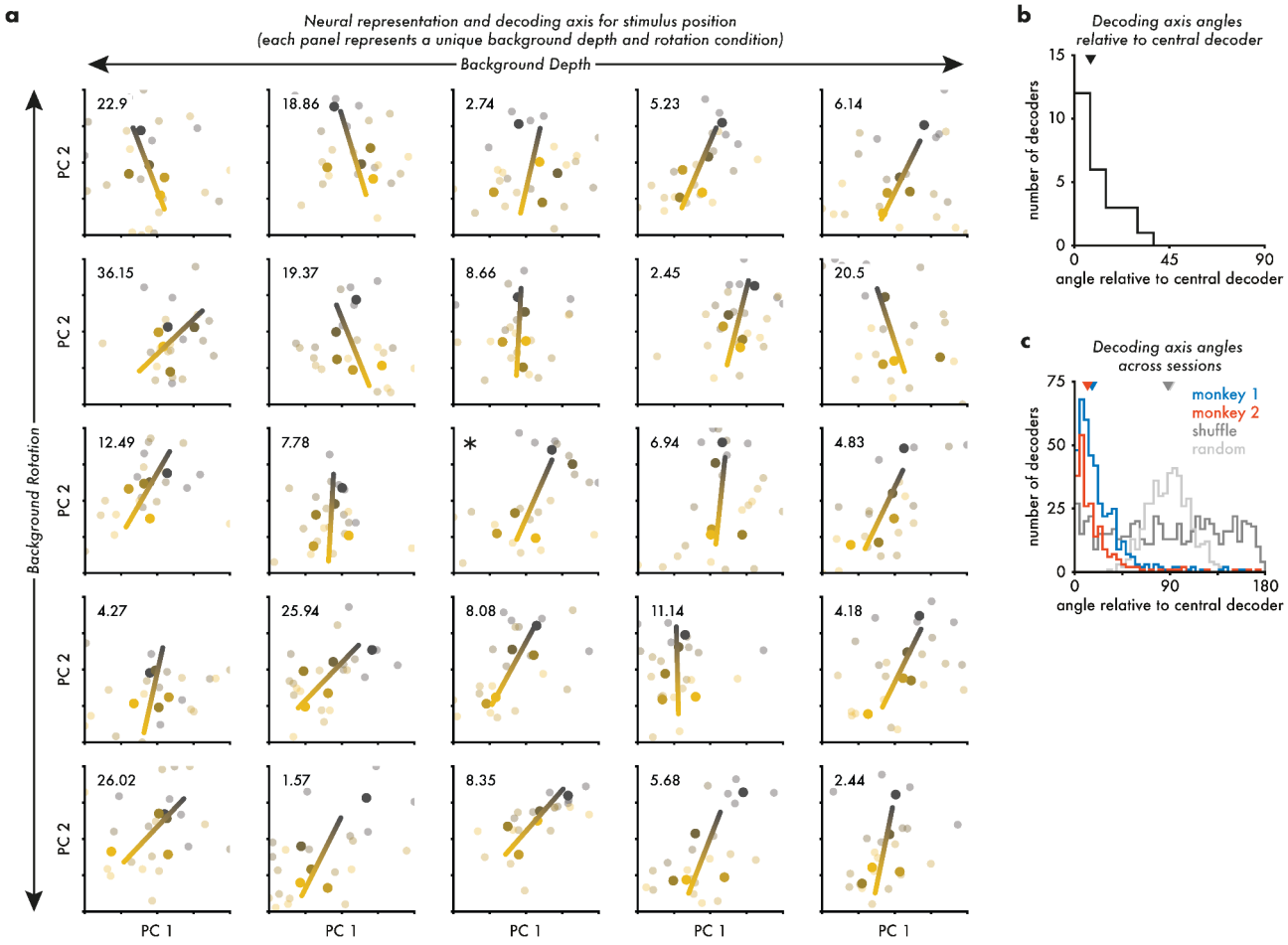

Fig. 3. Object position axes across background variations are aligned with each other. Since object position decoding is tolerant to background variations, we tested whether the linear decoding axes for each background configuration were aligned by visualizing the decoders in the first two principal components of the neural response space. These dimensions were computed for the full set of neural responses obtained in each session. (**a**) As with Fig. 2a, each panel represents a unique configuration of the background rotation and depth with rows representing variations in rotation and columns representing variations in depth. Each dim point represents a single image presentation, and bright points represent trial-averaged responses. Gray-to-yellow gradient represents monotonic variation in object position. A gradient line was fit to the responses for each background condition, shown here in two dimensions for illustration. The lines shown have been normalized to have the same length in the projected space shown. The text label at the top left represents the relative angle between each decoder and the central decoder (the middle background condition plot, marked with $*$) calculated in the full dimensional space of responses used for decoding. (**b**) Distribution of angles in A as a histogram. Arrow at the top represents the median angle for this session (7.78°). (**c**) Distribution of relative decoder angles across all object position decoders (like those in A) across sessions for both monkeys (blue and red distributions). Blue and red arrows represent the median of angles across sessions (16.4° for monkey 1, 12.07° for monkey 2). Dark gray distribution represents the angles of object position decoders after shuffling the position values for each trial (median 88.13°, shown as dark gray arrow). Light gray distribution represents the angles between randomly chosen vectors of the same dimensionality as the neural population space (median 89.99°, shown as light gray arrow).

variation in the background rotation and depth. Using a threshold paradigm, we tested this idea for stimulus step sizes that approach the limits of perception.

Human subjects viewed two images of the object separated by two different masks (Fig. 4a) and reported whether the object in the second image was positioned to the left or right of the object in the first presentation. The offset between the two object positions was varied systematically to allow us to calculate a discrimination threshold for each background variation condition. Across blocks of trials, we varied the amount of within-trial image-to-image variability in the background objects across the two presentations of the central object. The context (the rocky and grassy textures) was held consistent across images to give the variations in the object and background features a frame of reference. When there was background variation, intrusion of that variation on decoding of position would manifest as an elevated discrimination threshold if the representation of the object position was not orthogonal to that of the background features[26,27]. However, consistent with the idea that the

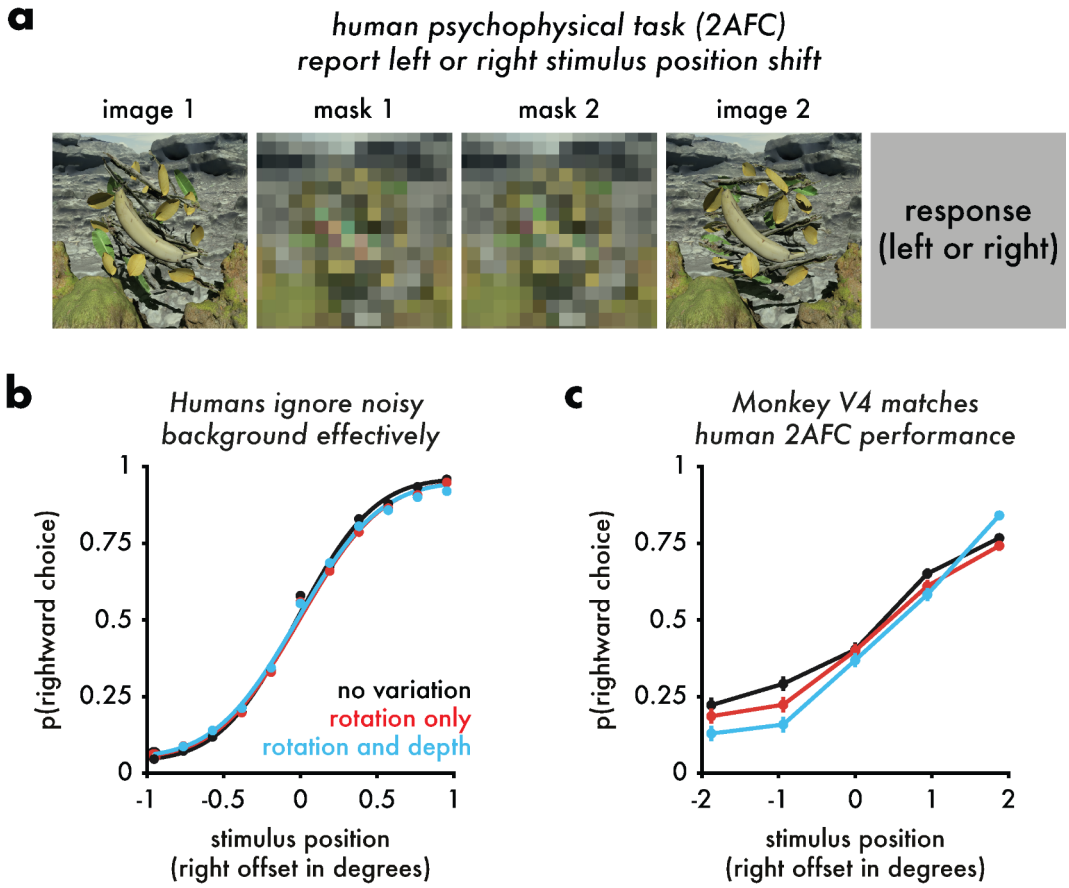

Fig. 4. Human psychophysics experiments suggest that object position discrimination is unaffected by stimulus background variation. Since object position representation in V4 neurons is robust to background variation, we tested whether changing background properties affects the decoding of object position in humans. (**a**) Human psychophysical task. Two images containing a object were presented, with two masks in between. Participants were instructed to report whether the relative position of the object in image 2 was left or right of that in image 1. In blocks, background rotation and/or depth were held constant or varied as described below. (**b**) Averaged (across participants, $N = 10$) psychometric functions for each background variation condition (individual participant performance shown in Supp. Fig. 5). Three background conditions were tested: black: no background variation between the two presentations of the object on each trial; red: background rotation changed randomly across the two presentations of the object, but background depth was held fixed; blue: both background rotation and depth were randomized across the two presentations of the object. Across participants, object position change detection performance was similar across the three background variation conditions (paired t-tests on thresholds between each pair, $p > 0.05$, Bonferroni corrected). Bayes Factor Analysis for the main effect of background variations suggests strong evidence for the absence of the effect of background variation in choices ($1/BF = 34.27$; see Methods for details). This value of 1/BF provides support for the absence of a background effect in the human psychophysics. (**c**) To compare human behavior and monkey electrophysiology, we selected stimulus presentations in the monkey experiments to approximate the three background variation levels used in the human psychophysics experiment (see Methods for details of sample matching). We trained linear discriminants to separate trials into right or left position shift, for each background variation condition. During training, the trials with the object in the central position were randomly assigned to be left or right. Classifier performance also did not differ substantially across the background variation conditions (paired t-tests on thresholds between each pair, $p > 0.05$, Bonferroni corrected). Similar to (b) above, Bayes Factor analysis suggests strong evidence for the absence of the effect of background ($1/BF = 26.25$ for monkey 1 and $1/BF = 8.99$ for monkey 2). These values of 1/BF provide support for the absence of a background effect of a background effect in the neural decoding.

orthogonal representations of object position and background features that we found in the neural recordings enable background-independent perception, introducing variability into the background did not significantly impact the position discrimination performance of human subjects (Fig. 4b and Supp. Fig. 5). To make a direct comparison between parameter decoding of neural representations and human psychophysical performance, we partitioned neural data into "no variation", "rotation variation only", and "depth and rotation variation" groups and trained linear discriminants to classify left or right position difference between pairs of presentations (Fig. 4c). As with the human psychophysical performance, the cross-validated discrimination performance of these

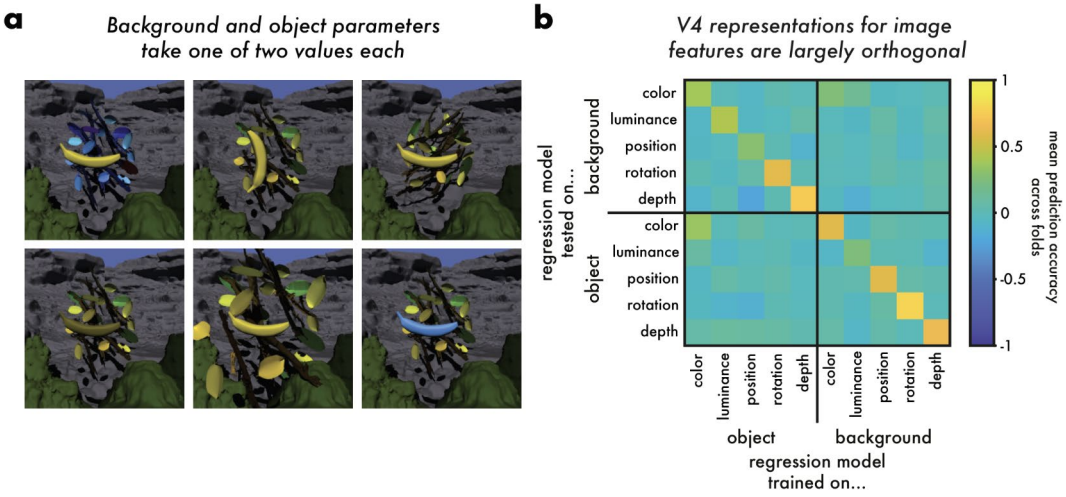

Fig. 5. Orthogonal representations for variations in up to 10 object and background features. We generated a large image set where the color, luminance, position, rotation, and depth of both the background and object each took one of two values yielding 2^{10} = 1,024 images. We collected V4 population responses to these images as in Supp. Fig. 1c. (**a**) Example images illustrating object and background parameter variation. (**b**) If two features are encoded orthogonally (independently) in neural population space, then a decoder trained on one feature should not support decoding of the other feature. We trained linear decoders of V4 responses for each object and background feature (x-axis) and tested the ability to decode each of the other features. The diagonal entries provide, for each feature, the correlation between the decoded and actual feature parameter values for a decoder trained on that feature. Correlations were obtained through cross-validation. Decoding performance was above chance (median correlation of 0.61) for all features ($p < 0.0001$; t-test across folds). The off-diagonal values depict the performance of a decoder trained on one feature (x-axis) for decoding another (y-axis). This cross-decoding performance was not distinguishable from chance (median correlation of 0.008) except in the case of the color of the object and background ($r = 0.36$).

classifiers did not differ across the three background variation groups. Thresholds for the neural classifiers are higher than those for the human subjects (note the difference in the x-axis scale between Fig. 4b and c), which is expected given that neuronal responses were measured to peripheral stimuli and that we recorded from a small subset of the neurons that could support psychophysical performance.

At least ten object and background features are represented approximately orthogonally in V4

To test the extent to which the orthogonality of representations of different features generalizes to other features of the object and background in our stimuli, we measured V4 responses to a large image set where the color, luminance, position, rotation, and depth of both the background and object each took one of two values (this yields 2^{10} = 1,024 unique images; Fig. 5a). If any two parameters are encoded orthogonally in neural population space, then it should be possible to linearly decode those parameters successfully despite the variation in the others. Conversely, a decoder trained on one parameter should not provide information about the others. To test these predictions, we trained linear decoders for each object or background feature and then tested our ability to decode each of the ten features with each decoder.

Each of the ten features was encoded in the V4 population despite the variation in the other features, meaning that the correlation between the actual value of the feature parameter and the value predicted by a cross-validated linear decoder was above chance (diagonals in Fig. 5b). In addition, the correlation between a given parameter and the value predicted by a decoder trained on a different parameter (off-diagonals in Fig. 5b) was indistinguishable from chance except in one case (the color of the central object and background objects). These observations suggest that a population of V4 neurons can encode a relatively large number of natural scene parameters independently, enabling observers to avoid distraction by task-irrelevant stimulus features. The observation that there is an interaction between central and background objects presents an opportunity for future work to test the prediction that task-irrelevant variation in background object color should affect psychophysical discrimination of central object color. This outcome would be consistent with the results of Singh et al.[27].

Object features represented orthogonal to a task-relevant feature do not affect behavioral estimation

The results of these first two experiments demonstrate that a feature of naturalistic scenes (object position) that is represented orthogonally to the features of its background can be perceptually estimated independently of those background features. To explore the generality of the results of the experiments presented above, we also analyzed published data from our lab to test the ideas in the context of within-object features. Specifically, we analyzed recordings from the same monkeys while they viewed isolated objects on a gray background (a detailed

explanation of the stimuli, task, and methods and a different analysis of a subset of these data are reported elsewhere[28]). We generated 50 three-dimensional shapes that varied in size, color, orientation, curvature, thickness, and several other features. (Supp. Fig. 6a). We flashed them while the monkeys fixated on a central dot (as with the images with the banana above). We analyzed the orthogonality of the shape features using cross-validated cross-decoding (Supp. Fig. 6b, compare to Fig. 5). We found that 116 of the 120 feature pairs were encoded orthogonally. The four feature pairs that could be cross-decoded were varied in the limited stimulus set in a correlated manner by chance (see Methods for details).

To test whether an orthogonally represented feature affects the monkeys' behavioral estimates of shape, we trained the monkeys to estimate the axial curvature of new 3D objects (Supp. Fig. 6c). The curvature was varied continuously, and the behavioral report was also on a continuous, analog scale. The monkeys were rewarded based on their estimation accuracy. Consistent with our hypothesis, the monkeys' curvature estimation behavior was invariant to task-irrelevant features that were represented orthogonally to curvature, including color, thickness, length, gloss, (Supp. Fig. 6d).

Discussion

Using a combination of multi-neuron electrophysiology in monkeys and human psychophysics, we tested the hypothesis that irrelevant visual features, whether in the object of interest or in the background of a scene, will not interfere with the perception of a target feature when their representations are orthogonal in visual cortex. We demonstrated that (1) in monkey area V4, the representation of object position is orthogonal to the representations of many irrelevant features of that object and the background, (2) the threshold for human observers to judge a change in object position was unaffected by the variations in the background stimulus that were shown neuronally to have orthogonal representations in monkey V4, and (3) the ability for monkey observers to estimate the curvature of an object was unaffected by irrelevant features of that object that were represented orthogonally in V4.

Mechanisms supporting orthogonality

Our study quantifies how neural populations represent multiple naturalistic stimulus variations, but it does not provide direct insight into how the encoding and processing of visual stimuli produce those representations. Under biologically realistic assumptions, simulations show that although it is possible to learn about the independence/orthogonality of feature representations within a population from small population recordings, it is generally not possible to characterize the role of each recorded neuron[10].

It is possible that other feature pairs that we did not study, such as purely lateral shifts in background objects relative to the central object, may interact either in behavior or neurally. In recent years, many studies have demonstrated that neural networks trained to categorize natural images produce representations that strongly resemble neural representations in the ventral visual stream[12,15,29–31]. These models provide an opportunity to understand the conditions under which aspects of natural stimuli are most likely to be represented orthogonally and which aspects might best be targeted in future studies to probe potential failures of orthogonality[25,32–37]. Additionally, our strong test of orthogonality based on cross-condition decoding revealed slight deviations in orthogonality of object position relative to background variations (Supp. Fig. 3). We predict that future studies on the capacity of a neural population for encoding naturalistic visual features (discussed further below) will shed light on the limits of orthogonal representations. We hope our results will produce a productive coupling of computational analysis of the mechanisms by which orthogonal representations emerge with behavioral experiments using the same parametrically varied computer graphics stimuli.

We also note that figure-ground segmentation is relevant to the ability to judge features of an object independent of the background[7,38,39]. Segregation of figure from ground does not guarantee that the representation of features of the ground does not influence the representation of the figure (or vice-versa). An exciting avenue for future work will be understanding, in a much wider set of natural scenes than we considered, the relationship between figure-ground segmentation, orthogonal representations of the features of a figure and ground, and perception of a figure independent of its background.

Relationship to the notion of untangling and representational geometry

The conditions under which objects can be disambiguated from neural population responses have been studied using the concept of *untangling*[40–43]. Untangling has been primarily discussed in the context of object classification. The hypothesis is that different objects can be appropriately classified (e.g. discriminating images of bananas from images of leaves) when the neural population representations of those objects are linearly separable in the face of irrelevant variations in the images (e.g. changes in position, orientation, size, or background). Support for this hypothesis comes from the observation that as one moves from early to late stages of the primate ventral visual stream, representations of different object categories become more linearly separable[25,33,42,44,45]. Progress has been made in understanding how the tuning functions and mixed selectivities of neurons support untangled population representations[46–48].

The untangling framework has been extended to address the structure of neural populations that represent object categories more generally by characterizing the geometry of the high-dimensional representational manifolds[34,35,49–52]. The capacity and dynamics of representational geometries in visual cortex correlate with classification behavior[42,53,54], in parietal and prefrontal cortices with perceptual decision-making[23,55–57], and in motor cortex with control of muscle activity[58–61].

Other studies consistent with this line of thinking have also considered features and have demonstrated that neural responses to relevant and distracting features of simple stimuli are linearly separable in the brain areas (or analogous layers of deep network models of vision) that are thought to mediate that aspect of vision[35,36,40,44,62]. Indeed, a previous study in our lab also found a relationship between our ability to linearly decode visual

information, the activity of neural populations in monkeys, and the ability of human observers to discriminate the same stimuli[32]. Our conclusions are also consistent with those reached in a study that considered neuronal representations in V4 and IT and behavioral estimates of the properties of objects presented against natural image backgrounds[25]. That study found an increase in the orthogonality of representation from V4 to IT and good behavioral estimation of the orthogonally represented properties. Given those results, it seems possible that our stimuli would have revealed increased orthogonality in areas further along the processing hierarchy than the V4 site of our electrode array; such a result would not change the general conclusions we draw about the relation between orthogonality and behavior performance.

Opportunities from studying parameterizable naturalistic images

The present study extends the measurements of the relationship between the neural untangling of lower-level features and visually guided behavior with respect to features in naturalistic images. For our naturalistic scenes, we did not find that variation in background object positions perturbed the perception representation of foreground object position. Interestingly, there are known cases with simple stimuli where the position of some scene elements influences the position (or motion) judgments of other elements. Both the Poggendorf[63] and the Flash Grab Effect[64] could be taken as of such interaction. What needs to be true of the scene for the visual system to generate maximally robust perceptual representations, particularly whether this is related to the statistical structure of natural scenes, remains an interesting and open question.

More generally, our view is that the perception of object features in complex natural images provides increased power for testing the untangling hypothesis in the context of feature decoding. Unlike the case of simpler stimuli, the number of task-irrelevant features available for manipulation is larger and is likely to more fully challenge the coding capacity of neural populations whose representations are of limited dimensionality[50]. Furthermore, visual distractors (like variation in the background) heavily influence scene categorization performance in artificial stimuli but not natural stimuli, suggesting that orthogonal feature representations in natural stimuli are more resilient to noise[34,65]. Studying the relationship between neurons and visually guided behavior using parameterizable naturalistic images solves many of the challenges inherent in using simple artificial stimuli on the one hand or natural images on the other[1,2,30,66–68]. The graphics-generated stimuli we employ strike a balance between the experimental control available through parameterization and the ability to measure principles governing neural responses to and perception of features of natural images that are difficult or impossible to glean using artificial stimuli.

Opportunities from cross-species investigations of visual perception

Our results highlight the power of pairing neural population recordings in animals with behavior in humans for understanding the neural basis of visual perception. We showed that the same principle (orthogonally represented features do not interfere perceptually) can be gleaned from cross-species approaches as from simultaneous recordings from behaving monkeys. Although simultaneously recording neurons and measuring behavior has many advantages, comparison with human performance provides some assurance that the neural results obtained in an animal model generalize to humans. In addition, our approach links observations from the more peripheral visual field locations, where, for technical reasons, the neural recordings are most often made, to the central visual field locations that are typically the focus of studies with human subjects.

Since the monkeys were rewarded simply for fixating during the recordings for our first two experiments, our experiments focus on neural population activity that is stimulus-driven rather than reflecting internally driven processes like attention or motivation. In future work, it will be interesting to merge our knowledge of how stimulus-driven and internal processes combine to influence neuronal responses and performance on visual tasks.

Conclusion

Our results provide behavioral and neurophysiological evidence supporting the powerful untangling hypothesis. They extend the study of untangling to representations of features of objects and backgrounds and demonstrate the value of parameterizable naturalistic images for studying the neural basis of visual perception. They also suggest a promising future for investigating the neural basis of perceptual and cognitive phenomena by leveraging the complementary strengths of multiple species.

Methods

Experimental models and subject details

Monkey electrophysiology

Two adult male rhesus monkeys (Macaca mulatta, 10 and 11 kg) were implanted with titanium head posts before behavioral training. Subsequently, multielectrode arrays were implanted in cortical area V4 identified by visualizing the sulci and using stereotactic coordinates. The monkeys were sourced from Alpha Genesis, Inc. All animal procedures were approved by the Institutional Animal Care and Use Committees of the University of Pittsburgh and Carnegie Mellon University, and all training, surgery, and experimentation methods were performed in accordance with the relevant guidelines and regulations. Additionally, this study is reported in accordance with ARRIVE animal use and reporting guidelines.

Human psychophysics

This study was preregistered at ClinicalTrials.gov, NCT number NCT05004649, https://clinicaltrials.gov/ct2/show/NCT05004649. The experimental protocols were approved by the University of Pennsylvania Institutional Review Board, and all recruitment and experimentation methods were performed in accordance with the

relevant guidelines and regulations. Participants were invited to volunteer to participate in this study. Participants provided informed consent and filled out a lab participant survey. We also screened for visual acuity using a Snellen eye chart and for color deficiencies using the Ishihara plate test. Participants were excluded prior to the experiment if their best-corrected visual acuity was worse than 20/40 in either eye or if they made any errors on the Ishihara plate test.

Participants were excluded after the conclusion of their first session if their horizontal position discrimination threshold in the no variation condition (see description of conditions below) was higher than 0.6° of visual angle, and participants excluded at this point did not participate in any further experimental sessions.

Experimental design

Image generation (for both human psychophysics and monkey electrophysiology)

All the stimuli were variants of the same natural visual scene: a square image with a central object (a banana) presented on an approximately circular array of overlapping background objects (made up of overlapping branches and leaves). These objects were rendered against a distant and static context (rocky and grassy textures) which served as a consistent cue to estimate the spatial parameters (position, orientation, depth). The central object and/or the background objects changed in horizontal position, rotation, and/or depth across different stimuli. In the larger set of stimuli (detailed below) the luminance and color of the central and background objects also changed. The central object and background objects are presented in the context of other objects (a rock ledge, a skyline, and three moss-covered stumps) that remain unchanged across all stimulus conditions. This natural visual scene was created using Blender, an open-source 3D creation suite (https://www.blender.org, Version 2.81a). The object and background parameters were varied using ISET3d, an open-source software package (https://github.com/ISET/iset3d) that works with a modified version of PBRT (https://github.com/scienstanford/pbrt-v3-spectral; unmodified version at https://github.com/mmp/pbrt-v3).

The images created using ISET3d were converted to RGB images using custom software (Natural Image Thresholds; https://github.com/AmyMNi/NaturalImageThresholds) written using MATLAB (MathWorks; Natick, MA) and based on the software package Virtual World Color Constancy (github.com/BrainardLab/VirtualWorldColorConstancy). Natural Image Thresholds is dependent on routines from the Psychophysics Toolbox (http://psychtoolbox.org), ISET3d (https://github.com/ISET/iset3d), ISETBio (http://github.com/isetbio/isetbio), PBRT (https://github.com/scienstanford/pbrt-v3-spectral; unmodified version at https://github.com/mmp/pbrt-v3), and the Palemedes Toolbox (palamedestoolbox.org).

To convert a hyperspectral image created using ISET3d to an RGB image for presentation on the calibrated monitor, the hyperspectral image data were first used to compute LMS cone excitations. The LMS cone excitations were converted to a metameric rendered image in the RGB color space of the monitor, based on the monitor calibration data. A scale factor was applied to this image so that its maximum RGB value was 1 and the image was then gamma corrected, again using monitor calibration data. This process was completed separately for the two different monitors used, one for the psychophysics and one for the neurophysiology.

Monkey electrophysiology

Array implantation, task parameters Both animals were implanted with titanium headposts before behavioral training. After training, microelectrode arrays were implanted in area V4 (96 recording sites; Blackrock Microsystems). Array placement was guided by stereotactic coordinates and visual inspection of the sulci and gyri. The monkeys were trained to perform a fixation task along with other behavioral tasks that were not relevant to this study. The stimulus images used in this study were not displayed outside of the context of this task. The monkeys fixated a central spot for a pre-stimulus blank period of 150–400 ms followed by stimulus presentations (200–250 ms) interleaved with blank intervals (200–250 ms). The stimuli were presented one at a time at a peripheral location that overlapped the receptive fields of the recorded neurons. In each trial, 6–8 stimuli were presented, after which the monkey received a liquid reward for having maintained fixation on the central spot until the end of the stimulus presentations. If the monkey broke fixation before the end of the stimulus presentations, the trial was terminated. The intertrial interval was at least 500 ms. The stimuli were presented pseudo-randomly.

The visual stimuli were presented on a calibrated (X-Rite calibrator) 24" ViewPixx LCD monitor (1920 × 1080 pixels; 120 Hz refresh rate) placed 54 cm (monkey 1) or 56 cm (monkey 2) from the monkey, using custom software written in MATLAB (Psychophysics Toolbox; Brainard, 1997; Pelli, 1997). Eye position was monitored using an infrared eye tracker (EyeLink 1000; SR Research). Eye position (1000 samples/s), neuronal activity (30,000 samples/s) and the signal from a photodiode was recorded to align neuronal responses to stimulus presentation times (30,000 samples/s) using Blackrock CerePlex hardware.

Neural responses The filtered electrical activity (bandpass 250–5000 Hz) was thresholded at 2–3% RMS value for each recording site and the threshold crossing timestamps were saved (along with the raw electrical signal, waveforms at each crossing, and other signals). Spikes were not sorted for these experiments, and 'unit' refers to the multiunit activity at each recording electrode. The stimulus-evoked firing rate of each V4 unit was calculated based on the spike count responses between 50 and 250 ms after stimulus onset to account for V4 response latency. The baseline firing rates were calculated based on the spike count responses in the 100 ms time period before the onset of the stimulus.

Neuron exclusion For each unit in an experimental session, the average stimulus-evoked responses across all stimuli were compared with the average baseline activity. The unit was included in further analyses if the average evoked activity was at least 1.1x the baseline activity. This lenient inclusion criterion was chosen because, for the chosen experimental design and stimuli, dimensionality-reduced decoding analyses are resilient to noise and

benefit from information distributed across many neurons. Each recording experiment yielded data from 90 to 95 units (mean 94.1).

Receptive field mapping A set of 2D closed contours, 3D solid objects, and black-and-white Gabor images were flashed as described above in the lower left quadrant of the screen. The positions and sizes were chosen manually across several experiments to home in on the receptive fields of each V4 recording site. Typically, a grid of 5×5 positions and two image sizes were chosen to overlap partially. The spikes were counted within a 50–250 ms window after stimulus onset, and a RF heat map was constructed for each site. The center of mass of this heat map was chosen as the center of the RF, and an ellipse was fit to circumscribe the central two standard deviations. This resulted in centers and extents of the RF of each recording site. The naturalistic image sets for the experiments described below were scaled such that the circular aperture within which the background objects were contained fully overlapped the population RF. This necessitated that the image boundary exceeded the RF of some neurons, but the image information outside of the circular aperture was held constant across images.

Experiment 1: Effect of task-irrelevant stimulus changes on the ability of V4 neurons to encode a feature of interest about the central object The first goal of the electrophysiology experiments was to determine if the information about the chosen parameter of the central object (banana position) interferes with information about distracting parameters (background object rotation and depth). To do this, we systematically varied the horizontal position of the object and the background parameters in an uncorrelated fashion. The values and ranges of the object and background parameters were customized for each monkey such that there was a differential response to each condition on average across all other conditions, i.e., a 3-way ANOVA for object position and the two background conditions all had a significant main effect ($p < 0.01$). Five values of object position, background depth, and background rotation were chosen and permuted, yielding 125 image stimuli. Further details of stimuli can be found in Fig. 1 and Supp. Figure 1, as well as the associated code and data repositories.

The data were collected in 26 recording experiments (17 sessions across 11 days from monkey 1 and 9 sessions across 8 days from monkey 2). Recording experiments with fewer than three repetitions per stimulus image were excluded. Therefore, each stimulus was presented between 3 and 16 times, yielding between 381 and 2084 presentations (mean 831).

Experiment 2: relationships between multiple object and background feature dimensions The second goal of the monkey electrophysiology was to determine whether different visual features are encoded orthogonally in neuronal population responses. Therefore, we measured responses to stimuli that varied many features of the central object (banana), including its horizontal position, depth, orientation, and two surface parameters (color and luminance). We also independently varied the same five features of the background objects (branches and leaves). We used two values for each of the ten features. We chose to make the ten features equally decodable by the population of V4 neurons (see Fig. 5). We measured responses to five repetitions of each of 2^{10}=1024 stimuli. Each stimulus image was repeated between two to three times. Because of the large dataset required for this experiment, the data analyzed in Fig. 5 were collected from one session from monkey 1.

Experiment 3: relationship between multiple object feature dimensions and their influence on behavior We repeated the fixation experiment in the same monkeys to display 3D objects that were parametrically generated and varied in up to 16 parameters. Details of shape generation have been published elsewhere[28]. Unlike in experiment 2, we did not generate all permutations of the 16 parameters; instead, we generated 50 objects with random values of those features (Supp. Figure 6a). Stimuli were repeated at least five times. We collected data from both monkeys in six and eight experimental sessions, respectively.

Please refer to the publication referred to above for details on the curvature estimation experiment. Briefly, we generated a base shape with a random set of features as above and displayed it in one of 20 values of axial curvature for 500–800 ms before displaying a 140° arc in the upper hemifield. When the fixation point disappeared, the monkey made a saccade to this arc to indicate the curvature estimate such that leftward saccades indicated straight and rightward curved reports. The monkey was rewarded based on the error in behavioral estimate. For each behavioral session, we tested up to four random base shapes simultaneously such that the shape varied in several features (including axial curvature) across trials. In subsets of sessions, we also tested a single base shape with variations in only one feature (in-plane orientation or color) across trials. Monkeys' curvature estimation behavior was not affected by trial-to-trial variations in single features or multiple features.

Human psychophysics

Apparatus A calibrated LCD color monitor (27-inch NEC MultiSync PA271Q QHD Color Critical Desktop W-LED Monitor with SpectraView Engine; NEC Display Solutions) displayed the stimuli in an otherwise dark room, after participants dark-adapted in the experimental room for a minimum of 5 min. The monitor was driven at a pixel resolution of 1920×1080, with a refresh rate of 60 Hz and with 8-bit resolution for each RGB channel. The host computer for this monitor was an Apple Macintosh with an Intel Core i7 processor. The head position of each participant was stabilized using a chin cup (Headspot, UHCOTech, Houston, TX). The participant's eyes were centered horizontally and vertically with respect to the monitor, which was 75 cm from the participant's eyes. The participant indicated their responses using a Logitech F310 gamepad controller.

Stimulus parameters The entire image subtended ~ 8° in width and height. The central object subtended ~ 4° in the longest dimension, and the circular array of background objects (branches and leaves) subtended ~ 5° of visual angle. The images were created using ISET3d at a resolution of 1920×1920 with 100 samples per pixel, at 31 equally spaced wavelengths between 400 nm and 700 nm.

Psychophysical task The psychophysical task was a two-interval forced choice (2AFC) task with one stimulus per interval. Each stimulus interval had a duration of 250 ms. Stimuli were presented at the center of the monitor. Between the two stimulus intervals, two masks were shown in succession at the center of the monitor (Fig. 5). Each mask was be presented for a duration of 400 ms, for a total interstimulus interval of 800 ms (see Session organization below for mask details). Display times are approximate as the actual display times were quantized by the hardware to integer multiples of the 16.67 ms frame rate.

The participant's task was to determine whether the central object presented in the second interval was to the left or to the right of the one presented in the first interval. Following the two intervals, the participant had an unlimited amount of time to press one of two response buttons on a gamepad to indicate their choice. Feedback was provided via auditory tones. Trials were separated by an intertrial interval of approximately one second.

The experimental programs can be found in the custom software package Natural Image Thresholds (https://github.com/AmyMNi/NaturalImageThresholds). They were written in MATLAB (MathWorks; Natick, MA) and were based on the software package Virtual World Color Constancy (github.com/BrainardLab/VirtualWorldColorConstancy). They rely on routines from the Psychophysics Toolbox (http://psychtoolbox.org) and mgl (http://justingardner.net/doku.php/mgl/overview).

Session organization The first session experimental session for each participant included participant enrollment procedures (informed consent, vision tests, etc.; see Participants above for details) as well as familiarization trials (see next paragraph) and lasted one and a half hours. The additional experimental sessions lasted approximately one hour each.

For the first session only, the participant began with 30 familiarization trials. The familiarization trials comprised, in order: 10 randomly selected easy trials (the largest position-change comparisons), 10 randomly selected medium-difficulty trials (the 4th and 5th largest position-change comparisons), and 10 randomly selected trials from all possible position-change comparisons. The familiarization trials did not include any task-irrelevant variability and data from these trials was not saved.

In each session, there were two reference positions for the object, and for each reference position there were 11 comparison positions: five comparison positions in the positive horizontal direction, five comparison positions in the negative horizontal direction, and a comparison position of 0 indicating no change. On each trial, one interval contained one of the two reference stimuli and the other interval will contain one of that reference stimulus's comparison stimuli. The order in which these two stimuli were presented within a trial was selected randomly per trial.

A block of trials consisted of presentation of the 11 comparison positions for each of the two reference positions for a total of 22 trials per block. The trials within a block were run in randomized order. Each was completed before the next block began. Each block was repeated 7 times in a run of trials, for a total of 154 trials per run.

Within each run of 154 trials, a single background variation condition was studied. There were three such conditions, as described in more detail below – "no variation", "rotation only", and "rotation and depth". Two runs for each of the three conditions was completed in each experimental session, and except as noted in the results, each subject completed 6 sessions. The six runs were conducted in random order within each session, and each run was separated by a break that lasted at least one minute and during which the participant was encouraged to stand or stretch as needed. After a minimum of one minute, the next run was initiated when the participant was ready.

Additionally, each session began with four practice trials (including in the first experimental session, where these practice trials were preceded by the familiarization trials as described). Each run after the first also started with one practice trial. The practice trials were all easy trials as described above and not include any task-irrelevant variability. The data from the practice trials was be saved. The maximum variation in background features was matched to the maximum variation in the neurophysiology experiments but sampled more finely as described for each of the variation blocks below.

For the "no variation" condition, there were not any changes to the background objects (the branches and leaves). This run determines the participant's threshold for discriminating the horizontal position of the central object without any task-irrelevant stimulus variation.

The "rotation only" run introduced task-irrelevant variability single task-irrelevant feature: rotation of the background objects. For each trial, a single rotation amount was drawn randomly from a pool of 51 rotations, and the background objects (leaves and sticks) in the stimulus were all rotated by that amount around their own centers. The rotation was drawn separately (randomly with replacement) for each of the two stimuli presented on a trial (the reference position stimulus and the comparison position stimulus). Thus, subjects had to judge the position of the central object across a change in the background, so that any effect of background variation on the positional representation of the central object would be expected to elevate threshold. The pool of 51 rotations comprised: a rotation of zero (no change to the background objects), 25 equally spaced rotations in the clockwise direction in 2° intervals, and 25 equally spaced rotation amounts in the counterclockwise direction in 2° intervals.

"Rotation and depth" runs had variation in two task-irrelevant features: rotation and depth of the background objects. For this run, there was a pool of 51 rotations, but along with the rotation of the background objects, these objects also varied in depth. There were 51 possible depth amounts (one depth amount of zero, 25 equally spaced depth amounts in the positive depth direction, and 25 equally spaced depth amounts in the negative direction; depth amounts ranged from −500 mm to 500 mm in the rendering scene space). One of the of images was a rotation of zero and a depth amount of zero. For the remaining 50 images in the pool, each of the remaining 50 rotation amounts was randomly assigned (without replacement) to one of the remaining 50 depth amounts. The

same depth shift was applied to each of the background objects. From this pool of 51 images, a single image was randomly drawn (with replacement) for each of the two stimuli presented in the trial.

Finally, as noted above (see Psychophysical task), two masks were shown per trial during the interstimulus interval. All masks across all background variation conditions were created from the same distribution of stimuli (stimuli with "no variation", thus containing no task-irrelevant noise). To create each of the two masks, first the central object positions in the first and second intervals of the trial were be determined. The two stimuli with that matched the central object positions in the first and second intervals were then used to create the trial masks. For each of these two stimuli, the average intensity was calculated in each RGB channel per 16×16 block of the stimulus. Next, each 16×16 block of a mask was randomly drawn from the mask corresponding to the two stimuli. Thus, the two masks shown per trial were each a random mixture of 16×16 blocks from stimuli with the two central object positions for that trial.

Statistical analysis and quantification

Monkey electrophysiology

Cross-validated parameter decoding (Fig. 2) First, the response matrix (multiunit spike rates for each site for each image stimulus presentation) was reduced to 10 dimensions of activity. This ensured sufficient dimensionality for decoding object and background parameters and explained between 87.8% and 94.8% (mean 91.2% across sessions) of the variance across stimulus responses. Parameter decoding without dimensionality reduction produced qualitatively similar results. Then, for each background condition (unique combination of background rotation and depth – or specific decoding), the object position in each presentation/trial was decoded from neural responses by learning regression weights from all other trials (leave-one-out cross-validation). Decoding accuracy was defined as the correlation between actual and decoded values so that perfect decoding would result in an accuracy of 1 and chance decoding accuracy of 0. We did not encounter decoding accuracies below 0. We also calculated other decoding performance measures like mean squared error, cosine distance, etc. While other measures provide more sensitivity in the specific kinds of decoding error, their estimate of aggregate performance was qualitatively similar to correlation-based measures.

The same procedure was repeated for general decoding, where the background parameters were ignored (Fig. 2b). We also trained general decoders using random subsets of trials across the dataset to match the training set of the specific decoders. Specifically, we subsampled the minimum number of trials that any of the 25 specific decoders were trained on from the full dataset regardless of background feature values, trained a position decoder, and tested it on the rest of the trials. We repeated this for 100 subsamples and averaged the decoded predictions across folds. Across folds, the performance of the trial-matched general decoders was not significantly different from either the general decoder trained on all the data at once in a leave-one-out fashion or the distribution of specific decoder performances (paired t-test, $p > 0.05$).

Error in decoding was defined as the difference between the predicted object position and the actual position (Fig. 2d). The same procedure for specific and general decoding was also repeated for each of the two background conditions (Supp. Figure 2). We also compared the linear weights of each neuron for each general decoder (Supp. Figure 3a-c). For this comparison, we normalized the range of feature values between 0 and 1 and averaged the weights for each neuron across folds. We also compared these weights to the feature sensitivity of each neuron (calculated in Supp. Figure 1d; plotted in Supp. Figure 3d-f).

To compare the performance of the specific decoders across conditions, we split the trials for each unique background condition into two sets. We trained and tested the specific decoders within (*self*) and across (*cross*) conditions and plotted the mean decoding performance difference across folds (Supp. Figure 3). We plotted these cross-decoding differences separately for background rotation and depth changes. We did not distinguish between increasing and decreasing values of background rotation or depth, i.e., a condition difference of 1 denotes that the decoder was tested on a background condition that was one away from the condition it was trained on. Since there were five levels of variation in each background condition, there were 5 pairs of conditions that were 0 levels away (self-decoder); 8 pairs, 1 away; 6 pairs, 2 away; 2 pairs, 3 away; 1 pair, 4 away. We observed slight deviations from the self-decoder as the level of variation increased.

Angle calculation (Fig. 3) To calculate the angle between the specific decoders, an n-dimensional line was fit to the dimensionality-reduced responses, and the unit vector was found. The angle between each specific decoder and the decoder for the central condition was calculated as the arc-cos of the dot product of the two unit vectors.

Linear discriminant analysis and comparison with human psychophysics (Fig. 4c) To directly compare human psychophysics discrimination accuracy with decoding results, we matched the three blocked conditions – no background variation, rotation only, and rotation and depth variation – by subsampling trials from the $5 \times 5 \times 5$ stimulus set from Experiment 1. For the three conditions, we either found all pairs of trials, pairs of trials that varied in rotation only (by holding background depth at the central value), or pairs of trials that varied in depth only (by holding background rotation at the central value). Then, for 200 folds, we sampled a maximum of 500 pairs of trials, and depending on the object position in those trials, we assigned a left or right choice. If the positions were identical, we randomly assigned the choice for that pair. We then collated the responses across the pairs of trials and a fit linear discriminant in a leave-one-out fashion to predict the correct choice. The classification prediction accuracy for each of the three blocked conditions was calculated independently.

Cross-decoding analysis (Fig. 5) For experiment 2, even though only two values were chosen for each of the five object and five background parameters, linear regression was chosen instead of classification using discriminant analysis for comparison with decoding analyses in the previous experiment. Even though each stimulus image was only repeated 2–3 times, since each parameter could take one of two values each, all unique pairs

of images would be informative about at least one parameter change. To enable cross-decoding, we altered the cross-validation procedure. For each parameter pair, for each of the 100 folds, we randomly split all image presentations evenly into training and testing sets (uneven splits also produced qualitatively similar results). We then trained a linear regression model for one parameter using the training trials and used it to predict the values of the other parameters for the held-out testing trials. The decoding accuracy was calculated as the average correlation across folds between the actual and decoded parameter values. Since each parameter decoder was trained while ignoring all other parameter variations, the diagonals in Fig. 5b are akin to the general decoder accuracy for those parameters, and the off diagonals correspond to how well those general decoders are aligned to the representations of the other parameters. The diagonal correlations were all significantly above 0 ($p < 10^{-80}$; t-test across folds), and none of the off-diagonal correlations were significant except the cross-decoding of background and object color. We repeated the same procedure for data from experiment 3 i.e., the responses to the 3D shapes from Srinath et al., 2024 for the 16 shape features. Unlike experiment 2, since the 16 features were not permuted but chosen randomly, of the 120 possible stimulus pairs, 4 pairs were correlated in the stimulus set. Therefore, the cross-decoding accuracy of the four sets of features that could be cross-decoded (Z orientation-color R at 0.24, curvature-surface gloss at 0.21, curvature-thickness X1 at 0.21, and thickness Y2-thickness Y3 at 0.23) can be trivially explained by correlations between those features in the stimulus set (0.29, 0.25, 0.43, and 0.43 respectively).

Human psychophysics (Fig. 4a-b)

Per session, the participant's threshold for discriminating object position was measured for each background variation condition. First, for each comparison position, the proportion of trials on which the participant responded that the comparison stimulus was located to the right of the reference stimulus was calculated. Next, the proportion of the comparison was chosen as rightwards was fit with a cumulative normal function using the Palamedes Toolbox (http://www.palamedestoolbox.org). To estimate all four parameters of the psychometric function (threshold, slope, lapse rate, and guess rate), the lapse rate was constrained to be equal to the guess rate and to be in the range [0, 0.05], and the maximum likelihood fit was determined. The threshold was calculated as the difference between the stimulus levels at performances equal to 0.7602 and 0.5 as determined by the cumulative normal fit.

We calculated the Bayes Factor using the MATLAB bayesFactor Toolbox[69] which calculates Bayes Factor (BF) for ANOVA designs detailed in Rouder et al.[70]. We followed the example to test the hypothesis that variations in background conditions significantly affect choices. First, we calculated the BF for the full model across subjects, i.e., using the main effect of object position and background variation conditions and the interaction effects (full model). Then, to isolate the effect of the background variation, we repeated the ANOVA while excluding background variations (restricted model). We then calculated the ratio of the BFs of the full model to the restricted models and inverted it to test for evidence of the absence of the effect. We also repeated this procedure for the monkey electrophysiology data that accompanies this analysis (Fig. 4).

Data availability

The data and code that generate the figures in this study have been deposited in a public Github repository https://github.com/ramanujansrinath/UntanglingBananas. MATLAB code for creating and displaying the images for human psychophysical experiments, as well as analyzing the raw data from these experiments, can be found at https://github.com/AmyMNi/NaturalImageThresholds. Request for further information should be directed to and will be fulfilled by the corresponding author David H. Brainard (brainard@psych.upenn.edu) in consultation with the other authors.

Received: 30 July 2024; Accepted: 31 January 2025
Published online: 13 February 2025

References

1. Martinez-Garcia, M., Bertalmío, M. & Malo, J. In praise of artifice reloaded: Caution with natural image databases in modeling vision. *Front. Neurosci.* **13**, 8 (2019).
2. Rust, N. C. & Movshon, J. A. In praise of artifice. *Nat. Neurosci.* **8**, 1647–1650 (2005).
3. Pasupathy, A. & Connor, C. E. Population coding of shape in area V4. *Nat. Neurosci.* **5**, 1332–1338 (2002).
4. Gallant, J. L., Braun, J. & Essen, D. V. Selectivity for polar, hyperbolic, and cartesian gratings in macaque visual cortex. *Science* **259**, 100–103 (1993).
5. Leopold, D. A. & Logothetis, N. K. Activity changes in early visual cortex reflect monkeys' percepts during binocular rivalry. *Nature* **379**, 549–553 (1996).
6. Peterhans, E. & von Heydt, R. Der. Subjective contours—Bridging the gap between psychophysics and physiology. *Trends Neurosci.* **14**, 112–119 (1991).
7. von der Heydt, R., Peterhans, E. & Baumgartner, G. Illusory contours and cortical neuron responses. *Science* **224**, 1260–1262 (1984).
8. Snow, J. C. & Culham, J. C. The treachery of images: How realism influences brain and behavior. *Trends Cogn. Sci.* **25**, 506–519 (2021).
9. Peters, B. & Kriegeskorte, N. Capturing the objects of vision with neural networks. *Nat. Hum. Behav.* **5**, 1127–1144 (2021).
10. Ruff, D. A., Ni, A. M. & Cohen, M. R. Cognition as a window into neuronal population space. *Annu. Rev. Neurosci.* **41**, 77–97 (2018).
11. Cadieu, C. et al. A model of V4 shape selectivity and invariance. *J. Neurophysiol.* **98**, 1733–1750 (2007).
12. Oleskiw, T. D., Nowack, A. & Pasupathy, A. Joint coding of shape and blur in area V4. *Nat. Commun.* **9**, 466 (2018).
13. Kim, T., Bair, W. & Pasupathy, A. Neural coding for shape and texture in Macaque Area V4. *J. Neurosci.* **39**, 4760–4774 (2019).
14. Yamane, Y. et al. Population coding of figure and ground in natural image patches by V4 neurons. *PLoS ONE* **15**, e0235128 (2020).
15. Srinath, R. et al. Early emergence of solid shape coding in natural and deep network vision. *Curr. Biol.* **31**, 51–65e5 (2021).

16. Hatanaka, G. et al. Processing of visual statistics of naturalistic videos in macaque visual areas V1 and V4. *Brain Struct. Funct.* **227**, 1385–1403 (2022).
17. Hubel, D. H. & Wiesel, T. N. Receptive fields and functional architecture of monkey striate cortex. *J. Physiol.* **195**, 215–243 (1968).
18. Born, R. T. & Tootell, R. B. Spatial frequency tuning of single units in macaque supragranular striate cortex. *Proc. Natl. Acad. Sci.* **88**, 7066–7070 (1991).
19. Nauhaus, I., Nielsen, K. J., Disney, A. A. & Callaway, E. M. Orthogonal micro-organization of orientation and spatial frequency in primate primary visual cortex. *Nat. Neurosci.* **15**, 1683–1690 (2012).
20. Everson, R. M. et al. Representation of spatial frequency and orientation in the visual cortex. *Proc. Natl. Acad. Sci.* **95**, 8334–8338 (1998).
21. Kohn, A. et al. Principles of corticocortical communication: Proposed schemes and design considerations. *Trends Neurosci.* **43**, 725–737 (2020).
22. Vyas, S., Golub, M. D., Sussillo, D. & Shenoy, K. V. Computation through neural population dynamics. *Annu. Rev. Neurosci.* **43**, 249–275 (2020).
23. Okazawa, G., Hatch, C. E., Mancoo, A., Machens, C. K. & Kiani, R. Representational geometry of perceptual decisions in the monkey parietal cortex. *Cell* **184**, 3748-3761.e18 (2021).
24. Misaki, M., Kim, Y., Bandettini, P. A. & Kriegeskorte, N. Comparison of multivariate classifiers and response normalizations for pattern-information fMRI. *NeuroImage* **53**, 103–118 (2010).
25. Hong, H., Yamins, D. L. K., Majaj, N. J. & DiCarlo, J. J. Explicit information for category-orthogonal object properties increases along the ventral stream. *Nat. Neurosci.* **19**, 613–622 (2016).
26. Reynolds, D. & Singh, V. Characterization of human lightness discrimination thresholds for independent spectral variations. 06.16.545355 Preprint at (2023). https://doi.org/10.1101/2023.06.16.545355 (2023).
27. Singh, V., Burge, J. & Brainard, D. H. Equivalent noise characterization of human lightness constancy. *J. Vis.* **22**, 2 (2022).
28. Srinath, R., Czarnik, M. M. & Cohen, M. R. Coordinated Response Modulations Enable Flexible Use of Visual Information. 07.10.602774 Preprint at (2024). https://doi.org/10.1101/2024.07.10.602774 (2024).
29. Bashivan, P., Kar, K. & DiCarlo, J. J. Neural population control via deep image synthesis. *Science* **364**, eaav9436 (2019).
30. Cowley, B. R., Stan, P. L., Pillow, J. W. & Smith, M. A. Compact deep neural network models of visual cortex. 11.22.568315 Preprint at (2023). https://doi.org/10.1101/2023.11.22.568315 (2023).
31. Pospisil, D. A., Pasupathy, A. & Bair, W. Artiphysiology' reveals V4-like shape tuning in a deep network trained for image classification. *eLife* **7**, e38242 (2018).
32. Kramer, L. E., Chen, Y. C., Long, B., Konkle, T. & Cohen, M. R. Contributions of early and mid-level visual cortex to high-level object categorization. 05.31.541514 Preprint at (2023). https://doi.org/10.1101/2023.05.31.541514 (2023).
33. Majaj, N. J., Hong, H., Solomon, E. A. & DiCarlo, J. J. Simple learned weighted sums of inferior temporal neuronal firing rates accurately predict human core object recognition performance. *J. Neurosci.* **35**, 13402–13418 (2015).
34. Chung, S., Lee, D. D. & Sompolinsky, H. Classification and geometry of general perceptual manifolds. *Phys. Rev. X* **8**, 031003 (2018).
35. Cohen, U., Chung, S., Lee, D. D. & Sompolinsky, H. Separability and geometry of object manifolds in deep neural networks. *Nat. Commun.* **11**, 746 (2020).
36. Chung, S., Lee, D. D. & Sompolinsky, H. Linear readout of object manifolds. *Phys. Rev. E* **93**, 060301 (2016).
37. Ni, A. M., Huang, C., Doiron, B. & Cohen, M. R. A general decoding strategy explains the relationship between behavior and correlated variability. *eLife* **11**, e67258 (2022).
38. Tsao, T. & Tsao, D. Y. A topological solution to object segmentation and tracking. *Proc. Natl. Acad. Sci. U. S. A.* **119**, e2204248119 (2022).
39. Luongo, F. J. et al. Mice and primates use distinct strategies for visual segmentation. *eLife* **12**, e74394 (2023).
40. DiCarlo, J. J. & Cox, D. D. Untangling invariant object recognition. *Trends Cognit. Sci.* **11**, 333–341 (2007).
41. Rust, N. C. & Dicarlo, J. J. Selectivity and tolerance ('invariance') both increase as visual information propagates from cortical area V4 to IT. *J. Neurosci.* **30**, 12978–12995 (2010).
42. DiCarlo, J. J., Zoccolan, D. & Rust, N. C. How does the brain solve visual object recognition? *Neuron* **73**, 415–434 (2012).
43. Pagan, M., Urban, L. S., Wohl, M. P. & Rust, N. C. Signals in inferotemporal and perirhinal cortex suggest an untangling of visual target information. *Nat. Neurosci.* **16**, 1132–1139 (2013).
44. Yamins, D. L. K. et al. Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proc. Natl. Acad. Sci. U. S. A.* **111**, 8619–8624 (2014).
45. Hénaff, O. J., Goris, R. L. T. & Simoncelli, E. P. Perceptual straightening of natural videos. *Nat. Neurosci.* **22**, 984–991 (2019).
46. Kriegeskorte, N. & Wei, X. X. Neural tuning and representational geometry. *Nat. Rev. Neurosci.* **22**, 703–718 (2021).
47. Fusi, S., Miller, E. K. & Rigotti, M. Why neurons mix: High dimensionality for higher cognition. *Curr. Opin. Neurobiol.* **37**, 66–74 (2016).
48. Rigotti, M. et al. The importance of mixed selectivity in complex cognitive tasks. *Nature* **497**, 585–590 (2013).
49. Saxena, S. & Cunningham, J. P. Towards the neural population doctrine. *Curr. Opin. Neurobiol.* **55**, 103–111 (2019).
50. Chung, S. & Abbott, L. F. Neural population geometry: An approach for understanding biological and artificial neural networks. *Curr. Opin. Neurobiol.* **70**, 137–144 (2021).
51. Jazayeri, M. & Ostojic, S. Interpreting neural computations by examining intrinsic and embedding dimensionality of neural activity. *Curr. Opin. Neurobiol.* **70**, 113–120 (2021).
52. Yuste, R. From the neuron doctrine to neural networks. *Nat. Rev. Neurosci.* **16**, 487–497 (2015).
53. Stringer, C., Pachitariu, M., Steinmetz, N., Carandini, M. & Harris, K. D. High-dimensional geometry of population responses in visual cortex. *Nature* **571**, 361–365 (2019).
54. Rajalingham, R. et al. Large-scale, high-resolution comparison of the core visual object recognition behavior of humans, monkeys, and state-of-the-art deep artificial neural networks. *J. Neurosci.* **38**, 7255–7269 (2018).
55. Ehrlich, D. B. & Murray, J. D. Geometry of neural computation unifies working memory and planning. *Proc. Natl. Acad. Sci.* **119**, e2115610119 (2022).
56. Mante, V., Sussillo, D., Shenoy, K. V. & Newsome, W. T. Context-dependent computation by recurrent dynamics in prefrontal cortex. *Nature* **503**, 78–84 (2013).
57. Bernardi, S. et al. The geometry of abstraction in the hippocampus and prefrontal cortex. *Cell* **183**, 954–967e21 (2020).
58. Russo, A. A. et al. Motor Cortex embeds muscle-like commands in an untangled population response. *Neuron* **97**, 953–966e8 (2018).
59. Gallego, J. A. et al. Cortical population activity within a preserved neural manifold underlies multiple motor behaviors. *Nat. Commun.* **9**, 4233 (2018).
60. Sussillo, D., Churchland, M. M., Kaufman, M. T. & Shenoy, K. V. A neural network that finds a naturalistic solution for the production of muscle activity. *Nat. Neurosci.* **18**, 1025–1033 (2015).
61. Shenoy, K. V., Sahani, M. & Churchland, M. M. Cortical control of arm movements: A dynamical systems perspective. *Annu. Rev. Neurosci.* **36**, 337–359 (2013).
62. Khaligh-Razavi, S. M. & Kriegeskorte, N. Deep supervised, but not unsupervised, models may explain IT cortical representation. *PLoS Comput. Biol.* **10**, e1003915 (2014).
63. Day, R. H. & Dickinson, R. G. The components of the Poggendorff Illusion. *Br. J. Psychol.* **67**, 537–552 (1976).

64. Cavanagh, P. & Anstis, S. The flash grab effect. *Vis. Res.* **91**, 8–20 (2013).
65. Zhou, H., Friedman, H. S. & Von Der Heydt, R. Coding of border ownership in monkey visual cortex. *J. Neurosci.* **20**, 6594–6611 (2000).
66. Maheswaranathan, N. et al. Interpreting the retinal neural code for natural scenes: From computations to neurons. *Neuron* **111**, 2742–2755e4 (2023).
67. Ding, X. et al. Information geometry of the retinal representation Manifold. *bioRxiv* 2023.05.17.541206 https://doi.org/10.1101/2023.05.17.541206 (2023).
68. Felsen, G. & Dan, Y. A natural approach to studying vision. *Nat. Neurosci.* **8**, 1643–1646 (2005).
69. Krekelberg, B. klabhub/bayesFactor: Bayes only. Zenodo (2024). https://doi.org/10.5281/zenodo.13744717
70. Rouder, J. N., Morey, R. D., Speckman, P. L. & Province, J. M. Default Bayes factors for ANOVA designs. *J. Math. Psychol.* **56**, 356–374 (2012).

Acknowledgements

We are grateful to K. McKracken for providing technical assistance and to Douglas Ruff, Cheng Xue, and Lily Kramer for comments on an earlier version of this manuscript and suggestions regarding data analysis. This work is supported by Eric and Wendy Schmidt AI in Science Postdoctoral Fellowship, a Schmidt Sciences, LLC program (to R.S.), the Simons Foundation (Simons Collaboration on the Global Brain award 542961SPI to M.R.C, postdoctoral fellowship to A.M.N.), the National Institutes of Health (awards R01EY022930, R01EY034723, and RF1NS121913 to M.R.C, K99NS118117 to A.M.N, K99EY035362 to R.S.)

Author contributions

Conceptualization - A.M.N., M.R.C., D.H.B.; Methodology - R.S., A.M.N., M.R.C., D.H.B.; Software - R.S., A.M.N., D.H.B; Formal Analysis - R.S., A.M.N., C.M. M.R.C., D.H.B.; Investigation - R.S., A.M.N., C.M.; Data curation - R.S., A.M.N., C.M.; Writing – Original Draft - R.S., M.R.C., D.H.B. ; Writing – Review & Editing - R.S., M.R.C., D.H.B.; Visualization - R.S.; Supervision - M.R.C., D.H.B.; Project Administration - M.R.C., D.H.B. ; Funding acquisition - R.S., A.M.N., M.R.C., D.H.B.

Declarations

Competing interests

The authors declare no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-025-88910-8.

Correspondence and requests for materials should be addressed to D.H.B.