What response properties do individual neurons need to underlie position and clutter "invariant" object recognition?

Nuo Li[1], David D. Cox[1,2], Davide Zoccolan[1,2,3], and James J. DiCarlo[1]*

[1] McGovern Institute for Brain Research and Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139

[2] The Rowland Institute at Harvard, Harvard University, Cambridge, Massachusetts 02142

[3] Neurobiology and Cognitive Neuroscience Sectors, International School for Advanced Studies (SISSA), 34014 Trieste, Italy

Number of figures: 7
Number of tables: 1

Abbreviated title: How do inferior temporal cortical neurons cope with clutter?

Keywords: object recognition, inferior temporal cortex, population code, classifier, translation, clutter, invariance

*Address for correspondence:

Dr. James DiCarlo (dicarlo@mit.edu)
McGovern Institute for Brain Research,
Massachusetts Institute of Technology,
77 Massachusetts Avenue,
Cambridge, MA 02139.

Abstract

Primates can easily identify visual objects over large changes in retinal position – a property commonly referred to as position “invariance”. This ability is widely assumed to depend on neurons in inferior temporal cortex (IT) that can respond selectively to isolated visual objects over similarly large ranges of retinal position. However, in the real world, objects rarely appear in isolation, and the interplay between position invariance and the representation of multiple objects (i.e. clutter) remains unresolved. At the heart of this issue is the intuition that the representations of nearby objects can interfere with one another, and that the large receptive fields needed for position invariance can exacerbate this problem by increasing the range over which interference acts. Indeed, most IT neurons’ responses are strongly affected by the presence of clutter. While external mechanisms (such as attention) are often invoked as a way out of the problem, we show (using recorded neuronal data and simulations) that the intrinsic properties of IT population responses, by themselves, can support object recognition in the face of limited clutter. Furthermore, we carried out extensive simulations of hypothetical neuronal populations to identify the essential individual-neuron ingredients of a good population representation. These simulations show that the crucial neuronal property to support recognition in clutter is not preservation of response magnitude, but preservation of each neuron’s rank-order object preference under identity-preserving image transformations (e.g. clutter). Since IT neuronal responses often exhibit that response property, while neurons in earlier visual areas (e.g. V1) do not, we suggest that preserving the rank-order object preference regardless of clutter, rather than the response magnitude, more precisely describes the goal of individual neurons at the top of the ventral visual stream.

Introduction

Primate brains have the remarkable ability to recognize visual objects across the wide range of retinal images that each object can produce – a property known as "invariance" or "tolerance" (see Discussion). To accomplish this task, the visual system must transform the object shape information acquired as a pixel-like image by the retina into a neuronal representation that is unaffected by identity-preserving changes in the image (due to variation in the object's position, size, pose, its illumination conditions, or the presence of other objects, i.e. "clutter"). This transformation is carried out along the hierarchal processing stages of the ventral visual stream that culminates in the inferior temporal (IT) cortex (Hung et al. 2005; Logothetis and Sheinberg 1996; Tanaka 1996).

Representation of multiple objects poses an especially difficult computational challenge. During natural vision, objects almost never appear in isolation and they appear on very different parts of the retina. This introduces two common identity-preserving image variations that our visual system must simultaneously deal with to recognize each object: 1) variability in object position and 2) the presence of visual clutter. Understanding the brain's solution to this problem is complicated by two observations. First, contemporary data reveal highly varied amounts of position sensitivity in individual IT neurons — each neuron's response magnitude can be strongly modulated by changes in object position; (Ito et al. 1995; Op de Beeck and Vogels 2000; Zoccolan et al. 2007), with IT receptive fields often spanning only a few degrees of visual angle (DiCarlo and Maunsell 2003). Second, IT neuronal responses to isolated objects are often highly sensitive to clutter – responses are powerfully reduced by the addition of other objects (Chelazzi et al. 1998b; Miller et al. 1993; Missal et al. 1999; Rolls et al. 2003; Rolls and Tovee 1995; Sato 1989; Sheinberg and Logothetis 2001; Zoccolan et al. 2005; Zoccolan et al. 2007), in some cases by as much as 50%.

In spite of these coding constraints at the neuronal level, humans and primates can effortlessly identify and categorize objects in natural scenes. This raises the question of what mechanisms allow the ventral stream to support position-invariant recognition in clutter. One possible explanation to deal with position invariance relies on the observation that IT neurons typically maintain their rank-order object selectivity within their receptive fields, even when the magnitude of their responses is strongly modulated by changes in object position (DiCarlo and Maunsell 2003; Ito et al. 1995; Logothetis and Sheinberg 1996; Op de Beeck and Vogels 2000; Tovée et al. 1994). Several authors have proposed that this property may allow a population of IT neurons to support position-invariant recognition (e.g. Gross et al. 1993; Logothetis and Sheinberg 1996; Vogels and Orban 1996). This is a reasonable but untested hypothesis, since no study has investigated whether preservation of object preference across position is sufficient to support position-invariant recognition. More importantly, the previous intuition applies to objects presented in isolation and may not extrapolate to more natural conditions in which multiple objects are present within a neuron's receptive field (i.e. clutter). In fact, several studies have proposed that additional mechanisms may be necessary to filter out the interference of clutter – e.g., shrinking of IT neurons' receptive fields (Rolls et al. 2003), or recruitment of attentional mechanisms to attenuate the suppressive effect of flanking objects (Chelazzi et al. 1998a; Moran and Desimone 1985; Sundberg et al. 2009).

In this study, we first asked if the intrinsic response properties of a small population of IT neurons (i.e. earliest part of response, no attentional cuing) could by themselves support object identification while tolerating some degree of clutter. Previous studies have shown that linear read-out of IT population can support position invariant recognition of isolated objects (Hung et al. 2005). Using similar techniques, we found that the IT population as a whole can readily support position-invariant recognition even when multiple objects are present (i.e., limited clutter).

These neuronal results demonstrate that clutter invariant recognition can be achieved through fast, feed-forward read-out of the IT neuronal representation (at least for limited clutter), and it led us to reconsider what individual-neuron response properties allowed IT to underlie such invariant object recognition from a population perspective. To do this, we simulated a wide range of potential neuronal populations with the goal of separating out the essential single-neuron ingredients of a "good" representation from those that are superfluous. We found that preservation of response magnitude in the face of position change (i.e., neurons with large receptive fields) or in the face of clutter – properties that individual IT neurons typically lack – are not necessary to robustly represent multiple objects in a neuronal population. Moreover, the lack of position-sensitivity in response magnitude can be detrimental in that it limits the flexibility of the representation to convey the necessary object position information to un-ambiguously represent multiple objects. Instead, we show that a much more important requirement is that individual neurons preserve their rank-order object selectivity across object position changes and clutter conditions. Indeed, IT neurons typically exhibit such a property, even when their response magnitude is highly sensitive to position and clutter (Brincat and Connor 2004; Ito et al. 1995; Logothetis and Sheinberg 1996; Zoccolan et al. 2005), while neurons in early visual areas (e.g. V1) do not.

Overall, these findings provide the first systematic demonstration of the key role played by preservation of rank-order selectivity in supporting invariant recognition – a notion that has been previously suggested (e.g. Gross et al. 1993; Logothetis and Sheinberg 1996; Vogels and Orban 1996), but never tested by decoding either recorded or simulated neuronal populations. More importantly, these results show that, at least under some conditions, clutter invariant recognition can be achieved through fast, feed-forward read-out of the IT neuronal representation, thus challenging the view that position-invariant recognition in clutter must be attained through attentional feedback,

Materials and Methods

Physiological recording

We recorded from well-isolated neurons in anterior IT in two rhesus macaque monkeys. Surgical procedures, eye monitoring, and recording methods were done using established techniques (DiCarlo and Maunsell 2000; Zoccolan et al. 2005), and were performed in accordance with the MIT Committee on Animal Care.

Visual stimulus displays ("scenes") consisted of combinations of three possible objects (star, triangle and cross shapes; 1.5 degree in size; solid white 57 Cd/m^2) that could appear in three possible locations (at the center of gaze, 2^o above, and 2^o below) on a uniform gray background (27 Cd/m^2; see Fig. 1). All combinations of:

a) one object in each possible position (9 scenes),

b) two objects (without duplicates, 18 scenes),

c) three objects (with no object repeated in the same scene, 6 scenes)

(33 scenes in total) were presented to the passively fixating monkeys with no attentional cuing to any object or retinal position. The scenes were presented at a rapid, but natural viewing rate (5 scenes/sec, 100 ms presentation followed by 100 ms blank; DiCarlo and Maunsell 2003), and randomly interleaved. For these reasons, as well as our previous detailed assessment of this issue (Zoccolan et al. 2005), we argue that attentional shifts do not contribute significantly to the results presented here.

Both monkeys had been previously trained to perform an identification task with the three objects appearing randomly interleaved in each of the three positions (in isolation), and both monkeys achieved greater than 90% accuracy in this task. Monkeys performed this identification task while we advanced the electrode, and all isolated neurons that were responsive during this task (t-test; $p<0.05$) were further studied with the 33 scenes under the fixation conditions described above. Between 10 and 30 repetitions of each scene were presented while recording from each IT neuron.

A total of 68 neurons were serially recorded (35 cells in monkey 1 and 33 in monkey 2). We took these units to be a reasonably unbiased sample of the IT population in that we only required good isolation and responsiveness. Because each of these neurons was tested with multiple repetitions of the exact same set of visual scenes, we could estimate the IT population response to each 100 ms glimpses of a scene by randomly drawing the response of each neuron during one presentation of that scene, (note that this procedure cannot optimize for any trial-by-trial correlation in the responses, see Discussion and Hung et al. 2005).

Data analysis

All analyses and simulations were done using in-house code developed in Matlab (Mathworks, Natick, MA) and publicly available Matlab SVM toolbox (http://www.isis.ecs.soton.ac.uk/isystems/kernel). We used classification analysis to assess neuronal population performance on two recognition tasks: 1) the "position-invariant" object recognition task and 2) the "position-specific" object recognition task, (see Fig. 1A). In its general form, classification analysis takes labeled multivariate data belonging to two classes (e.g. "The star is present" and "The star is not present") and seeks a decision boundary that best separates the two classes. Our goal was to measure the "goodness" of a neuronal population at conveying information that can be accessed by downstream areas using simple linear read-out mechanisms. Thus, we used linear discriminant analysis as a simple unbiased way of asking that question (Fisher 1936). Because each linear discriminant simply performs a weighted sum with a threshold (Gochin 1994), the use of linear classifiers allows us to assess what information in a neuronal population can be directly extracted by pooling mechanisms that roughly parallel those available to real downstream neurons. In other words, linear classifiers do not provide a total measure of information in the population, but instead provide a measure of the information explicitly available in the IT population to directly support a visual task (i.e. information available to a linear decoder).

Because each task had more than two possible answers (e.g. "Which of the *three* objects was present?"), overall performance was assessed using standard multi-classification methods in which multiple two-way linear classifiers were constructed (Hung et al. 2005; Rifkin et al. 2007); see below for details). Each two-way linear classifier had the form:

$$f(\mathbf{x}) = \mathbf{w}^{\mathbf{T}}\mathbf{x} + b \qquad (1)$$

where the classifier reported "object present" for $f(\mathbf{x}) \geq 0$ and "object not present" for $f(\mathbf{x}) < 0$. $\mathbf{x}$ is an N-dimensional column vector containing the responses of N neurons in a population to a given presentation (i.e., in a given trial) of a particular scene (spike counts in a small time window for real neurons or simulated response rates for simulated neurons). $\mathbf{w}$ is a N-dimensional column vector of weights, b is a constant threshold that, together, describe the position and orientation of the decision boundary. $\mathbf{w}$ and b were found using the standard method of Fisher linear discriminant using neuronal response data from a labeled training set (Duda et al. 2001). Performance testing was always carried out using data that was not included in the training set (data partitioning for training and testing is described in sections below).

$$\mathbf{w} = \hat{\mathbf{S}}^{-1}(\hat{\mu}_1 - \hat{\mu}_2) \qquad b = \frac{1}{2}(\hat{\mu}_1 + \hat{\mu}_2)\hat{\mathbf{S}}^{-1}(\hat{\mu}_2 - \hat{\mu}_1)$$

where

$$\hat{\mu}_i = \frac{1}{N_i}\sum_{n=1}^{N_i}\mathbf{x}_{i,n} \qquad \hat{\mathbf{S}} = \frac{1}{N_1 + N_2}\sum_{i=1}^{2}\sum_{n=1}^{K_i}(\mathbf{x}_{i,n} - \hat{\mu}_i)(\mathbf{x}_{i,n} - \hat{\mu}_i)^{\mathbf{T}}$$

$\hat{\mu}_1$ and $\hat{\mu}_2$ are the mean of all the training data belonging to each of the two classes ($\mathbf{x}_1$'s and $\mathbf{x}_2$'s) and $\hat{\mathbf{S}}$ is the total within-class covariance matrix (Fisher linear discriminant analysis assumes that the data belonging to two classes are identically distributed, $\mathbf{S}_1$=$\mathbf{S}_2$=$\hat{\mathbf{S}}$). K_i is the number of data points in each class used for classifier training.

How well the classifier learns the decision boundary from training data can impact classification performance -- more training data can lead to better estimate of the decision boundary and more advanced classifiers such as a support vector machines (SVM, Duda et al. 2001) are better at finding the optimal decision boundary. However, for the results presented here, linear classifier performance is almost entirely dependent on how well the data are formatted. (That is, how linearly separable are the two classes?) This was verified by using SVM classifiers in some tested conditions. Results obtained were qualitatively unaffected: SVM led to slightly better absolute performance, but the relative performance for the key comparisons was unaffected. Thus here, we equate goodness of a representation for a recognition task with linear separability of the data with respect to that task, and our methods are designed to measure this.

Recognition task performance of the real IT population

For each recorded IT neuron, we computed spike counts over the time window from 100 to 200 ms following the presentation onset of each scene. The start of this time window was based on the well-known latency of IT neurons (Baylis and Rolls 1987). The end of the window is well below the reaction times of the monkeys when performing an identification task with these objects (DiCarlo and Maunsell 2000), and is thus consistent with an integration window that could, in principle, be used by downstream neurons to support recognition. Previous work has shown that, although the length of this window can have small quantitative effects on performance, the ability of the IT population to support categorization and identification tasks using different portions of this window is qualitatively similar (Hung et al. 2005).

In the "*position-invariant task*", three binary linear classifiers (above) were trained to report if their particular object (e.g. “star”) was present or not in any position (i.e. one classifier for each of the three objects). The reported performance in the recognition task was the average performance across all three classifiers (Fig. 1B). In the "*position-specific task*", a binary classifier was trained to report if a particular object was present or not at a particular position (e.g. “star in the top position”). A total of nine such classifiers were built (3 objects x 3 positions), and the reported performance in the task was the average performance across all nine classifiers (Fig. 1B). Since each classifier was binary, chance performance for each was 50%.

The performance of each binary classifier was determined using leave-one-out cross-validation. For each question (e.g., of the sorts in Fig. 1A), the classifier performance was evaluated as following: spike-counts of individual neurons to a given scene were randomly drawn (without replacement) from the recorded set of presentations (trials) and were used to assemble a “single-trial” population response vector for that scene. Any scene presentations from one neuron could “go with” any particular scene presentation from another neuron. The final data set was obtained by repeating this procedure 10 times for each scene, yielding a labeled M × N matrix, where N is the number of neurons and M is the number of trials (10) times the number of visual scenes (33) that were presented (i.e., M = 330). Once the response matrix was created, we carried out classification (training and testing) on that matrix. Specifically, in every round of classification, we first left out one population responses vector (one row in the response matrix) for testing, the remaining trials were used to train the classifier. We repeat this procedure 330 times such that every trial (row) in the response matrix was tested once. Finally, the overall mean classifier performance and its standard error (obtained by bootstrap re-sampling) across all questions for a task were reported in Fig. 1B.

Recognition task performance of hypothetical neuronal populations

To explore hypothetical single-unit response properties for supporting the two recognition tasks, we created an abstract stimulus space that captured the essence of the recognition tasks, and allowed us to succinctly specify the responses of IT neurons in accordance with previous empirical results and variations of those results. Specifically, the abstract stimulus space has two continuous dimensions that formed the two axes of the space (object identity, $s \in [-1.0, 1.0]$; object position, $p \in [-1.0, 1.0]$) and provides a graphical perspective on the nature of the recognition tasks (Fig 2A). In this space, a single point represents a visual "scene" containing a single object. To establish a recognition task that is comparable to what was tested in the real IT population (above), three objects (A, B, C) and three positions (X, Y, Z) were selected, indicated by the nine square regions evenly placed as a 3x3 grid in this stimulus space (see Fig. 2A left column). We then generated a large class of hypothetical neuronal populations to differently represent this stimulus space (see below for detail), such that we could evaluate and compare them in the exact same recognition task with the goal of separating out the essential single-neuron ingredients of a "good" representation.

To determine the performance of a hypothetical population on a given recognition task, the following four steps were carried out in each simulation "run":

1) construct a population with particular single-unit parameters (our key independent variables),
2) simulate the population responses (i.e. the vectors $\mathbf{x}$, Eq. (1)) to a set of labeled stimulus "scenes",
3) use these responses to build classifiers for the recognition task (i.e. find $\mathbf{w}$ and b, Eq. (1)),
4) test the performance of those classifiers on the recognition task using an independent set of stimulus "scenes".

Because of variability in each simulated population and its responses (described below) as well as variability in the exact test stimuli, performance was reported as the mean and standard deviation of at least 15 such "runs" (in practice, variation in performance across runs was almost entirely the result of variability in the make-up of each population). Given a recognition task, the key manipulation was step 1 – the selection of single unit properties to construct a population. The details of steps 2-4 are described next; the details of step 1 are specific to the different types of population we simulated (IT, V1, "abstract") and are described at the end of the Methods.

For the "*position-invariant task*", we built three binary classifiers (one for each of the three objects; "A/not-A", "B/not-B", "C/not-C"). Correct performance with each visual "scene" required that all three classifiers were correct, regardless of how many objects were present. For example, if the "scene" consisted of only object A, the "A/not-A" classifier must report "yes", and the "B/not-B" and "C/not-C" classifiers must report "no", regardless of the object A's position. For the "*position-specific task*", we built three binary classifiers, one for each of the three objects at a given position (e.g. "A/not-A" at position X, “B/not-B” at position X, “C/not-C” at position X). If the “scene” did not contain any object at position X, all three classifiers must report “no”. For each classification task, the chance performance was established from "shuffle" runs, in which we tested the classifiers after having randomly shuffled the labeling of the training data. We ran a corresponding shuffle run for all the simulation runs and we plotted “shuffle” performance as the average of these runs.

In our simulations, we assessed the performance in recognition tasks with and without the presence of clutter. That is, we considered both the simple case in which all “scenes” contained only a single object, and the more natural case in which some “scenes” contained more than one object. Specifically, for the simulations “without clutter”, the labeled training data was 3000 single-object “scenes” (3000 points randomly selected from within the 9 square regions of the 2D stimulus space, see Fig 2A) and the test data was 300 single-object “scenes” randomly selected in the same manner. For the simulations “with clutter”, the labeled training data was a mixture of 1000 single-object “scenes”, 1000 two-object “scenes”, and 1000 three-object “scenes” (we ensured that no two objects occupied a single position), and the test data was a mixture of 100 single-object scenes, 100 two-object scenes, and 100 three-object scenes randomly selected in the same manner.

In summary, the testing of each hypothetical population on each recognition task ("*position-invariant task*" or "*position-specific task*") consisted of at least 15 simulation runs. For each run, a new population of neurons was randomly sampled from a prescribed distribution of single-unit response properties (details of these are described below). A set of classifiers was then trained and tested on the recognition tasks (e.g. Fig 2A). All performance was reported as the mean and standard deviation of the 15 runs.

Note that here, we are simply interested in investigating how well a representation can support the recognition tasks free of the limitations from the classifier training (e.g., learning from sparse training data). Therefore, we trained the classifiers using all the position and clutter conditions (including the conditions the classifier would be tested on later), and asked how well a representation could possibly support a task given all the benefits of experience. This approach sets an upper bound on the goodness of a representation, but does not address how well a representation allows the classifier to generalize outside the realm of its experience (see Discussion).

Simulating hypothetical neuronal populations

Each hypothetical population consisted of N single "neurons" (N was varied for some simulations, see Results) where we specified each neuron's response (*R*) to the visual scene (*v*) using a response function (*H*), a small non-zero response baseline (*c*), and trial-by-trial response variability (*Noise*).

$$R(v) = H(v) + c + Noise(v) \qquad (2)$$

Our main goal was to understand how differences in single unit response functions (*H*) lead to differences in population performance. The form of *H*(*v*) for IT, V1 and "abstract" populations is given below, as well as how it was varied (e.g. different hypothetical IT populations). The absolute value of *H*(*v*) is not important except insofar as it relates to the magnitude of *Noise*(*v*), which was proportional to *H*(*v*) (see below). In practice, each neuron's response function *H*(*v*) was scaled so that one of the single object conditions produced the maximum value of 1.0, and c was always set to 0.1.

A noise term was included in Eq. (2) to make the simulations roughly consistent with noise levels seen in spiking neurons. However, our goal was to achieve an understanding that was largely robust to the details of the spiking noise model. Since spike counts of real neurons are approximately Poisson (Shadlen and Newsome 1998; Tolhurst et al. 1983), we simply assumed that the response variability was proportional to the mean of the response. In practice, the *Noise*(*v*) in equation (1) was drawn from a normal distribution with mean zero and variance proportional to the neuron's response. That is:

$$Noise(v) \sim N(0, \rho \cdot (H(v) + c))$$

Thus, the response, *R*(*v*), of each unit approximates the averaged responses from a pool of *m* Poisson neurons, where ρ is smaller for larger *m*. Responses were cut off at zero. For all the simulation results presented in this paper, we set ρ to 0.25, such that each simulated neuron approximated the averaged responses from four Poisson neurons. Not surprisingly, the noise magnitude relative to the signal (ρ) and the number of neurons (*N*) in a population both had strong effects on *absolute* performance of simulated populations. The strategy of all our simulations was to hold these two parameters constant at reasonable values while varying the more interesting single-unit properties of the population. Indeed, we found that, other than floor and ceiling effects, changing the magnitude of ρ and N did not change the *relative* performance of any two populations (i.e. the key measure in our study).

Simulated IT populations

We simulated IT-like neuronal responses by first defining how a neuron responds to single objects (the condition for which the most data exists in the literature), and then defining how the responses to single objects are combined ("clutter rules"). We note that these IT "models" are not complete models (because they do not describe the response of each IT neuron to any possible real-world image), but are functional quantitative *descriptions* of IT neurons based on existing results (see Discussion).

The response to single objects was modeled using a 2D-Gaussian centered somewhere in the 2D stimulus space (Fig 2B), and we assumed independent tuning for shape and position. Though we assumed Gaussian tuning, our main results were qualitatively robust to this assumption (e.g. see Fig 5). Thus, each simulated neuron's response function (H) to single objects (single points in the 2D-stimulus space (s,p)) was:

$$H(v) = H(s,p) = G(\mu_s,\sigma_s) \cdot G(\mu_p,\sigma_p)$$

where G is a Gaussian profile. For each simulation run, each neuron's parameters were drawn as follows: the Gaussian center location (μ_s, μ_p) was randomly assigned within the stimulus space according to a uniform distribution. σ_s specified the standard deviation of a neuron's Gaussian tuning along the object identity axis and we will refer to it as the neurons' object (shape) selectivity. In all results presented in the main text, σ_s was kept constant at 0.3 (except in Fig 5, "IT" units $\sigma_s = 0.2$). σ_p specified the width of a neuron's tuning along the position axis. Therefore, the position-sensitivity, i.e. receptive field (RF) size, of all individual neurons could be manipulated by varying σ_p. In the reported results, each population had a single value of σ_p (i.e. the position sensitivity of all neurons in each population was identical). The tails of the Gaussian profiles were cut off at three standard deviations (value = 0.011). To avoid potential edge effects, the stimulus space was toroidal, i.e., each tuning function with a tail extending beyond one of the edges of the stimulus space was continued into the opposite side of the space by joining the two opposite edges of the space (see Fig. 2C). The uniform tiling of the receptive field (RF) centers along the position axis was chosen for simplicity, although it does not match the observed foveal bias in the position preference of real IT neurons (Op de Beeck and Vogels 2000). However, this departure from empirical observations does not affect the conclusions of our study, since variations in the density of the RFs over the stimulus space would not affect the relative classification performance of different simulated populations, as long as the training and testing stimuli were drawn from the same distribution for all the tested populations (as done in our simulations).

To simulate the IT responses to visual scenes containing multiple objects, we defined four different "clutter rules" (CCI, LIN, AVG, DIV, Fig 3D) specifying how a neuron's responses to multiple objects could be predicted from its responses to single objects (i.e. descriptive models). These rules were implemented as follows. If objects A and B elicited, respectively, neuronal responses H_a and H_b when presented in isolation (note that this is a function of both the object identity and its spatial position, defined by the Gaussian response functions described above), then the neuron's response to a visual scene (v) consisted of both A and B was:

1) CCI: the maximum of H_a and H_b (i.e. complete clutter invariance);

2) LIN: the sum of H_a and H_b (linear rule);

3) AVG: the average of H_a and H_b (average rule);

4) DIV: the divisive normalization of H_a and H_b (divisive normalization rule).

Divisive normalization was defined as:

$$\mathbf{H} = \frac{\mathbf{H}_a + \mathbf{H}_b}{\left\| \mathbf{H}_a + \mathbf{H}_b + \lambda \right\|}$$

The constant λ was small (0.01) and changing it did not qualitatively alter the simulation results. All of these clutter rules naturally extended to three or more objects. To ensure that the comparison between different clutter rules was not affected by signal-to-noise confounds, we normalized each neuron's responses to the mean of its responses to all the stimuli (including both the 3000 training and the 300 testing stimuli) presented within a simulation run. Conceptually, this normalization roughly equated populations following different rules in terms of averaged number of spikes produced. Without such normalization, neurons obeying to the LIN rule would be more active, on average, than neurons obeying to the AVG rule, resulting in better signal-to-noise. In practice, the normalization similarly affected the absolute performance obtained by simulating the different clutter rules, with only a minor impact on their relative magnitude – see, for instance, the performance on the "position-invariant" task shown in the inset of Fig 3C: with normalization (shown): CCI 75%, LIN 76%, AVG 67%, DIV 73%; without normalization: CCI 62%, LIN 62%, AVG 53%, DIV 55%.

In sum, there were five parameters for each simulated IT neuron: 1) μ_i, the preferred object (the center of the Gaussian on the object identity dimension); 2) μ_p, the preferred position (the center of the Gaussian on the position axis); 3) σ_s, (inverse of) sensitivity to object identity; 4) σ_p, position sensitivity; and 5) the "clutter rule" – how the response to multiple objects was predicted from the responses to the single objects. To isolate the effects of the two main parameters of interest (single-unit position sensitivity, $\sigma_{p,}$ and single unit "clutter rule") while counterbalancing across the exact Gaussian center locations (μ_i and μ_p), we simulated many different populations in which the center values of the Gaussians were randomly generated within the stimulus space (see example in Fig 2C). All the results presented in the main text of the paper were obtained by averaging the performance on visual tasks over sets of at least 15 such simulated population runs, where each run in a set contained neurons with the same values of the parameters (σ_p, σ_s, and "clutter rule"), but different random Gaussian centers. To further facilitate the comparison of the effect of different "clutter rules", the same sets of randomly generated Gaussian centers were used while the clutter rule was varied (Fig. 3C, D).

Simulated V1 population to compare with the recorded IT population

To compare the recorded IT population results with a meaningful baseline (Fig. 1B), we simulated populations of V1 simple cell like units (n=68, matched to the IT population in the number of recorded trials and Poisson-like noise within a 100ms spike-count window) in response to the same set of visual scenes that were presented to the animals during IT recording (e.g. 450x150 pixels image containing "stars" and "triangles"). We simulated each V1 unit as a 2D Gabor operator on the images, qualitatively consistent with current empirical results (DeAngelis et al. 1993; Jones and Palmer 1987), and the response of each V1 neuron to a visual "scene" was the thresholded dot product of its Gabor function applied to the “scene”. To synthesize a V1 population, we randomly draw each V1 unit's receptive field position, size (20x20~80x80 pixels), orientation (0~180°), spatial frequency (0.05~0.20 cycles/pixel), and phase (0~180°) from uniform distributions. A different set of V1 units (new random draws) were chosen for each simulation run, and the performance we report in Fig 1B was the average performance over at least 15 such runs (15 different V1 populations). Though the random sampling of the V1 units' parameters may introduce variability in the V1 classification performance, this variability was small relative to the absolute performance (error bars in Fig. 1B show standard deviations).

Simulated V1 population to compare with simulated IT population

To compare the simulated IT populations with a meaningful baseline (Fig. 5, 6), we again simulated populations of V1 units. In this case, we simulated each V1 unit’s 2D response function spanning a discretized stimulus space (n objects x n positions) that was roughly matched to the continuous stimulus space we defined for the simulated IT population. We used images containing 64 2D white silhouettes shapes (Zoccolan et al. 2005) on a constant gray background and we computed each unit’s responses to images of each white shape at 64 azimuth positions (64 objects x 64 positions = a total of 4096 images). On average, the objects were ~3 times the size of the V1 receptive fields in diameter. Our main conclusion was not dependent on the exact parameterization of the stimulus space or the shape of the V1 response functions in this stimulus space. This was verified by simulating the V1 response functions on 64 natural objects on gray backgrounds, yielding similar classification performance.

Simulated “abstract” populations

We explored classes of hypothetical neuronal populations consisting of neurons with more abstract response functions in the 2D stimulus space than the 2D Gaussians used to model IT units (a diverse range of response function shapes was used). Some of these populations were built such that the rank-order object selectivity of individual neurons was preserved across position changes and clutter conditions, while other populations, by construction, lacked this property (Fig 5; $(i)_P$-$(v)_P$, $(i)_C$-$(iv)_C$). The populations with response functions that preserved the rank-order selectivity across the position axis were constructed as following (see Fig. 5C, right):

i)p position-invariant response and narrow Gaussian sensitivity along the identity axis;

ii)p wide Gaussian sensitivity along the position axis and narrow Gaussian sensitivity along the identity axis;

iii)p position-invariant response and sinusoidal sensitivity along the identity axis;

iv)p multi-lobed Gaussian sensitivity along both the position and identity axis;

v)p random tuning profile. The random 2D response function was created by multiplying two independently drawn, random 1D response functions (smoothed), specifying the selectivity profile along each of the two stimulus axes.

By construction, these response functions maintained their rank-order object preferences across position changes (Fig. 5C right panel), so that the response modulations resulting from position changes did not impact the object preference rank-order. To simulate their counter parts, (similar response functions but with rank-order not preserved, Fig. 5C left panel), response functions $(i)_p$~$(iv)_p$ above were simply rotated in the stimulus space for an arbitrary angle (±30~60°). The rotations created diagonals in the response matrix over the stimulus space, thus the neurons' rank-order object preference was no longer preserved under position variations. The random response functions with non-preserved rank-order object preference, $(v)_p$, were created by smoothing matrices of random numbers.

When multiple objects were present in the visual scene, the stimulus space became n-dimensional representing each object's position and identity (n = 2 times the number of objects). For the purpose of simplicity, in Figure 5 and 6, we only considered visual scenes with two objects and object position was ignored. Therefore, in this reduced formulation, the stimulus space was only 2-dimensional, representing the identity of the two objects (such a simplification does not limit the generality of our conclusions). Within the stimulus space, response functions produced by all the systematic clutter rules (CCI, LIN, AVG, and DIV) maintained their rank-order object preference across clutter conditions. That is, if a neuron preferred object "A" over "B", the neuron would maintain that preference when another object "X" was added (i.e. "AX" ≥ "BX"), regardless of the identity of the distractor "X", (e.g. see AVG in Fig. 5D). To contrast, we simulated four other response functions (Fig. 5 $(i)_C$-$(iv)_C$) that did not maintain this rank-order object preference. That is, adding specific "X" reversed the neuron's response preference for "A" over "B" (i.e. "AX" ≤ "BX" in certain cases). The details of these other response functions are not of critical importance other than the fact that they exhibited distinct shapes and covered a range of single-neuron clutter sensitivity. In practice, they were generated as following:

i)c we first established a CCI response function inside the 2-dimensional stimulus space (object-object). Each neuron had a Gaussian tuning along the object identity axis, and its conjoint tuning in the object-object stimulus space was established by taking the maximum between two Gaussian tunings along the individual stimulus axes. The final response function had the shape of a "cross" centered on the preferred object of the neuron paired with itself. Once the CCI response function was establish, we then rotated (±30~60°) the response function inside the stimulus space to create diagonals (such as what was done for $(i)_P$-$(iv)_P$).

ii)c rotated version of LIN response function;

iii)c sum of two different CCI response functions with their centers some distance apart within the stimulus space (at least 0.3 of the width of the stimulus space);

iv)c we first established a CCI response function. We then added a separate Gaussian lobe, of variable width, to the CCI response function.

Single-neuron metrics: position sensitivity, clutter sensitivity, and rank order

Relating the goodness of a population (i.e. classifier performance) to single-neuron properties, we contrasted different populations by three different single-neuron metrics: position sensitivity, clutter sensitivity, and rank-order of object selectivity.

To quantify different populations' position sensitivity (see Fig. 6A), we carried out a position sensitivity "experiment" on each neuron. We first found its most preferred object and preferred position by finding the peak of its 2D response function. Using this preferred object, we measured the neuron's responses to 1D changes in object position, and the magnitude of the neuron's position sensitivity was quantified as the area under this 1D response function (this is equivalent to mapping a neuron's receptive field with its most preferred object, analogous to standard measurements of position tolerance; Zoccolan et al. 2007). This position sensitivity index was normalized so it ranged from 0 to 1 for each neuron. The position sensitivity of a population was the average of all the individual neurons' position sensitivity indices.

To compute the magnitude of each population's clutter sensitivity (see Fig. 6B), we first found each neuron's peak along the diagonal of the stimulus space (i.e. most preferred object paired with itself), its clutter sensitivity index was then computed as the averaged reduction in response from this maximum response when this preferred object was paired with other objects. The clutter sensitivity index was normalized so it ranged from 0 to 1, (analogous to standard measurements of clutter tolerance; Zoccolan et al. 2007).

To quantify how well a population's neurons maintained their rank-order object preference in the face of transformations, we employed commonly used separability index ((Brincat and Connor 2004; Janssen et al. 2008), see Fig. 4B & D). The separability index computes the correlation between a neuron's actual responses and the predicted responses assuming independent tunings along the object and transformation axis (i.e. a neuron's response is characterized by the product of its tuning along the object and transformation axis). The separability index ranged from -1 to 1, and was computed for the recorded IT population and the simulated V1 population as following: for position transformations, a neuron's responses were assembled in a 3 x 3 response matrix, ***M***, (there were 3 object presented at 3 positions in the experiment). For clutter transformation, the response matrix ***M*** was 2 x 6 (2 objects under 6 clutter conditions, e.g. Fig 4C). The predicted response was computed by first taking the singular value decomposition of ***M (M = USV')***, then reconstructing the predicted response from the first principle component (i.e. product of the first columns of ***U*** and ***V***). To avoid bias, each neuron's data was split in half: one half was used to generate the predicted response, the other half used to compute the correlation with the prediction (i.e. the separability index). Only the selective neurons were included in this analysis (Fig. 4): to be deemed selective across position, neurons need to pass an one-way ANOVA test across object identity ($p<0.05$; 32 neurons in total); to be deemed selective across clutter conditions, neurons need to pass an one-way ANVOA test across clutter conditions ($p<0.05$; 25 neurons). For clutter, each neuron could contribute multiple separability index values depending on the precise configuration of the stimulus display (e.g. of the sorts shown in Fig 4C lower panel). In total, there were 63 cases from the IT population and 68 cases from the V1 population in Figure 4D.

Results

The first goal of this study was to examine the ability of a recorded IT neuronal population to support object identification tasks in the face of object position variation and clutter. These two types of image variation are intimately related in that, when images contain multiple objects (cluttered images), those objects invariably occupy different retinal positions. Thus, a neuronal representation that signals object identity must overcome both types of variation simultaneously. The second, related goal was to examine simulated IT populations with different single-unit response properties, in order to understand the relationship between single-unit IT response properties and population performance in those tasks. To accomplish these two goals, we constructed visual "scenes" that are simpler than those typically encountered in the real world, but that engage the computational crux of object recognition – object identification in the face of image variation. Specifically, we tested the populations' ability to support two types of tasks: 1) identify objects irrespective of their position and the presence of other objects ("position-invariant recognition"); 2) identify objects at specific positions irrespective of the presence of other objects ("position-specific recognition"; See Fig 1A).

We used linear classifiers to test the capability of the recorded and simulated populations to support the two recognition tasks, and we took the classifiers' performance as a measure of the goodness of the representations provided by the populations. Successful performance on both tasks means that the population representation can support clutter invariant recognition and it can simultaneously represent multiple objects (at least up to the number of objects tested; see Discussion). The justification for such an approach and the implementation details of the classifiers are provided in the Methods section.

1. The Primate IT Neuronal Population

To test the basic ability of primate IT to directly support position- and clutter-invariant object recognition (identification), we recorded the responses of a population of monkey IT neurons (n=68) to a set of 33 simple visual scenes. Each scene was constructed from three possible objects (star, triangle, cross) and three possible retinal positions (-2°, 0°, +2° to the center of gaze; see Materials and Methods for details). Some scenes contained only single objects in isolation, while others contained those same objects in the presence of other objects (two or three objects in a scene, Fig. 1A; see Methods).

Task 1: Position-invariant identification: What object(s) are in the scene?

We began our analysis of these IT population data by examining the simple situation in which each presented visual scene contained just one of the three possible objects in any one of the three possible retinal positions (9 of the 33 scenes). By restricting to these scenes only, we could ask how well the IT population could support position-invariant object identification without visual clutter. Specifically, for correct performance, each linear classifier had to respond only when its preferred object was present regardless of the object's position (see Materials and Methods). Consistent with previous work (Hung et al., 2005), we found that even a small IT population (n=68) can support this task well above chance (Fig. 1B, mean 69.1%, $p<<10^{-6}$, chance = 50%), even though most neurons are highly sensitivity to changes in object position (median response reduction of 35.9% going from preferred object in the best position to worst position, within 2 deg of fovea). Moreover, we found no systematic relationship between the magnitude of a neuron's position sensitivity and its contributions to task performance (i.e. weight in the classifier; correlation=0.19, p=0.3).

We next considered a more complex situation in which we asked if the IT population could directly support object identification even in the face of limited clutter (other objects in the scene; see Fig. 1A, upper panel). The task of the linear classifiers was the same as above, except that we now included scenes in which multiple objects were present (two or three objects, 33 scenes total, see Methods). The presence of such additional objects often strongly suppresses the responses of individual IT neurons (Chelazzi et al. 1998b; Miller et al. 1993; Missal et al. 1999; Rolls et al. 2003; Rolls and Tovee 1995; Sato 1989; Sheinberg and Logothetis 2001; Zoccolan et al. 2005; Zoccolan et al. 2007), and for this set of IT neurons, the median reduction in response to the most preferred object was 36.4%. Thus we asked whether the IT population performance would be similarly affected in this task. However, we found performance well above chance (Fig 1B, mean: 68.9%, $p<<10^{-6}$), and only slightly degraded from that observed with single objects (the performance in the two cases was not significantly different, p=0.59, two-tailed t-test). This shows that the ability of the IT population to support position-invariant object identification is largely unaffected by the presence of limited visual clutter, even when individual IT neuronal responses are strongly affected by that clutter. We found no systematic relationship between the magnitude of a neuron's clutter sensitivity and its contributions to task performance (correlation=0.19, p=0.3).

Task 2: Position-specific identification: What object is located at each position?

We have shown above that the IT population can directly report object identities regardless of object positions, even when the scene contains multiple objects (at least under the limited conditions tested here, see Discussion). This result implies that the IT population can simultaneously represent the identity of multiple objects. However, to represent multiple objects unambiguously, the population should directly represent not only "what" objects are present (i.e. Task 1 above), but also "where" they are. Although this question touches on deep theoretical issues and possibilities about how such information is "bound" together (Riesenhuber and Poggio 1999; Roudi and Treves 2008; Treisman 1999), we here ask a very basic question: can the IT population report object identity at specific positions? To do this, we used the same set of visual scenes (containing both single and multiple objects) and neuronal population response data, and we built linear discriminant classifiers to perform the same object identification task at each of the three possible positions (see Fig. 1A bottom panel). At each of these positions, we found that such classifiers performed even better than the position-invariant classifiers (mean: 73.6%, Fig. 1B). This means that downstream neurons could, in parallel, reliably report the identity and position of each object in the image from the IT population response (at least up to the limited clutter conditions tested here).

It is well known that population size can strongly influence the reliability of signals and thus increase the total amount of information that is conveyed by neuronal representations. It is also known that cortical neurons can integrate information over a number of synaptic inputs (~10,000; Braitenberg 1978) that is much larger than the number of IT neurons that can be reasonably be recorded with current techniques. To overcome this limitation, we used the linear discriminant approach to estimate how the amount of information conveyed by an IT neuronal population would scale with the number of units in the population. To this aim, we synthesized larger populations of Poisson-spiking neurons from the response profiles of the measured IT population. This procedure does not assume any stimulus selectivity that was not already in the population (because all synthesized neurons were copies of one of the original 68 neurons), but it does allow for moderate amounts of pooling to overcome the high trial-to-trial variability of cortical neurons (Shadlen and Newsome 1998), thus increasing the information that can be extracted from the IT population on a single trial. We found that the performance on both recognition tasks scaled at a very similar rate as the population size grew (Supplemental Fig. 1). Notably, the absolute performance saturated at very high levels for population sizes that were similar to those postulated to support visual discrimination tasks in other visual areas ((Shadlen et al. 1996); >80% correct for a population of several hundred neurons. Here, "position-specific" task: >85%; "position-invariant" task: >80%, n=680).

Overall these data show that, although individual IT neuronal responses are often highly sensitive to object position (DiCarlo and Maunsell 2003; Op de Beeck and Vogels 2000; Zoccolan et al. 2007) and to the presence of visual clutter (Chelazzi et al. 1998b; Miller et al. 1993; Missal et al. 1999; Rolls et al. 2003; Rolls and Tovee 1995; Sato 1989; Sheinberg and Logothetis 2001; Zoccolan et al. 2005; Zoccolan et al. 2007), the IT population was able to overcome the inadequacy of single IT neurons -- object identity can be extracted invariant of retinal position and the presence of clutter (up to a certain degree, Fig 1B). Notably, the performance of the IT population on all of these tasks is greater than that expected of a comparably sized population of V1 neurons (simple cell simulation; see Methods; Fig 1B; this is not simply explained by smaller V1 RF size or lack of coverage, see Fig 5, 6C). Thus, motivated by these findings with the recorded IT population, we sought to understand what single-neuron response properties are most important in providing a population representation that robustly supports position- and clutter-invariant object identification ("What"; Fig. 1B 1st and 2nd bar), yet can also support position-specific object identification ("Where"; Fig. 1B 3rd bar).

2. Simulated "IT" Neuronal Populations

While our empirical IT data provide a basic "proof of existence" that position- and clutter-sensitive neurons can support invariant recognition, they provide little insight into what single unit properties are important to this ability. To explore this issue further, we simulated artificial populations of neurons with different position and clutter sensitivity, as a tool to ask what kind of single unit response properties are more or less important for a population of such neurons to support position- and clutter- invariant object recognition.

To do this, we created an abstract two-dimensional (2D) stimulus space with object identity (e.g. shape) on one axis and retinal position (e.g. azimuth) on the other axis. A neuron's response to a single object (i.e., a point in the 2D stimulus space) was determined by a 2D Gaussian tuning function over the stimulus space (Fig. 2B; see Methods). Its center specified the neuron's preferred stimulus (i.e., the preferred shape and position), its standard deviation along the shape axis (σ_s) controlled it selectivity for object shape (i.e., a lower σ_s results in a sharper shape tuning), and its standard deviation along the position axis (σ_p) controlled its sensitivity to changes in object position (i.e., the size of its receptive field). In the presence of multiple objects, we constructed different "clutter rules" specifying how a neuron's response to multiple objects depended on the responses to single objects. Briefly, the response to multiple objects was defined as either: the maximum (CCI), the sum (LIN), the average (AVG), or the divisive normalization (DIV) of the neuron's responses to the constituent objects in isolation (Fig. 3D). We also included a negative control where the neuron's response in clutter was not systematically related to the responses to the constituent objects (RAND). These different clutter rules specified different amounts of individual-neuron clutter sensitivity. The main results of the paper are not limited to these initial assumptions, as we also explored other (more general) types of representations (see Fig. 5).

The aim of these simulations was to create artificial neuronal populations with different kinds of single-unit response functions (described below). Linear classifiers were then used to assess the "goodness" these populations in supporting position- and clutter- invariant recognition. This allowed us to evaluate the relative pros and cons of different single-unit response properties in the context of a population code. Note that this framework is agnostic about: the number of shape dimensions, what aspect(s) of visual shape are coded along one of those shape dimensions, and what exact visual stimuli real IT neurons are tuned for. Instead, it is simply a tool to facilitate thinking about the best way to represent information about object identity and position using populations of neurons.

Effect of varying the position and clutter sensitivity of individual neurons

Similarly to what was done for the recorded IT population, we examined how well a simulated population can identify objects in visual scenes containing multiple objects. In such a context, we varied the individual neuronal position and clutter sensitivity and examined their effects on a population's ability to support the recognition tasks (Fig. 3). To do this, we synthesized neuronal populations in which all single neurons in each population had the same position sensitivity (σ_p) and clutter sensitivity (clutter rule), but across a series of these populations, we systematically varied the single-neuron position (Fig 3B) and clutter sensitivity (Fig 3D).

Figure 3A shows how the performance in the recognition tasks depends on the position sensitivity of the individual neurons. In fact, varying individual neuronal position sensitivity over a wide range produced little effect on the populations' ability to support position-invariant task (Fig 3A dashed line). At the same time, only populations of position sensitive neurons conveyed the necessary position information to support the position-specific task (Fig 3A, solid line). Trivially, the population performance on both recognition tasks rapidly decreased if the single neurons were made too position sensitive due to the populations' loss of coverage of the stimulus space (Fig. 3A, small σ_p). So far, these results are only a confirmation of the rather intuitive expectation that one can take position-sensitive neurons and combine them in a population to support position-invariant task. However, with these same methods and intuition in hand, we next go on to show that the same conclusion on single-neuron position sensitivity holds when multiple objects ("clutter") are present in the scene.

Figures 3C shows the classifier performance in the position-invariant and position-specific task in the presence of multiple objects, when the simulated neuronal populations followed different "clutter rules" (stimuli consisted of single, double, and triplet objects). Surprisingly, populations of non-clutter-invariant neurons (LIN, AVG, DIV) performed comparably well to populations of complete-clutter-invariant neurons (CCI) (Fig. 3C, insets). Performance was substantially reduced only when neurons followed the random "clutter rule" (RAND; black bars in the insets) in which the responses to multiple objects were not predictable from the responses to single objects. In fact, the choice of the "clutter rule" had relatively little effect on population performance even though individual neurons behaved very differently under different "rules" (Fig 3D). Furthermore, the different populations conveyed object identity information in similar format, (correlations among linear classifier weights within clutter rule: 0.988; across clutter rule: 0.975; Table 1). Thus, for a downstream observer, the amount of individual-neuron clutter sensitivity did not matter, to the extent that the object identity information can be read out in nearly identical fashion (albeit with different classifier thresholds).

Together, these results show that single-neuron properties previously assumed to be important (i.e., response magnitude that is largely maintained across transformations) only minimally impact the goodness of the representation (but see Discussion for possible limitations of such a conclusion). Furthermore, in the case of position, the high sensitivity often observed in individual IT neurons should be viewed as a desirable property for a representation capable of directly supporting a range of recognition tasks (also see DiCarlo and Cox 2007).

What response property of individual IT neurons enables populations of such neurons to support object recognition?

If the amount of position and clutter sensitivity only has a small impact on a representation's ability to support invariant recognition tasks, a fundamental question then arises: what key single-neuron property has the visual system achieved in IT that is not present in early visual areas (e.g. V1)? Or, to put it another way, given that V1 neurons have high position-sensitivity (i.e., small receptive fields), which is a potentially useful property (as shown in Fig. 3A, C), what property do individual V1 neurons lack that makes the V1 population inferior to the IT population for object recognition (Fig. 1)?

A distinguishing hallmark of IT is that neurons' preference among different objects is often preserved across image transformations (at least with respect to position and size; Brincat and Connor 2004; Ito et al. 1995; Schwartz et al. 1983) despite variations in the receptive field sizes. This was true in our recorded IT population as well. An example IT neuron's object preference across position is shown in Fig 4A. Conversely, when we simulated V1 neuronal responses (spatially local Gabor operators on the same visual scenes, see Methods), we found that the rank-order of their object selectivity was not preserved, because of the interaction of object parts with the neurons' receptive fields (e.g. Fig. 4A). To quantify the preservation of the rank-order object preference across the population, we used a standard separability metric (see Materials and Methods; Brincat and Connor 2004; Janssen et al. 2008). On average, we found that the IT neurons had much higher separability from the simulated V1 units (Fig. 4B, $p<10^{-14}$, two-tailed t-test). More interestingly, we also noted that neuronal responses under clutter could be interpreted in the same framework. When we plotted the IT neurons' responses to different objects under the same clutter conditions (i.e. when paired with the same distractor at the same position), most single IT neurons showed preservation of their object preference rank order (see example in Fig. 4C) and the IT population showed much more separable responses than the simulated V1 population

(Fig. 4D, $p<10^{-22}$, two-tailed t-test).

The preservation of the rank-order object selectivity over position changes has been previously suggested to be important for achieving a position-invariant object representation (Gross et al. 1993; Logothetis and Sheinberg 1996; Vogels and Orban 1996), but, to our knowledge, has not been systematically evaluated and demonstrated. Furthermore, the notion that preserving the rank-order of object selectivity in clutter can result in a clutter-invariant population representation has never been proposed. Instead, it is commonly assumed that attentional control is necessary to overcome clutter given the clutter sensitivity of single IT neurons (Desimone and Duncan 1995; Reynolds and Chelazzi 2004; Serre et al. 2007). Is preservation of the rank-order selectivity in clutter important to achieve a clutter-invariant representation and can such a property overcome the coding issues associated with the presence of clutter? To validate these intuitions and clarify the relationship between single neuron response properties and goodness of a population representation (e.g. Fig. 1B and Fig. 4), we directly examined the importance of rank-order preservation as a key single-neuron response property.

Effect of varying the preservation of the rank-order selectivity of individual neurons

We simulated many different neuronal populations consisting of neurons with abstract response functions (i.e., unlike V1 and IT, generated without regard for experimental data). We chose these abstract response functions such that some preserved the object rank-order across transformations while others did not (e.g. Fig 5C). In addition, their response magnitude spanned a wide range of sensitivity to position and clutter (measured by appropriate indices, see Methods). This allowed us to assess what single-unit response property is a good predictor of the population performance on the invariant recognition tasks. To minimize other confounding differences between these response functions, all of the populations were matched in terms of number of neurons and approximate coverage of the stimulus space.

We first concentrated on the position aspect of the recognition task by only using visual scenes of single objects. As shown in Fig 5B, we found that populations of neurons that preserved the rank-order of their object preference across positions (see example in Fig. 5A, right) performed much better on the position invariant recognition task than populations of neurons that did not (see example in Fig. 5A, left). We also found that some populations of neurons, whose response functions were not Gaussian, but nevertheless, preserved the rank-order of object preference across positions (e.g., Fig 5C, plot labeled $(iii)_P$), performed nearly as well as the population of neurons with Gaussian tuning functions (Fig 5C, plot labeled "*IT*"). This implies that, at a purely computational level of information representation, Gaussian response functions are not required to support position-invariant recognition.

Next, we showed how a similar rationale could explain the high performance achieved by all the systematic clutter rules when multiple objects were present (c.f. Fig 3C). In fact, the systematic clutter rules we simulated (Fig. 3D) all produced rank-order preservation of object preference across clutter conditions (only one of those rules – CCI – also yielded clutter invariance). Except for the RAND rule, each neuron maintained its relative object preference (rank-order) even though its absolute firing rate could change dramatically in the face of clutter (an example neuron following the AVG rule is shown in Fig. 5D, right). To confirm that this preservation of rank-order selectivity across clutter conditions underlies the high classification performance, we simulated neuronal populations that did not maintain the rank-order of their object preference in clutter (e.g. Fig. 5D, left). Indeed, the performance on the clutter-invariant recognition task was much lower for the latter populations (compare gray to black bars in Fig. 5E). This directly demonstrated that, to achieve clutter-invariant recognition, the degree of clutter sensitivity of individual neuronal responses is not critical. Instead, it is more important that neurons maintain the rank-order of their object selectivity in the face of clutter.

At a more quantitative level, the preservation of the rank-order selectivity at the single-unit level was a good predictor of the population performance across all the populations we simulated, while standard measures of single neuron sensitivity to transformations were not (Fig. 6A). We also tested whether strict separability of tuning along the identity and position dimensions yielded higher recognition performance as compared to the less strict requirement of preserving the rank-order object selectivity. Tuning in a multi-dimensional stimulus space is separable if it is the product of the tuning along individual stimulus dimensions, and there are reasons to believe that separable tuning curves could be mathematically optimal for creating a representation where multiple stimulus attributes need to be read out with linear tools (Ma et al. 2006; Sanger 2003). We found that these two alternative coding schemes both yielded equally high recognition performance (Fig. 6B), but this negative result does not fully settle the issue because the difficulty of our object recognition tests may not have been powerful enough to distinguish among these alternatives. Finally, the better performance achieved by the populations preserving the rank-order selectivity (e.g., IT versus V1) cannot be accounted for by the degree of coverage of the stimulus space, since coverage was approximately equated across the tested populations. To further confirm this, we varied the number of units in the simulated IT and V1 populations and examined their performance on the position invariant task (Fig 6C). We found that even when a high number of units was simulated (to rule out any possible coverage differences between the V1 and IT populations), V1 performance quickly saturated to a much lower value than IT performance, failing to succeed in the simple invariant task asked here.

In summary, response functions that preserved the rank-order of object selectivity across position changes and clutter led to neuronal populations that were highly robust in supporting invariant recognition, regardless of the specific shape of the neuronal tuning curves or their degree of sensitivity to the tested transformations.

Discussion

Most studies aimed at understanding invariant object representation in IT have understandably concentrated on measuring the responses of single IT neurons to preferred objects presented over transformations (e.g., addition of "distractor" objects to the image). Although perhaps at odds with colloquial thinking about IT, that work has shown that single IT neurons' firing rates can be quite sensitive to these identity-preserving image changes, often much more sensitive than behavioral recognition (Aggelopoulos and Rolls 2005; DiCarlo and Maunsell 2003; Op de Beeck and Vogels 2000; Tovée et al. 1994; Zoccolan et al. 2007). To consolidate this discrepancy, it is tempting to conclude that this sensitivity in single IT neurons reflects inadequacies of those neurons to achieve invariance in natural vision (where multiple objects are constantly present), and that the visual system must engage additional mechanisms (e.g. attention) to overcome the interference of visual clutter. However, these explanations assume a straightforward relationship between the response properties of individual neurons and the behavior of populations of such neurons. The primary goal of this study was to examine that assumption.

By gathering neuronal data from IT and “reading-out” the population using biologically plausible mechanisms (linear classifiers), we report that intrinsic response properties of a small population of IT neurons (i.e. earliest part of response in the absence attention) already supports object identification while tolerating a moderate degree of clutter. This is true even when multiple objects and their positions must be reported. However, this leaves open the possibility that the IT population would be able to do even better if the individual neurons were somehow less position- and clutter-sensitive. We tested this possibility by carrying out simulations that showed that: 1) low sensitivity to position changes in individual neurons is not needed to support position-invariant recognition, 2) low sensitivity to clutter in individual neurons is not needed to support clutter-invariant recognition, 3) position-sensitive neurons are advantageous, since they allow the unambiguous representation of object position, and 4) preservation of the rank-order of object selectivity is a single neuron response property that is highly predictive of good population recognition performance (see summary in Fig 7).

At a first level, our results are a reminder that even simple rate codes in populations of neurons can convey information that is not readily apparent from the responses of single-units (Kohn and Movshon 2004; Riesenhuber and Poggio 1999). For example, a strong interpretation of "what" and "where" pathways creates a so called "binding problem" (Treisman 1999) in that IT is assumed to represent only object identity, and this has led to a number of speculations as to how that the object identity and position can be bound back together (Reynolds and Desimone 1999; Riesenhuber and Poggio 1999; Shadlen and Movshon 1999). However, at least with respect to the object position and identity, direct examination of IT population responses shows that this particular form of the binding problem does not exist, since object identity is represented jointly with object position in IT (Fig 1B), as previously suggested (DiCarlo and Cox 2007; Edelman and Intrator 2003; Riesenhuber and Poggio 1999; Roudi and Treves 2008, Serre et al. 2007; Hung et al. 2005). At a deeper level, our results show that single-neuron properties previously assumed to be important (i.e., response magnitude that is largely maintained across transformations) only minimally impact the goodness of the representation (but see below for possible limitations to such a conclusion), and that the sensitivity to transformations often observed in individual IT neurons (i.e. “tolerant” IT neurons, see Fig. 7) should not be viewed as a failure to achieve perfection, but a desirable property for a representation capable of directly supporting a range of recognition tasks (also see DiCarlo and Cox 2007).

Ideal single neuron response properties?

If transformation-sensitive responses in individual neurons are not a limiting factor in creating a population that can support highly invariant recognition, what single-neuron property is required? This problem is ill defined because, in general, no individual neuron will dominate the performance of a population (e.g., its limitations can always be compensated by other neurons). However, if one assumes that all neurons in the population have similar response functions but with different preferred shapes (objects) and positions (i.e., different Gaussian centers in our 2D stimulus space), we showed that populations of neurons with rank-order object preference that is preserved across image transformations (here, position and clutter, but this could be size, pose, etc.) form much more powerful object representations than populations of neurons that lack this property (Fig. 5, 6). The potential importance of preserving the rank-order object selectivity over preserving the magnitude of neuronal responses (in the face of transformations) has previously been suggested in the literature with respect to position and size (Gross et al. 1993; Logothetis and Sheinberg 1996; Vogels and Orban 1996). Here we provided direct confirmation of this: by simulating abstract neuronal response functions, we found that rank-order preservation of object selectivity in individual neurons was a very good predictor of population performance, while the extent to which neuronal response magnitude was preserved was a poor predictor (Fig. 6A). Interestingly, unlike V1 neurons, IT neurons appear to preserve their rank-order selectivity over changes in object position (DiCarlo and Maunsell 2003; Ito et al. 1995; Logothetis and Sheinberg 1996; Op de Beeck and Vogels 2000; Tovée et al. 1994), even when their receptive field size is small (DiCarlo and Maunsell 2003). This single-unit response pattern in IT neurons has been termed selectivity "tolerance" (e.g. tolerance to position), and it explains why IT populations perform better than (e.g.) V1 populations on object recognition tasks, even if both have small single-unit receptive fields (also see DiCarlo and Cox 2007).

Furthermore, we also demonstrated that the same rationale explains why single neurons following any of the systematic clutter rules (i.e., all rules in Fig. 3C, D, except RAND) performed well as a population in recognition tasks under cluttered conditions (Fig. 3C). In particular, even though each systematic clutter rule produced different amounts of clutter sensitivity in individual neurons (Fig 3D), all of these rules had a more important feature in common – they each preserved the rank-order object preference, in the sense that each neuron always responded more to its preferred object than non-preferred objects, even in the presence of other clutter objects (provided that these clutter objects were the same and were present at the same positions in both cases). Again, it is not the amount of clutter sensitivity that matters, but the preservation of the rank-order selectivity that is guaranteed by these clutter rules. And again, although single IT neuron responses are strongly altered by clutter (Chelazzi et al. 1998b; Miller et al. 1993; Missal et al. 1999; Rolls et al. 2003; Rolls and Tovee 1995; Sato 1989; Sheinberg and Logothetis 2001; Zoccolan et al. 2005; Zoccolan et al. 2007), the rank-order selectivity of those neurons appears to be maintained in clutter (Zoccolan et al. 2005; Zoccolan et al. 2007). Even though our neuronal data were collected at short eccentricities (-2° to +2°), this message applies to larger eccentricities because the simulations are scale independent, and IT clutter suppression has been found to be virtually identical at short eccentricities (Zoccolan et al, 2005) and mid-range eccentricities (i.e., 4-7º from fovea; Chelazzi et al, 1998). Naturally, the conclusion also applies to any population representation, whose units behave similarly to IT neurons in clutter, including a class of object recognition models (Zoccolan et al. 2007) that are able to support recognition of multiple objects (Serre et al. 2007). The summary “take home” message of our work is given in graphical form in Figure 7.

More broadly, the response functions with preserved rank-order selectivity performed well because this class of response functions is well matched to the desired output function the classifier is trying the construct (Ma et al. 2006; Poggio 1990; Salinas 2006) – the independent read-out of object identity and image transformations (but see also limitations below).

Limitations

An overriding theme of this paper is that task constraints dictate ideal response functions (Salinas 2006), but not always in an obvious way. Here, we explored a range of descriptive models of single-unit IT neuronal response functions, and determined which performed best at the population level. While this approach gives insight into which response properties are important, it does not provide guidance on how to construct mechanistic models of such IT neurons (i.e., models that operate directly on the visual image). In addition, although our descriptive models of IT show how the task constraints of object recognition imply that IT neurons should have sensitivity to object identity that is preserved across position and clutter, this still allows a large number of possible descriptive models. That is, there are a large number of response functions with rank-order preserved that are capable of supporting the recognition tasks (e.g. Fig 5C). Other constraints such as wiring limitations (i.e. number of afferent connection per neuron allowed) and the number of neurons in a population will further constrain the ideal set of descriptive IT models. Further simulations could address these issues.

Our classification analysis shows how downstream neurons can utilize the same IT population to achieve good performance on at least two different visual recognition tasks. For the brain to utilize a fixed IT representation in this flexible manner requires different downstream readouts of IT – different weightings across the IT populations. Although each such readout is biological plausible (see Methods), this study cannot address how different readouts are mechanistically achieved in the brain. However, given the flexibility of our behavior and the number of tasks we are capable of performing, we speculate that these different readouts are not hard-wired, but might be dynamically invoked in the frontal cortices where decisions and associations are rapidly made (Freedman and Assad 2006; Freedman et al. 2001).

As shown in Fig. 3C, although populations of neurons with different clutter rules gave approximately equal recognition performance, populations of neurons that were insensitive to both position and clutter (CCI) provided the best performance in the position-invariant tasks in clutter. This suggests that, at least for certain recognition tasks, high position and clutter invariance may be desirable (note that such "invariance" is an extreme form of tolerance, see Fig. 7). More generally, we are not ruling out possible advantages of having single-unit responses that are insensitive to image transformations. For example, here we focus on the ability of a population to present well-formatted information to downstream neurons, but we do not address the problem of how the downstream neurons find that information (computationally, how the linear classifiers find the best set of weights on the IT population without the benefit of visual experience with at least some similar conditions). That is, we do not explore the representation's ability to generalize well outside its realm of experience. In particular, it is reasonable to expect that position-insensitive single neurons would facilitate generalization over position, (e.g. identifying an object at a position in which it has never been seen), and clutter-insensitive single neurons would facilitate identifying a familiar object among novel objects in a cluttered scene. That is, ideal properties depend on one's goal: at a descriptive level, transformation-sensitive neurons are more desirable for supporting a range of recognition tasks, and transformation-insensitive neurons may be more desirable for generalization (Fig 7). This might explain why the IT population contains a mix of highly transformation-sensitive and insensitive neurons (Zoccolan et al. 2007), but this still leaves open the mechanistic question of how those neurons are created (again, how they find the conditions to generalize over). Generalization is especially challenging in the case of clutter given the virtually infinite number of clutter conditions that can be encountered in natural vision. This may explain why the brain employs attentional resources to achieve higher clutter-invariance at the level of individual ventral stream neurons (Chelazzi et al. 1998b; Moran and Desimone 1985; Reynolds and Chelazzi 2004; Reynolds and Desimone 1999).

In this paper, we restricted ourselves to analysis on the recorded data, and followed by simulated data that mimics the real data but allowed us to systematically vary particular parameters of interest and examine their impact on population performance. While this approach gives us good understanding about the behavior of a particular type of neural code, it lacks a deep theoretical foundation. For example, all the clutter rules achieved similar performance because all the clutter rules produced response functions that preserved the rank-order of object selectivity, yet it remains unclear which class of response functions is mathematically optimal. Very likely, the preservation of rank-order object selectivity is not the sole attribute that determines the goodness of a representation to support object recognition. Probably, many attributes of a neuron's response function determine how closely they match the output response function (e.g. response function shape, size, etc). Here our results showed that among those different attributes, preservation of rank-order object preference is the most important. With assumptions about: 1) the tasks a representation supports; 2) the neuronal noise characteristic (e.g. Poisson); and 3) the readout mechanisms, the problem might be formalized mathematically (Ma et al. 2006; Salinas 2006). However, formalizing this in a theoretical framework is beyond the scope of this paper.

Moving forward

Minimally, we hope that the results and simulations presented here clarify the single-unit properties that enable a neuronal population to support different kinds of recognition tasks. We believe that these results offer at least two avenues of forward guidance. First, we predict that, if one estimates each neuron's rank-order preservation of object preferences in the face of image transformations (such as position and clutter), that property will gradually increase along the ventral visual hierarchy. This may be true even though the RF sizes or clutter sensitivity may vary widely (e.g., some IT neurons have smaller RFs than some V4 neurons). Future physiology studies should be geared more toward measuring selectivity across transformations rather than measuring response magnitude alone. These data are already available for some transformations, such as positions and size (Brincat and Connor 2004; Gross et al. 1993; Ito et al. 1995; Janssen et al. 2008), visual cue (Sary et al. 1993), occlusion (Kovacs et al. 1995) and clutter (Zoccolan et al. 2005), but more systematic measurements and comparisons across visual areas are needed. In particular, preservation of rank-order selectivity could potentially be used as a metric to probe the complexity of tuning for each representation (e.g., V1 neurons probably have good rank-order preservation for Gabor patch stimuli, but not for object stimuli, even if those objects are small enough to fit within their RF). Second, in contrast to preservation of the response magnitude, preservation of the rank-order object selectivity is a more precise and parsimonious goal for computational approaches aimed at capturing the mechanisms underlying a powerful object representation in the brain. The key question then is to understand how the ventral stream takes the initial response functions with little rank-order preservation (Fig 5A, V1 units) and achieves rank-order preservation at its highest stages (Fig 5A IT units). Understanding this is the crux of understanding how invariant object recognition is achieved.

Acknowledgements

This work was supported by National Eye Institute (NIH-R01-EY014970), the Pew Charitable Trusts (PEW University of California, San Francisco 2893sc) and the McKnight Foundation. N.L. was supported by a NEI Integrative Training Grant for Vision. D.D.C. was supported by a National Defense Science and Engineering Graduate Fellowship. D.Z. was supported by a Postdoctoral Fellowship of the International Human Frontier Science Program Organization and a by a Charles A. King Trust Postdoctoral Research Fellowship.

References

Aggelopoulos NC, and Rolls ET. Scene perception: inferior temporal cortex neurons encode the positions of different objects in the scene. *Eur J Neurosci* 22: 2903-2916, 2005.

Baylis GC, and Rolls ET. Responses of neurons in the inferior temporal cortex in short term and serial recognition memory tasks. *Experimental Brain Research* 65: 614-622, 1987.

Braitenberg V. *Cortical Architectonics: General and Areal. In Architectonics of the Cerebral Cortex*. New York: Raven, 1978.

Brincat SL, and Connor CE. Underlying principles of visual shape selectivity in posterior inferotemporal cortex. *Nat Neurosci* 7: 880-886, 2004.

Chelazzi L, Duncan J, Miller EK, and Desimone R. Responses of neurons in inferior temporal cortex during memory-guided visual search. *J Neurophysiol* 80: 2918-2940, 1998a.

Chelazzi L, Duncan J, Miller EK, and Desimone R. Responses of neurons in inferior temporal cortex during memory-guided visual search. *J Neurophysiology* 80: 2918-2940, 1998b.

DeAngelis GC, Ohzawa I, and Freeman RD. Spatiotemporal organization of simple-cell receptive fields in the cat's striate cortex. I. General characteristics and postnatal development. *J Neurophysiol* 69: 1091-1117, 1993.

Desimone R, and Duncan J. Neural mechanisms of selective visual attention. *Annual Review of Neuroscience* 18: 193-222, 1995.

DiCarlo JJ, and Cox DD. Untangling invariant object recognition. *Trends in Cognitive Sciences* 11: 333-341, 2007.

DiCarlo JJ, and Maunsell JHR. Anterior Inferotemporal Neurons of Monkeys Engaged in Object Recognition Can be Highly Sensitive to Object Retinal Position. *J Neurophysiol* 89: 3264-3278, 2003.

DiCarlo JJ, and Maunsell JHR. Form representation in monkey inferotemporal cortex is virtually unaltered by free viewing. *Nat Neurosci* 3: 814-821, 2000.

Duda RO, Hart PE, and Stork DG. *Pattern Classification*. New York: Wiley-Interscience, 2001.

Edelman S, and Intrator N. Towards structural systematicity in distributed, statically bound visual representations. *Cognitive Science* 27: 73-110, 2003.

Fisher R. The use of multiple measurements in taxonomic problems. *Annals of Eugenics* 7: 179-188, 1936.

Freedman DJ, and Assad JA. Experience-dependent representation of visual categories in parietal cortex. *Nature* 443: 85-88, 2006.

Freedman DJ, Riesenhuber M, Poggio T, and Miller EK. Categorical representation of visual stimuli in the primate prefrontal cortex. *Science* 291: 312-316., 2001.

Gochin PM. Properties of simulated neurons from a model of primate inferior temporal cortex. *Cereb Cortex* 4: 532-543, 1994.

Gross CG, Rodman HR, Gochin PM, and Colombo MW. Inferior Temporal Cortex as a Pattern Recognition Device. In: *Computational Learning & Cognition*, edited by Baum EBSoc for Industrial & Applied Math, 1993, p. 44-73.

Hung CP, Kreiman G, Poggio T, and DiCarlo JJ. Fast readout of object identity from macaque inferior temporal cortex. *Science* 310: 863-866, 2005.

Ito M, Tamura H, Fujita I, and Tanaka K. Size and position invariance of neuronal responses in monkey inferotemporal cortex. *Journal of Neurophysiology* 73: 218-226, 1995.

Janssen P, Srivastava S, Ombelet S, and Orban GA. Coding of shape and position in macaque lateral intraparietal area. *J Neurosci* 28: 6679-6690, 2008.

Jones JP, and Palmer LA. An evaluation of the two-dimensional Gabor filter model of simple receptive fields in cat striate cortex. *J Neurophysiol* 58: 1233-1258, 1987.

Kohn A, and Movshon JA. Adaptation changes the direction tuning of macaque MT neurons. *Nat Neurosci* 7: 764-772, 2004.

Kovacs G, Vogels R, and Orban GA. Cortical correlate of pattern backward masking. *Proc Natl Acad Sci U S A* 92: 5587-5591., 1995.

Logothetis NK, and Sheinberg DL. Visual object recognition. *Ann Rev Neurosci* 19: 577-621, 1996.

Ma WJ, Beck JM, Latham PE, and Pouget A. Bayesian inference with probabilistic population codes. *Nat Neurosci* 9: 1432-1438, 2006.

Miller EK, Gochin PM, and Gross CG. Suppression of visual responses of neurons in inferior temporal cortex of the awake macaque by addition of a second stimulus. *Brain Res* 616: 25-29, 1993.

Missal M, Vogels R, Li C, and Orban GA. Shape interactions in macaque inferior temporal neurons. *Journal of Neurophysiology* 82: 131-142, 1999.

Moran J, and Desimone R. Selective attention gates visual processing in the extrastriate cortex. *Science* 229: 782-784, 1985.

Op de Beeck H, and Vogels R. Spatial sensitivity of macaque inferior temporal neurons. *J Comp Neurol* 426: 505-518., 2000.

Poggio T. A theory of how the brain might work. *Cold Spring Harb Symp Quant Biol* 55: 899-910, 1990.

Reynolds JH, and Chelazzi L. Attentional modulation of visual processing. *Annu Rev Neurosci* 27: 611-647, 2004.

Reynolds JH, and Desimone R. The role of neural mechanisms of attention in solving the binding problem. *Neuron* 24: 19-29, 111-125, 1999.

Riesenhuber M, and Poggio T. Are cortical models really bound by the "binding problem"? *Neuron* 24: 87-93, 111-125., 1999.

Rifkin R, Bouvrie J, Schutte K, Chikkerur S, Kouh M, Ezzat T, and Poggio T. Phonetic classification using linear regularized least squares and second-order features. *CBCL Paper/AI Technical Report, Massachusetts Institute of Technology* 2007-019: 2007.

Rolls ET, Aggelopoulos NC, and Zheng F. The receptive fields of inferior temporal cortex neurons in natural scenes. *J Neurosci* 23: 339-348, 2003.

Rolls ET, and Tovee MJ. The responses of single neurons in the temporal visual cortical areas of the macaque when more than one stimulus is present in the receptive field. *Exp Brain Res* 103: 409-420, 1995.

Roudi Y, and Treves A. Representing where along with what information in a model of a cortical patch. *PLoS Comput Biol* 4: e1000012, 2008.

Salinas E. How behavioral constraints may determine optimal sensory representations. *PLoS Biol* 4: e387, 2006.

Sanger TD. Neural population codes. *Curr Opin Neurobiol* 13: 238-249, 2003.

Sary G, Vogels R, and Orban GA. Cue-invariant shape selectivity of macaque inferior temporal neurons. *Science* 260: 995-997, 1993.

Sato T. Interactions of visual stimuli in the receptive fields of inferior temporal neurons in awake macaques. *Experimental Brain Research* 77: 23-30, 1989.

Schwartz EL, Desimone R, Albright TD, and Gross CG. Shape recognition and inferior temporal neurons. *Proceedings of the National Academy of Science (USA)* 80: 5776-5778, 1983.

Serre T, Kreiman G, Kouh M, Cadieu C, Knoblich U, and Poggio T. A quantitative theory of immediate visual recognition. *Prog Brain Res* 165: 33-56, 2007.

Shadlen MN, Britten KH, Newsome WT, and Movshon JA. A computational analysis of the relationship between neuronal and behavioral responses to visual motion. *J Neurosci* 16: 1486-1510, 1996.

Shadlen MN, and Movshon JA. Synchrony unbound: A critical evaluation of the temporal binding hypothesis. 1999.

Shadlen MN, and Newsome WT. The variable discharge of cortical neurons: Implications for connectivity, computation and information coding. *J Neuroscience* 18: 3870-3896, 1998.

Sheinberg DL, and Logothetis NK. Noticing familiar objects in real world scenes: the role of temporal cortical neurons in natural vision. *J Neurosci* 21: 1340-1350., 2001.

Sundberg KA, Mitchell JF, and Reynolds JH. Spatial attention modulates center-surround interactions in macaque visual area v4. *Neuron* 61: 952-963, 2009.

Tanaka K. Inferotemporal cortex and object vision. *Annual Review of Neuroscience* 19: 109-139, 1996.

Tolhurst DJ, Movshon JA, and Dean AF. The statistical reliability of signals in single neurons in cat and monkey visual cortex. *Vision Res* 23: 775-785, 1983.

Tovée MJ, Rolls ET, and Azzopardi P. Translation invariance in the responses to faces of single neurons in the temporal visual cortical areas of the alert monkey. *Journal of Neurophysiology* 72: 1049-1060, 1994.

Treisman A. Solutions to the binding problem: progress through controversy and convergence. *Neuron* 24: 105-110, 111-125, 1999.

Vogels R, and Orban GA. Coding of stimulus invariances by inferior temporal neurons. *Prog Brain Res* 112: 195-211, 1996.

Zoccolan D, Cox DD, and DiCarlo JJ. Multiple object response normalization in monkey inferotemporal cortex. *J Neurosci* 25: 8150-8164, 2005.

Zoccolan D, Kouh M, Poggio T, and DiCarlo JJ. Trade-off between object selectivity and tolerance in monkey inferotemporal cortex. *J Neurosci* 27: 12292-12307, 2007.

Figure Legend

Figure 1 **(A)** Visual recognition tasks. Three objects (star, triangle, cross) were shown at three possible positions (-2°, 0°, and +2° relative to the fovea) either in isolation or in combinations of pairs or triplets. Using the IT population response data to each visual “scene”, linear discriminant classifiers were used to measure how well the population had solved two different visual recognition tasks. One task required the linear discriminants to classify object identity irrespective of its position, (“position-invariant task”). In the particular example illustrated, the classifier was asked to classify the presence of a star (report "yes" to all visual displays that contain a star regardless of the star’s position). In the other task, the classifier had to report object identity at a particular position, (“position-specific task”). In the example illustrated, the classifier had to report "yes" only to the visual scenes in which the star was present in the top position while disregarding other displays (even those in which the star was present in another position). **(B)** Classification performance for a real IT population and a simulated V1 population on the "position-invariant" and "position-specific" tasks. All performance was averaged performance using "leave-one-out" cross validation procedure (See details in Methods).

Figure 2 A schematic drawing of the simulation design and tasks. **(A)** The two recognition tasks that each simulated population was asked to solve. The tasks are analogous to those tested for the real IT population (c.f. Fig. 1A). On the left, the “2D stimulus space” is displayed: the y-axis shows a dimension of object shape (identity) and the x-axis shows a dimension of retinal position, and a point in

the space corresponds to the presence of a single visual object at some position (in a scene). One example question for each task is illustrated by a black rectangular region. For these questions, visual scenes that contain a point within the black region should be reported as "yes". To approximate the three objects and three positions used during the collection of the real IT data (Fig. 1A), all scenes were drawn to contain points only within the nine dotted squares regions (objects A,B,C; positions X,Y,Z). The tasks are re-displayed on the right in the same format as Figure 1A. **(B)** The response profile of an example simulated IT unit in the 2D stimulus space. **(C)** An example simulated IT population (i.e. a set of simulated units like that in (B), but with randomly chosen center positions, see Methods for details). Each colored circle indicates one unit. The color indicates the strength of spiking response.

Figure 3 The effect of single-neuron position and clutter sensitivity on population recognition performance. **(A)** Population performance on the recognition tasks with visual scenes containing single objects. Performance was averaged over multiple simulation runs; error bars indicate standard deviation. The dash line indicates the performance from shuffled runs (i.e. chance). The performance of the invariant populations performed above chance on the "position-specific" task because the neurons were sensitive to object identity and therefore conveyed some information about this conjoint identity and position task. **(B)** Example populations illustrating different amount of single-neuron position sensitivity. Each column is an example population consisted of neurons with a particular σ_p. Within each column, the top plot shows the responses of all the units to their most preferred object across changes in that object's position. The bottom panel shows the responses of an example unit to three different objects. The shape selectivity of all neurons was the same (i.e. same σ_s). **(C)** Population performance on visual scenes containing multiple objects. Different colors represent data from populations with different single-neuron clutter sensitivity (blue, CCI; red, LIN; green, AVG; magenta, DIV). Because the simulation parameters and populations were not exactly matched, one should not make direct comparison of the absolute performance between (A) and (C). The performance obtained using σ_p= 0.3 is shown in insert for better comparison. (**D**) An illustration of a single-unit's responses to single objects and pairs of objects, under different clutter rules.

Figure 4 Real IT neurons show more preserved rank-order object preference than simulated V1 units.

(**A**) Neuronal responses to two objects at three positions for an example simulated V1 unit and a real IT neuron. (**B**) Distributions of rank order preservation across position for the V1 and IT population. The rank order preservation was quantified using a standard separability index metric (Brincat and Connor 2004; Janssen et al. 2008). The distributions contain 68 cases for V1 and 32 cases for IT. (**C**) Neuronal responses to two objects across different clutter conditions. (**D**) Distributions of rank order preservation across clutter conditions. The distributions contain 68 cases for V1 and 63 cases for IT, see Methods.

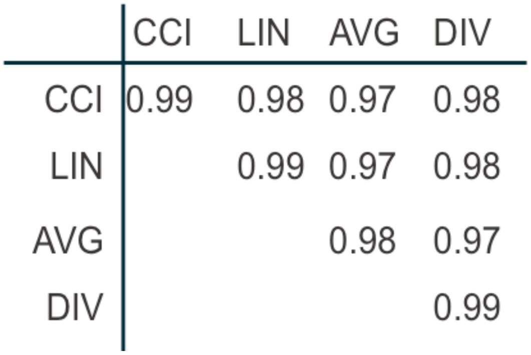
Figure 5 The effect of single-unit rank-order preservation on population recognition performance. (**A**) Example single-units that maintained rank-order object preference (e.g. IT) or not (e.g. V1) across position. (**B**) Averaged population performance on position-invariant task. (**C**) Example units from the populations in (B). All units in each population had similarly "shaped" response functions, but positioned randomly to cover the 2D stimulus space, see Methods. (**D**) Example single-units that maintained rank-order object preference (e.g. AVG) or not (e.g. $(iv)_c$) across clutter conditions. (**E**) Averaged population performance on clutter-invariant task, same as (B).

Figure 6 Rank-order preservation of single-units, not sensitivity to transformations, predicts population performance on invariant recognition tasks. **(A)** Combined average performance on the invariant recognition tasks across all the simulated populations when they are sorted by their single-unit rank-order preservation or single-unit sensitivity to position or clutter. The degree of sensitivity was a measure of a neuron's average response reduction from its preferred stimulus, (see Methods). Each dot on the plot is the performance from one population. The performance of simulated IT (blue) and V1 (red) populations is highlighted in the plots on the position-invariant recognition task. The more clutter sensitive populations appear to perform slightly better than the less clutter sensitive populations because LIN, AVG, and DIV all qualified as clutter sensitive when sorted by their clutter sensitivity. (**B**) Combined average performance on the recognition task in clutter when rank-order preserved populations were further sorted based on their single-unit separability. A joint tuning is strictly separable (i.e. independent) if it is the product of the tuning along individual stimulus dimensions. Some rank-order preserved populations could have non-independent tunings (e.g. CCI, DIV). (**C**) IT and V1 population performance on the position-invariant recognition task (single object) as a function of unit number.

Error bars in (B) and (C) indicate standard deviations.

Figure 7 Summary of the goodness of different single-unit response properties for supporting invariant object recognition tasks at the population level. In each subplot, the x-axis shows the values of some identity-preserving transformation (e.g. object retinal position, or the presence of different distractor objects, see Fig. 5); the y-axis shows the response of hypothetical single neurons to three different objects (red, blue and green; red is the "preferred" object in all cases). Major y-axis: single neurons can have a range of response sensitivity to a particular transformation X (e.g. for position, the receptive field size in response to the preferred object). Major x-axis: neurons may also preserve or not preserve their object rank-order. Among these single-unit response properties, rank-order preservation is much more predictive of the population's ability to support invariant recognition (assuming equal numbers of neurons in each population; see Fig. 6 and Methods). For example, neurons can be largely insensitive to both position and clutter, yet form an inadequate population (see Fig. 6). Conversely, neurons can be highly sensitive to both position and clutter and still form a very good population representation. (* Note, large RF is bad for position-specific recognition, but potentially useful for generalization over position). The term "invariant" has been used to describe the idealized neuron in the upper right plot, and the term "tolerant" to reflect the limited invariance of real neurons and real behavior.

Table



	CCI	LIN	AVG	DIV
CCI	0.99	0.98	0.97	0.98
LIN		0.99	0.97	0.98
AVG			0.98	0.97
DIV				0.99

Table 1 Correlations between the discriminant weights used to read-out populations implementing different clutter "rules". The diagonal in the table is the correlation of the weights vectors for the same populations obtained across different simulation runs, thus the values on the diagonal is an estimate of the upper-bound on the correlation values given the noise.

Figure 1

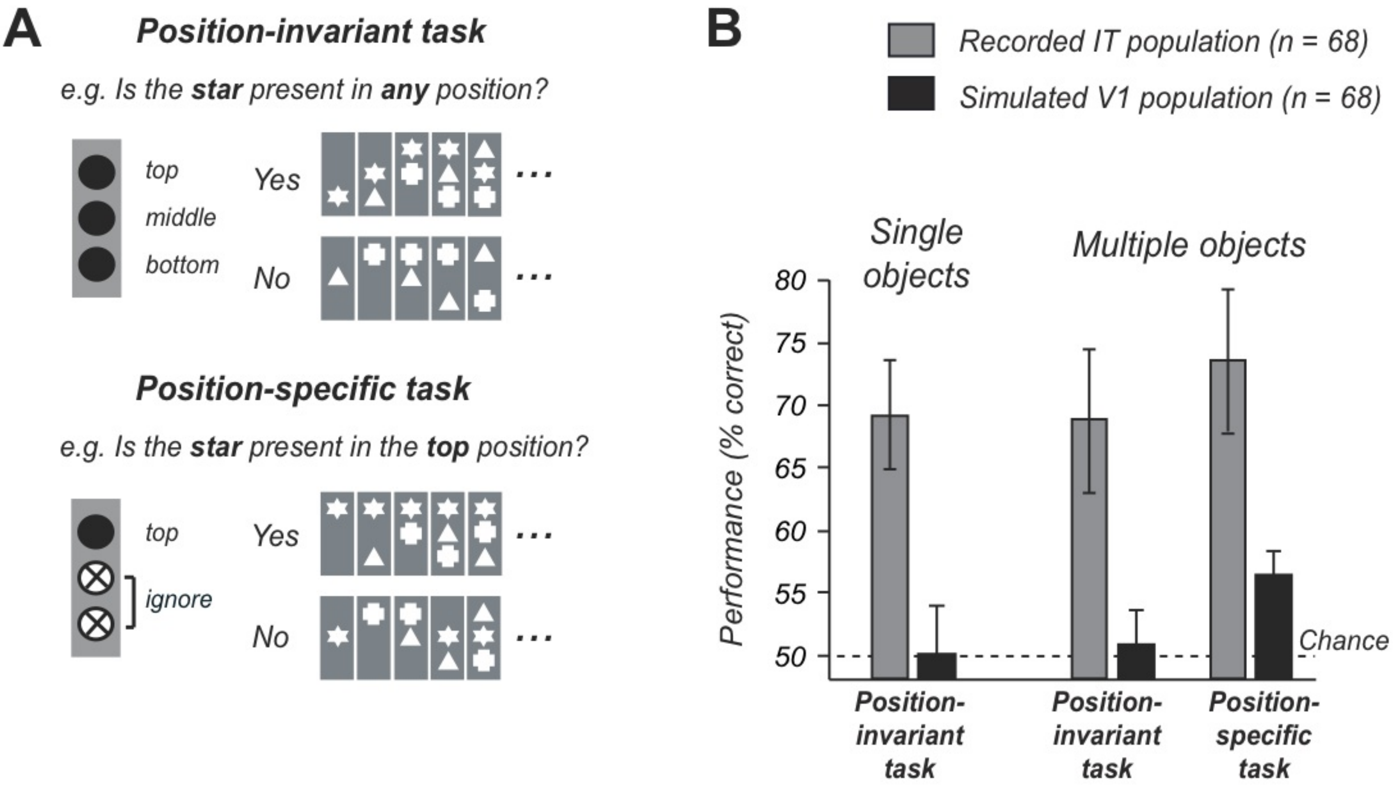

A
Position-invariant task
e.g. Is the star present in any position?
top
middle
bottom
Yes
No
. . .
Position-specific task
e.g. Is the star present in the top position?
top
ignore
Yes
No
. . .
B
Recorded IT population (n = 68)
Simulated V1 population (n = 68)
Single objects
Multiple objects
Performance (% correct)
80
75
70
65
60
55
50
Chance
Position-invariant task
Position-invariant task
Position-specific task

Figure 2

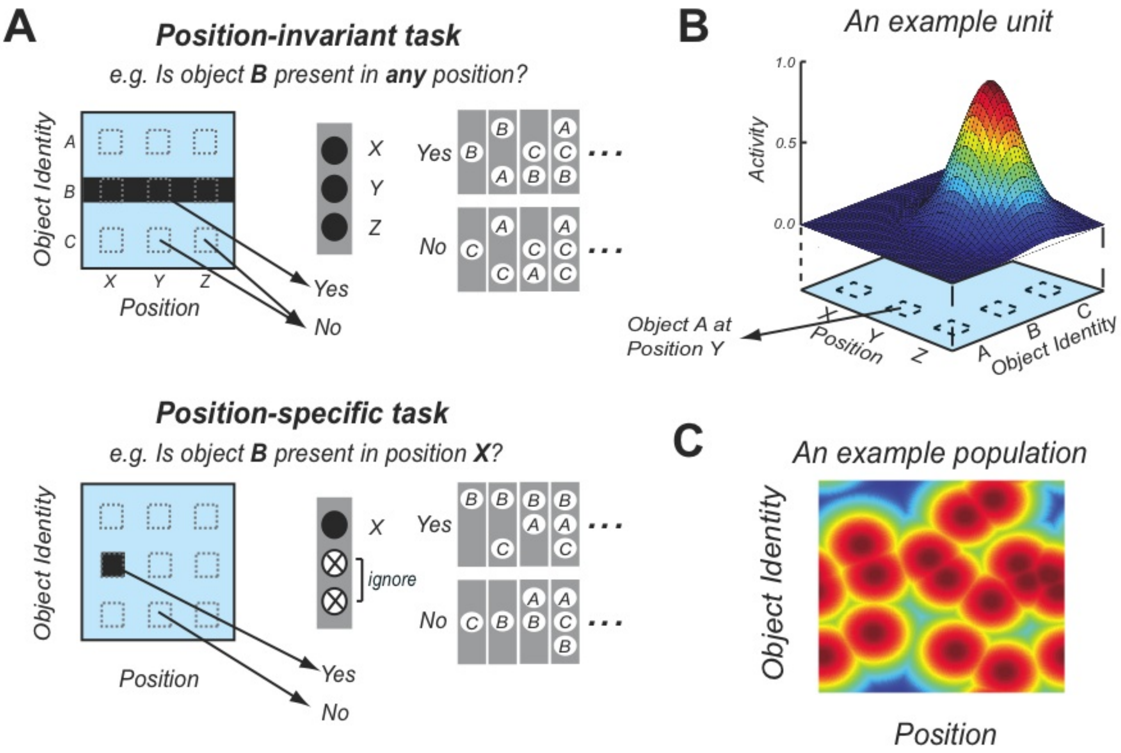

A
Position-invariant task
e.g. Is object B present in any position?
Object Identity
A
B
C
X
Y
Z
Position
X
Y
Z
Yes
No
Yes
B
B
A
C
B
A
C
B
No
C
A
C
C
A
A
C
C
...
...
Position-specific task
e.g. Is object B present in position X?
Object Identity
Position
X
ignore
Yes
No
Yes
B
B
C
B
A
B
A
C
No
C
B
A
B
A
C
B
...
...
B
An example unit
Activity
1.0
0.5
0.0
Object A at Position Y
X
Y
Z
Position
A
B
C
Object Identity
C
An example population
Object Identity
Position

Figure 3

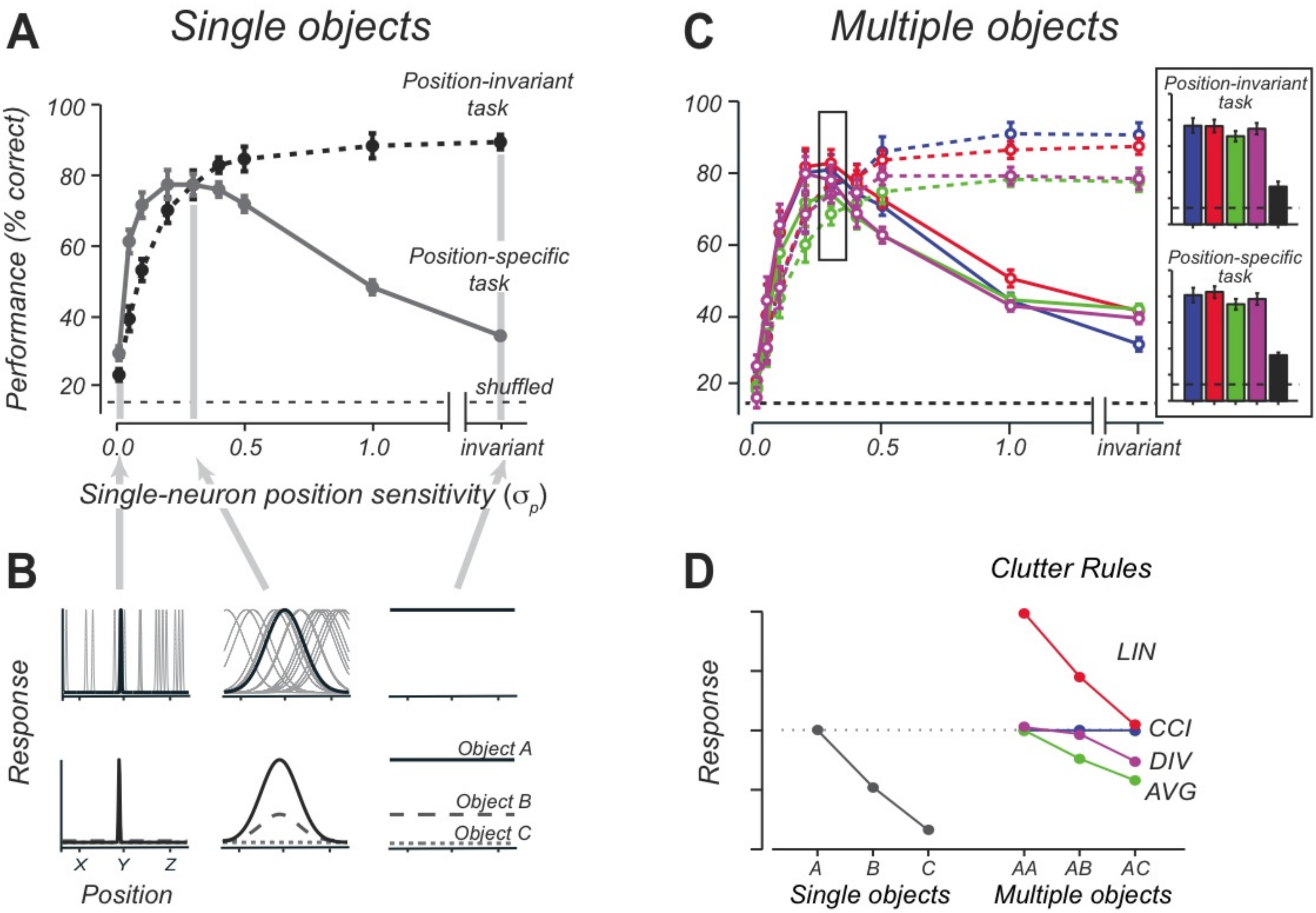

A
Single objects
Performance (% correct)
100
80
60
40
20
Position-invariant task
Position-specific task
shuffled
0.0
0.5
1.0
invariant
Single-neuron position sensitivity (σ_p)
B
Response
Object A
Object B
Object C
X
Y
Z
Position
C
Multiple objects
100
80
60
40
20
0.0
0.5
1.0
invariant
Position-invariant task
Position-specific task
D
Clutter Rules
Response
LIN
CCI
DIV
AVG
A
B
C
AA
AB
AC
Single objects
Multiple objects

Figure 4

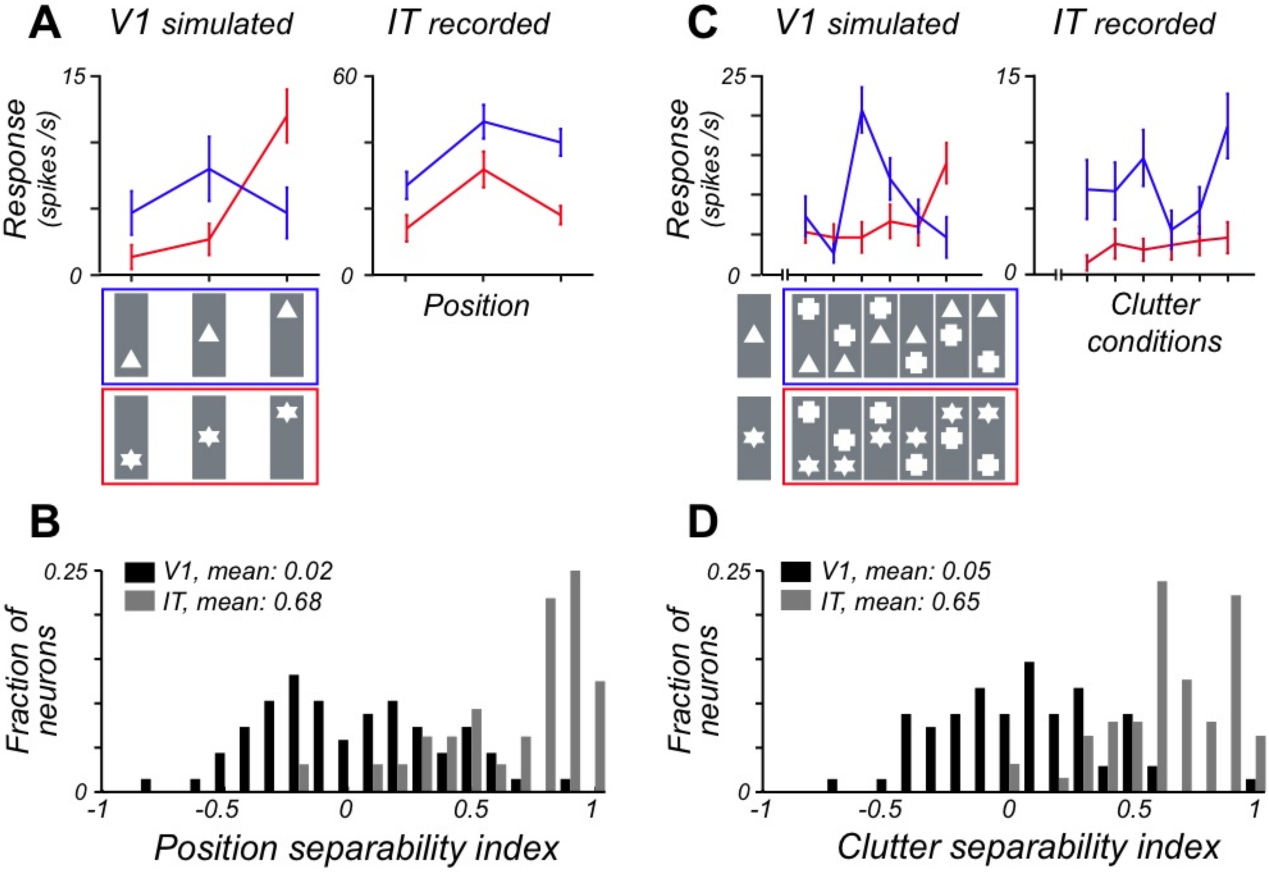

A
V1 simulated
IT recorded
Response (spikes /s)
15
0
60
0
Position
B
V1, mean: 0.02
IT, mean: 0.68
Fraction of neurons
0.25
0
-1
-0.5
0
0.5
1
Position separability index
C
V1 simulated
IT recorded
Response (spikes /s)
25
0
15
0
Clutter conditions
D
V1, mean: 0.05
IT, mean: 0.65
Fraction of neurons
0.25
0
-1
-0.5
0
0.5
1
Clutter separability index

Figure 5

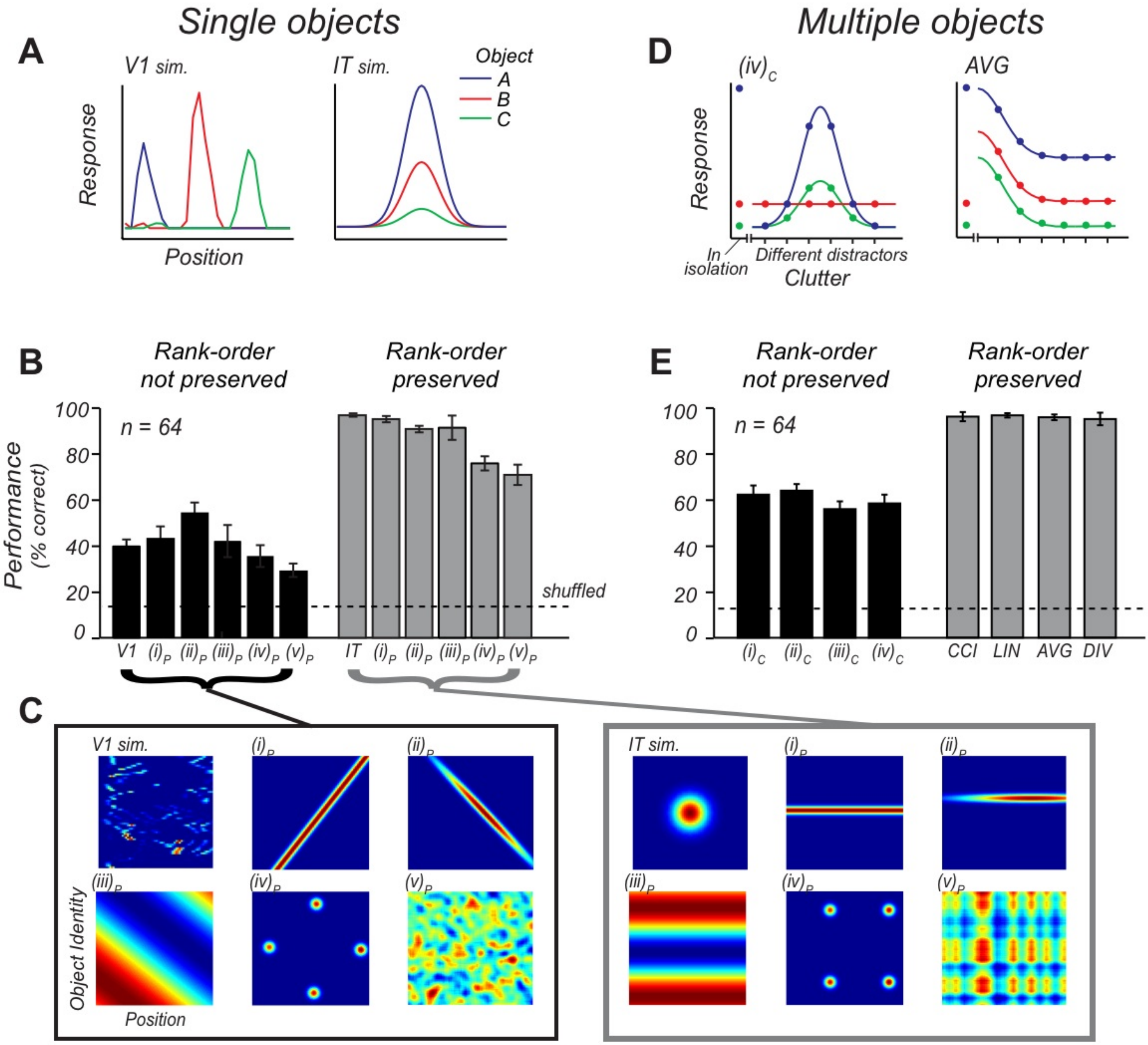
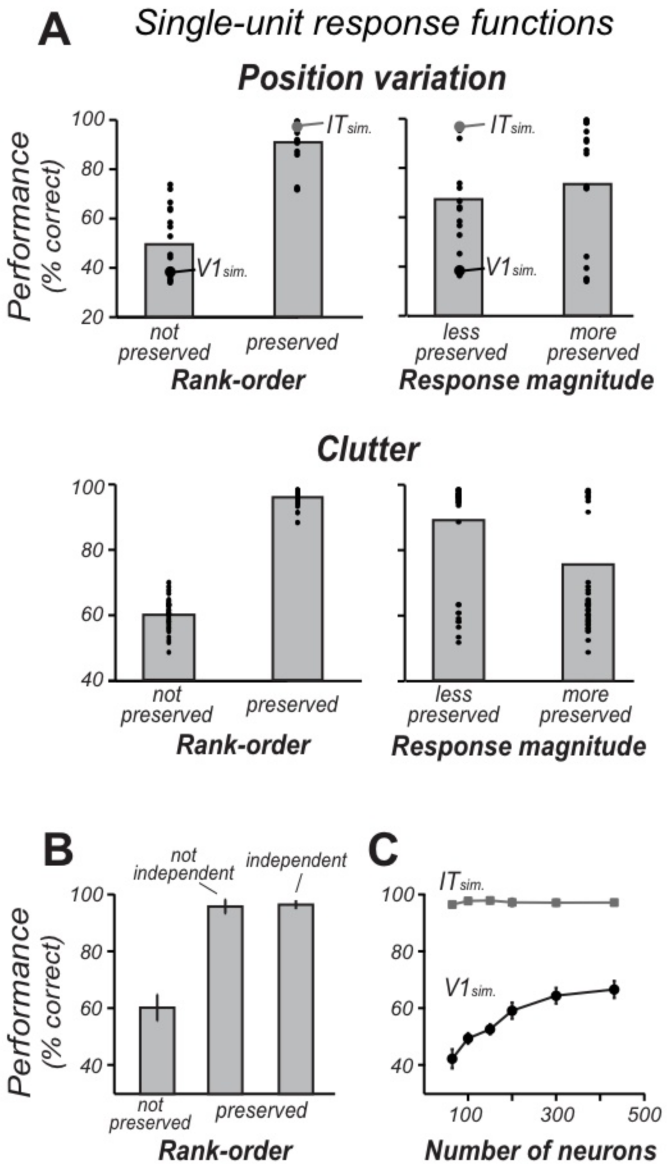
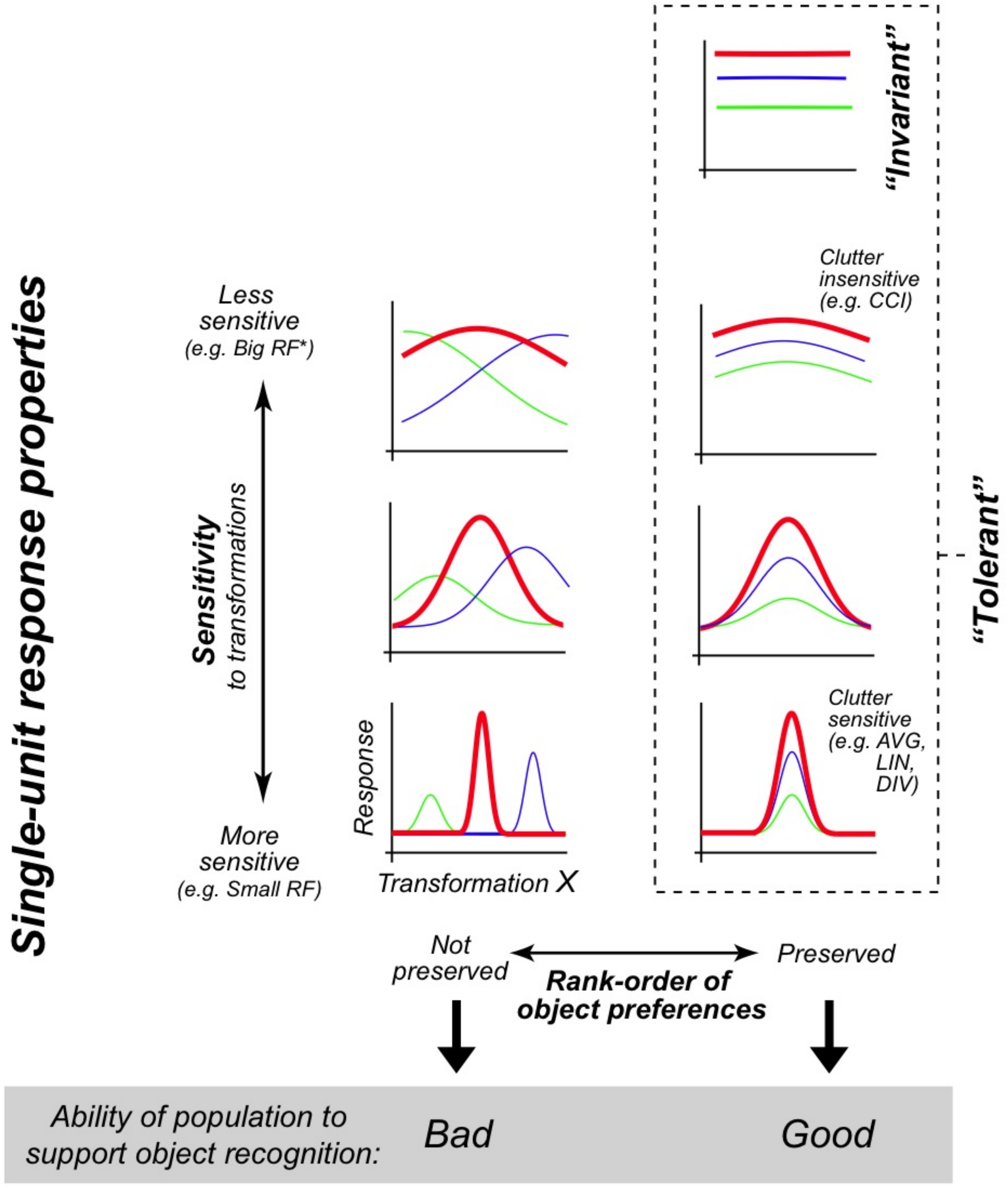

Single objects
Multiple objects
A
V1 sim.
IT sim.
Object
A
B
C
Response
Position
D
(iv)C
AVG
Response
In isolation
Different distractors
Clutter
B
Rank-order not preserved
Rank-order preserved
100
80
60
40
20
0
n = 64
Performance (% correct)
shuffled
V1
(i)P
(ii)P
(iii)P
(iv)P
(v)P
IT
(i)P
(ii)P
(iii)P
(iv)P
(v)P
E
Rank-order not preserved
Rank-order preserved
100
80
60
40
20
0
n = 64
(i)C
(ii)C
(iii)C
(iv)C
CCI
LIN
AVG
DIV
C
V1 sim.
(i)P
(ii)P
(iii)P
(iv)P
(v)P
Object Identity
Position
IT sim.
(i)P
(ii)P
(iii)P
(iv)P
(v)P

Figure 6

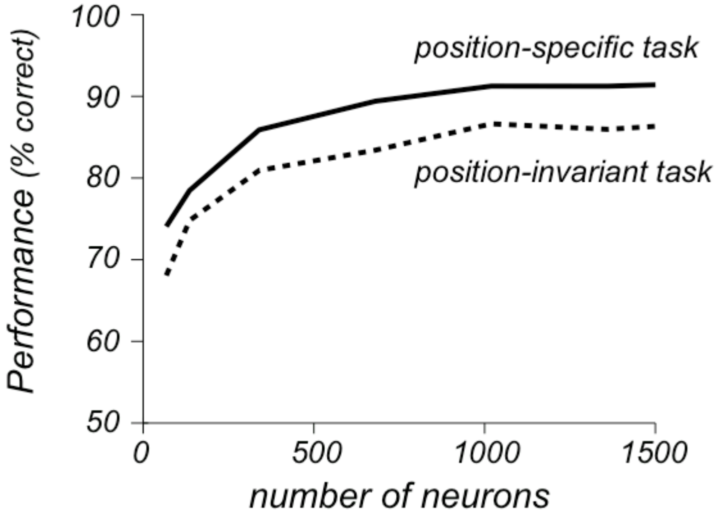


A
Single-unit response functions
Position variation
Performance (% correct)
100
80
60
40
20
IT sim.
V1 sim.
not preserved
preserved
Rank-order
IT sim.
V1 sim.
less preserved
more preserved
Response magnitude
Clutter
100
80
60
40
not preserved
preserved
Rank-order
less preserved
more preserved
Response magnitude
B
Performance (% correct)
not independent
independent
100
80
60
40
not preserved
preserved
Rank-order
C
IT sim.
V1 sim.
100
80
60
40
100
300
500
Number of neurons

Figure 7


Single-unit response properties
"Invariant"
Less sensitive (e.g. Big RF*)
Clutter insensitive (e.g. CCI)
Sensitivity to transformations
"Tolerant"
More sensitive (e.g. Small RF)
Response
Transformation X
Clutter sensitive (e.g. AVG, LIN, DIV)
Not preserved
Rank-order of object preferences
Preserved
Ability of population to support object recognition:
Bad
Good

Supplemental figure 1



To overcome the inherent trial-to-trial variability in the recorded neurons and estimate the absolute performance that can be achieved with our recorded selectivity, we synthesized larger populations of Poisson-spiking neurons from the response profiles of the measured IT population (n=68). Note that this procedure does not assume any selectivity that is not already in the recorded data since the synthesized neuron are just copies of the one of the original 68 neurons. However, increasing the population size does allow for pooling of the responses to overcome the response variability, thus increasing the fidelity of the readout. The plot shows the readout performance on the two tasks as we increased the number of neurons.