Unsupervised learning of invariant object representation in primate visual cortex

by

Nuo Li

B.S. Biomedical Engineering
Washington University, 2005

SUBMITTED TO THE DEPARTMENT OF BRAIN AND COGNITIVE SCIENCES IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY
AT THE
MASSACHUSETTS INSTITUTE OF TECHNOLOGY

JUNE 2011

Signature of Author: ____
Department of Brain and Cognitive Sciences
June, 2011

Certified by: ________________________
James J. DiCarlo
Associate Professor of Neuroscience
s Supervisor

Accepted by: ________________________
Earl K. Miller
Picower Professor of Neuroscience
Director, BCS Graduate Program

Unsupervised learning of invariant object representation in primate visual cortex

by
Nuo Li

SUBMITTED TO THE DEPARTMENT OF BRAIN AND COGNITIVE SCIENCES IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DOCTORAL DEGREE

ABSTRACT

Visual object recognition (categorization and identification) is one of the most fundamental cognitive functions for our survival. Our visual system has the remarkable ability to convey to us visual object and category information in a manner that is largely tolerant ("invariant") to the exact position, size, pose of the object, illumination, and clutter. The ventral visual stream in non-human primate has solved this problem. At the highest stage of the visual hierarchy, the inferior temporal cortex (IT), neurons have selectivity for objects and maintain that selectivity across variations in the images. A reasonably sized population of these tolerant neurons can support object recognition. However, we do not yet understand how IT neurons construct this neuronal tolerance.

The aim of this thesis is to tackle this question and to examine the hypothesis that the ventral visual stream may leverage experience to build its neuronal tolerance. One potentially powerful idea is that time can act as an implicit teacher, in that each object's identity tends to remain temporally stable, thus different retinal images of the same object are temporally contiguous. In theory, the ventral stream could take advantage of this natural tendency and learn to associate together the neuronal representations of temporally contiguous retinal images to yield tolerant object selectivity in IT cortex.

In this thesis, I report neuronal support for this hypothesis in IT of non-human primates. First, targeted alteration of temporally contiguous experience with object images at different retinal positions rapidly reshaped IT neurons' position tolerance. Second, similar temporal contiguity manipulation of experience with object images at different sizes similarly reshaped IT size tolerance. These instances of experience-induced effect were similar in magnitude, grew gradually stronger with increasing visual experience, and the size of the effect was large. Taken together, these studies show that unsupervised, temporally contiguous experience can reshape and build at least two types of IT tolerance, and that they can do so under a wide range of spatiotemporal regimes encountered during natural visual exploration. These results suggest that the ventral visual stream uses temporal contiguity visual experience with a general unsupervised tolerance learning (UTL) mechanism to build its invariant object representation.

Thesis Supervisor: James J. DiCarlo
Title: Associate Professor of Neuroscience

ACKNOWLEDGEMENTS

During graduate school, I repeatedly asked myself what I was doing and what I truly want in life. At the end of it, it is by no means that I have the answers. But the process of asking these questions and making feeble attempts to try to peek into the answer has shaped my thinking and forced me to be explicit about what I value, enjoy, and appreciate in life.

Scientific pursuit has always been a significant part of this journey of discovery. And during times when goals and purposes of life are less clear, this pursuit has kept me focused and motivated to be ready for another day. This thesis captures only a glimpse of my academic life at MIT and this acknowledgement section cannot contain all my thanks to the people and events that had helped me along the way. Mere words are not enough to convey the full depth of it:

I am most grateful to my PhD advisor, Jim DiCarlo, who created an academic home for me. His character, both as a scientist and a teacher, taught me how to approach and handle problems not only in science, but also much beyond. I am grateful to the members of the DiCarlo lab, past and current. I would not be able to navigate the practice of scientific research without their support and teaching. I am grateful to my PhD committee, Dr. Nancy Kanwisher, Dr. Earl Miller, and Dr. John Assad, for their helpful comments and feedback on my work. I am grateful to the MIT community and my classmates for creating a sense of community and home for me. Finally, to all my friends and family members who are not a part of my academic circle, Jimmy, Yingying, Toby, Stacy, John, and many others.

To my father, a biochemist and a very brave man, who came to this country to foster a new home for my mom and I. He inspired me and gave me the opportunity to take the path I am taking today.

To my mother, a very strong woman, whose extraordinary wisdom shines through her ordinary life. She has always been there for me unconditionally.

To my fiancée, Shan, for accepting me in full and giving me the answers to some of the most important questions in life.

To them, I dedicate this thesis.

TABLE OF CONTENT

Chapter 1

Object recognition and its neuronal substrate

1. 1 The "invariance problem"

Our visual system can quickly parse a visual image into recognizable objects to guild decisions, actions, and memory formation (Potter 1976; Thorpe et al. 1996). The computations that underlie this ability are remarkable in that our object percept is little affected by (often dramatic) variations in the images. That is, each object may appear at different positions on the retina, in different scales and poses, and on different backgrounds. Each of these transformations of the object will lead to a very different pattern of photoreceptor activation on the retina, yet our visual system somehow represents those retinal images in a way that is high tolerant to any object-identity-preserving transformations. The process for achieving that tolerance is commonly known as solving the "invariance problem" (DiCarlo and Cox, 2007; Riesenhuber and Poggio 2000).

The ventral visual stream in non-human primates solves the "invariance problem" and underlies the visual recognition behavior (Fig. 1-1, Logothetis and Sheinberg 1996; Miyashita 1993; Tanaka 1996; DiCarlo and Cox 2007). The involvement of the ventral stream in recognition behavior is inferred from lesion and micro-stimulation studies and is supported by electrophysiology studies. At the highest stage of the stream, the inferior temporal cortex (IT), neurons are selective among complex objects and that selectivity is maintained within the neurons' receptive fields (Desimone, Albright, Gross, Bruce, 1984; Ito et al 1995; Brincat and Connor 2004; Hung et al. 2005); small populations of IT neurons can be read out to support invariant recognition tasks (Hung et al. 2005; Li et al. 2009); lesions of IT lead to deficits in

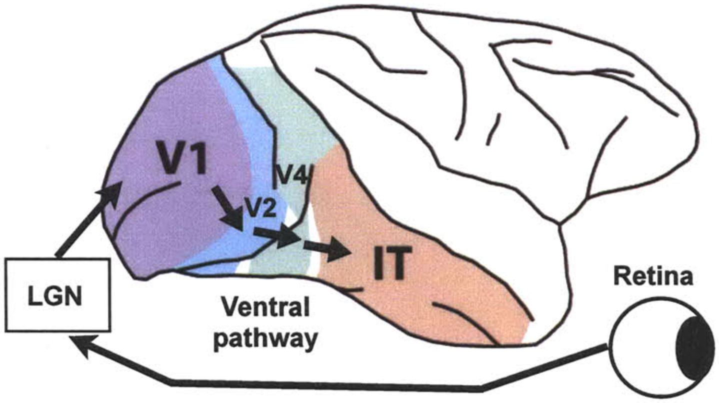


Figure 1-1: The ventral visual stream in non-human primates.

primates' recognition behavior (Dean 1982; Weiskrantz and Saunders 1984); artificial activation of IT neuronal ensembles can bias animals' perceptual reports in object discrimination tasks (Afraz et al. 2006).

The study of the "invariance problem" has attracted interest from neuroscientists, cognitive scientists, as well as the computer vision community, for both engineering and scientific reasons (Riesenhuber and Poggio 2000; DiCarlo and Cox, 2007; Pinto et al. 2008, 2010). From an engineering perspective, people hope to harness what they can learn from the brain to aid the designing of better computer vision systems. Recently, building of biologically-inspired computer vision systems has enjoyed great success (Serre and Poggio 2000; Pinto et al, 2009), but there is still a long way to go as state-of-the-art computer vision systems consistently underperform on difficult object recognition tasks when compared to human observers (Pinto et al, 2010). From a scientific perspective, because IT object representation is achieved through a series of cortical transformations along the ventral visual stream (i.e. V1--> V2--> V4--> IT; Felleman and Van Essen, 1991; Kobetake and Tanaka 1994; Rust and DiCarlo, 2009), given the repetitive nature of this hierarchical architecture, a detailed understanding of these transformations will likely teach us key principles of cortical computation. IT projects directly to brain areas responsible for decision, action, and memory; thus an understanding of IT representation allows us to understand the basic building blocks of memory and cognition. Indeed, IT is postulated to play a role in visual memory and short-term memory (Miller and Desimone 1994; Miller et al. 1991; Miyashita 1993; 1988; Sakai and Miyashita 1991). IT has potentially analogous structures in the human brain: the Lateral Occipital Complex (LOC), Fusiform Face Area (FFA), and Parahippocampal Place Area (PPA) underlie the processing of objects, faces and places (Kanwisher et al. 1997; Epstein and Kanwisher 1998; Grill-Spector et al. 2001); thus an understanding of object processing in monkeys will likely elucidate functions of higher-order visual processing in humans. Deficits in object recognition are symptoms of many neurological disorders (e.g. agnosia, Alzheimer's, autism, dyslexia); thus an understanding of the underlying circuitry for recognition may have clinical relevance as well.

1. 2 Object representation in the monkey ventral visual stream

Around the early 1970s, Charlie Gross made the seminal discovery that certain neurons in the temporal lobe of non-human primates respond to the sight of a hand and other complex shapes (Gross et al 1969, 1972). This was followed by numerous attempts to characterize the response properties of these temporal lobe neurons (for review, see Tanaka 1996, Logothetis and Sheinberg, 1996, Gross CG, 2002). Early neurophysiological studies of inferior temporal cortex (IT) led to the textbook notion that an "ideal" IT contains a set of shape detectors that each conveys information about object identity (shape selectivity) while being highly invariant to variations in the images. This gave rise to the popular (and somewhat misinterpreted) notion of the "grandmother cell", a neuron that only responds when the person's grandmother comes into sight. The "grandmother cell" was derived from the observation that some neurons in IT exhibit a high degree of selectivity and tolerance, and more recently, similar findings were also made in the medial temporal lobe of the human brain (Quiroga et al, 2005).

However, the "grandmother cell" idea suffers from a number of shortcomings, among which the most serious challenges are: 1) the problem of combinatorial explosion (that is, we need neurons to code for all possible objects we know, which also presents issues for downstream processing because of the huge dimensionality), and 2) the fact that we can easily recognize "novel" objects that we never encountered before. Contemporary neurophysiological data in non-human primates is also at odds with the "grandmother cell" theory. Several studies show that, when probed with a large set of visual stimuli, most individual IT neurons' responses exhibit a wide range of selectivity, and information about objects is conveyed by distributed patterns of activity across populations of neurons (Tanaka 2003; Pasupathy and Connor, 2002; Hung et al, 2005; Zoccolan et al 2007; Rust and DiCarlo, 2009). Furthermore, most individual IT neurons are not as tolerant as previously described (i.e., with receptive fields spanning tens of degrees in visual angle; Gross et al, 1969; Tovee et al, 1994); rather, most individual IT neurons only show limited tolerance in their firing rates to image transformations such as object position shift (with receptive fields that can be as small as a couple of degrees, DiCarlo and Maunsell, 2003), view change (Logothetis et al, 1994, 1995; Freedman et al, 2006), and are often strongly suppressed by objects in the background (i.e., clutter, Chelazzi et al. 1998; 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). Overall, the tolerance observed in single IT neurons' firing rates are often much more limited than the recognition behavior itself.

Over the years, the notion of a single neuron "feature detector" became a straw man or foil and gave way to a more distributed coding scheme. Previous investigators took notice of the fact that though IT neurons' receptive field size varies dramatically across studies (perhaps partly due to differences in the choice of stimuli), most of the IT neurons, nevertheless, preserve their rank-order object preference across object position and size variation (Ito et al. 1995; Logothetis and Sheinberg 1996; Op de Beeck and Vogels 2000; Tovée et al. 1994; DiCarlo and Maunsell 2003), and it was proposed that a population of such neurons can support position and size tolerant recognition (e.g., Gross et al. 1993; Logothetis and Sheinberg 1996; Vogels and Orban 1996). This intuition, though never rigorously tested, is fairly straightforward to understand and appreciate for object position and size variation. However, such intuition becomes less clear for more challenging transformations such as clutter, and begins to break down for object pose variation. This is mostly due to a poor understanding of the basic feature elements that are encoded by IT neurons (i.e., IT basis set).

In the first part of this thesis (see Specific Aims below), I will test this previously proposed intuition and extend it to more difficult transformations such as clutter. We begin by making the observation that the intrinsic response properties of a small population of IT neurons (i.e., earliest part of response, no attentional cuing) can by themselves support object identification while tolerating some degree of position variation and clutter. We then use a combination of computer simulation and characterization of the real IT neuronal responses to try to identity what single neuron response properties are more or less important to enable the IT population to support such invariant recognition. The outcome from this approach will consolidate the existing single-unit data, outline neuronal measures of interest for future physiology experiments, and define the computational goals for building a "good" object representation to support invariant recognition tasks. However, it does not yet inform us how those desired response properties are built by the ventral visual stream.

1. 3 Building tolerant object representation from temporal contiguity of natural visual experience

Understanding the IT object representation will require solving the mystery of how IT attains its neuronal tolerance. There is debate over the role of experience (vs. innate mechanisms) in building tolerance in object recognition. Some work shows that experience is important (Bulthoff and Edelman 1992; Bulthoff et al. 1995; Dill and Edelman 2001; Dill and Fahle 1998; 1997; Foster and Kahn 1985; Hayward and Tarr 1997; Logothetis et al. 1995; Logothetis and Pauls 1995; Nazir and O'Regan 1990; Poggio 1990; Tarr and Gauthier 1998; Vetter et al. 1995) while other work shows that adult recognition can be tolerant without further experience (Biederman 1987; Biederman and Cooper 1991; Biederman and Gerhardstein 1993; Cooper et al. 1992; Ellis et al. 1989; Wang et al. 2005). These results are not incompatible with each other, and different findings are at least partly explained by the objects that are tested (e.g., novel conjunctions vs. commonly-occurring features, Tarr and Bulthoff 1998).

Previous computational work has provided a range of ideas as to how might the visual system acquire its tolerance from visual experience (e.g. Poggio 1990; Olshausen et al. 1993; Riesenhuber and Poggio 1999; Ullman and Soloviev 1999; Ringach and Shapley 2004; etc.). One powerful idea is that time can act as an implicit teacher, and the *temporal contiguity* of object features during natural visual experience can instruct the learning of tolerance, potentially in an unsupervised (bottom-up) manner (Foldiak 1991). The overarching logic is as follows: during natural visual experience, objects tend to remain present for seconds or more, while object

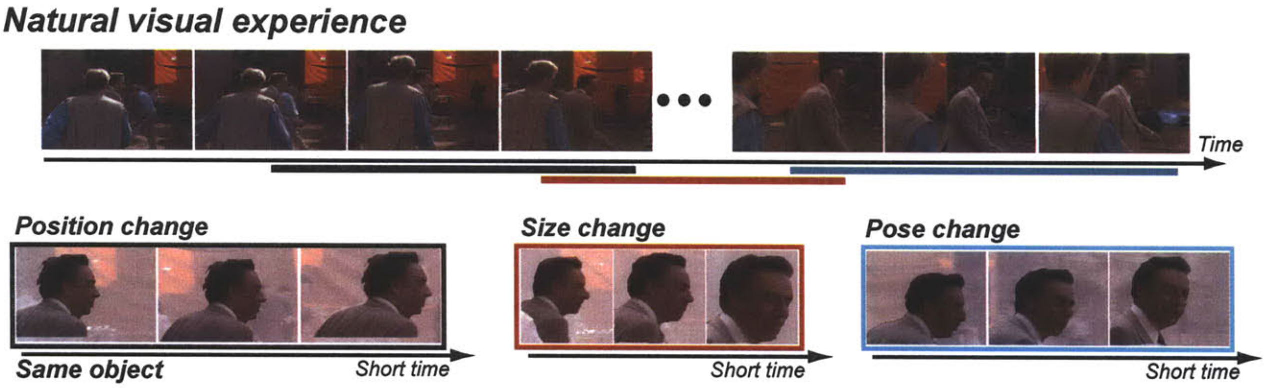


Figure 1-2: Temporal contiguity in natural visual experience. During natural visual experience, objects tend to persist for seconds or longer, thus different images acquired on the retina over short time periods tend to contain the same object. *Top*, a video sequence typical of what the visual system might naturally encounter. *Bottom*, segments of the video clip that unfold over short time periods to highlight the different types of image variation that could occur.

motion or viewer motion (e.g. eye movement) tends to cause rapid changes in the retinal image cast by each object over shorter time intervals (hundreds of ms). In theory, IT could construct a tolerant object representation by taking advantage of this natural tendency for temporally contiguous retinal images to belong to the same object (Fig. 1-2).

A number of studies have explored this idea in computational models of the ventral visual stream and show that learning to extract slowly-varying features across time can produce tolerant feature representations with units that mimic the response properties of the ventral stream neurons (Wiskott and Sejnowski 2002; Masquelier and Thorpe 2007; Sprekeler et al. 2007; Wyss et al. 2006). Furthermore, a number of psychophysical studies have shown that human object perception depends on the spatiotemporal contiguity of visual experience, and manipulating the spatiotemporal statistics of visual experience can alter the tolerance of human object perception (Edelman and Duvdevani-Bar 1997; Wallis and Bulthoff 1999; Cox et al. 2005).

There are also hints in the literature that neuronal signatures of temporal contiguity based learning may be present in the monkey ventral visual stream. In the late 1980s, Miyashita (later followed by other groups) found that certain IT and perirhinal neurons could learn to give similar responses to temporally nearby stimuli when instructed by reward (i.e., so-called "paired associates" learning; Messinger et al., 2001; Miyashita, 1988; Sakai and Miyashita, 1991), or sometimes, even in the absence of reward (Erickson and Desimone, 1999). Though these studies were originally motivated in the context of visual memory (Miyashita, 1993) and used visual presentation rates of seconds or more, it was recognized that the same associational learning across time might also be used to learn invariant visual features for object recognition (e.g. Foldiak, 1991; Stryker, 1991; Wallis, 1998; Wiskott and Sejnowski, 2002).

Motivated by the temporal contiguity learning theory, previous modeling, human psychophysics, and electrophysiology work, this thesis tests a form of the temporal contiguity hypothesis. I first measure baseline IT neuronal tolerance from awake behaving monkeys, I then provide appropriate unsupervised visual experience to those monkeys, which allows us to manipulate the spatiotemporal contiguity statistics of that experience. I then re-measure the IT neuronal tolerance after this experience and any change in the neuronal tolerance is compared to what would be predicted under the temporal contiguity hypothesis. The results from these experiments add to our understanding of how cortical representations are actively maintained by the sensory environment. This will set the stage for long-term studies on the fundamental question of how the ventral stream initially set up its object representations at multiple levels during the course of visual development.

1. 4 Specific Aims

The goal of my doctoral research is to contribute to our understanding of how the ventral visual stream constructs its tolerant object representation. To tackle this problem, I use a combination of experimental and computational approaches to examine two complementary issues: 1) what are the necessary response properties of single neurons to make up a "good" object representation (*aim 1*); and 2) what computational principles do the ventral stream neurons rely on to achieve those properties (*aim 2 & 3*). Below, I briefly summarize the key results borne out from experiments performed under each aim. These results are presented in detail in chapters 2, 3, and 4 of this thesis.

1. 4. 1 Aim 1. What single neuron responses property makes a "good" object representation? (position and clutter tolerance)

In *aim 1,* I try to infer what individual neuron response properties allow IT to underlie invariant object recognition from a population perspective. To do this, I explore a large number of potential population representations with different single-neuron makeup by means of computer simulation. Using linear read out tools to test these populations' ability to support invariant recognition tasks, these simulations allow us to ask what single-neuron properties best correlate with population performance. These simulations show that the crucial neuronal property to support recognition is not preservation of response magnitude, but preservation of each neuron's rank-order object preference under identity-preserving image transformations (e.g. position, clutter). This preservation of rank-order selectivity is consistent with the response properties observed in real IT neurons, whereas neurons in early visual areas (e.g. V1) lack it. Thus, we suggest this more precisely describes the goal of individual neurons at the top of the ventral visual stream. *This work was published in Li et al, J Neurophysiol 2009.*

1. 4. 2 Aim 2. How do ventral stream neurons acquire their tolerant response property? (position tolerance)

The approach described in *aim 1* offers useful description of neuronal data but does not tell us how those desired response properties are built by the ventral visual stream. To approach this question, I record from IT neurons in awake non-human primates to test computationally motivated hypotheses on how IT tolerance might be built. One powerful idea is that time can act as an implicit teacher, and the *temporal contiguity* of object features during natural visual experience can instruct the learning of tolerance, potentially in an unsupervised (bottom-up) manner (see section 1.3 above; Foldiak, 1991; Wiskott and Sejnowski, 2002; Masquelier et al, 2007). In theory, the ventral stream could construct a tolerant object representation by taking advantage of the natural tendency for temporally contiguous retinal images to belong to the same object and learn to associate the neuronal representations that occur closely in time. I recently found a neuronal signature of such learning in IT: temporally contiguous experience with different object images at different retinal positions can robustly reshape ("break") IT position tolerance, producing a tendency to confuse the identities of the temporally coupled objects across their manipulated positions. This unsupervised temporal tolerance learning (UTL) is substantial, increases with experience, and is significant in single IT neurons after just one hour of experience. *This work was published in Li & DiCarlo, Science 2008.*

1. 4. 3 Aim 3. Does unsupervised temporal slowness learning reflect a general mechanism for invariance learning? (size tolerance)

Does this newly uncovered IT neuronal learning (UTL) reflect a general unsupervised learning mechanism the ventral stream relies on to build and maintain its tolerance to all types of image transformations (e.g. to object size and pose variation)? In particular, the previous manipulation was deployed in the context of an eye movement, while eye movements drive a great amount of image statistics relevant for learning position tolerance, in the spatiotemporal regimes of natural vision, objects often undergo transformation as a result of object motion. This raises the question as to whether the involvement of eye movement is required for the learning to occur. To answer these questions, I tested UTL for an different image transformation: changes in object size. Extending the paradigm from our previous position tolerance work, the same type of unsupervised experience that reshapes IT position tolerance also predictably reshapes IT size tolerance, and the magnitude of reshaping is quantitatively similar. This tolerance reshaping can be induced under naturally occurring dynamic visual experience, even without eye movements. Furthermore, unsupervised temporally contiguous experience can build new neuronal tolerance. Taken together, these studies show that unsupervised, temporally contiguous experience can reshape and build at least two types of IT tolerance, and that they can do so under a wide range of spatiotemporal regimes encountered during natural visual exploration. *This work was published in Li & DiCarlo, Neuron 2010.*

1. 5 References

Afraz S, Kiani R, and Esteky H. Microstimulation of inferotemporal cortex influences face categorization. Nature 2006.

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.

Bi GQ, and Poo MM. Synaptic modifications in cultured hippocampal neurons: dependence on spike timing, synaptic strength, and postsynaptic cell type. J Neurosci 18: 10464-10472, 1998.

Biederman I. Recognition-by-components: a theory of human image understanding. Psychol Rev 94: 115-147, 1987.

Biederman I, and Cooper EE. Evidence for complete translational and reflectional invariance in visual object priming. Perception 20: 585-593, 1991.

Biederman I, and Gerhardstein PC. Recognizing depth-rotated objects: evidence and conditions for three- dimensional viewpoint invariance [published erratum appears in J Exp Psychol Hum Percept Perform 1994 Feb;20(1):80] [see comments]. J Exp Psychol Hum Percept Perform 19: 1162-1182, 1993.

Brady TF, and Oliva A. Statistical learning using real-world scenes: extracting categorical regularities without conscious intent. Psychol Sci 19: 678-685, 2008.

Bulthoff HH, and Edelman S. Psychophysical support for a two-dimensional view interpolation theory of object recognition. Proc Natl Acad Sci U S A 89: 60-64, 1992.

Bulthoff HH, Edelman S, and Tarr MJ. How are three-dimensional objects represented in the brain? Cerebral Cortex 3: 247-260, 1995.

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, 1998.

Cooper EE, Biederman I, and Hummel JE. Metric invariance in object recognition: a review and further evidence. Can J Psychol 46: 191-214., 1992.

Cox DD, Meier P, Oertelt N, and DiCarlo JJ. 'Breaking' position-invariant object recognition. Nat Neurosci 8: 1145-1147, 2005.

Dean P. Visual behavior in monkeys with inferotemporal lesions. In: Analysis of Visual Behavior, edited by Ingle D, Goodale M, and Mansfield JMIT Press, 1982, p. 587-627.

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.

DiCarlo JJ, Cox DD. Untangling invariant object recognition. Trends Cogn Sci 11:333-341, 2007.

Dill M, and Edelman S. Imperfect invariance to object translation in the discrimination of complex shapes. Perception 30: 707-724, 2001.

Dill M, and Fahle M. Limited translation invariance of human pattern recognition. Perception & Psychophysics 60: 65-81, 1998.

Dill M, and Fahle M. The role of visual field position in pattern-discrimination learning. Proc R Soc Lond B Biol Sci 264: 1031-1036, 1997.

Edelman S, and Duvdevani-Bar S. Similarity, connectionism, and the problem of representation in vision. Neural Comput 9: 701-720, 1997.

Ellis R, Allport DA, Humphreys GW, and Collis J. Varieties of object constancy. Q J Exp Psychol A 41: 775-796., 1989.

Epstein R, and Kanwisher N. A cortical representation of the local visual environment. Nature 392: 598-601, 1998.

Erickson CA, and Desimone R. Responses of macaque perirhinal neurons during and after visual stimulus association learning. J Neurosci 19: 10404-10416, 1999.

Felleman DJ and Van Essen DC. Distributed hierarchical processing in the primate cerebral cortex. Cereb Cortex 1: 1-47, 1991;

Fiser J, and Aslin RN. Statistical learning of higher-order temporal structure from visual shape sequences. J Exp Psychol Learn Mem Cogn 28: 458-467, 2002a.

Fiser J, and Aslin RN. Statistical learning of new visual feature combinations by infants. Proc Natl Acad Sci U S A 99: 15822-15826, 2002b.

Fiser J, and Aslin RN. Unsupervised statistical learning of higher-order spatial structures from visual scenes. Psychol Sci 12: 499-504, 2001.

Foldiak P. Learning invariance from transformation sequences. Neural Computation 3: 194-200, 1991.

Foster DH, and Kahn JI. Internal representations and operations in the visual comparison of transformed patterns: effects of pattern point-inversion, position symmetry, and separation. Biol Cybern 51: 305-312, 1985.

Freedman DJ, Riesenhuber M, Poggio T, and Miller EK. A comparison of primate prefrontal and iTemporal cortices during visual categorization. J Neurosci 23: 5235-5246, 2003.

Freedman DJ, Riesenhuber M, Poggio T, and Miller EK. Experience-dependent sharpening of visual shape selectivity in inferior temporal cortex. Cereb Cortex 16: 1631-1644, 2006.

Fu YX, Djupsund K, Gao H, Hayden B, Shen K, and Dan Y. Temporal specificity in the cortical plasticity of visual space representation. Science 296: 1999-2003, 2002.

Grill-Spector K, Kourtzi Z, and Kanwisher N. The lateral occipital complex and its role in object recognition. Vision Res 41: 1409-1422, 2001.

Gross CG, Bender DB, Rocha-Miranda CE. Visual receptive fields of neurons in inferotemporal cortex of the monkey. Science 166: 1303-1306, 1969.

Gross CG, Rocha-Miranda CE, Bender DB. Visual properties of neurons in the inferotemporal cortex of the Macaque. J Neurophysiol 35: 96-111, 1972.

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.

Gross CG. Genealogy of the "grandmother cell". Neuroscientist 8: 512-520, 2002.

Hayward WG, and Tarr MJ. Testing conditions for viewpoint invariance in object recognition. J Exp Psychol Hum Percept Perform 23: 1511-1521., 1997.

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.

Kanwisher N, McDermott J, and Chun MM. The fusiform face area: a module in human extrastriate cortex specialized for face perception. J Neurosci 17: 4302-4311, 1997.

Keysers C, Xiao DK, Foldiak P, and Perrett DI. The speed of sight. J Cogn Neurosci 13: 90-101., 2001.

Kobatake E, Tanaka K. Neuronal selectivities to complex object features in the ventral visual pathway of the macaque cerebral cortex. J Neurophysiol 71: 856-867, 1994.

Li N, DiCarlo JJ. Unsupervised natural experience rapidly alters invariant object representation in visual cortex. Science 321: 1502-1507, 2008.

Li N, Cox DD, Zoccolan D, DiCarlo JJ. What response properties do individual neurons need to underlie position and clutter "invariant" object recognition? J Neurophysiol 102: 360-376, 2009.

Li N, and DiCarlo JJ. Unsupervised natural visual experience rapidly reshapes size-invariant object representation in inferior temporal cortex. *Neuron* 67: 1062-1075, 2010.

Logothetis NK, Pauls J, Bulthoff HH, and Poggio T. View-dependent object recognition by monkeys. Curr Biol 4: 401-414, 1994.

Logothetis NK, Pauls J, and Poggio T. Shape representation in the inferior temporal cortex of monkeys. Curr Biol 5: 552-563, 1995.

Logothetis NK, and Pauls JP. Psychophysical and physiological evidence for viewer-centered object representation in the primate. Cerebral Cortex 5: 270-288, 1995.

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

Markram H, Lubke J, Frotscher M, and Sakmann B. Regulation of synaptic efficacy by coincidence of postsynaptic APs and EPSPs. Science 275: 213-215, 1997.

Masquelier T, and Thorpe SJ. Unsupervised learning of visual features through spike timing dependent plasticity. PLoS Comput Biol 3: e31, 2007.

Messinger A, Squire LR, Zola SM, and Albright TD. Neuronal representations of stimulus associations develop in the temporal lobe during learning. Proc Natl Acad Sci U S A 98: 12239-12244, 2001.

Miller EK, and Desimone R. Parallel neuronal mechanisms for short-term memory. Science 263: 520-522, 1994.

Miller EK, Li L, and Desimone R. A neural mechanism for working and recognition memory in inferior temporal cortex. Science 254: 1377-1379, 1991.

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

Miyashita Y. Inferior temporal cortex: where visual perception meets memory. Annual Review of Neuroscience 16: 245-263, 1993.

Miyashita Y. Neuronal correlate of visual associative long-term memory in the primate temporal cortex. Nature 335: 817-820, 1988.

Nazir TA, and O'Regan JK. Some results on translation invariance in the human visual system. Spat Vis 5: 81-100, 1990.

Olshausen BA, Anderson CH, and Van Essen DC. A neurobiological model of visual attention and invariant pattern recognition based on dynamic routing of information. Journal of

Neuroscience 13: 4700-4719, 1993.

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

Pasupathy A, Connor CE. Population coding of shape in area V4. Nat Neurosci 5:1332-1338, 2002.

Pinto N, Cox DD, and DiCarlo JJ. Why is real-world visual object recognition hard? PLoS Comput Biol 4: e27, 2008.

Pinto N, Doukhan D, DiCarlo JJ, and Cox DD. A high-throughput screening approach to discovering good forms of biologically inspired visual representation. PLoS Comput Biol 5: e1000579, 2009.

Pinto N, Majaj NJ, Barhomi Y, Solomon EA, Cox DD, and DiCarlo JJ. Human versus machine: comparing visual object recognition systems on a level playing field. COSYNE, Salt Lake City, UT, 2010.

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

Potter MC. Short-term conceptual memory for pictures. J Exp Psychol [Hum Learn] 2: 509-522, 1976.

Quiroga RQ, Reddy L, Kreiman G, Koch C, Fried I. Invariant visual representation by single neurons in the human brain. Nature 435: 1102-1107, 2005.

Riesenhuber M, and Poggio T. Hierarchical models of object recognition in cortex. Nat Neurosci 2: 1019-1025, 1999.

Riesenhuber M, and Poggio T. Models of object recognition. Nat Neurosci 3 Suppl: 1199-1204., 2000.

Ringach DL, and Shapley R. Reverse correlation in neurophysiology. Cognitive Science 28: 147-166, 2004.

Robinson DA. IEEE Transactions on Biomedical Engineering 101, 131 (1963).

Rolls ET, Baylis GC, Hasselmo ME, and Nalwa V. The effect of learning on the face selective responses of neurons in the cortex in the superior temporal sulcus of the monkey. Exp Brain Res 76: 153-164, 1989.

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.

Sakai K, and Miyashita Y. Neural organization for the long-term memory of paired associates. Nature 354: 152-155, 1991.

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.

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.

Sjostrom PJ, and Nelson SB. Spike timing, calcium signals and synaptic plasticity. Curr Opin Neurobiol 12: 305-314, 2002.

Song S, Miller KD, and Abbott LF. Competitive Hebbian learning through spike-timing-dependent synaptic plasticity. Nat Neurosci 3: 919-926, 2000.

Sprekeler H, Michaelis C, and Wiskott L. Slowness: an objective for spike-timing-dependent plasticity? PLoS Comput Biol 3: e112, 2007.

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

Tanaka K. Columns for complex visual object features in the inferotemporal cortex: clustering of cells with similar but slightly different stimulus selectivities. Cereb Cortex 13: 90-99, 2003.

Tarr MJ, and Bulthoff HH. Image-based object recognition in man, monkey and machine [In Process Citation]. Cognition 67: 1-20, 1998.

Tarr MJ, and Gauthier I. Do viewpoint-dependent mechanisms generalize across members of a class? [In Process Citation]. Cognition 67: 73-110, 1998.

Thorpe S, Fize D, and Marlot C. Speed of processing in the human visual system. Nature 381: 520-522, 1996.

Tovee MJ, Rolls ET, Azzopardi P. Translation invariance in the responses to faces of single neurons in the temporal visual cortical areas of the alert macaque. J Neurophysiol 72: 1049-1060, 1994.

Turk-Browne NB, Junge J, and Scholl BJ. The automaticity of visual statistical learning. J Exp Psychol Gen 134: 552-564, 2005.

Ullman S, and Soloviev S. Computation of pattern invariance in brain-like structures. Neural

Netw 12: 1021-1036, 1999.

Vetter T, Hurlbert A, and Poggio T. View-based models of 3D object recognition: Invariance to imaging transformations. Cerebral Cortex 3: 261-269, 1995.

Vogels R, Sáry G, and Orban GA. How task-related are the responses of inferior temporal neurons? Visual Neuroscience 12: 207-214, 1995.

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

Wallis G. Spatio-temporal influences at the neural level of object recognition. Network 9, 265-278, 1998.

Wallis G, and Bulthoff H. Learning to recognize objects. Trends Cogn Sci 3: 22-31, 1999.

Wallis G, and Rolls ET. Invariant face and object recognition in the visual system. Progress in Neurobiology 51: 167-194, 1997.

Wang G, Obama S, Yamashita W, Sugihara T, and Tanaka K. Prior experience of rotation is not required for recognizing objects seen from different angles. Nat Neurosci 8: 1768-1775, 2005.

Weiskrantz L, and Saunders RC. Impairments of visual object transforms in monkeys. Brain 107: 1033-1072, 1984.

Wiskott L, and Sejnowski TJ. Slow feature analysis: unsupervised learning of invariances. Neural Comput 14: 715-770, 2002.

Wyss R, Konig P, and Verschure PF. A model of the ventral visual system based on temporal stability and local memory. PLoS Biol 4: e120, 2006.

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.

Chapter 2

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

2. 1 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.

2. 2 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 unambiguously 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.

2. 3 Materials and Methods

2. 3. 1. 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

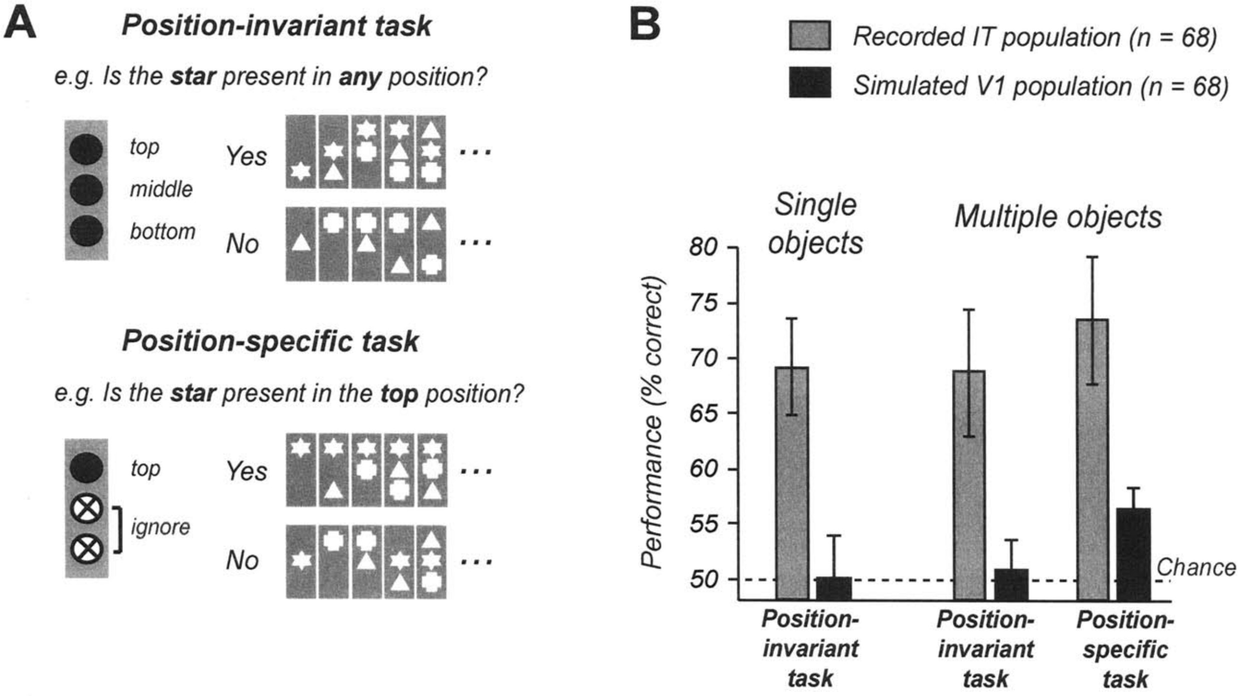


Figure 2-1: (A) Visual recognition tasks. Three objects (star, triangle, cross) were shown at three possible positions (-2°, 0°, and +2° relative to the fovea) either in isolation or in combinations of pairs or triplets. Using the IT population response data to each visual "scene", linear discriminant 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).

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° above, and 2° below) on a uniform gray background (27 Cd/m^2; see Fig. 2-1). 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).

2. 3. 2. 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. 2-1A). 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.

2. 3. 3. 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. 2-1B). 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. 2-1B). 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. 2-1A), 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. 2-1B.

2. 3. 4. Recognition task performance of hypothetical neuronal populations

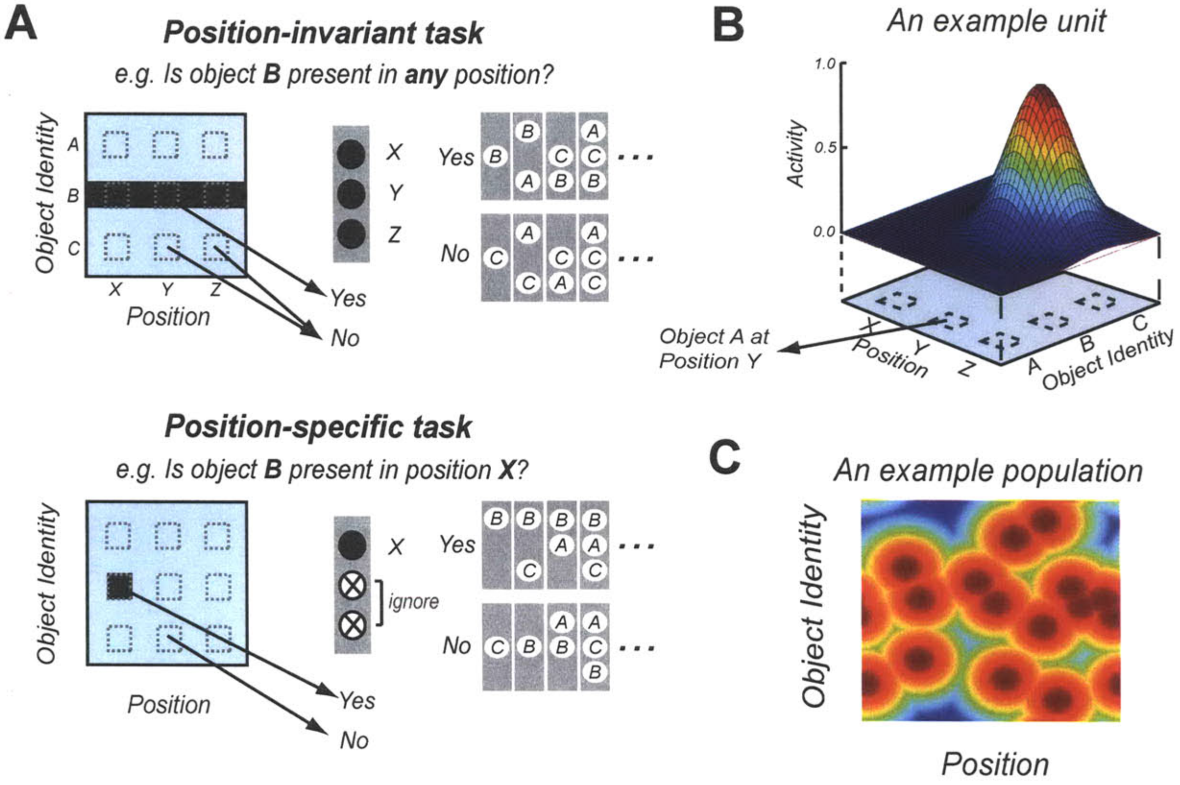


Figure 2-2: A schematic drawing of the simulation design and tasks. **(A)** The two recognition tasks that each simulated population was asked to solve. The tasks are analogous to those tested for the real IT population (c.f. Fig. 2-1A). On the left, the "2D stimulus space" is displayed: the y-axis shows 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. 2-1A), 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 2-1A. **(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.

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 2-2A). 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. 2-2A 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 **x**, Eq. (1)) to a set of labeled stimulus "scenes",

3) use these responses to build classifiers for the recognition task (i.e. find **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. 2-2A) 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. 2-2A). 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).

2. 3. 5. 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*).

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

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).

2. 3. 6. 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. 2-2B), 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. 2-5). 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. 2-5, "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. 2-2C). 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. 2-3D) 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. 2-3C: 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, σ_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. 2-2C). 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. 2-3C, D).

2. 3. 7. Simulated V1 population to compare with the recorded IT population

To compare the recorded IT population results with a meaningful baseline (Fig. 2-1B), 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. 2-1B 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. 2-1B show standard deviations).

2. 3. 8. Simulated V1 population to compare with simulated IT population

To compare the simulated IT populations with a meaningful baseline (Fig. 2-5, 2-6), 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.

2. 3. 9. 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. 2-5; $(i)_P$-$(v)_P$, $(i)_C$-$(iv)_C$). The populations with response functions that preserved the rank-order selectivity across the position axis were constructed as following (see Fig. 2-5C, 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. 2-5C 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. 2-5C 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 2-5 and 2-6, 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. 2-5D). To contrast, we simulated four other response functions (Fig. 2-5 $(i)_C$-$(iv)_C$) 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.

2. 3. 10. 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. 2-6A), 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. 2-6B), 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. 2-4B & 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. 2-4C). 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. 2-4): 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. 2-4C lower panel). In total, there were 63 cases from the IT population and 68 cases from the V1 population in Figure 2-4D.

2. 4 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. 2-1A).

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.

2. 4. 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. 2-1A; see Methods).

2. 4. 1. 1 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. 2-1B, 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. 2-1A, 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. 2-1B, 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).

2. 4. 1. 2 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. 2-1A bottom panel). At each of these positions, we found that such classifiers performed even better than the position-invariant classifiers (mean: 73.6%, Fig. 2-1B). 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. 2-S1). 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. 2-1B). 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. 2-1B; this is not simply explained by smaller V1 RF size or lack of coverage, see Fig. 2-5, 2-6C). 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. 2-1B 1st and 2nd bar), yet can also support position-specific object identification ("Where"; Fig. 2-1B 3rd bar).

2. 4. 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. 2-2B; 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. 2-3D). 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. 2-5).

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.

2. 4. 3 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. 2-3). 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. 2-3B) and clutter sensitivity (Fig. 2-3D).

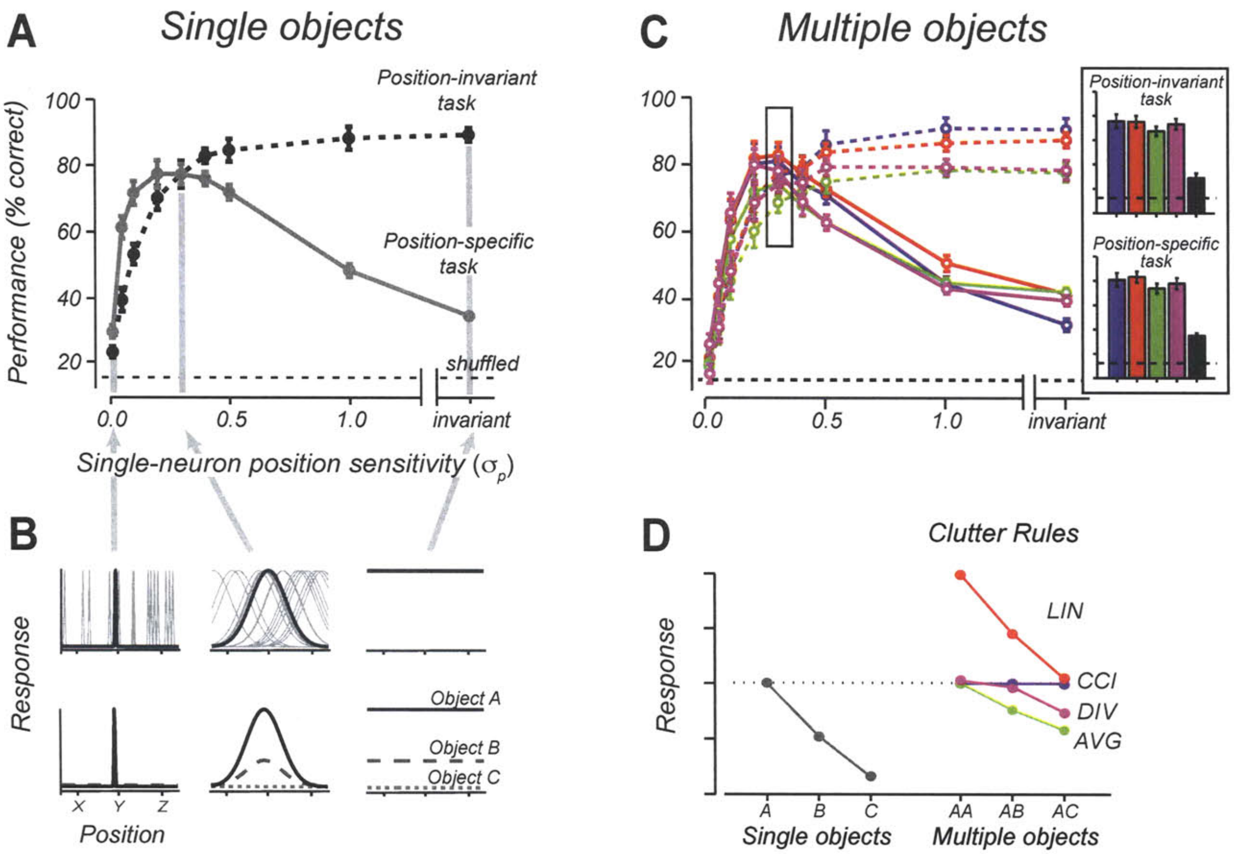
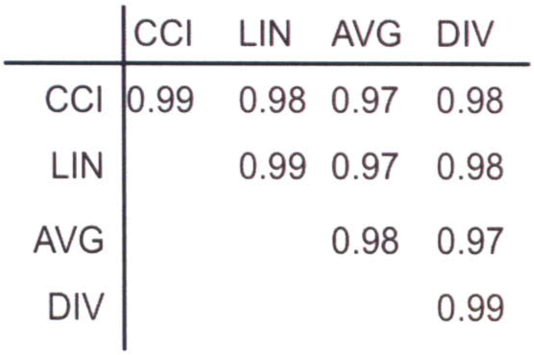


Figure 2-3: The effect of single-neuron position and clutter sensitivity on population recognition performance. **(A)** Population performance on the recognition tasks with visual scenes containing single objects. Performance was averaged over multiple simulation runs; error bars indicate standard deviation. 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 2-3A 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. 2-3A dashed line). At the same time, only populations of position sensitive neurons conveyed the necessary position information to support the position-specific task (Fig. 2-3A, 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. 2-3A, 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 2-3C 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. 2-3C, 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. 2-3D). 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 2-1). 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



	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 2-1: 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.

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).

2. 4. 4 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. 2-3A, C), what property do individual V1 neurons lack that makes the V1 population inferior to the IT population for object recognition (Fig. 2-1)?

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. 2-4A. 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

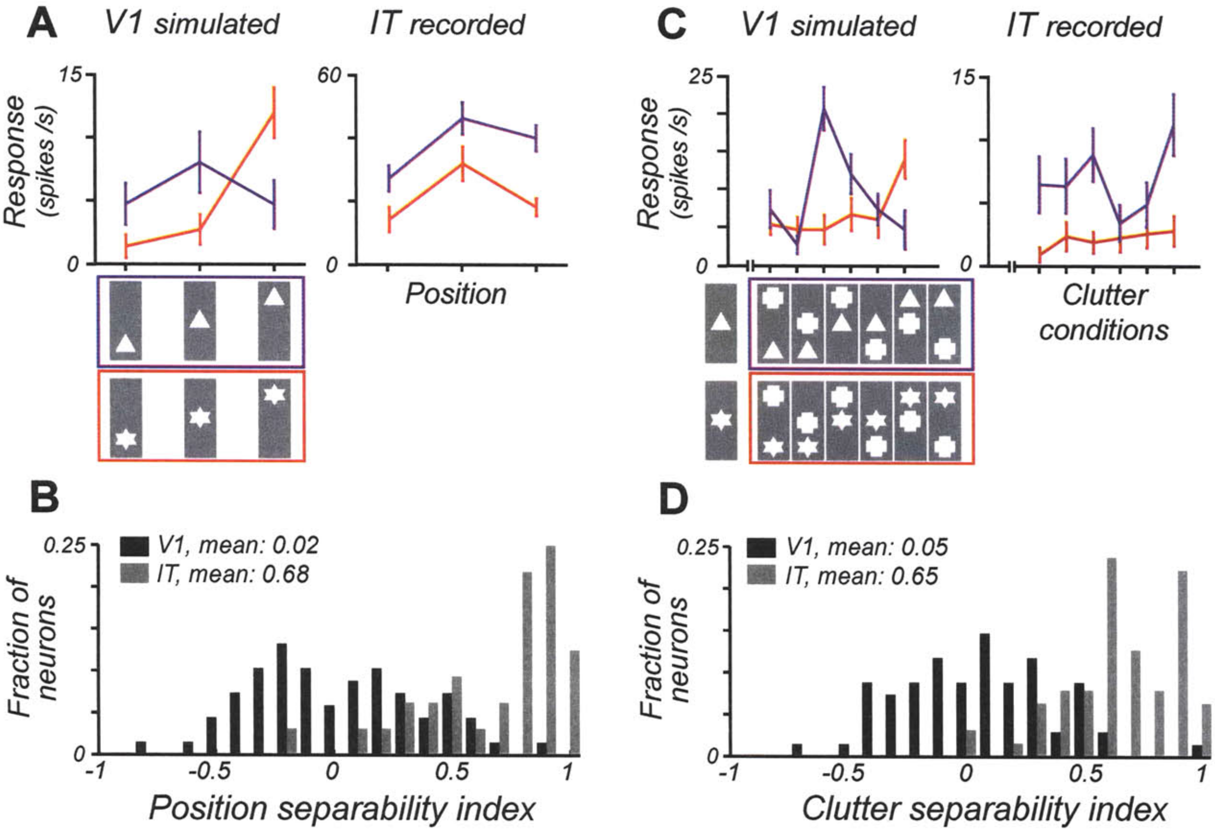


Figure 2-4: Real IT neurons show more preserved rank-order object preference than simulated V1 units. (**A**) Neuronal responses to two objects at three positions for an example simulated V1 unit and a real IT neuron. (**B**) Distributions of rank order preservation across position for the V1 and IT 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.

with the neurons' receptive fields (e.g. Fig. 2-4A). 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. 2-4B, $p<10^{-14}$, 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. 2-4C) and the IT population showed much more separable responses than the simulated V1 population (Fig. 2-4D, $p<10^{-22}$, 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. 2-1B and Fig. 2-4), we directly examined the importance of rank-order preservation as a key single-neuron response property.

2. 4. 5 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. 2-5C). 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

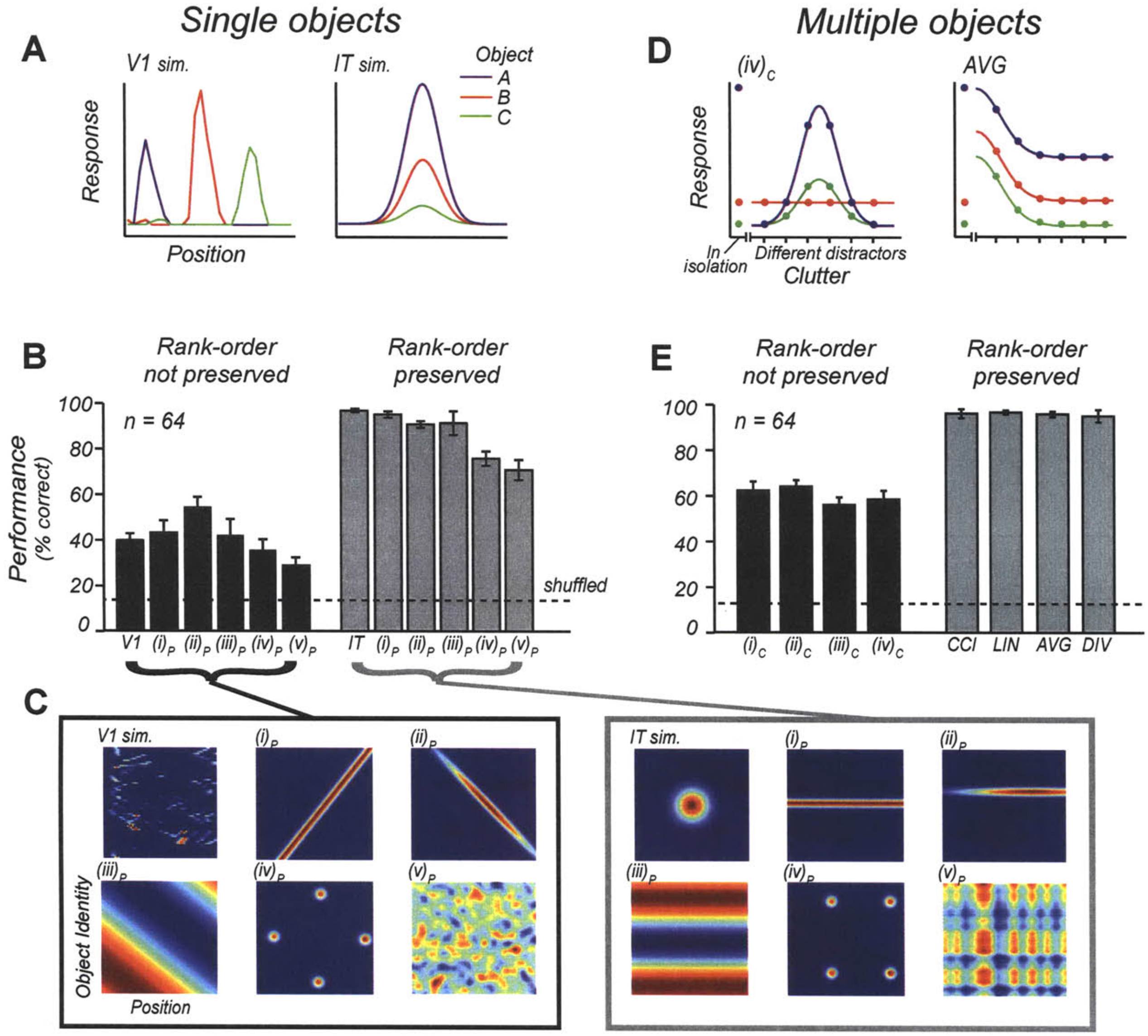


Figure 2-5: The effect of single-unit rank-order preservation on population recognition performance. (**A**) Example single-units that maintained rank-order object preference (e.g. IT) or not (e.g. V1) across position. (**B**) Averaged population performance on position-invariant task. (**C**) Example 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).

of single objects. As shown in Fig. 2-5B, we found that populations of neurons that preserved the rank-order of their object preference across positions (see example in Fig. 2-5A, right) performed much better on the position invariant recognition task than populations of neurons that did not (see example in Fig. 2-5A, 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. 2-5C, plot labeled $(iii)_P$), performed nearly as well as the population of neurons with Gaussian tuning functions (Fig. 2-5C, 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. 2-3C). In fact, the systematic clutter rules we simulated (Fig. 2-3D) 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. 2-5D, 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. 2-5D, left). Indeed, the performance on the clutter-invariant recognition task was much lower for the latter populations (compare gray to black bars in Fig. 2-5E). 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. 2-6A). 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.

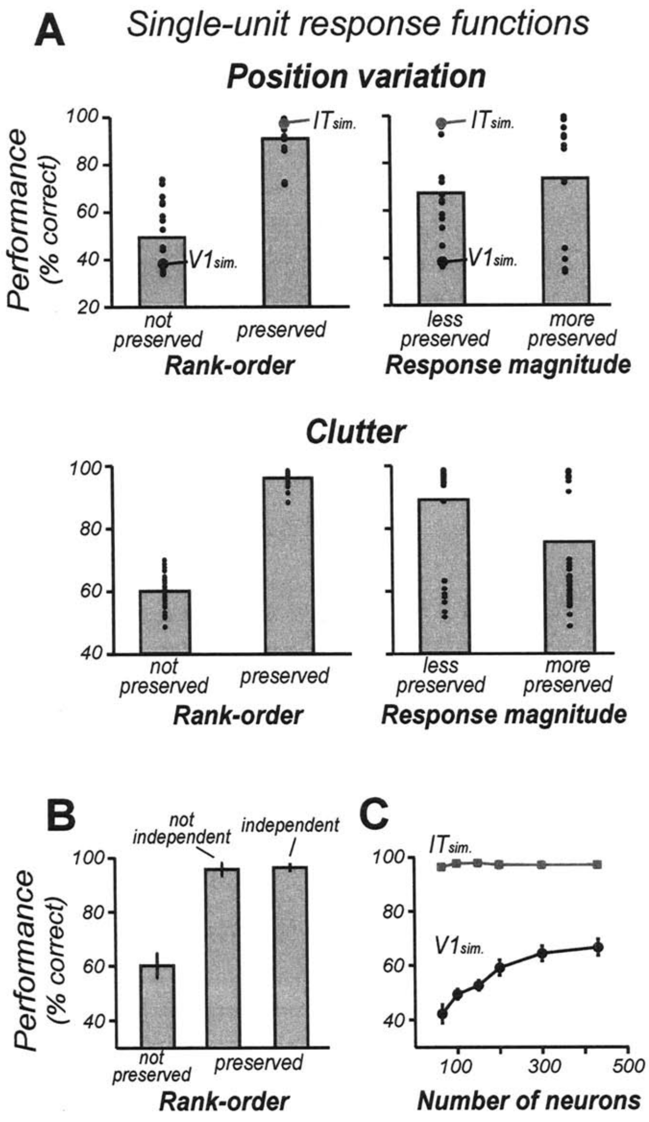


Figure 2-6: Rank-order preservation of single-units, not sensitivity to transformations, predicts population performance on invariant recognition tasks. **(A)** Combined average performance on the invariant recognition tasks across all the simulated populations when they are sorted by their single-unit 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.

2006; Sanger 2003). We found that these two alternative coding schemes both yielded equally high recognition performance (Fig. 2-6B), 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. 2-6C). 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.

2. 5 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. 2-7).

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. 2-1B), 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. 2-7) 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

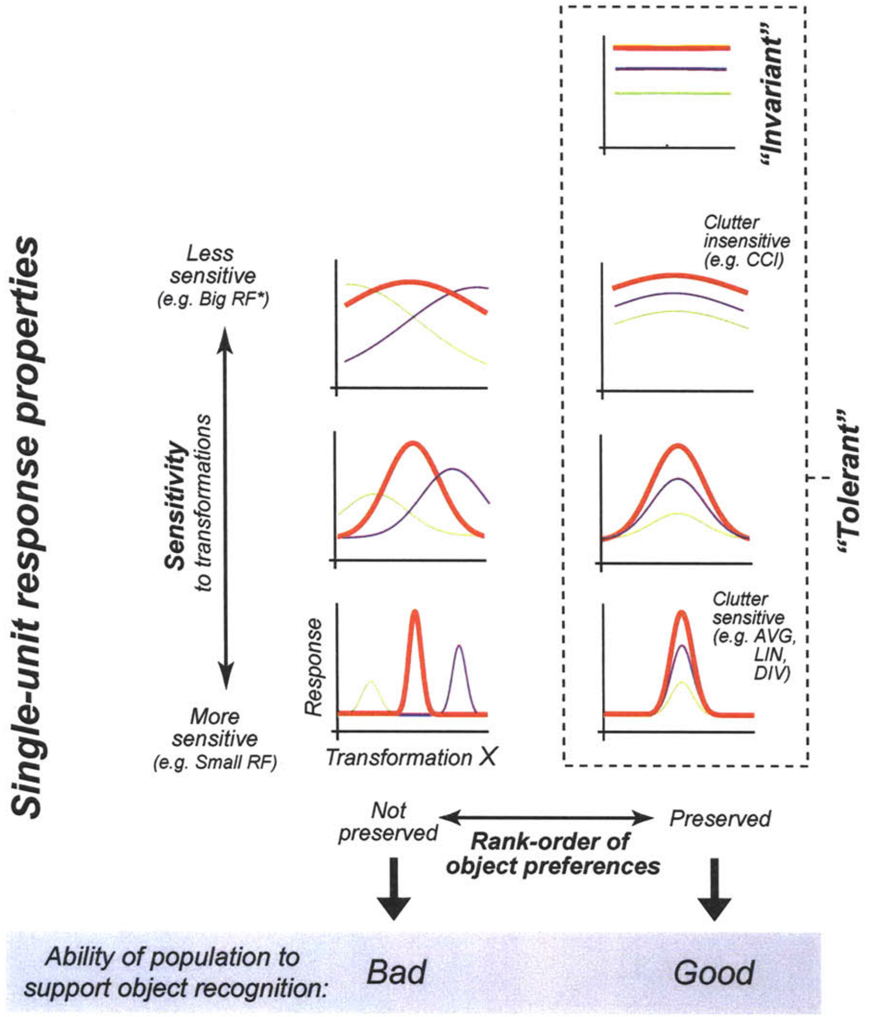


Figure 2-7: Summary of the goodness of different single-unit response properties for supporting invariant object recognition tasks at the population level. In each subplot, the x-axis shows the values of some identity-preserving transformation (e.g. object retinal position, or the presence of different distractor objects, see Fig. 2-5); 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. 2-6 and Methods). For example, neurons can be largely insensitive to both position and clutter, yet form an inadequate population (see Fig. 2-6). 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.

Cox 2007).

2. 5. 1 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. 2-5, 2-6). 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. 2-6A). 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. 2-3C, D, except RAND) performed well as a population in recognition tasks under cluttered conditions (Fig. 2-3C). In particular, even though each systematic clutter rule produced different amounts of clutter sensitivity in individual neurons (Fig. 2-3D), 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., 4-7° 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 2-7.

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).

2. 5. 2 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. 2-5C). 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. 2-3C, 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. 2-7). 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. 2-7). 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.

2. 5. 3 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. 2-5A, V1 units) and achieves rank-order preservation at its highest stages (Fig. 2-5A IT units). Understanding this is the crux of understanding how invariant object recognition is achieved.

2. 6 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.

2. 7 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.

2. 8 Supplemental figure

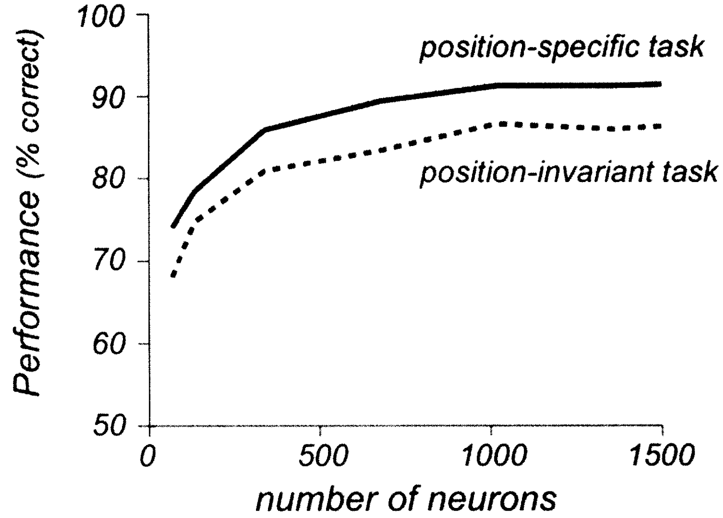


Figure 2-S1: To overcome the inherent trial-to-trial variability in the recorded neurons and estimate the absolute performance that can be achieved with our recorded selectivity, we synthesized larger populations of Poisson-spiking neurons from the response profiles of the measured IT population (n=68). 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.

Chapter 3

Unsupervised Natural Visual Experience Rapidly Reshapes Position Invariant Object Representation in Inferior Temporal Cortex

3. 1 Abstract

Object recognition is challenging because each object produces myriad retinal images. Responses of neurons from the inferior temporal cortex (IT) are selective to different objects, yet tolerant ("invariant") to changes in object position, scale, pose, etc.. How does the brain construct this neuronal tolerance? Here we report a form of neuronal learning that suggests the underlying solution. Targeted alteration of the natural temporal contiguity of visual experience caused specific changes in IT position tolerance. This unsupervised temporal slowness learning (UTL) is substantial, increases with experience, and is significant in single IT neurons after just one hour. Together with previous theoretical work and human object perception experiments, we speculate that UTL may reflect the mechanism by which the visual stream builds and maintains tolerant object representations.

3. 2 Introduction

When presented with a visual image, primates can rapidly (<200 ms) recognize objects in spite of large variations in object position, scale, pose, etc. (*1, 2*). This ability likely derives from the responses of neurons at high levels of the primate ventral visual stream

(*3-5*). But how are these powerful "invariant" neuronal object representations built by the visual system? Based on theoretical (*6-10, 32*) and behavioral (*11, 12*) work, one possibility is that tolerance ("invariance") is learned from the temporal contiguity of object features during natural visual experience, potentially in an unsupervised manner. Specifically, during natural visual experience, objects tend to remain present for seconds or more, while object motion or viewer motion (e.g. eye movements) tends to cause rapid changes in the retinal image cast by each object over shorter time intervals (hundreds of ms). The ventral visual stream could construct a tolerant object representation by taking advantage of this natural tendency for temporally contiguous retinal images to belong to the same object. If this hypothesis is correct, it might be possible to uncover a neuronal signature of the underlying learning by using targeted alteration of those spatiotemporal statistics (*11, 12*).

To look for such a signature, we focused on position tolerance. If two objects consistently swapped identity across temporally contiguous changes in retinal position then, following sufficient experience in this "altered" visual world, the visual system might incorrectly associate the neural representations of those objects viewed at different positions into a single object representation (*11, 12*). We focused on the top level of the primate ventral visual stream (IT), where many individual neurons possess position tolerance – they respond preferentially to different objects, and that selectivity is largely maintained across changes in object retinal position, even when images are simply presented to a fixating animal (*13, 14*).

3. 3 Results and discussion

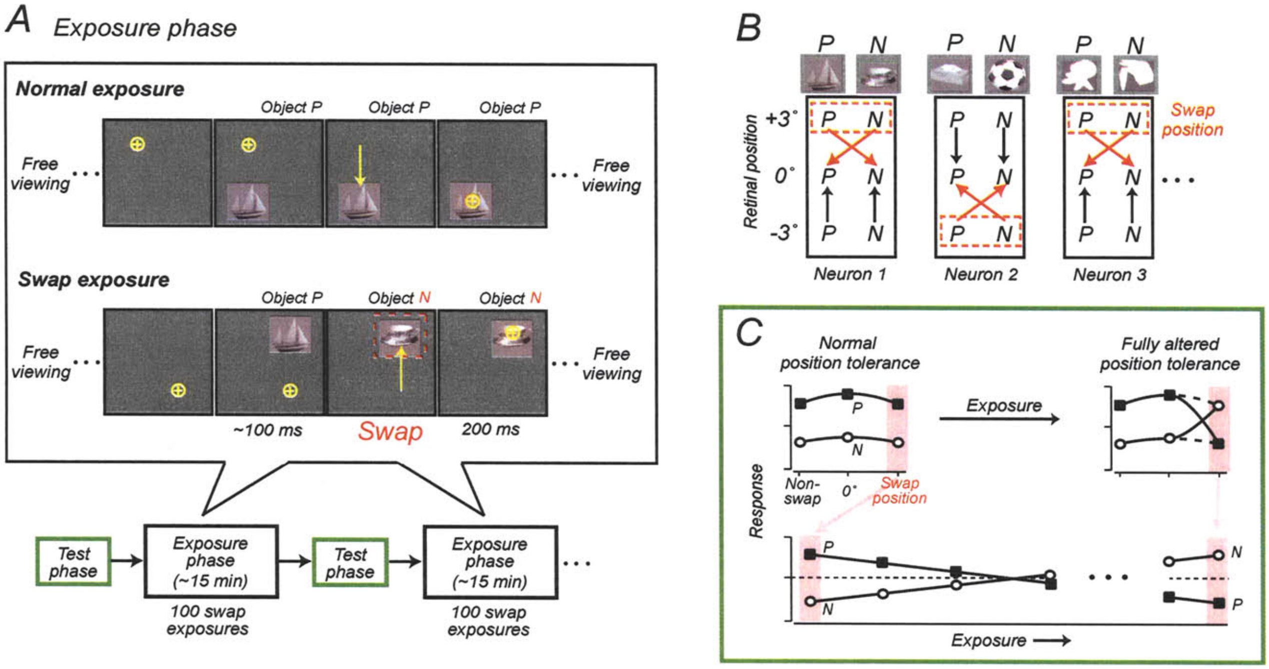


Fig. 3-1: Experimental design and predictions. (**A**) IT responses were tested in *Test Phases* (green boxes, see text) which alternated with *Exposures Phases*. Each *Exposure Phase* consisted of 100 normal exposures (50 P->P, 50 N->N) and 100 swap exposures (50 P->N, 50 N->P). Stimulus size was 1.5°. (*15*) (**B**) Each box shows the *Exposure Phase* design for a single neuron. Arrows show the saccade-induced temporal contiguity of retinal images (arrow heads point to the retinal images occurring later in time, i.e. at the end of the saccade). The swap position was strictly alternated (neuron-by-neuron) so that it was counterbalanced across neurons. (**C**) Prediction for responses collected in the *Test Phase*: if the visual system builds tolerance using temporal contiguity (here driven by saccades), the swap exposure should cause incorrect grouping of two different object images (here P and N). Thus, the predicted effect is a decrease in object selectivity at the swap position that increases with increasing exposure (in the limit, reversing object preference), and little or no change in object selectivity at the non-swap position.

We tested a strong, "online" form of the temporal contiguity hypothesis – two monkeys visually explored an altered visual world (Fig. 3-1A, *Exposure Phase*), and we paused every ~15 minutes to test each IT neuron for any change in position tolerance produced by that altered experience (Fig. 3-1A, *Test Phase*). We concentrated on each neuron's responses to two objects that elicited strong (object "P", preferred) and moderate (object "N", non-preferred) responses, and we tested the position tolerance of that object selectivity by briefly presenting each object at 3° above, below, or at the center of gaze (see (*15*) and fig. 3-S1). All neuronal data reported in this study were obtained in these *Test Phase*: animal tasks unrelated to the test stimuli, no attentional cuing, and completely randomized, brief presentations of test stimuli (*15*). We alternated between these two phases (*Test Phase* ~5 min; *Exposure Phase* ~15 minutes) until neuronal isolation was lost.

To create the altered visual world (*Exposure Phase* in Fig. 3-1A), each monkey freely viewed the video monitor on which isolated objects appeared intermittently, and its only task was to freely look at each object. This exposure "task" is a natural, automatic primate behavior in that it requires no training. However, by using real-time eye-tracking (*16*), the images that played out on the monkey's retina during exploration of this world were under precise experimental control (*15*). The objects were placed on the video monitor so as to (initially) cast their image at one of two possible *retinal* positions (+3° or -3°). One of these retinal positions was pre-chosen for targeted alteration in visual experience (the "swap" position; counterbalanced across neurons, see Fig. 3-1B and (*15*)); the other position acted as a control (the "non-swap" position). The monkey quickly saccaded to each object (mean: 108 ms after object appearance), which rapidly brought the object image to the center of its retina (mean saccade duration 23 ms). When the object had appeared at the "non-swap" position, its identity remained stable as the monkey saccaded to it, typical of real-world visual experience ("Normal exposure", Fig. 3-1A, see (*15*)). However, when the object had appeared at the "swap" position, it was always replaced by the other object (e.g. P->N) as the monkey saccaded to it (Fig. 3-1A, "Swap exposure"). This experience manipulation took advantage of the fact that primates are effectively blind during the brief time it takes to complete a saccade (*17*). It consistently made the image of one object at a peripheral retinal position ("swap" position) temporally contiguous with the retinal image of the other object at the center of the retina (Fig. 3-1).

We recorded from 101 IT neurons while the monkeys were exposed to this altered visual world (isolation held for at least two *Test Phases;* n = 50 in Monkey 1; 51 in Monkey 2). For each neuron, we measured its object selectivity at each position as the difference in response to the two objects (P-N; all key effects were also found with a contrast index of selectivity, fig. 3-S6). We found that, at the "swap" position, IT neurons (on average) decreased their initial object selectivity for P over N, and this change in object selectivity grew monotonically stronger with increasing numbers of "swap" exposure trials (Fig. 3-2A, C). However, the same IT neurons

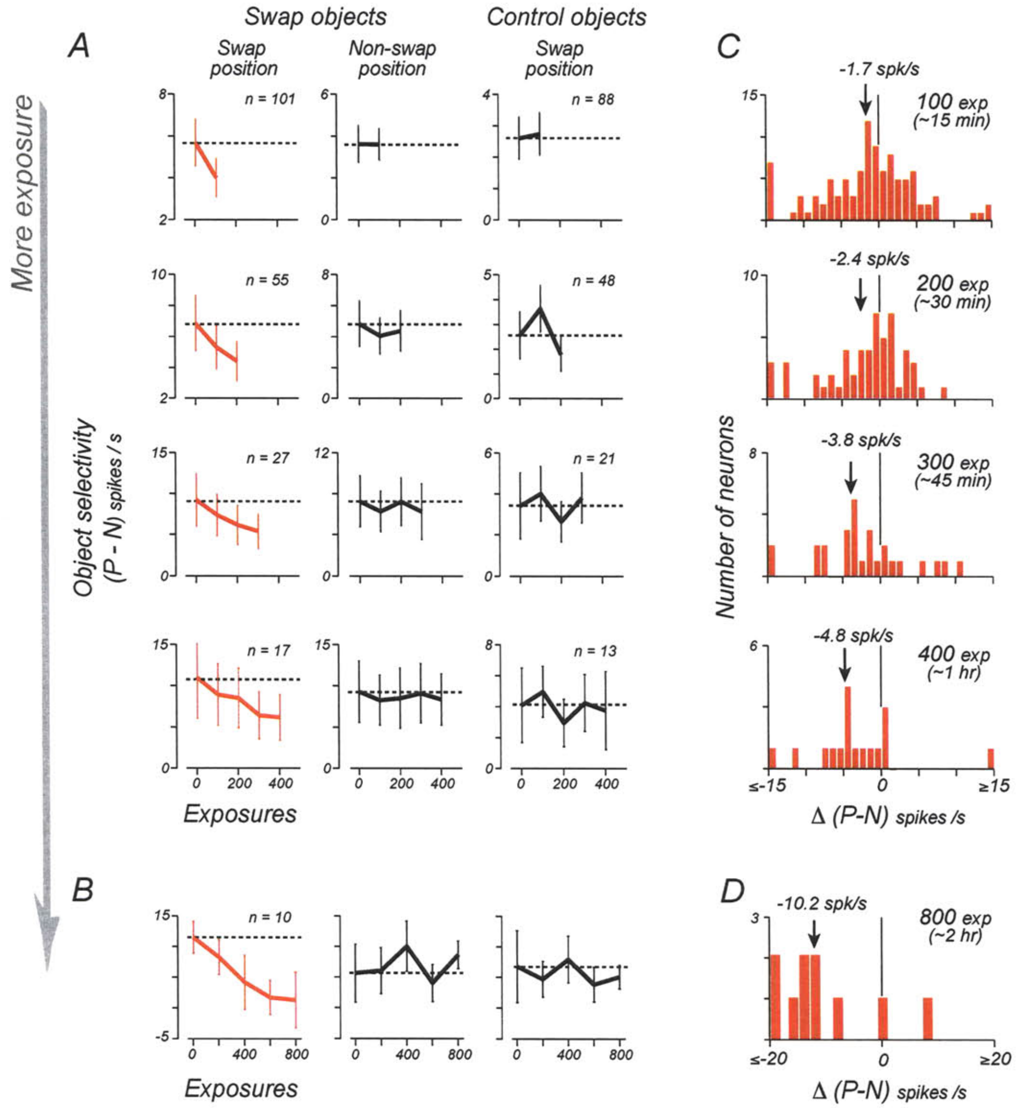


Fig. 3-2: Change in the population object selectivity. (**A**) Mean population object selectivity at the swap and (equally eccentric) non-swap position, and for control objects at the swap position. Each row of plots shows effect among all neurons held for at least the indicated amount of exposure (e.g. top row shows all neurons held for over 100 swap exposures -- including the neurons from the lower rows). The object selectivity for each neuron was the difference in its response to object P and N. To avoid any bias in this estimate, for each neuron we defined the labels "P" (preferred) and "N" by using a portion of the pre-exposure data (10 repetitions) to determine these labels, and the reminder to compute the displayed results in all analyses using these labels. Though there was, by chance, slightly greater initial selectivity at the swap position, this cannot explain the position-specificity of the observed change in selectivity (table S2). (**B**) Mean population object selectivity of 10 multi-unit sites. Error bars (**A, B**) are standard error of the mean. (**C**) Histograms of the object selectivity change at the swap position, **Δ**(P-N) = $(P-N)_{post-exposure} - (P-N)_{pre-exposure}$. The arrows indicate the means of the distributions. The mean **Δ**(P-N) at the non-swap position was: 0.01, -0.5, -0.9, -0.9 spikes/s respectively. The variability around that mean (i.e. distribution along the x-axis) is commensurate with repeated measurements in the face of known Poisson spiking variability (fig. 3-S11). (**D**) Object selectivity changes at the multi-unit sites. The mean **Δ**(P-N) at the non-swap position was 1.6 spikes/s.

showed (Fig. 3-2A) no average change in their object selectivity at the equally eccentric control position ("non-swap" position), and little change in their object selectivity among two other (non-exposed) control objects (see below).

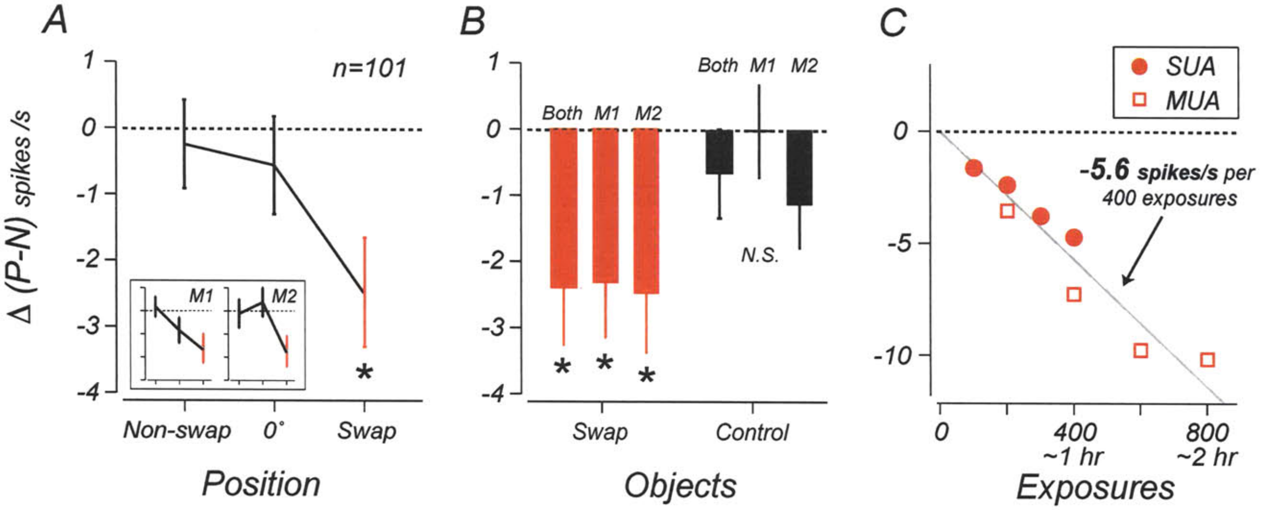


Fig. 3-3: Position-specificity, object-specificity, and time course. (**A**) Mean object selectivity change, **Δ**(P-N), at the swap, non-swap, and central (0°) retinal position. **Δ**(P-N) was computed as in Fig. 3-2C from each neuron's first and last available *Test Phase* (mean ~200 swap exposures). The insets show the same analysis performed separately for each monkey. (**B**) Mean object selectivity change for the (exposed) swap objects and (non-exposed) control objects at the swap position. Error bars (**A, B**) are standard error of the mean. The swap object selectivity change at the swap position is statistically significant (*) in the pooled data as well as in individual animals ($p < 0.05$, one-tailed t-test against 0). (**C**) Mean object selectivity change as a function of the number of swap exposures for all single-units ($n = 101$) and multi-unit sites ($n = 10$). Each data point shows the average across all the neurons/sites held for a particular amount of time. Gray line is best linear fit with a zero intercept; slope is mean effect size: -5.6 spikes/s per 400 exposures. The slope at the non-swap position using the same analysis was 0.6 spikes/s (not shown).

Because each IT neuron was tested for different amounts of exposure time, we first computed a net object selectivity change, Δ(P-N), in the IT population by using the first and last available *Test Phase* data for each neuron. The prediction was that Δ(P-N) should be negative (i.e. in the direction of object preference reversal), and greatest at the "swap" position (Fig. 3-1C). This prediction was born out (Fig. 3-3A). The position-specificity of the experience-induced changes in object selectivity was confirmed by two different statistical approaches: 1) a direct comparison of Δ(P-N) between the swap and non-swap position (n = 101; p = 0.005, one tailed paired t-test); 2) a significant interaction between position and exposure – that is, object selectivity decreased at the swap position with increasing amounts of exposure (p = 0.009 by one-tailed bootstrap; p = 0.007 by one-tailed permutation test; tests were done on (P-N)).

The changes in object selectivity at the swap position were also largely shape-specific. For 88 of the 101 neurons, we monitored the neuron's selectivity among two control objects not shown to the animals during the *Exposure Phase* (chosen in a similar way as the P and N objects, fully-interleaved testing in each *Test Phase*; see (*15*)). Across the IT population, control object selectivity at the swap position did not significantly change (Fig. 3-2A), and the swap object selectivity changed significantly more than the control object selectivity (Fig. 3-3B; n = 88, p = 0.009 one tailed paired t-test of swap vs. control objects at the swap position).

These changes in object selectivity were substantial in magnitude (average change of ~5 spikes/ s per 400 exposures at the swap position; Figs. 3-2C, 3-3C), and were visibly obvious and highly significant at the population level. In the face of well known Poisson spiking variability (*18, 19*), these effects were only weakly visible in most single IT neurons recorded for short durations, but were much more apparent over the maximal one hour exposure time that we could hold isolation (Fig. 3-2C, lower panels). To ask if the object selectivity changes continued to grow even larger with longer periods of exposure, we next recorded multi-unit activity (MUA) in one animal (Monkey 2), which allowed us to record from a number of (non-isolated) neurons around the electrode tip (which all tend to have similar selectivity (*20, 21*)) while the monkey was exposed to the altered visual world for the entire experimental session (~2+ hrs) (*15*). The MUA data replicated the single-unit results -- a change in object selectivity only at the swap position (Fig. 3-2C; "position x exposure" interaction: p = 0.03, one-tailed bootstrap; p = 0.014, one-tailed permutation test; n = 10). Furthermore, the MUA object selectivity change at the swap position continued to increase as the animal received even more exposure to the altered visual world, followed a very similar time course in the *rate* of object selectivity change (~5 spikes/s per 400 exposures; see Fig. 3-3C), and even showed a slight reversal in object selectivity (N>P in Fig. 3-4D).

Our main results were similar in magnitude (Fig. 3-3A, B) and statistically significant in each of the two monkeys (Monkey 1: $p = 0.019$; Monkey 2: $p = 0.0192$; one-tailed t-test). Each monkey performed a different task during the *Test Phase* (*15*), suggesting that these neuronal changes are not task dependent.

Because we selected the objects P and N so that they both tended to drive the neuron (*15*), the population distribution of selectivity for P and N at each position was very broad (95% range: [-5.7 spikes/s to 31.0 spikes/s] pooled across position; $n = 101$). However, our prediction assumes that the IT neurons were initially object-selective (i.e. response to object P greater than object N). Consistent with this, neurons in our population with no initial object selectivity at the center of gaze showed little average change in object selectivity at the swap position with exposure (fig. 3-S5). To test the learning effect in the most selective IT neurons, we selected the neurons with significant object selectivity ($n = 52/101$; two-way ANOVA test (2 objects x 3 positions), $p < 0.05$, significant main object effect or interaction). Among this smaller number of object-selective neurons, the learning effect remained highly significant and still specific to the swap position ($p = 0.002$ by t-test; $p = 0.009$ by bootstrap; $p = 0.004$ by permutation test; see above).

To further characterize the response changes to individual objects, we closely examined the selective neurons held for at least 300 exposures ($n = 28/52$) and the multi-unit sites ($n = 10$). For each neuron/site, we used linear regression to measure any trend in the neuron's response to each object as a function of exposure time (Fig 3-4A). Changes in response to P and N at the swap position were visibly obvious in a fraction of single neurons/sites (Fig. 3-4A), and statistically significant object selectivity change was encountered in 32% of instances (12/38; Fig. 3-4C) (*15*). Across our neuronal population, the change in object selectivity at the swap position was due to both a decreased response to object P and an increased response to object N (approximately equal change; Fig. 3-4B). These response changes were highly visible in the single-units and multi-units held for the longest exposure times (Fig. 3-4D).

These changes in the position profile of IT object selectivity (i.e. position tolerance) cannot be explained by changes in attention or by adaptation (also see Fig. 3-S10). First, a simple fatigue-

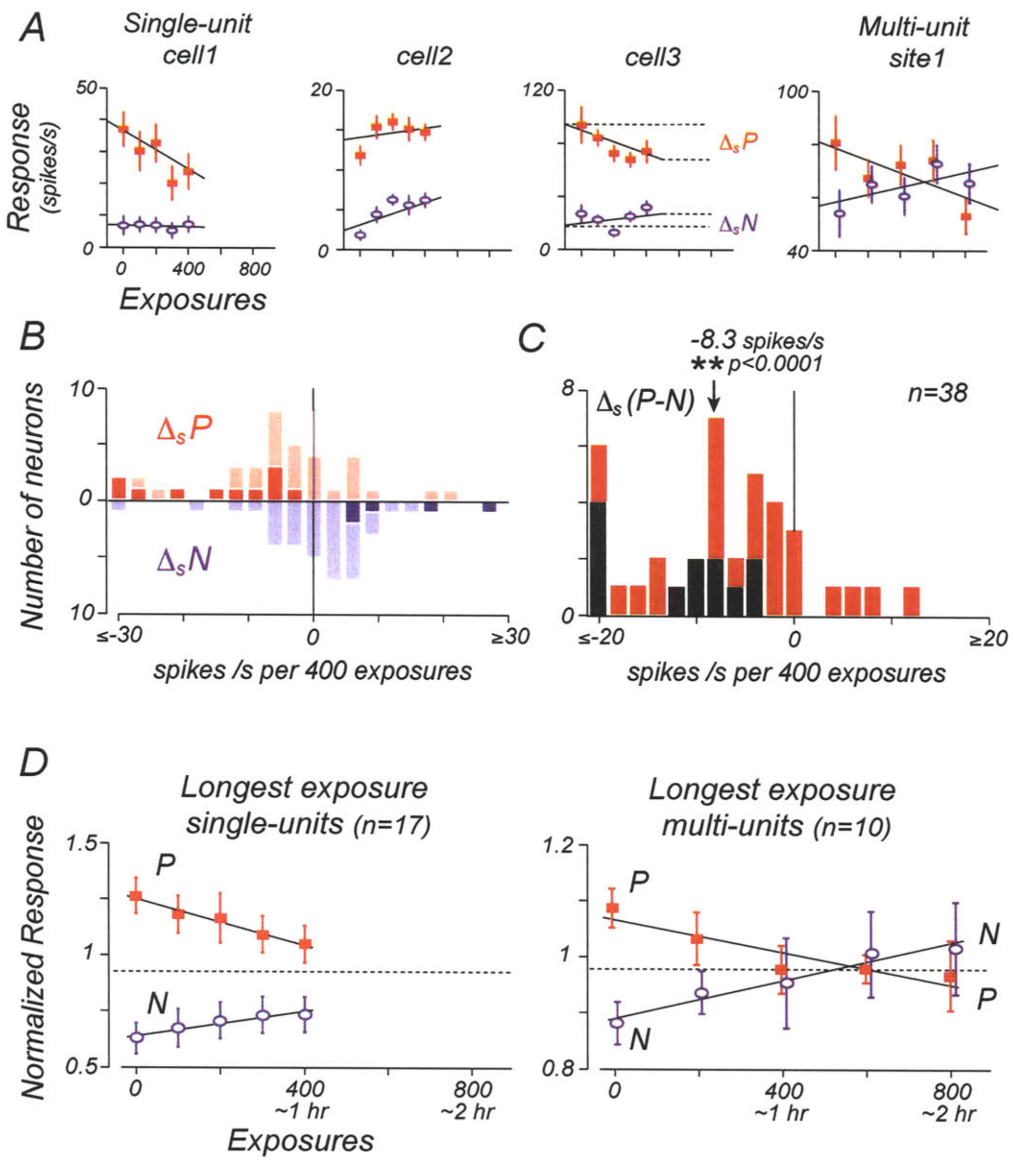


Fig. 3-4. Responses to objects P and N. (**A**) Response data to object P and N at the swap position for three example neurons and one multi-unit sites as a function of exposure time. The solid line is standard linear regression. The slope of each line (**Δs**) provides a measure of the response change to object P and N for each neuron. Some neurons showed a response decrease to P, some showed a response enhancement to N, while others showed both (see examples). (**B**) Histograms of the slopes obtained for the object selective neurons/sites tested for at least 300 exposures. The dark-colored bars indicate neurons with significant change by a permutation test ($p < 0.05$; see (15)). (**C**) Histograms of the slopes from linear regression fits to object selectivity (P-N) as a function of exposure time, same units as in (**B**). Arrow indicates the mean of the distribution, (the mean **Δs**(P-N) at the non-swap position was -1.7 spikes/s, $p = 0.38$). The black bars indicate instances (32%; 12/38) that showed a significant change in object selectivity by permutation test ($p < 0.05$). Results were very similar when we discarded neurons/sites with greater initial selectivity at the swap position (fig 3-S8). (**D**) Data from all the neurons/sites tested for the longest exposure time. The plot shows the mean normalized response to object P and N as a function of exposure time (c.f. Fig. 3-1C; see fig. 3-S3 for data at the non-swap position and for control objects). Error bars (**A, D**) are standard error of the mean.

adaptation model cannot explain the position-specificity of the changes because, during the recording of each neuron, each object was experienced equally often at the swap and non-swap positions (also see table 3-S2). Second, we measured these object selectivity changes with briefly presented, fully randomized stimuli while the monkeys performed tasks unrelated to the stimuli (*15*), arguing against an attentional account. Third, both of these explanations predict response decrease to *all* objects at the swap position, yet we found that the change in object selectivity at the swap position was due to an *increase* in response to object N (+2.3 spikes/s per 400 swap exposures) as well as a decrease in response to object P (3.0 spikes/s per 400 swap exposures; Fig. 3-4). Fourth, neither possibility can explain the shape-specificity of the changes.

We term this effect "unsupervised temporal slowness learning" (UTL), because the selectivity changes depend on the temporal contiguity of object images on the retina, and are consistent with the hypothesis that the natural stability (slowness) of object identity instructs the learning without external supervision (*6-10, 32*). Our current data as well as previous human object perception experiments (*11*) cannot rule out the possibility that the brain's saccade generation mechanisms or the associated attentional mechanisms (*22, 23*) may also be needed. Indeed, eye-movement signals are present in the ventral stream (*24, 25*). The relatively fast time-scale and unsupervised nature of UTL may allow rapid advances in answering these questions, systematically characterizing the spatiotemporal sensory statistics that drive it, and understanding if and how it extends to other types of image tolerance (e.g. changes in object scale, pose (*26, 27*)).

IT neurons "learn" to give similar responses to different visual shapes ("paired associates") when reward is used to *explicitly* teach monkeys to associate those shapes over long time scales (1-5 sec between images, e.g. (*28, 29*)), but sometimes without explicit instruction (*30, 31*). Here, a top down explanation of the neuronal selectivity changes is unlikely because animals performed tasks that were unrelated to the object images when the selectivity was probed, and the selectivity changes were present in the earliest part of the IT responses (~100 ms; fig 3-S4). UTL could be an instance of the same underlying plasticity mechanisms: here the "associations" are between object images at different retinal positions (which, in the real world, are typically images of the same object). However, UTL may be qualitatively different because: 1) the learning is retinal position-specific, 2) it operates over the much shorter time scales of natural visual exploration (~200 ms), 3) it is unsupervised in that, besides the visual world, no external "teacher" was used to direct the learning (e.g. no association-contingent reward was used, but we do not rule out the role of internal "teachers" such as efferent eye-movement signals). These distinctions are important because we naturally receive orders of magnitude more such experience (e.g. ~10^8 unsupervised temporal-contiguity saccadic "experiences" per year of life).

Our results show that targeted alteration of natural, unsupervised visual experience changes the position tolerance of IT neurons as predicted by the hypothesis that the brain employs a temporal contiguity learning strategy to build that tolerance in the first place. Several computational models show how such strategies can build tolerance (*6-10, 32*), and such models can be implemented using Hebbian-like learning rules (*8, 47*) that are consistent with spike-timing-dependent plasticity (*33*). One can imagine IT neurons using almost-temporally-coincident activity to learn which sets of its afferents correspond to features of the same object at different positions. The time-course and task-independence of UTL is consistent with synaptic plasticity (*34, 35*), but our data do not constrain the locus of plasticity, and changes at multiple levels of the ventral visual stream are likely (*36, 37*).

We do not yet know if UTL reflects mechanisms than are necessary for building tolerant representations. But these same experience manipulations change the position tolerance of human object perception -- producing a tendency to (e.g.) perceive one object to be the same identity as another object across a "swap" position (*11*). Moreover, given that the animals had a lifetime of visual experience to potentially build their IT position tolerance, the strength of UTL is substantial (~5 spikes/s change per hour) -- just one hour of UTL is comparable to attentional effect sizes (*38*), and is more than double that observed in previous IT learning studies over much longer training intervals (*39-41*). We do not yet know how far we can push this learning, but we see that just two hours of (highly targeted) unsupervised experience begins to reverse the object preferences of IT neurons (Fig. 3-4D). This discovery re-emphasizes the importance of plasticity in vision (*4, 31, 34, 36, 39, 40, 42, 43*) by showing that it extends to a bedrock property of the adult ventral visual stream -- position tolerant object selectivity (*44-46*), and studies along the post-natal developmental time line are now needed.

3. 4 References and notes

1. S. Thorpe, D. Fize, C. Marlot, *Nature* **381**, 520 (1996).
2. M. C. Potter, *J Exp Psychol* **2**, 509 (1976).
3. C. P. Hung, G. Kreiman, T. Poggio, J. J. DiCarlo, *Science* **310**, 863 (2005).
4. N. K. Logothetis, D. L. Sheinberg, *Ann. Rev. Neurosci.* **19**, 577 (1996).
5. R. Q. Quiroga, L. Reddy, G. Kreiman, C. Koch, I. Fried, *Nature* **435**, 1102 (2005).
6. L. Wiskott, T. J. Sejnowski, *Neural Computation* **14**, 715 (2002).
7. P. Foldiak, *Neural Computation* **3**, 194 (1991).
8. G. Wallis, E. T. Rolls, *Progress in Neurobiology* **51**, 167 (1997).
9. R. Wyss, P. Konig, P. F. Verschure, *PLoS Biol.* **4**, e120 (2006).
10. T. Masquelier, S. J. Thorpe, *PLoS Comp. Biol.* **3**, e31 (2007).
11. D. D. Cox, P. Meier, N. Oertelt, J. J. DiCarlo, *Nat. Neurosci.* **8**, 1145 (2005).
12. G. Wallis, H. H. Bulthoff, *Proc. Natl. Acad. Sci. U S A* **98**, 4800 (2001).
13. M. Ito, H. Tamura, I. Fujita, K. Tanaka, *J. Neurophysiol.* **73**, 218 (1995).
14. H. Op de Beeck, R. Vogels, *J. Comp. Neurol.* **426**, 505 (2000).
15. Materials and methods are available as supporting material at Science Online.
16. J. J. DiCarlo, J. H. R. Maunsell, *Nat. Neurosci.* **3**, 814 (2000).
17. J. Ross, M. C. Morrone, M. E. Goldberg, D. C. Burr, *Trends Neurosci.* **24**, 113 (2001).
18. D. J. Tolhurst, J. A. Movshon, A. F. Dean, *Vision Res.* **23**, 775 (1983).
19. M. N. Shadlen, W. T. Newsome, *J. Neurosci.* **18**, 3870 (1998).
20. K. Tanaka, *Cereb Cortex* **13**, 90 (2003).
21. G. Kreiman *et al.*, *Neuron* **49**, 433 (2006).
22. T. Moore, M. Fallah, *Proc. Natl. Acad. Sci. U S A* **98**, 1273 (2001).
23. E. Kowler, E. Anderson, B. Dosher, E. Blaser, *Vision Res.* **35**, 1897 (1995).
24. J. L. Ringo, S. Sobotka, M. D. Diltz, C. M. Bunce, *J. Neurophysiol.* **71**, 1285 (1994).
25. T. Moore, A. S. Tolias, P. H. Schiller, *Proc. Natl. Acad. Sci. U S A* **95**, 8981 (1998).
26. S. Edelman, S. Duvdevani-Bar, *Neural Computation* **9**, 701 (1997).
27. G. Wallis, H. Bulthoff, *Trends Cogn. Sci.* **3**, 22 (1999).

28. K. Sakai, Y. Miyashita, *Nature* **354**, 152 (1991).

29. A. Messinger, L. R. Squire, S. M. Zola, T. D. Albright, *Proc. Natl. Acad. Sci. U S A* **98**, 12239 (2001).

30. Y. Miyashita, *Nature* **335**, 817 (1988).

31. C. A. Erickson, R. Desimone, *J. Neurosci.* **19**, 10404 (1999).

32. T. Masquelier, T. Serre, S. J. Thorpe, T. Poggio, CBCL *Tech. Report #269,* Massachusetts Institute of Technology (2007).

33. H. Sprekeler, C. Michaelis, L. Wiskott, *PLoS Comp. Biol.* **3**, e112 (2007).

34. C. D. Meliza, Y. Dan, *Neuron* **49**, 183 (2006).

35. H. Markram, J. Lubke, M. Frotscher, B. Sakmann, *Science* **275**, 213 (1997).

36. T. Yang, J. H. Maunsell, *J. Neurosci.* **24**, 1617 (2004).

37. Z. Kourtzi, J. J. DiCarlo, *Curr. Opin. Neurobiol.* **16**, 152 (2006).

38. J. H. R. Maunsell, E. P. Cook, *Philos. Trans. R. Soc. Lond. B. Biol. Sci.* **357**, 1063 (2002).

39. C. I. Baker, M. Behrmann, C. R. Olson, *Nat. Neurosci.* **5**, 1210 (2002).

40. E. Kobatake, G. Wang, K. Tanaka, *J. Neurophysiol.* **80**, 324 (1998).

41. N. Sigala, N. K. Logothetis, *Nature* **415**, 318 (2002).

42. E. T. Rolls, G. C. Baylis, M. E. Hasselmo, V. Nalwa, *Exp. Brain Res.* **76**, 153 (1989).

43. A. Seitz, T. Watanabe, *Trends Cogn. Sci.* **9**, 329 (2005).

44. M. Dill, M. Fahle, *Perception & Psychophysics* **60**, 65 (1998).

45. M. Dill, S. Edelman, *Perception* **30**, 707 (2001).

46. T. A. Nazir, J. K. O'Regan, *Spat. Vis.* **5**, 81 (1990).

47. W. Gerstner, R. Kempter, J.L. van Hemmen, H. Wagner, *Nature* **383**, 76 (1996).

48. We thank D. Cox, R. Desimone, N. Kanwisher, J. Maunsell and N. Rust for helpful comments and discussion, and J. Deutsch, B. Kennedy, M. Maloof and R. Marini for technical support. This work was supported by the National Institutes of Health (R01-EY014970) and The McKnight Endowment Fund for Neuroscience.

3. 5 Materials and methods

3. 5. 1 Animals and surgery

Experiments were performed on two male rhesus monkeys (Macaca mulatta, 5.6 and 4.3 kg). Aseptic surgery was performed to implant a head post and a scleral search coil. After brief behavioral training (1-3 months), a second surgery was performed to place a recording chamber (18 mm diameter) to reach the anterior half of the temporal lobe. All animal procedures were performed in accordance with National Institute of Health guidelines and the Massachusetts Institute of Technology Committee on Animal Care.

3. 5. 2 Stimuli presentation and behavioral task

Each recorded neuron was probed in a dimly lit room with a set of 100 achromatic images of isolated objects (52 natural objects and 48 silhouette white shapes; Fig. 3-S2) presented on a gray background (21" CRT monitor, 85 Hz refresh rate, ~48cm away, background gray luminance: 22 Cd/m^2, max white: 46 Cd/m^2). All objects subtended ~1.5° (average of a bounding square and the filled area). Custom software controlled the stimulus presentation and behavioral monitoring. Eye position was monitored in nearly real-time using standard sclera coil technique (2), and saccades >0.2° were reliably detected (*1*) .

During each *Test Phase* (~4 minutes), neuronal selectivity at three retinal positions (-3°, 0°, +3° elevation; all 0° azimuth) was probed in two different tasks. Monkey 1 freely searched an array of eight small dots (size 0.2°) vertically arranged 3° apart (Fig. 3-S1). The dots never changed in appearance, but on each "trial", one dot was randomly baited in that a juice reward was given when the animal foveated that dot, and the next "trial" continued uninterrupted. Typically, the monkey saccaded from one dot to another (not always the closest dot) looking for the hidden reward. During this task, objects were presented (100 ms duration) at controlled retinal positions (onset time was the detected end of a saccade; approximately one such presentation every other saccade, never back-to-back saccades). The monkey's task was unrelated to these test stimuli.

To limit unwanted experience across retinal positions, each such presented object was immediately removed upon detection of any saccade, and these aborted presentations were not included in the analyses. Monkey 2 performed a more standard fixation task in which it foveated a single, central dot (size 0.2°, ±1.5° fixation window) while object images were presented at a natural, rapid rate (5 images/s; 100 ms duration, 100 ms blank intervals). Reward was given at trial end (5-8 images presented per trial). Upon any break in fixation, any currently present object image was immediately removed (and not included in the analyses), and the trial aborted. The animal successfully maintained fixation in 75% of the trials. The presentations before the broken fixation were included in the analyses. To address possible adaptation concerns, we re-performed the key analysis after discarding the first image presentation in each fixation trial, and the result was essentially unchanged (also see Fig, 3-S8). Aside from the task differences (free-viewing search vs. fixation), retinal stimulation in the two tasks was essentially identical in that each tested object image was presented 3° above, below or on the current center of gaze, for 100 ms duration (20-30 pseudo-randomly interleaved repetitions of each). Given equivalent retinal stimulation, IT neuronal responsivity and selectivity are virtually identical when measured during free viewing and fixation (*1*). Consistent with this, we found comparable exposure-induced changes in IT object selectivity in each animal (e.g. see Fig. 3-3).

During each *Exposure Phase* (~15 minutes), animals freely viewed the monitor while object images were intermittently (~13 per minute) presented (pseudo-randomly) at +3° and -3° (always relative to the current center of gaze). The onset of each object was ~65 ms (randomly chosen between 30-100 ms) after the end of a saccade (defined by eye velocity < 10°/s). Because foveating a suddenly appearing object is a natural, automatic behavior, essentially no training was required, and the monkeys almost always saccaded directly to the object (>90%) within 108 ms (median; range: 66-205 ms) and foveated it for >200 ms (after which it was removed and the monkeys received a drop of juice). For objects presented at the "swap" position (+3° or -3°; strictly alternated neuron-by-neuron), the object (e.g. "P") was consistently swapped by another object (e.g. "N") upon the detection of saccade onset (eye speed > 60°/s). We took great care to ensure the spatiotemporal precision of the stimuli delivery. The to-be-swapped object was always successfully removed before the end of the saccade, and the new object was present at the to-be-center of the retina within 1 ms of the saccade end (mean; worst case was 10 ms after saccade end). To prevent unintended retinal experience, the object image was automatically taken away if the saccade was not directed toward the object, or if the eye landed more than 1.5° away from its center.

3. 5. 3 Neuronal recordings

The extra-cellular spiking activity of single, well-isolated IT neurons was recorded using standard microelectrode methods (*1*). 101 neurons were randomly sampled over a ~4x6 mm area of the ventral STS and ventral surface lateral to the AMTS (Horsey-Clark coordinates: AP 11-15 mm; ML 15-21 mm at recording depth) from the right hemisphere of Monkey 1 and left hemisphere of Monkey 2. In each daily recording session, we advanced a microelectrode while the 100 object images (Fig. 3-S2) were pseudo-randomly presented at the center of gaze while the monkey performed either the free-viewing search task or the fixation task (see *Test Phase* above). All responsive neurons with a well-isolated waveform were further probed with the same object set (initial screening of 100 objects, ~5 repetitions per object, all presented at the center of gaze).

Main test objects ("swap" objects): Among the objects that drove the neuron "significantly" above its background response (t-test against 50 ms epoch before stimuli onset, $p < 0.05$, not corrected for multiple tests), the most preferred (P) and least preferred (N) objects were chosen as a pair for the *Exposure Phase* ("swap objects") subject to the condition that both objects were either from the "natural" object set or the "silhouette" object set (see Fig. 3-S2). This object selection procedure aims to find conditions in which *both* objects drive the neuron, and object P drives the neuron more strongly than object N. For most neurons, object N was not the second most preferred object (IT neurons typically respond well to ~10% of object images, which is consistent with our online observation that N was roughly the tenth most preferred object in the set of 100 tested objects). Note that, because the procedure for choosing objects P and N was based on limited initial testing, it does not fully guarantee that the selectivity will be found with further testing (see main text). Furthermore, because the initial testing was at the center-of-gaze position and IT neurons are not uniformly position tolerant, the procedure also does not guarantee that the response to P is greater than N at all three tested positions (i.e. possible "negative" object selectivity at some positions). We use post-hoc analysis and screening to examine any unexpected effects of this screening procedure (e.g. post-hoc removal of neurons with low or negative selectivity, table 3-S1). Post-hoc analyses also showed that roughly equal numbers of neurons were recorded using swap objects from each set (57 natural, 44 silhouette) and neurons showed virtually the same reported changes in object selectivity when sorted by object set type.)

Control objects: For each recorded neuron, we also used the initial response testing (above) to choose a pair of control objects. Our goal was to choose these two objects among which the neuron was selective, but were very distant from the "swap objects" in shape space. Because we do not know the dimensions of shape space, we cannot strictly enforce this. In practice, we simply insured that the control objects were always chosen from the object set that was not used for the swap objects (i.e. when two objects from the "natural" set were used as "swap" objects, the two control objects were from the silhouette set, see Fig. 3-S2). Within this constraint, the control objects were chosen using the exact same responsivity and selectivity criteria as the test objects (above).

Once the initial screening and object selection were completed, we carried out the *Test* and *Exposure Phase* in alternation for as long as we could maintain isolate of the neuron's waveform. Both swap objects and control objects were presented (tested) at all three positions during each *Test Phase* but only the swap objects were shown and manipulated during each *Exposure Phase.*

In addition, multiple-unit activity (MUA) was collected from 10 sites on the IT ventral surface of Monkey 2 on 10 different experimental sessions (days). Nearby IT neuron have similar object selectivity (3) and, consistent with this, we have previously shown that MUA is shape selective and moderately position tolerant (4). MUA was defined as all the signal waveforms in the spike band (300 Hz – 7 kHz) that crossed a threshold set to ~2 s.d. of the background activity. The threshold was held constant for the entire session. All other recording procedures were identical to the recording procedure used for the single-unit recording except: 1) during the *Test Phase,* more repetitions (~50) were collected per object image at each position; 2) each *Exposure Phase* was approximately twice as long (e.g. 200 "swap" exposures instead of 100).

3. 5. 4 Data analysis

To get the most statistical power from the data, average firing rates were computed over a time window optimally chosen for each neuron by an algorithm that estimated neuronal response onset and offset time (relative to stimulus onset) using all stimuli (see below). Analyses for the single-unit data in the main text were performed using such neuron-specific spike count windows (described next). All analyses were also repeated using a standard, fixed spike count window (onset 100 ms; offset 200 ms) with very similar results (Δ_S(P-N) = -5.4 spikes/s per 400 exposures at the swap position; -0.3 at the non-swap position; $p < 0.01$, "position x exposure" interaction by permutation test).

Neuronal response window: Specifically, for the single-unit data, each neuron's responses to each stimulus condition were computed over the same spike count time window. That window was optimally chosen for each neuron using the following algorithm (5). Given a neuron, we first computed its average firing rate profiles FRobj(t) in overlapping time bins of 25 ms shifted in time steps of 1 ms. This averaged firing rate was computed with all the response data to all the objects presented in all *Test Phases* for the neuron. We also computed the neuron's background rate FRbk as the mean spike rate between 0 and 50 ms after stimulus onset. Because multiple images were presented on each trial at a relatively rapid rate (see above), the more standard epoch before each stimulus onset was not ideal for computing background because it was occupied by the neuron's response to the previous stimulus. We previously found that the short epoch starting at stimulus onset and ending before the neuronal response (i.e. before the known IT latency) provides the most reliable estimate of a "background" firing rate (5). Then, by subtracting this background rate from the averaged object response rate profile, we obtained an averaged driven rate profile FRdriven(t) = FRobj(t) - FRbk. Finally, we identified the samples for which FRdriven(t) was at least 20% of its peak value. The largest continuous interval of samples fulfilling this requirement was always centered on the peak of the neuronal response. If no other samples, outside this main "peak" interval, fulfilled the requirement, the extremes of the interval were chosen as the extremes of the optimal spike count window for that neuron. If the firing profile exceeded 20% of its peak in other regions of the time axis, these were merged with the main interval only if they were within 25 ms from it. In this case, the extremes of the merged interval were chosen as the extremes of the optimal spike count window. In principle this algorithm could yield a very small response window, so we limited that possibility by imposing a minimum response window size of 50 ms at the estimated onset latency. (In practice, only two neurons out of 101 had this minimum window size imposed; neuron1: 30ms; neuron2: 40ms)._All analyses presented in the main text were carried out by counting spikes in these neuron-specific optimized time windows. The mean window start time (± s.d.) was 119.5 ± 38 ms, the mean window end time was 244 ± 76 ms, and the median duration was 110 ± 50 ms. These time windows are consistent with previous work (*6*) and with animal reaction times in recognition tasks (*7*).

For the multi-unit data, we used a standard, fixed (100-200 ms) spike count time window for all recording sites.

For all the results presented in the main text, the object selectivity for an IT neuron was computed as the difference in its response to object P and N. To avoid any bias in this estimate of selectivity, for each neuron we defined the labels "P" and "N" by splitting the pre-exposure response data and used a portion of it (10 response repetitions to each object at each position) to determine these labels ("P" is the preferred object that elicited a bigger overall response pooled across position). The label "P" and "N" for the neuron was then held fixed across positions and later *Test Phases*. All remaining data were used to compute the selectivity results in the main text using these labels. This procedure ensured that any observed response difference between object P and N reflected true selectivity for a neuron, not selection bias.

In cases when neuronal response data is normalized and combined (e.g. Figs. 3-4D, 3-S3, 3-S9), we used the same normalization scheme through out all the analyses. Specifically, each neuron's response from each *Test Phase* was normalized to its mean response to all objects at all positions in that *Test Phase*.

3. 5. 5 Statistical tests for the "Position x Exposure" interaction

A key part of the prediction is that any change in object selectivity should be found predominantly at the swap position (Fig. 3-1C). Individual t-tests show a highly significant effect at the swap position, but no significant effect at the non-swap position (see Fig. 3-3A). However, to directly test for an interaction between position and our independent variable (exposure), we applied a general linear model to the response difference (in firing rate) between the object P and N. The formulation is similar to the analysis of variance tests (ANOVA). However, it is not subject to assumptions about the noise distributions.

The model had the following form:

$$(P-N)_{\substack{neuron=n \\ position=p \\ \exp osure=e}} = a_n + (b_1 \cdot p) + (b_2 \cdot e) + (b_3 \cdot (p \cdot e))$$

The three independent variables of the model are: position (p), exposure (e), and their interaction (i.e. their product, $p{\cdot}e$). The position factor has two levels (i.e. $p = -1$ for swap position, 1 for non-swap position) the exposure factor has up to 5 levels depending how long a neuron was held, (i.e. $e = 0$ for pre-exposure, and can be up to 400 exposures in increments of 100's,). Each a_n is the selectivity offset specific to each neuron; b_1, b_2, and b_3 are slope parameters that are shared among all the neurons. Thus, the complete model for our population of 101 neurons contained a total of 104 parameters (101 a_n's, b_1, b_2, and b_3) that were fitted simultaneously to our entire data set. The a_n's absorb neuron-by-neuron selectivity differences that are not of interest here, and the remaining three parameters describe the main effects in the population, with b_3 of primary interest (interaction).

To test for a statistically significant interaction, we fit the linear model to the data (standard least squares), and then asked if the observed value of the interaction parameter (b_3) was statistically different from 0. The validity of the significance was verified by two different approaches: 1) bootstrap and 2) permutation test.

Bootstrap is widely used to provide confidence intervals on parameter estimates. However, the bootstrap confidence interval can also be used to provide a significance level for a hypothesis test (*8*). To do this, we estimated the distribution of the b_3 estimate via a bootstrap over both neurons and repetitions of each neuron's response data. The exact procedure was done as follows: for each round of bootstrap over neurons, we randomly selected (with replacement) 101 neurons from our recorded 101 neurons, so a neuron could potentially enter the analysis multiple times. Once a set of neuron was selected, we then randomly selected (with replacement) the response repetitions included of each neuron (our unit of data is a scalar spike rate in response to a single repetition of one object at one position). Each neuron's (P-N) was computed from its selected response repetitions. The linear model was then fit to the data at the end of these two random samples to obtain a new b_3 value. This procedure was repeated 1000 times yielding a distribution of b_3 values, and the final p-value was computed as the fraction of times that distribution is less than 0 ($p = 0.009$, 1000 samples). This p-value is interpreted as: if we were to repeat this experiment, with both the variability observed in the neuronal responses as well as the variability in which neurons were sampled, what is the chance that we would *not* see the interaction observed here? In effect, the bootstrap analysis determined the confidence interval around our originally observed b_3 value, and the duality of confidence intervals and hypotheses testing allowed us to report that confidence interval as a p value (*8*). When the same analysis was applied to the pair of control objects that was included in the *Test Phase* but was not shown during the *Exposure Phase,* we did not observe a significant interaction ($p > 0.05$). Simulations with Poisson spiking neurons have confirmed the correctness of our analysis code.

To further confirm statistical significance level of the interaction (position x exposure) by permutation test, we created a null distribution by randomly permuting the position labels of the original data points (swap vs. non-swap) from each exposure level. Following each permutation, we fit the linear model to the permuted response data. Repeating this, we determined the fraction of times that the interaction term (b_3) was greater the observed value and took this as the p value ($p = 0.007$, one-tailed test, 1000 samples). Again, when we applied the same test to the response data to the control objects, we did not observe a significant interaction ($p > 0.05$).

Together, our statistical tests point to a position-specific change in object selectivity at the swap position that increases with amount of exposure. The same statistical tests were applied to the multi-unit data ($n = 10$ sites), also yielding a significant interaction between position and exposure ($p = 0.03$ bootstrap, $p = 0.014$ permutation test) for the swap objects. Again no such interaction was observed for the control objects ($p > 0.05$ for both bootstrap and permutation test).

3. 5. 6 Statistical tests for the response change in single neurons/sites

In Fig. 3-4B and C, we evaluated each single-unit-neuron/multi-unit-site's response change to P, N or (P-N) by fitting linear regression as a function of exposure time to obtain a slope (Δ_SP, Δ_SN, or Δ_S(P-N)). The statistical significance of the response change for each single-unit neuron or multi-unit site was evaluated by permutation test. Specifically, for each neuron, we randomly permutated the exposure level label for each response sample, (i.e. which *Test Phase* each sample of P and N belongs to). We then re-fit the linear regression to the permuted data and repeated the same procedure 1000 times yielding a distribution of slopes ("null distribution"). The p value was determined by counting the faction of time the null distribution exceeded the linear regression slope obtained from the data. All neuron/sites with $p < 0.05$ was deemed significant (dark colored bars in Fig. 3-4B and C).

3. 6 Supporting text

3. 6. 1 Relationship between learning effect size and IT neurons' object selectivity at the center of gaze

Implicit in our experimental design and hypothesis is that the IT neurons prefer object P over N. Indeed, object selectivity at the center of gaze may provide the "driving force" for the temporal-continuity learning in our experiment. Thus, if all our IT neurons had no selectivity among the objects P and N, then the temporal-continuity learning hypothesis makes no clear prediction in our data. On average, our recorded IT neurons did have the desired net positive object selectivity (Table 3-S1). However, even though we aimed to pick objects so that P > N, we used an initial screen that mainly sought to insure that *both* objects P and N tended to drive the neuron (see main text and Materials and methods, above). As a result, the population distribution of selectivity for P and N at each position was very broad, and some neurons did not have clear object selectivity at the center of gaze. Here we consider subsets of IT neurons with ever-more stringent positive object selectivity at the center of gaze by re-performing our key statistical analyses on these subsets. In sum, the main results are unchanged: 1) as with the entire population, these subsets of selective neurons show a significant interaction between position and exposure time, (see details of this statistical test in Materials and methods, above); 2) as with the entire population, these subsets of selective neurons show a significant shift in their mean object selectivity (P-N) at the swap position compared to the non-swap position (pair-wise t-test). These results are summarized in Table 3-S1.

To examine whether a relationship exist between the IT neuron's selectivity at the center of gaze and the magnitude of the learning effect, we divided the neurons (n = 101) into sub-populations based on each neuron's selectivity for P and N at the center of gaze (Fig 3-S5). Examining the magnitude of the learning effect across these sub-populations of neurons revealed two things: 1) neurons with no or negative selectivity, on average, produced no learning effect (effect size = -0.8 spikes/s per 400 exposures); 2) effect size grew increasingly stronger among neurons with higher initial object selectivity at the center of gaze. This is consistent with the notion that selectivity at the center of gaze may be a good estimate of the "driving force" for the learning effect. Expected Poisson spiking variability in the neuronal responses prevented us from determining if this measure is a perfect predictor of the learning effect size, but we speculate that the learning effect size in each neuron cannot be simply predicted from object selectivity at the center of gaze alone, but probably also depends on (e.g.) its initial response magnitudes at the swap position.

3. 6. 2 The object selectivity change is robust to the choice of selectivity metric

Is the change in object selectivity reported here robust to the choice of metric quantifying it? In the main text, we performed all the analyses by quantifying the object selectivity in raw response difference (spikes/s). Here we explore another standard selectivity metric and show the reported results is un-affected by the choice of metric.

Specifically, for each neuron we computed a "contrast" object preference index (OPI) at each position using a standard metric (*9, 10*) :

$$OPI = \frac{P - N}{P + N}$$

where P and N is the neuron's response to object P and N. OPI ranges from -1 to +1 and 0 is no object selectivity. We computed each neuron's change in object selectivity magnitude as:

$$\Delta OPI = OPI_{post-\exp osure} - OPI_{pre-\exp osure}$$

The re-make of main text Fig. 3-3 using the OPI metric is shown in Fig. 3-S6. To weight all neurons approximately equally before pooling, the ΔOPI in Fig. 3-S6 was normalized for each neuron to its pre-exposure object selectivity:

$$\text{Normalized } \Delta OPI = \frac{\Delta OPI}{(OPI_{pre-\exp osure} + 1)}$$

Because the OPI metric ([-1 1]) can have occasionally near-zero or negative values, the addition of one in the denominator was used to regularize the normalization.

3. 6. 3 Dependence of effect size on initial object selectivity

We considered the possibility that the observed change in object selectivity at the swap position (but not at the non-swap position) might somehow have resulted from larger (or smaller) initial object selectivity magnitude at that position (relative to the non-swap position). Because we changed the swap and non-swap position across each recorded neuron, position is counter-balanced across neurons so, in the limit, no difference in initial object selectivity magnitude is expected. Indeed, the initial object selectivity was very closely matched between the swap and non-swap positions across the neuronal population (Fig 3-S7).

However, even with the counterbalance, it turned out that our recorded population had a slightly greater initial selectivity at the swap position (by ~ 2 spikes/s in (P-N); Table 3-S2). This difference between the swap and non-swap position was not significant ($p > 0.05$, two-tailed t-test). Nevertheless, to fully control for any effect of this slight difference on our results, we performed a post-hoc analysis to completely match the initial object selectivity. Specifically, we discarded neurons with significantly greater selectivity (P-N) at the swap position ($p < 0.05$, one-tailed permutation test), and then re-performed the key analyses on the remaining population. This re-analysis was done for the population of all 101 neurons in Fig. 3-2 & 3-3 and for the population of 38 highly-object-selective neurons/sites in Fig 3-4C. The results showed that our post-hoc selection had completely eliminated the small bias in initial selectivity at the two positions, but the effect size at the swap position in the "equalized" populations was almost as the original populations (and still no effect at the non-swap position; see Table 3-S2 and Fig. 3-S8).

Finally, by sorting the neurons based on their response profile across positions, we found that even neurons that initially were less responsive at the swap position showed a change in object selectivity (Fig 3-S9 – especially see panel B).

3. 6. 4 The object selectivity change cannot be explained by "adaptation"

When repeatedly probed with visual stimuli, neurons in ventral visual stream have been shown to reduce their evoked responses, a phenomenon referred to as "adaptation" (*11-15*). Our IT neuronal data do show evidence of "adaptation" (outlined below). However, the key changes in IT selectivity we report in this manuscript cannot be explained by "adaptation". First, although each object was shown equally often at the two key positions (swap and non-swap positions), the selectivity change we found was specific to the swap position. This specificity cannot be explained by any standard model of "adaptation". Second, approximately half of the selectivity change was due to an enhanced response to object N (see main text), which is inconsistent with any fatigue-adaptation model. Third, the selectivity change we found continued to grow larger and larger for as long as we could measure it (up to two hours),which is suggestive of plasticity, rather than cortical fatigue.

Consistent with previous reports on IT "adaptation" (*11-15*), we observed a reduction in the neurons' responses over short time scales (i.e. within a trial, ~1 sec) and intermediate time scales (i.e. within a *Test Phase*, ~4 min). Such reduction was not specific to objects or position (Fig 3-S10A and B). However, the change in selectivity we report here emerged slowly over a much longer times scale (across multiple *Test Phases*), and was specific to both the swap position and the object (Fig 3-S10 C). Interestingly, there was virtually no change in response across the longer time scale over which the learning effect emerged (e.g. non-swap position, Fig. 3-S10 right panel of row C).

3. 6. 5 Monte Carlo simulations with Poisson spiking neurons

Beyond the main (net) change in object selectivity reported here, our figures show that many individual neurons *appear* to undergo changes in object selectivity in both the predicted and non-predicted directions (e.g. scatter of the individual neurons in Fig 3-2C). Is this non-predicted variability in the observed object selectivity accounted for by Poisson variability (i.e. noise effects on repeat measurements)?

To address this question, we ran Monte Carlo simulations using Poisson spiking statistics. We first computed each neuron's mean firing rates to each object at each position before any exposure. Using these estimated firing rates, we simulated Poisson spiking neurons. We then took repeated measurements from this simulated population of neurons. 30-50 response repetitions were collected from each simulated neuron to constitute a simulated *Test Phase*, (matched to the number of response repetitions collected from real neurons/sites). Each simulated neuron was then tested across the same amount of exposure time (i.e. number of *Test Phase*) as the real neurons.

In the simulations, we assumed that the object selectivity for P and N at the swap position was changing at the rate estimated from the data (5.6 spikes/s per 400 exposures, see Fig. 3-3C) and not changing at the non-swap position. That is, the simulated neurons' firing rates at the non-swap position were held fixed at each simulated Test Phase, while the firing rates at the swap position were undergoing changes at -0.7 spikes/s for object P and +0.7 spikes/s for object N across each *Test Phase* (100 exposures).

This simulation allowed us to determine the expected variability in a real neuronal population's selectivity under Poisson firing statistics. Finally, we compared this expected variability to that observed in our data by plotting the results together (Fig. 3-S11). These Monte Carlo simulations showed that the magnitude of the non-predicted changes in object selectivity is comparable to that expected from well-established cortical neuron Poisson spiking statistics (*16, 17*).

3. 7 Supporting figures

3. 7. 1 Supporting figure 3-S1

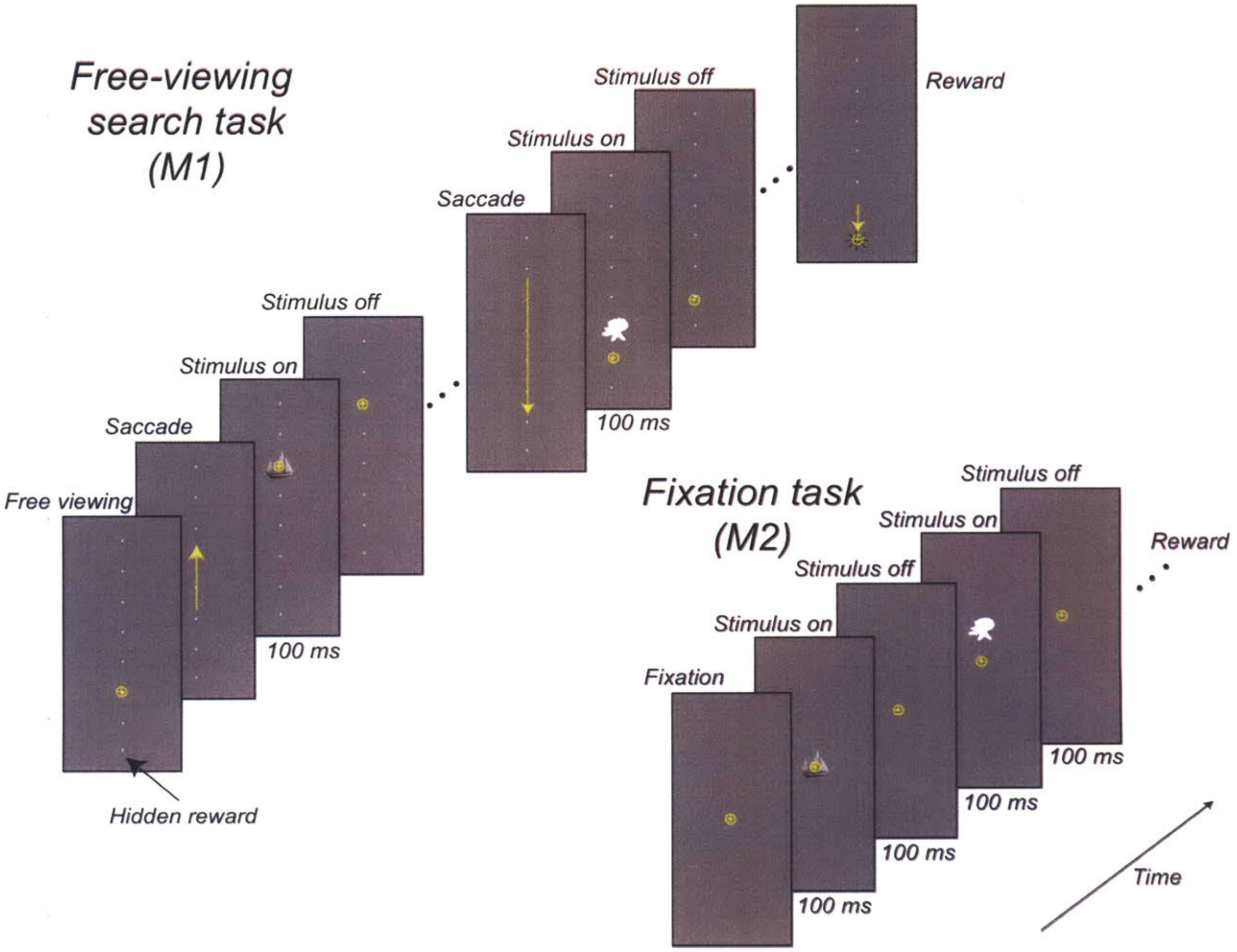


Fig. 3-S1. Monkey tasks during neuronal testing. During the *Test Phase* of the experiment, Monkey 1 performed a free-viewing search task, searching for a reward "hidden" beneath one of eight, spatially fixed dots; Monkey 2 performed a standard, fixation task while stimuli were presented at a rate of 5 per second (100 ms duration, 100 ms gaps). In both cases, retinal stimulation for the presentation of each object image was identical in that object images were presented 3° above, 3° below, or at the center of gaze for 100 ms. In both cases, these object

image presentations were fully randomized and unrelated to the animals' tasks. Each animal also performed its *Test Phase* task while we advanced a microelectrode to search for neurons, but object images were only presented at the center of gaze in that case.

3. 7. 2 Supporting figure 3-S2

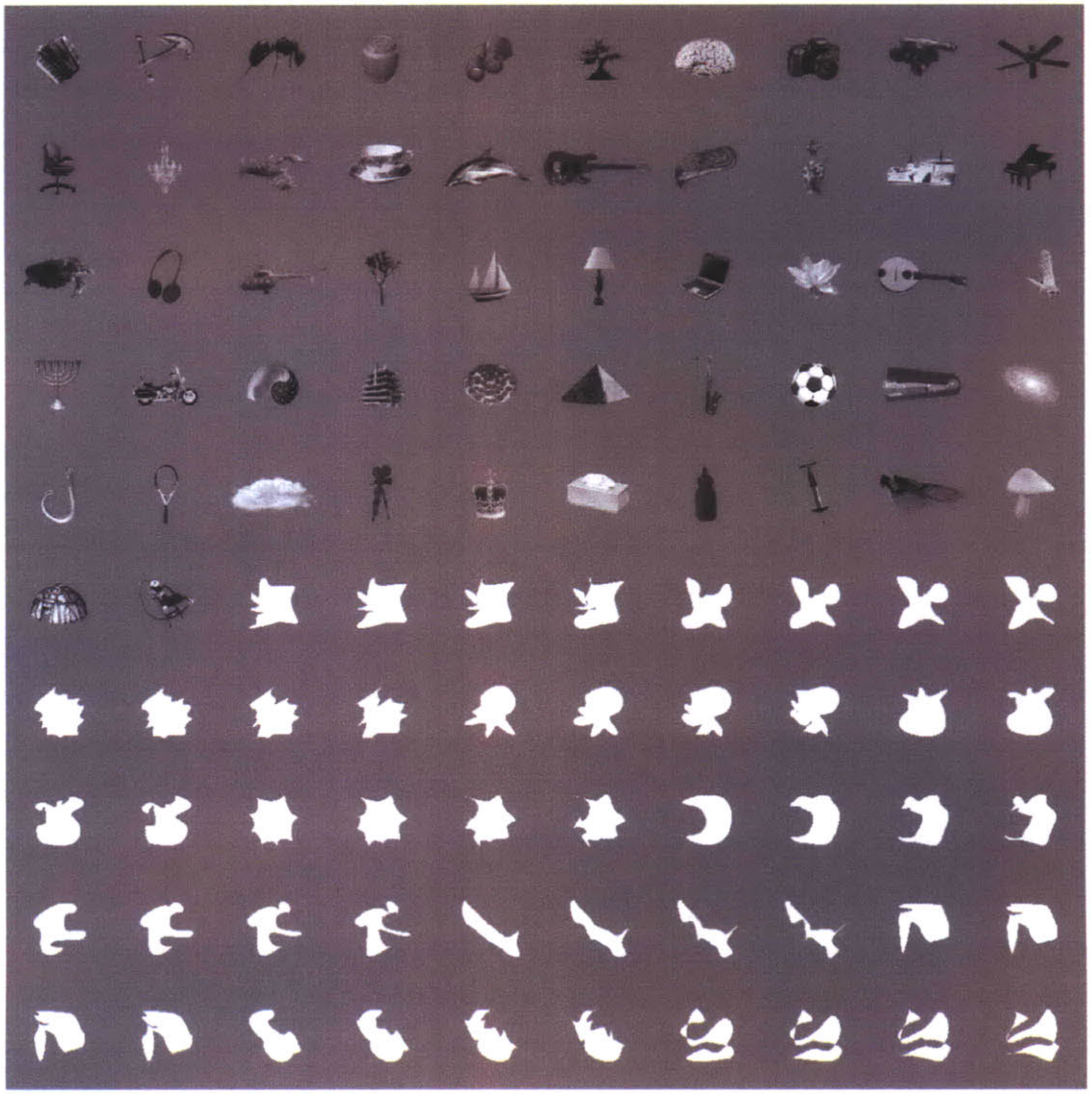


Fig. 3-S2. Set of 100 object stimuli. The figure reflects the true relative size of the objects shown in the experiment. The first 52 images are referred to as the "natural" object set, the latter 48 images as the "silhouette" object set. We only swapped pairs of objects draw from within the same set (see Materials and methods).

3. 7. 3 Supporting figure 3-S3

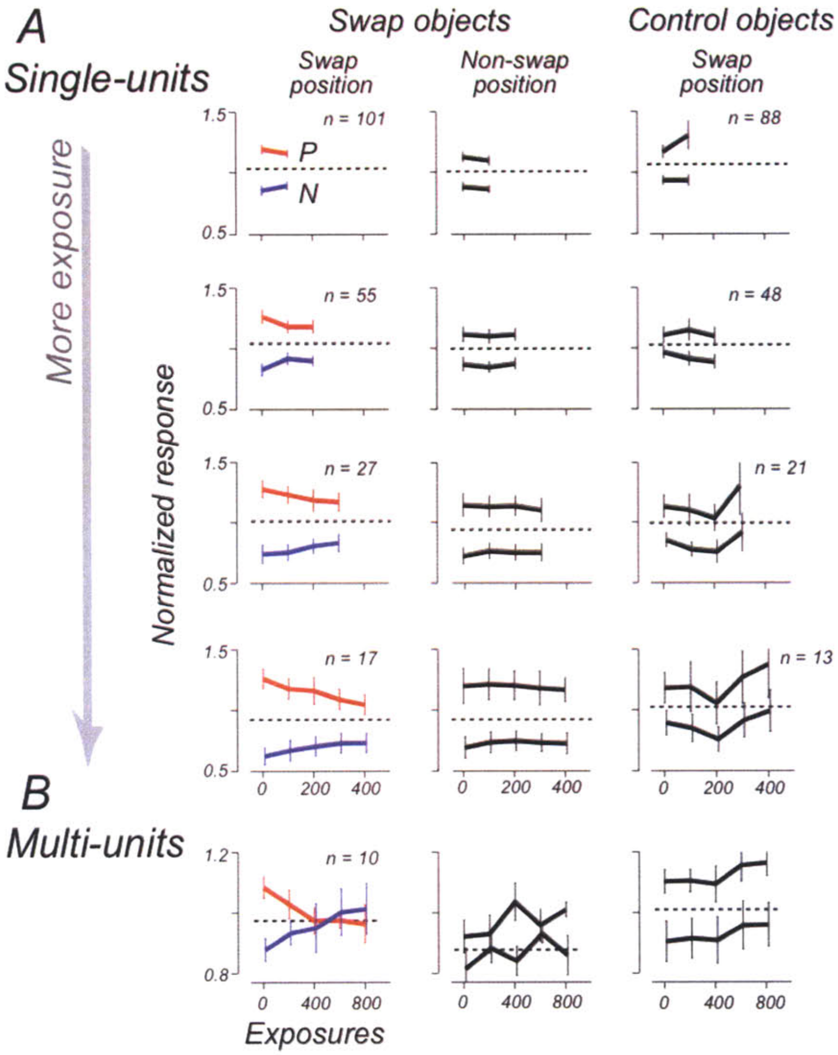


Fig. 3-S3. Mean normalized population response to object P and N as a function of exposure time. (**A**) Single-unit population data. Each row of plots show neurons held for different amounts of time, (e.g. the top row shows all neurons held for over 100 swap exposures -- including the neurons from the lower rows; the second row shows the neurons held for over 200 exposures; etc). Each neuron's response from each *Test Phase* was normalized to its mean response to all objects at all positions in that *Test Phase*. (**B**) Multi-unit population data. Note the scale change on the x-axes between (A) and (B). Though, by chance, there turned out to be slightly greater initial selectivity at the swap position than the non-swap position, our reported effect was still strongly present even when post-hoc analysis was used to eliminate this initial selectivity difference (see Table 3-S2 and Fig 3-S8).

3. 7. 4 Supporting figure 3-S4

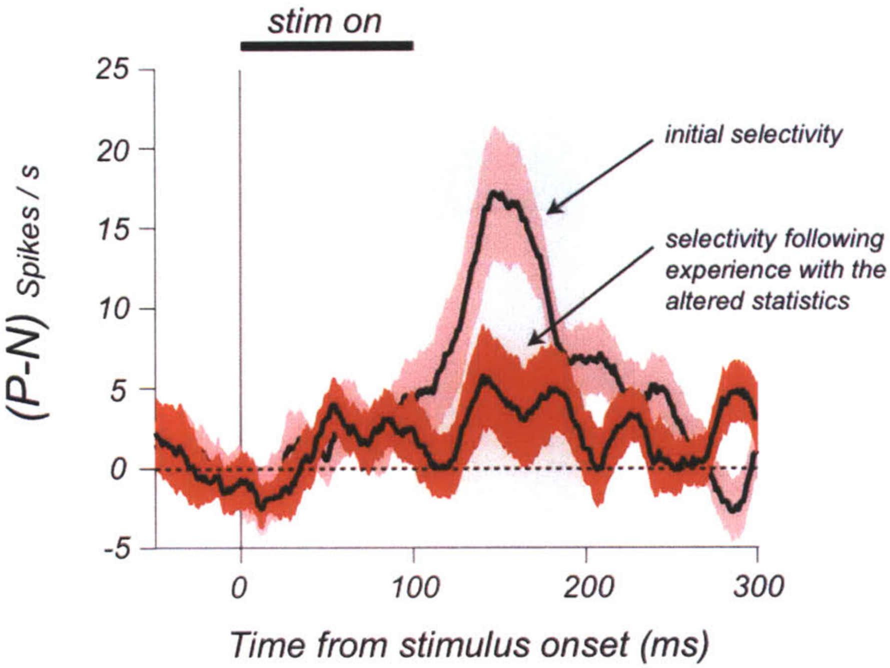


Fig. 3-S4. The response time course of the IT selectivity (measured in the *Test Phase,* see Materials and methods) before and after at least one hour of experience in the altered visual world. The plot shows the averaged PSTH difference between object P and N. Colored area indicates the standard error of the mean. Only the neurons (n = 17) and multi-unit sites (n = 10) held for the longest exposure time are included in the plot. Each neuron's PSTH was computed by smoothing the response data with a Gaussian window (s.d. 10 ms). The responses are aligned at the onset of the stimulus. The black bar on the top indicates the duration of the stimulus. Note that the change in selectivity following experience was present even in the earliest part of the response (~100 ms). The same neurons showed no change for objects presented at the equally eccentric (non-swap) retinal position (time course not shown; see Figs. 3-2, 3-3).

3. 7. 5 Supporting figure 3-S5

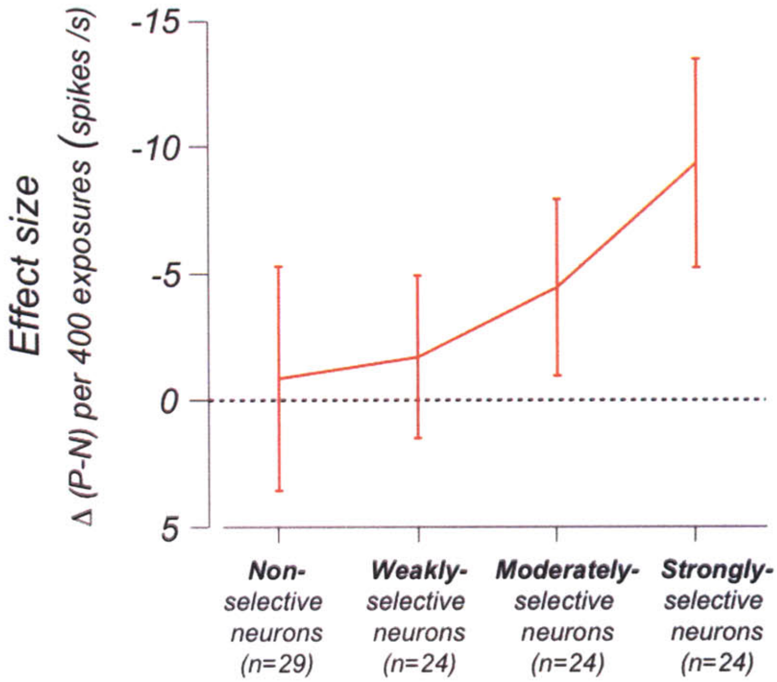


Fig. 3-S5. Relationship between the magnitude of the learning effect and the IT neurons' object selectivity at the center of gaze. The abscissa shows neurons grouped by the amount of object selectivity at the center of gaze ("Non-selective neurons" are those with object selectivity (P-N) less than 1 spike/s; the remaining neurons were split into three even groups). The mean selectivity among the particular tested two objects (P and N) in the three groups was: 2, 5, and 15 spikes/s. The ordinate shows the mean experience-dependent effect size for the neurons in each group. Effect size was estimated for each neuron using regression analysis that leverages all the available data for each neuron (same as in Fig. 3-4C), so it reflects an unbiased estimate of effect size that is not confounded by any differences in total duration of exposure. There was no correlation between this effect size estimate and the duration of exposure (r = 0.05, p = 0.62). The plot shows the mean effect size in each group of neurons and s.e.m. When the same analysis was carried out at the non-swap position, the effect sizes were near zero for all four groups of neurons (not shown).

3. 7. 6 Supporting figure 3-S6

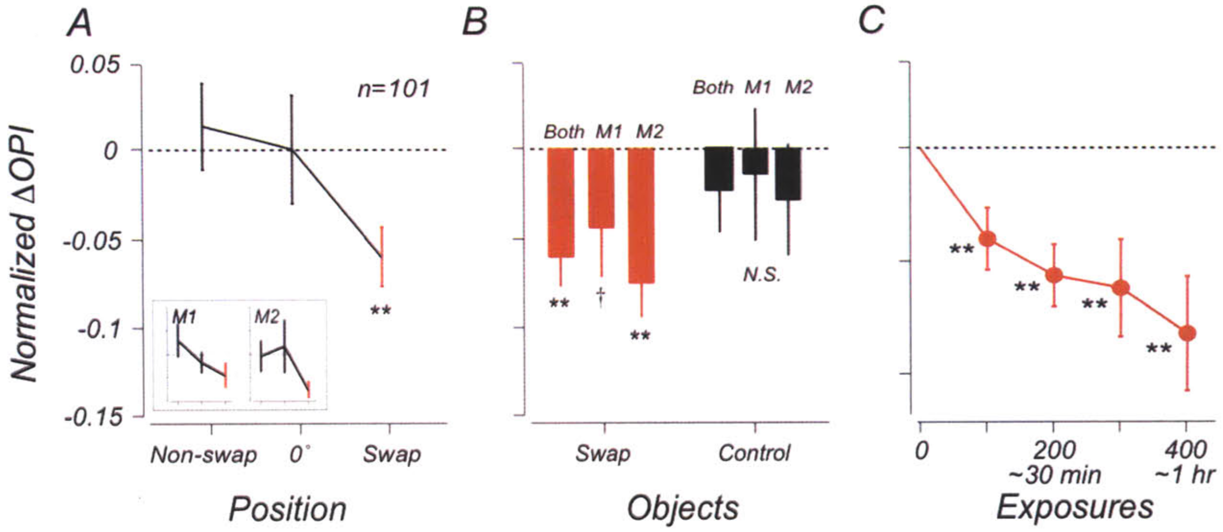


Fig. 3-S6. Results in another selectivity metric. Here, results from all the neurons are re-plotted in the format of Fig. 3-3 but in normalized ΔOPI metric. (** significantly less than 0 at p < 0.001, one tailed t-test; N.S. p > 0.05; † approaching significance, p = 0.058).

3. 7. 7 Supporting figure 3-S7

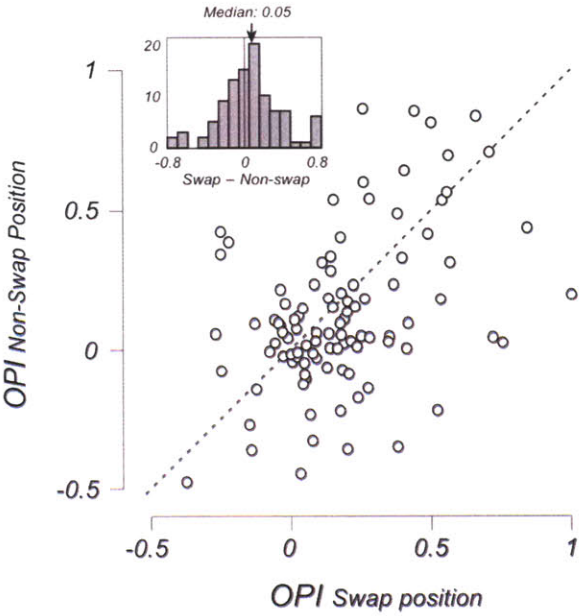


Fig. 3-S7. Initial object selectivity (OPI) of all the neurons at the swap position and non-swap position. Because the object P and N are determined from the summed response across all positions using separate data, and IT neurons are not perfectly position-tolerant, the object

selectivity at any given position (swap or non-swap) could have initially negative values. The inset shows the histogram of the object selectivity difference between the swap and non-swap positions. Though, by chance, there turned out to be slightly greater initial selectivity at the swap position than the non-swap position, our reported effect was still strongly present even when post-hoc analysis was used to eliminate this initial selectivity difference (see Table 3-S2 and Fig 3-S8). This plot also illustrates the broad distribution of selectivity in the full population, including some neurons with weak or negative selectivity (see Fig. 3-S5 above).

3. 7. 8 Supporting figure 3-S8

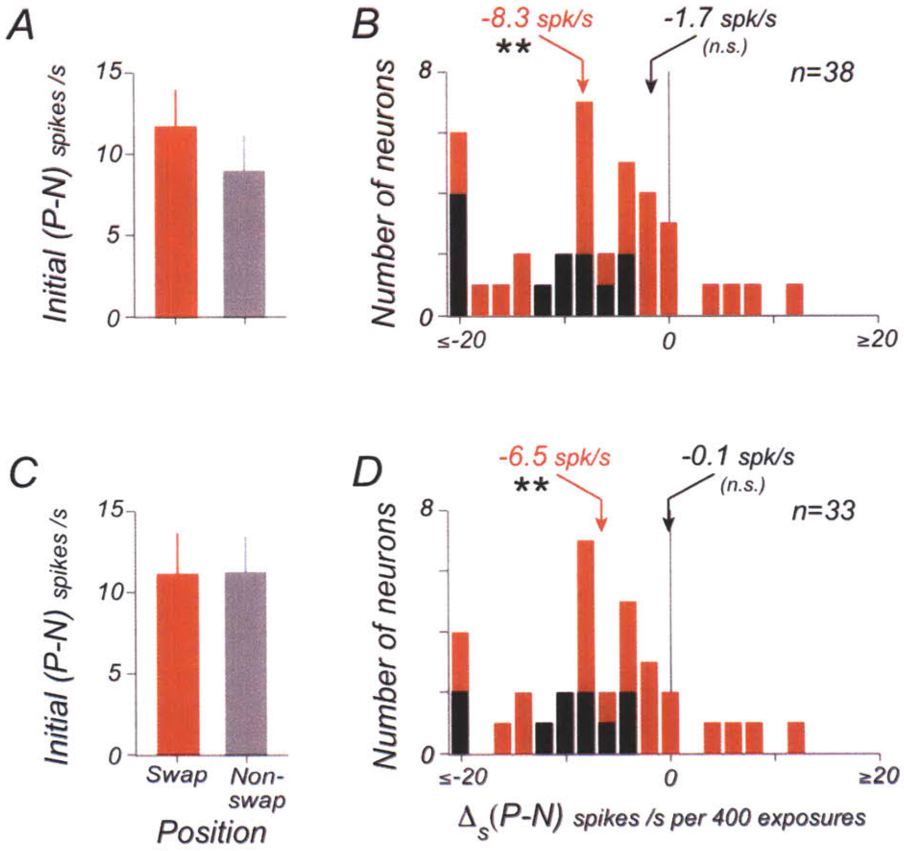


Fig. 3-S8. Post-hoc selection to eliminate the initially greater selectivity at the swap position among the 38 highly object selective neurons/sites tested for at least 300 exposures (see main text). **(A)(B)** Data from the original set of 38 neurons. **(A)** mean selectivity at the swap and non-swap position before exposure. **(B)** is a re-plot of Fig 3-4C. Red arrow indicates the mean Δ_S(P-N). Black arrow indicates the mean Δ_S(P-N) from the non-swap position (individual neural data not shown). (** significantly less than 0 at $p < 0.001$, one tailed t-test; n.s. $p > 0.05$;) **(C)(D)** Re-make of **(A)** and **(B)** after neurons/sites with significantly greater selectivity at the swap position were discarded.

3. 7. 9 Supporting figure 3-S9

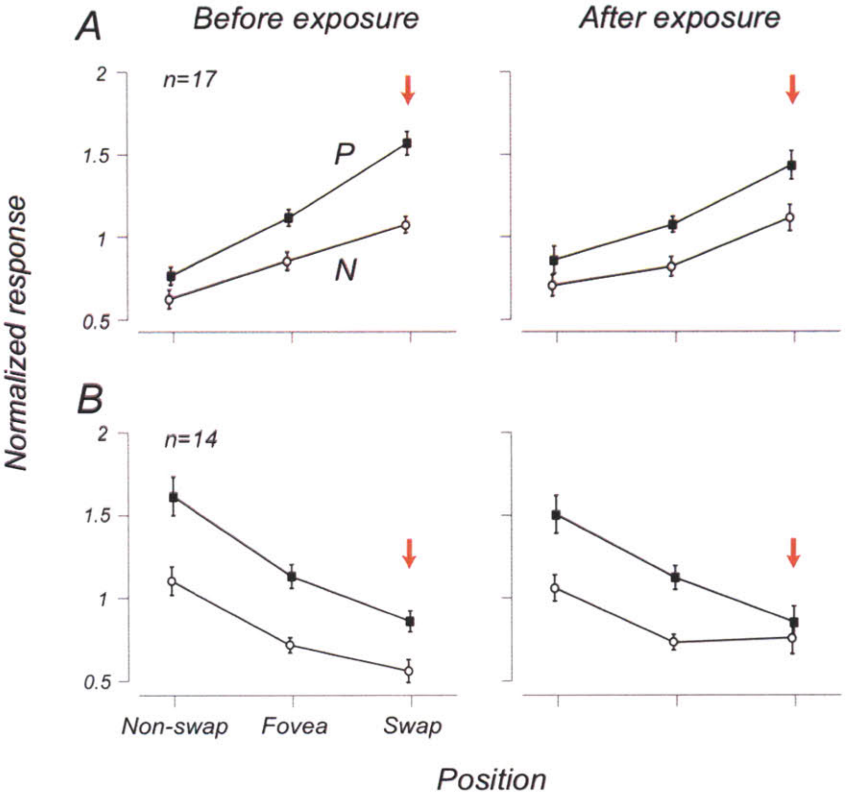


Fig. 3-S9. Population response averages before and after exposure for those neurons preferring either the swap or non-swap position. Neurons were sorted by their receptive field profiles. Neurons in (**A**) were selected as those for which their maximum response (among either object) was evoked at the swap position and their minimum response at the non-swap position (n = 14; max and min taken among all three tested positions). Neurons in (**B**) were selected as those that had a maximum response at the non-swap position and minimum response at the swap position (n = 17). Neurons in (**A**) and (**B**) underwent, on average, ~200 exposures (~30 min). Each neuron's response was normalized to its averaged response to all objects at all positions before being combined in the group average. Error bars indicate the standard error of the mean.

3. 7. 10 Supporting figure 3-S10

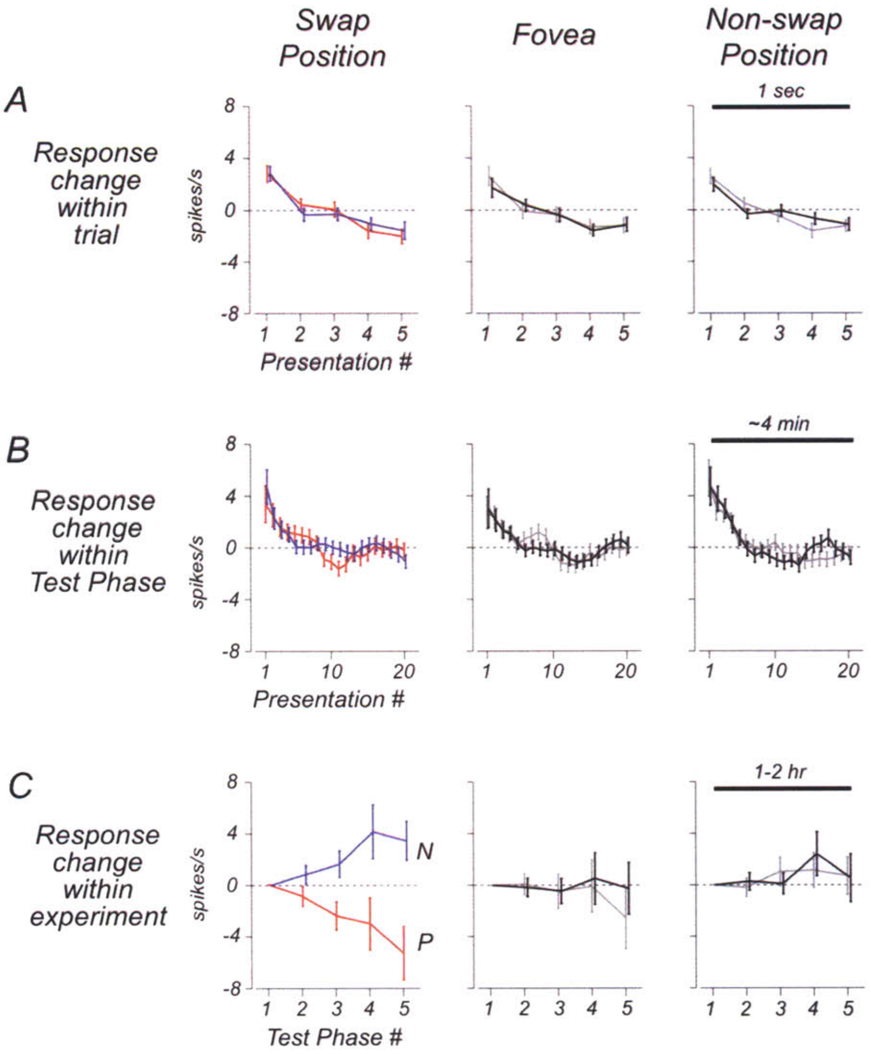


Fig. 3-S10. Response changes across different time scales for the swap objects. (**A**) Population mean response changes within a trial (n=111, single-unit data and multi-unit data combined). For each neuron, the mean responses to each object at each position were subtracted before averaging. Responses were binned by the order in which the image (object at a position) appeared in the sequence of stimuli shown in the trial. (**B**) Population mean response changes within a *Test Phase*. For each neuron, the mean responses to each object at each position from the *Test Phases* were subtracted before averaging. Responses were binned by the order they appear in a *Test Phase* (20 - 30 repeats of each object at each position were tested during each *Test Phase*; see Materials and methods). These plots are smoothed with a sliding window (4 data point wide). (**C**) Population mean response changes across multiple *Test Phases (*the left plot in this row illustrates one view of our reported learning effect). For each neuron, the mean responses from the first *Test Phase* (pre-exposure) were subtracted before averaging. Red and gray traces show responses to object P; blue and black traces show responses to object N.

3. 7. 11 Supporting figure 3-S11

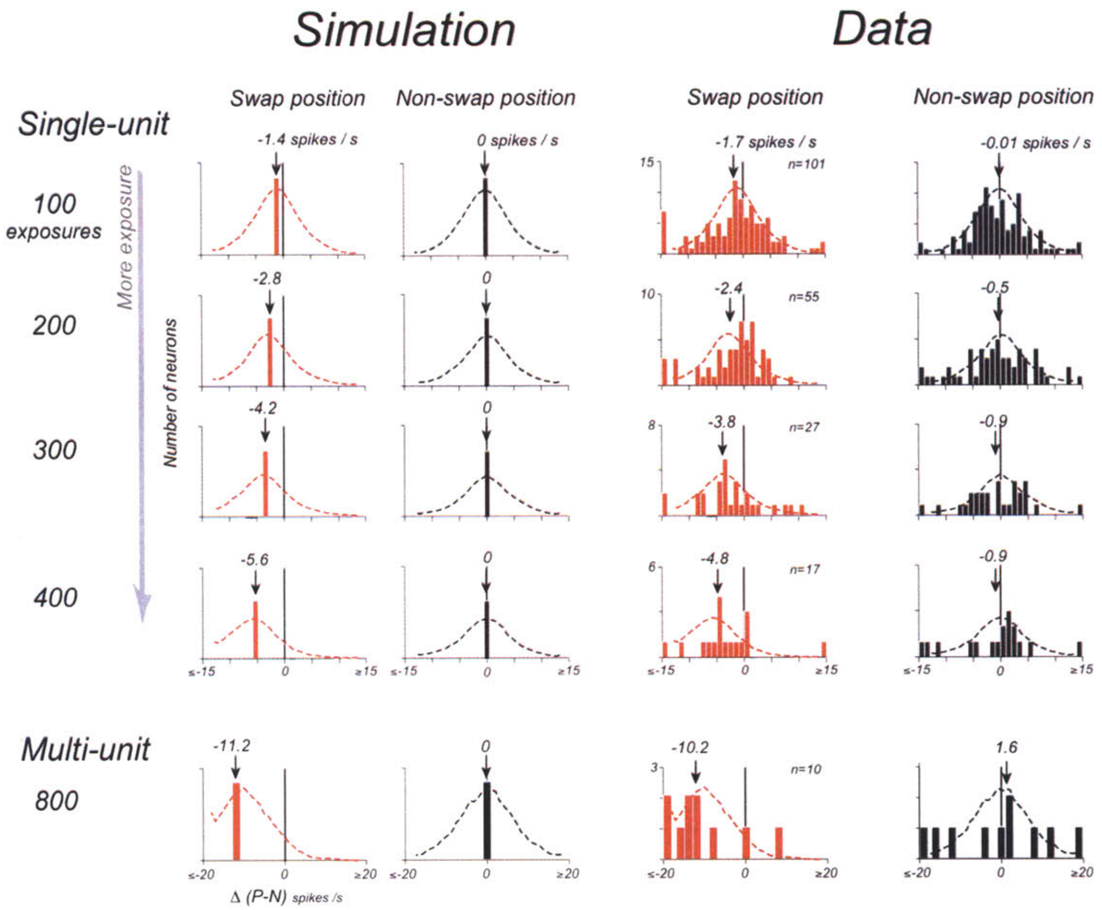


Fig. 3-S11. Simulations to compare the expectations of Poisson spiking variability predictions with our experimental observed variation in object selectivity. The left panels show the average effect size (change in object selectivity) at each time point estimated from our data at the swap position (red bars and arrows, based on 5.6 spikes/s per 400 exposures, see Fig. 3-3C) and at the non-swap position (black bars and arrows, assuming no-change, see Fig. 3-2A). The smooth curves show the averaged histogram expected under the assumption of Poisson spiking statistics (using 100 Monte Carlo runs assuming: the same distribution of mean firing rates as that observed in the recorded population, the same number of response repetitions as that collected in our experiments: 30 for single-unit, 50 multi-units). The right panel shows the data from the recorded population. (The red histograms for the single-unit data is a re-plot of Fig. 3-2C). Arrows indicate the mean object selectivity changes observed in the data. The over-laid dash lines are the distributions generated from the simulations (left), simply re-plotted on top of the data. Note that the dashed curves are quite broad (the effect of Poisson spiking "noise") and are approximately matched to the empirical distributions (solid histograms).

3. 7. 12 Supporting figure 3-S12

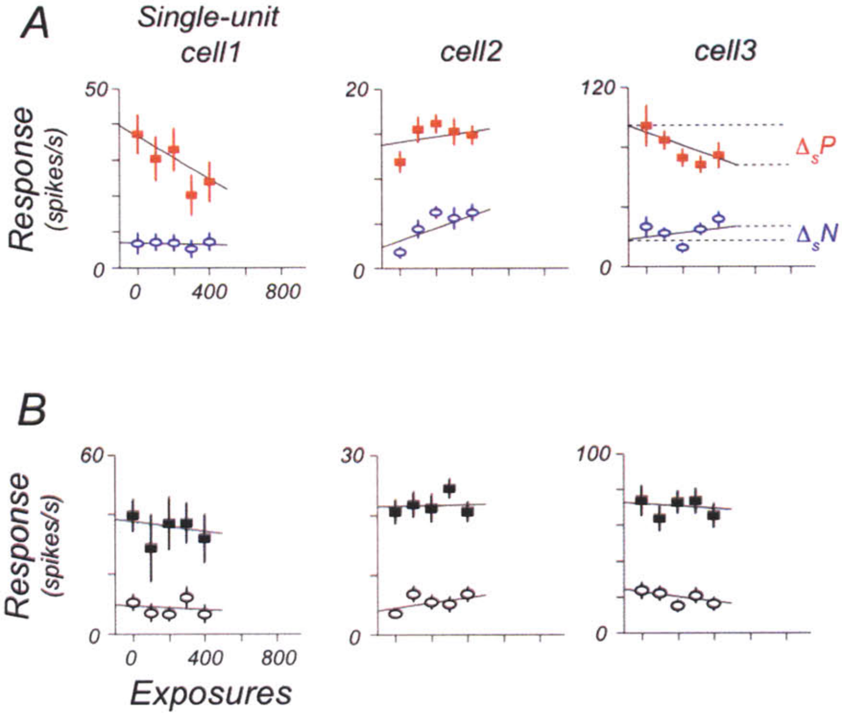


Fig. 3-S12. Responses to objects P and N from example neurons in Fig. 3-4A. (**A**) Response data to object P and N at the swap position (This is a re-plot of the neurons in Fig. 3-4A). (**B**) Response data from the non-swap position for these neurons. The object P and N are determined from the summed response across all positions using separate data before exposure. In both panels, the solid lines are the best-fit linear regression.

3. 8 Supporting tables

3. 8. 1 Supporting table 3-S1

Screen criteria based on responses at the center of gaze	Mean response rates under each screen (spikes/s)						Permutation Test for "position x exposure" interaction	t-Test swap vs. non-swap position
	Center of gaze		Swap position		Non-swap position			
	P	N	P	N	P	N		
Full population (no screen) n=101	20	14	21	16	20	16	*p = 0.007	*p = 0.005
Statistically selective (i.e. P>N at p<0.05) n=56	25	16	27	19	25	19	*p = 0.010	*p = 0.023

Statistically selective AND (P-N)>3 spikes/s n=48	29	18	29	20	27	20	*p = 0.016	*p = 0.041
Statistically selective AND (P-N)>5 spikes/s n=38	33	20	32	21	30	22	*p = 0.039	*p = 0.025

Table 3-S1. Summary of key statistical tests on different subsets of center-of-gaze-selective neurons. Left table columns show the mean responses of all the neurons in each subset to objects P and N. The selectivity screens were done after the "P" and "N" labels were determined using separate data to avoid bias (that is, a positive response difference implies real selectivity for P over N, not selection bias). The P-N differences are greater at the swap position for all groups of neurons because a few most foveal selective neurons (3/38) had much greater selectivity at the swap position. The right columns show the outcome of the two key statistical analyses. These results are robust to removal of outlier points (Fig 3-2C top panel, left-most bar). Though the overall mean effect size per neuron is larger for more selective neurons (i.e. going from the top to the bottom of the table, see Fig. 3-S5), the statistical significance level (p values) is slightly smaller (slightly larger p value) due to the drop in the number of neurons.

3. 8. 2 Supporting table 3-S2

	Difference in **(P-N) pre-exposure** Swap – Non-swap (spikes/s)	**Effect size** (spikes/s)	
		Swap position	**Non-Swap position**
Full population in Fig. 3-2, 3-3 (n=101)	2.05	-5.6 *	0.9 $^{\text{(n.s.)}}$
Matched (P-N) pre-exposure (n=84)	-0.01	-4.2 *	0.73 $^{\text{(n.s.)}}$

	Difference in **(P-N) pre-exposure** Swap – Non-swap (spikes/s)	**Effect size** (spikes/s)	
		Swap position	**Non-Swap position**
Object-selective neurons/sites in Fig. 3-4C (n=38)	2.70	-8.3 **	-1.7 $^{\text{(n.s.)}}$
Matched (P-N) pre-exposure (n=33)	-0.34	-6.5 **	-0.1 $^{\text{(n.s.)}}$

Table 3-S2. Summary of the key results after the initially greater selectivity at the swap position were adjusted for post-hoc. Effect size, Δ_S(P–N), was estimated for each neuron using regression analysis that included all the available data for each neuron (same as in Fig. 3-4C).

Δ_S(P–N) was significantly different from 0 only at the swap position (* $p<0.05$; ** $p<0.001$; one tailed t-Test).

3. 9 Supporting references and notes

1. J. J. DiCarlo, J. H. R. Maunsell, *Nat. Neurosci.* **3**, 814 (2000).
2. D. A. Robinson, *IEEE Transactions on Biomedical Engineering* **101**, 131 (1963).
3. K. Tanaka, *Cereb Cortex* **13**, 90 (2003).
4. G. Kreiman *et al., Neuron* **49**, 433 (2006).
5. D. Zoccolan, M. Kouh, T. Poggio, J. J. DiCarlo, *J. Neurosci.* **27**, 12292 (2007).
6. G. C. Baylis, E. T. Rolls, C. M. Leonard, *J. Neurosci.* **7**, 330 (1987).
7. J. J. DiCarlo, J. H. R. Maunsell, *J. Neurophysiol.* **89**, 3264 (2003).
8. B. Efron, R. J. Tibshirani, *An Introduction to the Bootstrap.* (Chapman & Hall, 2003)
9. C. I. Baker, M. Behrmann, C. R. Olson, *Nat. Neurosci.* **5**, 1210 (2002).
10. N. Sigala, N. K. Logothetis, *Nature* **415**, 318 (2002).
11. H. Sawamura, G. A. Orban, R. Vogels, *Neuron* **49**, 307 (2006).
12. E. K. Miller, L. Li, R. Desimone, *Science* **254**, 1377 (1991).
13. S. Sobotka, J. L. Ringo, *Exp. Brain. Res.* **96**, 28 (1993).
14. C. G. Gross, P. H. Schiller, C. Wells, G. L. Gerstein, *J. Neurophysiol.* **30**, 833 (1967).
15. G. C. Baylis, E. T. Rolls, *Exp. Brain. Res.* **65**, 614 (1987).
16. D. J. Tolhurst, J. A. Movshon, A. F. Dean, *Vision Res.* **23**, 775 (1983).
17. M. N. Shadlen, W. T. Newsome, *J. Neurosc.* **18**, 3870 (1998).

Chapter 4

Unsupervised Natural Visual Experience Rapidly Reshapes Size Invariant Object Representation in Inferior Temporal Cortex

4. 1 Abstract

We easily recognize objects and faces across a myriad of retinal images produced by each object. One hypothesis is that this tolerance (a.k.a. "invariance") is learned by relying on the fact that object identities are temporally stable. While we previously found neuronal evidence supporting this idea at the top of the non-human primate ventral visual stream (inferior temporal cortex, IT), we here test if this is a general tolerance learning mechanism. First, we found that the same type of unsupervised experience that reshaped IT position tolerance also predictably reshaped IT size tolerance, and the magnitude of reshaping was quantitatively similar. Second, this tolerance reshaping can be induced under naturally occurring dynamic visual experience, even without eye movements. Third, unsupervised temporal contiguous experience can build new neuronal tolerance. These results suggest that the ventral visual stream uses a general unsupervised tolerance learning (UTL) algorithm to build its invariant object representation.

4. 2 Introduction

Our ability to recognize objects and faces is remarkably tolerant to variation in the retinal images produced by each object. That is, we can easily recognize each object even though it can appear at different positions, sizes, poses, etc. In the primate brain, the solution to this "invariance" problem is thought to be achieved through a series of transformations along the ventral visual stream. At the highest stage of this stream, the inferior temporal cortex (IT), a tolerant object representation is obtained in which individual IT neurons have a preference for some objects ("selectivity") over others and that rank-order preference is largely maintained across identity preserving image transformations (Ito et al., 1995; Logothetis and Sheinberg, 1996; Tanaka, 1996; Vogels and Orban, 1996). Though most IT neurons are not strictly "invariant" (DiCarlo and Maunsell, 2003; Ito et al., 1995; Logothetis and Sheinberg, 1996; Vogels and Orban, 1996), reasonably sized populations of these so-called "tolerant" neurons can support object recognition tasks (Afraz et al., 2006; Hung et al., 2005; Li et al., 2009). However, we do not yet understand how IT neurons construct this tolerant response phenomenology.

One potentially powerful idea is that time can act as an implicit teacher, in that the temporal contiguity of object features during natural visual experience can instruct the learning of tolerance, potentially in an unsupervised manner (Foldiak, 1991; Masquelier et al., 2007; Masquelier and Thorpe, 2007; Sprekeler et al., 2007; Stryker, 1991; Wiskott and Sejnowski, 2002; Wyss et al., 2006). The overarching logic is as follows: during natural visual experience, objects tend to remain present for seconds or more, while object motion or viewer motion (e.g. eye movements) tend to cause rapid changes in the retinal image cast by each object over shorter time intervals (hundreds of ms). In theory, the ventral stream could construct a tolerant object representation by taking advantage of this natural tendency for temporally contiguous retinal images to belong to the same object, thus yielding tolerant object selectivity in IT cortex. A recent experimental result in adult non-human primate IT has provided some neuronal support for this temporal contiguity hypothesis (Li and DiCarlo, 2008). Specifically, we found that alterations of unsupervised experience of temporally contiguous object image changes across saccadic eye movements can induce rapid reshaping (within hours) of IT neuronal position tolerance (i.e. a reshaping of each IT neuron's ability to respond with consistent object selectivity across the retina). This IT neuronal learning likely has perceptual consequences because similar temporal contiguity manipulations of eye-movement-driven position experience can produce qualitatively similar changes in the position tolerance of human object perception (Cox et al., 2005).

However, these previous studies have two key limitations. First, they only uncovered evidence for temporal contiguity learning under a very restricted set of conditions: they showed learning effects only in the context of eye movements, and they only tested one type of tolerance -- position tolerance. Because eye movements drive a great deal of the image statistics relevant only to position tolerance (temporally-contiguous image translations), the previous results could reflect only a special case of tolerance learning. Second, the previous studies did not directly show that temporally-contiguous image statistics can *build* new tolerance, but only showed that alterations of those statistics can disrupt normal tolerance. Because of these limitations, we do not know if the naive ventral stream uses a general, temporal contiguity driven, learning mechanism to construct its tolerance to all types of image variation.

Here, we set out to test the temporal contiguity hypothesis in three new ways. First, we reasoned that, if the ventral stream is using temporal contiguity to drive a general tolerance-building mechanism, alterations in that temporal contiguity should reshape other types of tolerance (e.g. size tolerance, pose tolerance, illumination tolerance), and the magnitude of that reshaping should be similar to that found for position tolerance. We decided to test size tolerance, because normal size tolerance in IT is much better described (Brincat and Connor, 2004; Ito et al., 1995; Logothetis and Sheinberg, 1996; Vogels and Orban, 1996) than pose or illumination tolerance. Our experimental logic follows our previous work on position tolerance (Cox et al., 2005; Li and DiCarlo, 2008). Specifically, when an adult animal with a mature (e.g. size-tolerant) object representation is exposed to an altered visual world in which object identity is consistently swapped across object size change, its visual system should learn from those image statistics such that it predictably "breaks" the size tolerance of that mature object representation. Assuming IT conveys this object representation (Afraz et al., 2006; Hung et al., 2005; Logothetis and Sheinberg, 1996; Tanaka, 1996), that learning should result in a specific change in the size tolerance of mature IT neurons (Figure 4-1).

Second, many types of identity-preserving image transformations in natural vision do not involve intervening eye movements (e.g. object motion producing a change in object image size). If the ventral stream is using a general tolerance-building mechanism, we should be able to find size tolerance reshaping even without intervening eye movements, and we should also be able to find size tolerance reshaping when the dynamics of the image statistics mimic naturally-occurring image dynamics.

Third, our previous studies (Cox et al., 2005; Li and DiCarlo, 2008) and our first two aims above use the "breaking" of naturally-occurring image statistics to try to "break" the normal tolerance observed in IT (Figure 4-1, i.e. to weaken existing IT object selectivity in a position- or size-specific manner). Such results support the inference that naturally occurring image statistics instruct the "building" of that tolerance in the naive ventral stream. However, we also sought to test that inference more directly by looking for evidence that temporally contiguous image statistics can *build* new tolerance in IT neurons with immature tolerance (i.e. can produce an *increase* in existing IT object selectivity in a position- or size-specific manner).

Our results showed that targeted alterations in the temporal contiguity of visual experience robustly and predictably reshaped IT neuronal size tolerance over a period of hours. This change in size tolerance grew gradually stronger with increasing visual experience and the rate of reshaping was very similar to previously reported position tolerance reshaping (Li and DiCarlo, 2008). Second, we found that the size tolerance reshaping occurred without eye movements, and it occurred when the dynamics of the image statistics mimicked naturally-occurring dynamics. Third, we found that exposure to "broken" temporal contiguity image statistics could weaken and even reverse the previously normal IT object selectivity at a specific position or size (i.e. "break" old correct tolerance and "build" new "incorrect" tolerance), and that naturally occurring temporal contiguity image statistics could build new, correct position or size tolerance. Taken together with previous work, these results argue that the ventral stream uses unsupervised, natural visual experience and a common learning mechanism (a.k.a. "unsupervised temporal tolerance learning", UTL) to build and maintain its tolerant ("invariant") object representation.

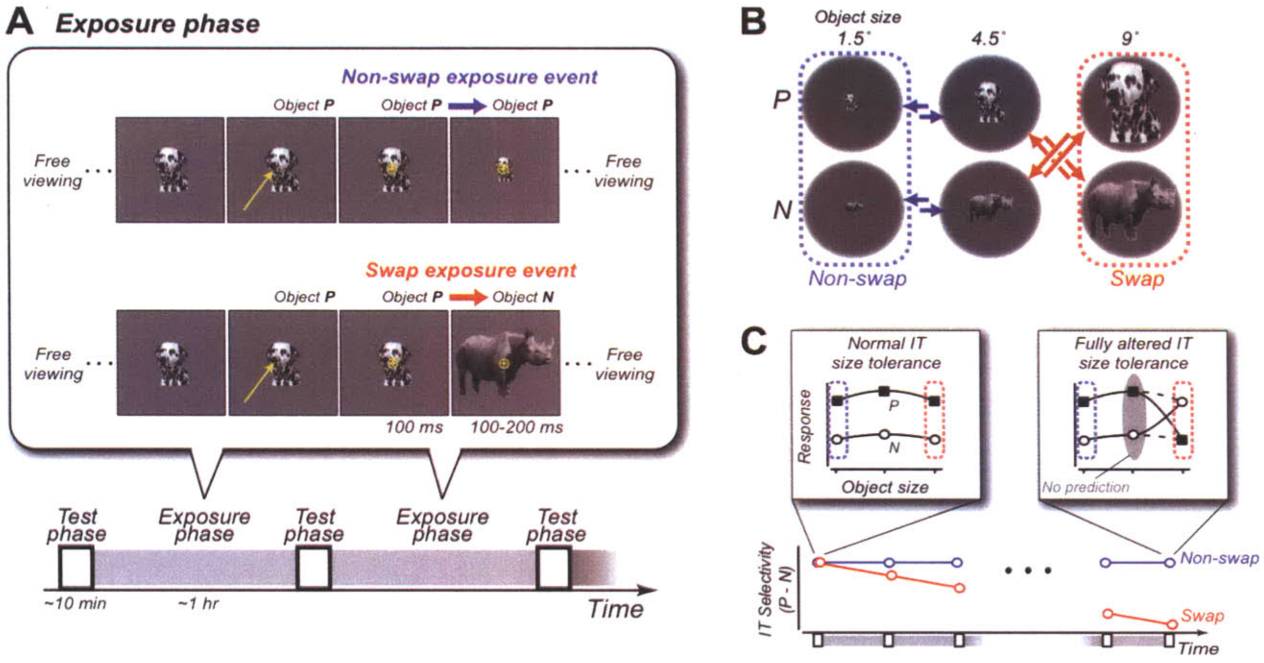


Figure 4-1. Experimental Design and Prediction.

(A) IT selectivity was tested in the *Test Phases* while animals received experience in the altered visual world in the *Exposure Phases*

(B) The chart shows the full exposure design for a single IT site in Experiment I. Arrows show the temporal contiguity experience of retinal images (arrow heads point to the retinal images occurring later in time, e.g. panel A). Each arrow shows a particular exposure event type (i.e. temporally-linked images shown to the animal), and all 8 exposure event types were shown equally often (randomly interleaved) in each *Exposure Phase*.

(C) Prediction for IT responses collected in the *Test Phase*: If the visual system builds size tolerance using temporal contiguity, the swap exposure should cause incorrect grouping of two different object images (P and N). The qualitative prediction is a decrease in object selectivity at the swap size (images and data points outlined in red) that grows stronger with increasing exposure (in the limit, reversing object preference as illustrated schematically here), and little or no change in object selectivity at the non-swap size. The experiment makes no quantitative prediction for the selectivity at the medium size (gray oval, see text).

4. 3 Results

In three separate experiments (Experiments I, II, III), two unsupervised non-human primates (Rhesus Macaque) were exposed to altered visual worlds in which we manipulated the temporal contiguity statistics of the animals' visual experience with object size (Figure 4-1A, *Exposure Phases*). In each experiment, we recorded multi-unit activity in an unbiased sample of recording sites in the anterior region of IT to monitor any experience-induced change (Figure 4-1A, *Test Phases*). Specifically, for each IT site, a preferred object (P) and a less preferred object (N) were chosen based on testing of a set of 96 objects (Figure 4-1B). We then measured the baseline IT neuronal selectivity for P and N at three retinal sizes (1.5°, 4.5°, and 9°) in a *Test Phase* (~10 min) by presenting the object images in a rapid, but naturally-paced sequence (5 images/sec) on the animals' center of gaze. For all the results below, we report selectivity values determined from these *Test Phases*, which we conducted both before and after experience manipulations. Thus, all response data shown in the results below were collected during orthogonal behavioral tasks in which object identity and size were irrelevant (Supplemental Experimental Procedures).

Consistent with previous reports (Kreiman et al., 2006), the initial *Test Phase* data showed that each IT site tended to maintain its preference for object P over object N at each size tested here (Figure 4-3 and Supplemental Figure 4-S3). That is, most IT sites showed good, baseline size tolerance. Following the logic outlined in the Introduction, the goal of Experiments I-III was to determine if consistently-applied unsupervised experience manipulations would predictably reshape that baseline size tolerance of each IT site (see Figure 4-1 for the basic prediction). In particular, we monitored changes in each IT site's preference for object P over N at each of the three objects sizes, and any change in that selectivity following experience that was not seen in control conditions was taken as evidence for an experience-induced reshaping of IT size tolerance.

In each Experiment (I-III), the key experience manipulation was deployed in one or more *Exposure Phases* which were all under precise, automated computer-display control to implement spatiotemporally reliable experience manipulations (see Methods). Specifically, during each *Exposure Phase* the animals freely viewed a gray display monitor on which images of object P or N intermittently appeared at a randomly-chosen retinal positions away from the center of gaze (object size: 1.5°, 4.5°, or 9°). The animals almost always looked to foveate each object (>95% of object appearances) within ~124 ms (mean; median, 109 ms), placing the object image on the center of gaze. Following that object acquisition saccade, we reliably manipulated the visual experience of the animals over the next 200-300 ms. The details of the experience manipulation (i.e. which object sizes where shown and the timing of those object images) were different in the three experiments, but all three experiments used the same basic logic outlined in the Introduction and Figure 4-1.

4. 3. 1 Experiment I: Does Unsupervised Visual Experience Reshape IT Size Tolerance?

In Experiment I, following the object acquisition saccade, we left the newly-foveated object image unchanged for 100 ms, and then we changed the size of the object image (while its retinal position remained on the animal's center of gaze) for the next 100 ms (Figure 4-1A). We reasoned that this creates a temporal experience linkage ("exposure event") between one object image at one size and another object image at another size. Importantly, on half of the exposure events, one object was swapped out for the other object: for example, a medium-sized (4.5°) object P would become a big (9°) object N (Figure 4-1A, "swap exposure event"). As one key control, we also exposed the animal to more normal exposure events in which object identity did not change during the size change (Figure 4-1A, "non-swap exposure event"). The full exposure design for one IT site is shown in Figure 4-1B, the animal received 800-1600 swap exposures within the time period of 2-3 hours. Each day, we made continuous recordings from a single IT site, and we always deployed the swap exposure at a particular object size (either 1.5° or 9°, i.e. swap size) while keeping the other size as a control (i.e. non-swap size). Across different IT sites (i.e. different recordings days), we strictly alternated the object size at which swap manipulation took place so that object size was counter-balanced across our recorded IT population (n= 27).

Unsupervised temporal tolerance leaning (UTL) theory makes the qualitative prediction that the altered experience will induce a size-specific confusion of object identity in the IT response as the ventral stream learns to associate the temporally-linked images. In particular, our exposure design should cause the IT site to reduce its original selectivity for images of object P and N at the swap size (perhaps even reversing that selectivity in the limit of large amounts of experience, Figure 4-1C, red). UTL is not currently specific enough to make a quantitative prediction of what this altered experience should do for selectivity among the medium object size images because those images were temporally-paired in two ways: with images at the swap size (altered visual experience) and with the images at the non-swap size (normal visual experience). Thus, our key experimental prediction and planned comparison is between the selectivity (P vs. N) at the swap and non-swap size: we predict a selectivity decrease at the swap size that should be much larger than any selectivity change at the non-swap object size (Figure 4-1C, blue).

This key prediction was born out by the data: as the animals received experience in the altered visual world, IT selectivity among objects P and N began to decrease at the swap size, but not at the control size. This change in selectivity grew stronger with increasing experience over the time course of 2-3 hours (Figure 4-2A). To quantify the selectivity change, for each IT site, we took the difference between the selectivity (P-N, response difference in units of spikes/s, see Experimental Procedures) in the first (pre-exposure) and last *Test Phase*. This Δ(P-N) sought to quantify the total amount of selectivity change for each IT site induced by our experience manipulation. On average, there was a significant decrease in selectivity at the swap size (Figure 4-2B, $p<0.0001$, two-tailed t-test against 0) and no significant change at the non-swap control size (Figure 4-2B, $p=0.89$). Incidentally, we also observed a significant decrease in selectivity at the medium size ($p=0.002$). This is not surprising given the images at the medium object size was exposed to the altered statistics half of the time when it was temporally paired with the images at the swap size. Because no prediction was made about the selectivity change at the medium size, we below concentrate on the planned comparison between the swap and non-swap size. We statistically confirmed the size specificity of the experience-induced decrease in selectivity by two different approaches: 1) a direct t-test on the Δ(P-N) between the swap and non-swap size ($p<0.001$, two-tailed); 2) a significant interaction of "exposure x object size" on the raw selectivity measurements (P-N) -- that is, IT selectivity was decreased by exposure only at the swap size ($p=0.0018$, repeated measures ANOVA; $p=0.006$, bootstrap, see Supplemental Experimental Procedures).

To ask if the experience induced selectivity change was specific to the manipulated objects or the features contained in those objects, we also tested each IT site's responses to a second pair of objects (P′ and N′, control objects; see Experimental Procedures). Images of these control objects at three sizes were tested together with the swap objects during all *Test Phases* (randomly interleaved), but they were not shown during the *Exposure Phase*. On average, we observed no change in IT selectivity among these un-exposed control objects (Supplemental Figure 4-S4). This shows that that the experience-induced reshaping of IT size tolerance has at least some specificity for the experienced objects or the features contained in those objects.

We next set out to quantify the amount of IT size tolerance reshaping induced by the altered visual experience. Because each IT site was tested for different amounts of exposure time (due to experimental time constraints), we wanted to control for this and still leverage all the data for each site to gain maximal power. To do so, we fit linear regressions to the (P-N) selectivity of individual sites at each object size (Figure 4-2C, insert). The slope of the line fit, which we will refer to as Δs(P-N), provided us with a sensitive, unbiased measure of the amount of selectivity change that normalizes the amount of exposure experience. The Δs(P-N) for the swap size and non-swap size is shown in Figure 4-2C and 4-2D, which qualitatively confirmed the result obtained in Figure 4-2B (using the simple measure of selectivity change), and showed a mean selectivity change of -9.2 spikes/s for every 800 swap exposure events.

Importantly, we note that this reshaping of IT tolerance was induced by unsupervised exposure to temporal-linked images that did not include a saccadic eye movement to make that link (Figure 4-1A). We also considered the possibility that small intervening microsaccades might still have been present, but found that they cannot account for the reshaping (Supplemental Figure 4-S7). The size specificity of the selectivity change also rules out alternative explanations such as adaptation, which would not predict this specificity (because our exposure design equated the amount of exposure for both the swap and non-swap size). We also found the same amount of tolerance reshaping when the sites were grouped by the physical object size at which we deployed the swap (1.5° vs. 9°, p=0.26, t-test). Thus the learning is independent of low-level factors like the total luminance of the swapped objects. In sum, we found that unsupervised, temporally-linked experience with object images across object size change can reshape IT size tolerance.

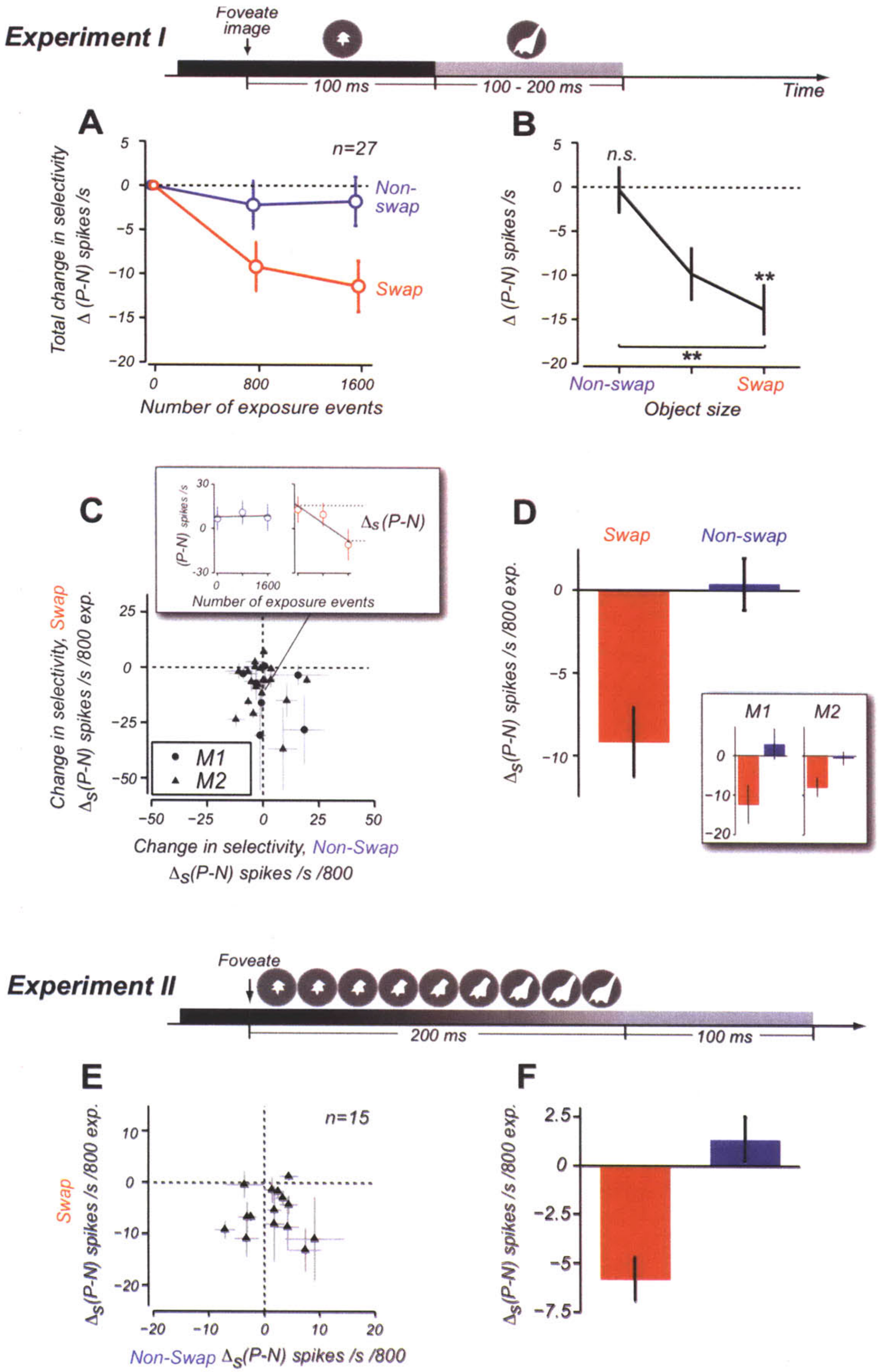

Experiment I
Foveate image
100 ms
100 - 200 ms
Time
A
n=27
Total change in selectivity
Δ (P-N) spikes /s
5
0
-5
-10
-15
-20
Non-swap
Swap
0
800
1600
Number of exposure events
B
n.s.
**
**
Δ (P-N) spikes /s
Non-swap
Swap
Object size
C
(P-N) spikes /s
30
0
-30
0
1600
Δs (P-N)
Number of exposure events
Change in selectivity, Swap
Δs(P-N) spikes /s /800 exp.
25
0
-25
-50
● M1
▲ M2
-50
-25
0
25
50
Change in selectivity, Non-Swap
Δs(P-N) spikes /s /800
D
Swap
Non-swap
Δs(P-N) spikes /s /800 exp.
0
-5
-10
M1
M2
0
-10
-20
Experiment II
Foveate
200 ms
100 ms
E
n=15
Swap
Δs(P-N) spikes /s /800 exp.
10
0
-10
-20
-20
-10
0
10
20
Non-Swap Δs(P-N) spikes /s /800
F
Δs(P-N) spikes /s /800 exp.
2.5
0
-2.5
-5
-7.5

Figure 4-2. Experimental I and II Key Results.

(A) Mean ± SEM IT object selectivity change, Δ(P-N), from the first *Test Phase* as a function of the number of exposure events. Each data point shows the average across all the sites tested for that particular amount of experience (n=27, 800 exposure events; n=22, 1600 exposure events).

(B) Mean ± SEM selectivity change at the swap, non-swap, and medium size (4.5°). For each IT site (n=27), total Δ(P-N) was computed using the data from the first and last *Test Phase*, excluding any middle *Test Phase* data. Hence, not all data from (A) were included. * $p<0.05$ by two tailed t-test; ** $p<0.01$; n.s. $p>0.05$.

(C) For each IT site (n=27), we fit a line (linear regression) to the (P-N) data as a function of the number of exposure events (insert). We used the slope of the line fit, Δs(P-N), to quantify the selectivity change. The Δs(P-N) is a measure that leverages all our data while normalizing out the variable of exposure amount (for sites with only two *Test Phases*, Δs(P-N) equals Δ(P-N)). Δs(P-N) was normalized to show selectivity change per 800 exposure events. Error bars indicate the standard error of the procedure to compute selectivity (Supplemental Experimental Procedures). M1, monkey 1; M2, monkey 2.

(D) Mean Δs(P-N) at the swap and non-swap size (n=27 IT sites; M1: 7, M2: 20). Error bars indicate SEM over neuronal sites.

(E) Change in selectivity, Δs(P-N), of all IT sites from Experiment II at the swap and non-swap size.

(F) Mean ± SEM Δs(P-N) at the swap and non-swap size.

4. 3. 2 Experiment II: Does Size Tolerance Learning Generalize to the "Natural" Visual World?

In the natural world, objects tend to undergo size change smoothly on our retina as result of object motion or viewer motion, but, in Experiment I (above), the object size changes we deployed were discontinuous: one image of an object was immediately replaced by an image of another object with no smooth transition (Figure 4-2, top). Therefore, although those results show that unsupervised experience with object images at different sizes, linked in time could induce the predicted IT selectivity change, we wanted to know if that learning was also found during exposure to more natural (i.e. temporally-smooth) image dynamics.

To answer this question, we carried out a second experiment (Experiment II) in which we deployed essentially the same manipulation as Experiment I (object identity changes during object size changes, no intervening eye movement), but with natural (i.e. smooth-varying) stimulus sequences. The dynamics in these movie stimuli were closely modeled after the kind of dynamics that our visual system encounters daily in the natural environment (Supplemental Figure 4-S2). To create smooth-varying object identity changes over object size changes, we created morph lines between pairs of objects we swapped in Experiment I (P and N). This allowed us to parametrically transform the shape of the objects (Figure 4-2, bottom). All other experimental procedures were identical to Experiment I except, in the *Exposure Phases*, objects underwent size change smoothly while changing identity ("swap exposure") or preserving identity ("non-swap exposure", Supplemental Figure 4-S2).

When we carried out this temporally-smooth experience manipulation on a new population of IT sites (n=15), we replicated the Experiment I results (Figure 4-2E and F): there was a predicted decrease in IT selectivity at the swap size and not at the non-swap control size. This size specificity of the effect was, again, confirmed statistically by: 1) direct t-test on the total selectivity change, Δ(P-N), between the swap and non-swap size (Δ(P-N)= -10.3 spikes/s at swap size, +2.8 at non-swap size; $p<0.0001$, two-tailed t-test); 2) a significant interaction of "exposure x object size" on the raw selectivity measurements (P-N) ($p<0.001$, repeated measures ANOVA; $p=0.001$, bootstrap). This result suggests that image linking across time is sufficient to induce tolerance learning in IT and is robust to the temporal details of that image linking (at least over the ~200 ms time windows of linking used here). More importantly, Experiment II shows that unsupervised size tolerance learning occurs in a spatiotemporal image regime encountered in real-world vision.

4. 3. 3 Size Tolerance Learning: Observations and Effect Size Comparison

Despite a wide diversity in the initial tuning of the recorded IT multi-unit sites, our experience manipulation induced a predictable selectivity change that was large enough to be observed in individual IT sites: 40% (17/42 sites, Experiment I and II data combined) of the individual IT sites showed a significant selectivity decrease at the swap size within a single recording session

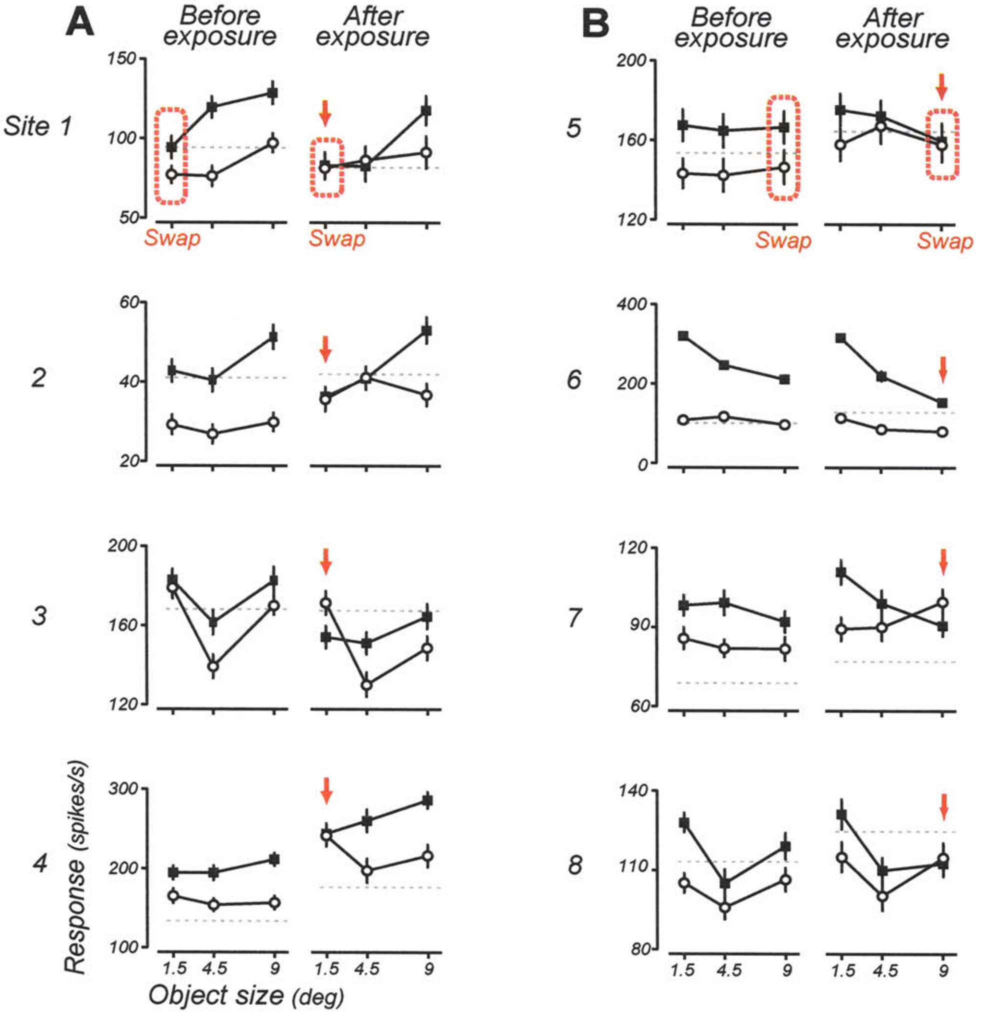


Figure 4-3. Example Single IT Sites.

Mean ± SEM IT response to P (solid square) and N (open circle) as a function of object size for eight example IT sites (from both Experiment I and II). The data shown are from the first ("before exposure") and last *Test Phase* ("after exposure"). (A) swap size, 1.5°; (B) swap size, 9° (highlighted by red boxes and arrows). Gray dotted lines show the baseline response to a blank image (interleaved with the test images).

(only 7% of sites showed significant selectivity decrease at the non-swap size, which is essentially the fraction expected by chance; 3/42 sites, p<0.05, permutation test, see Supplemental Experimental Procedures). Eight example sites are shown in Figure 4-3.

We found that the magnitude of size-tolerance reshaping depended on the initial selectivity at the medium object size, 4.5° (Pearson correlation, r= 0.54, p<0.01). That is, on average, IT sites that we initially encountered with greater object selectivity at the medium size underwent greater exposure-induced selectivity change at the swap size. This correlation is not simply explained by the hypothesis that it is easier to "break" highly-selective neurons (e.g. due to factors that might have nothing to do with neuronal learning, such as loss of isolation, etc.), because the correlation was not seen for changes in selectivity at the non-swapped size (r= -0.16, p= 0.35) and we found no average change in selectivity at the non-swapped size (Figure 4-2 and statistics above). Instead, this observation is consistent with the overarching hypothesis of this study: the initial image selectivity at the medium object size provides (at least part of) the driving force for selectivity learning because those images are temporally-linked with the swapped images at the swap size.

The change in selectivity produced by the experience manipulation was found throughout the entire time period of the IT response, including the earliest part of that period where IT neurons are just beginning to respond above baseline (~100 ms from stimulus onset, Supplemental Figure 4-S5). This shows the experience-induced change in IT selectivity cannot be explained by changes in long lag feedback alone (>100 ms; also see Discussion). On average, the selectivity change at the swap size resulted from both a decrease in the response to the image of the preferred object (P), and an increase in the response to the less preferred object (N). Consistent with this, we found that the experience manipulation produced no average change in the IT sites' mean response rate (Supplemental Figure 4-S5).

In this study, we concentrated on multi-unit response data because it had a clear advantage as a direct test of our hypothesis -- it allowed us to longitudinally track IT selectivity during altered visual experience across the entirety of each experimental session. We also examined the underlying single-unit data and found results that were consistent with the multi-unit data. Figure 4-4A shows an example of a rare single-unit IT neuronal recording that we were able to track across an entire recording session (~3 hr). The confidence that we were recording from the same unit comes from the consistency of the unit's waveform and its consistent pattern of response among the non-exposed control object images (Figure 4-4B). During this stable recording, the (P-N) selectivity at the swap size gradually decreased while the selectivity at the non-swap size remained stable, perfectly mirroring the multi-unit results described above.

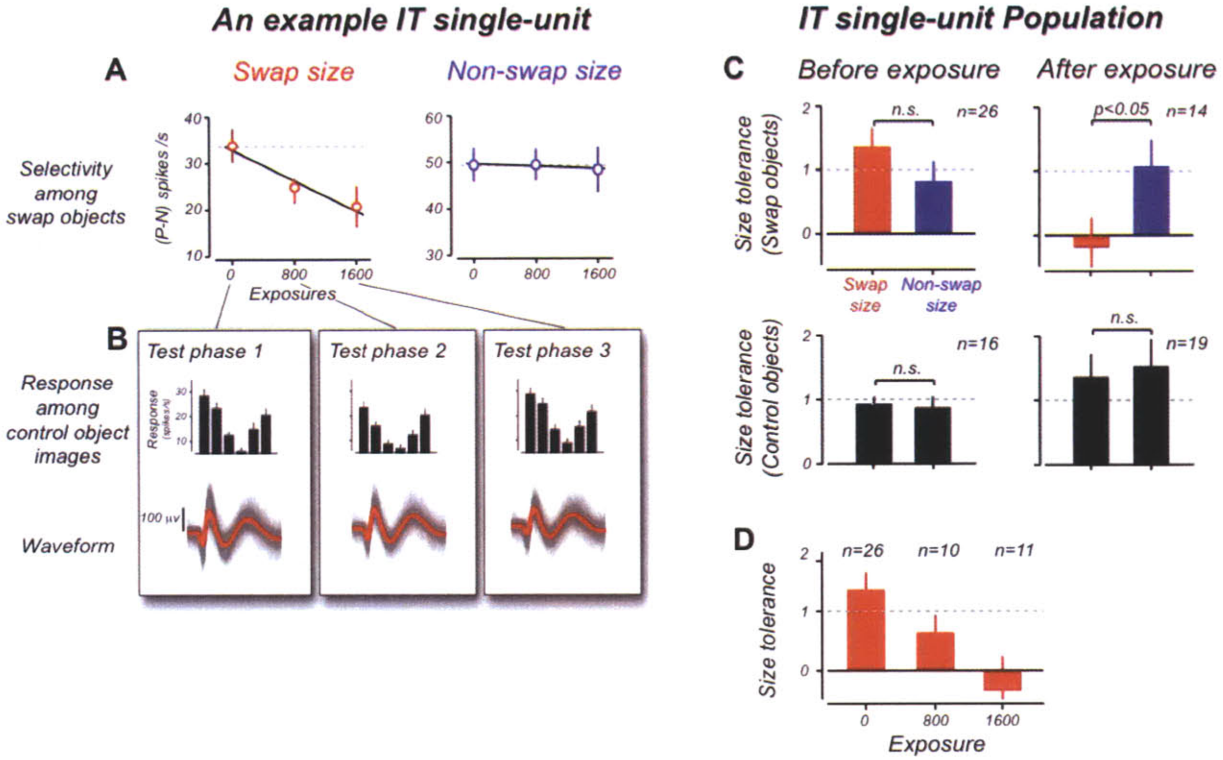


Figure 4-4. Single-unit Results.

(A) P vs. N selectivity of a rare single-unit IT neuron that was isolated across an entire recording session (~3 hr).

(B) The example single-unit's response to the six control object images during each *Test Phase* and its waveforms (gray: all traces from a *Test Phase*; red: mean).

(C) Mean ± SEM size tolerance at the swap (red) and non-swap (blue) size for single-units obtained before and after exposure. Size tolerance for the control objects is also shown at these two sizes (black). Each neuron's size tolerance was computed as $(P\text{-}N)/(P\text{-}N)_{medium}$, where (P-N) is the selectivity at the tested size and $(P\text{-}N)_{medium}$ is the selectivity at the medium object size. Only units that showed selectivity at the medium size were included ($(P\text{-}N)_{medium}>1$ spikes/s). The top and bottom panels include neurons that had selectivity for the swap objects, the control objects, or both. Thus they show different but overlapping populations of neurons. The result is unchanged if we only examine populations for which each neuron has selectivity for both the swap and control objects (i.e. the intersections of the neuronal populations in top and bottom panels, Supplemental Figure 4-S6).

(D) Mean ± SEM size tolerance at the swap size further broken out by the amount of exposure to the altered visual statistics. To quantify the change in IT size tolerance, we performed linear regression of the size tolerance as a function of the amount of experience. Consistent with the multi-unit results, we found a significant negative slope (Δ size tolerance = -0.84 per 800 exposure; p=0.002, bootstrap; c.f. -0.42 for multi-unit, Supplemental Figure 4-S6). No decrease in size tolerance was observed at the non-swap control size (Δ size tolerance = 0.30; c.f. 0.12 for multi-unit).

However these ~3 hr single-unit recordings were very rare because single-units have limited hold-time in the awake primate physiology preparation. Thus we took a more standard population approach to analyze the single-unit data (Baker et al., 2002; Kobatake et al., 1998; Sakai and Miyashita, 1991; Sigala et al., 2002). Specifically, we performed spike-sorting analyses to obtain clear single-units from each *Test Phase* (Experimental Procedures). We considered each single-unit obtained from each *Test Phase* as a sample of the IT population, taken either before or after the experience in the altered visual world. This analysis does not require that the sampled units were the same neurons. The prediction is that IT single-units sampled after exposure (i.e. at the last *Test Phase* of each day) would be less size tolerant at the swap size than at the non-swap size. This prediction was clearly observed in our single-unit data (Figure 4-4C, after exposure, p<0.05; for reference, the size tolerance before the exposure is also shown and we observed no difference between the swap and non-swap size). The result was robust to the choice of the criteria to define "single-units" (Supplemental Figure 4-S6). Similarly, we found that each single-unit population sampled after successively more exposure showed a successively larger change in size tolerance (Figure 4-4D).

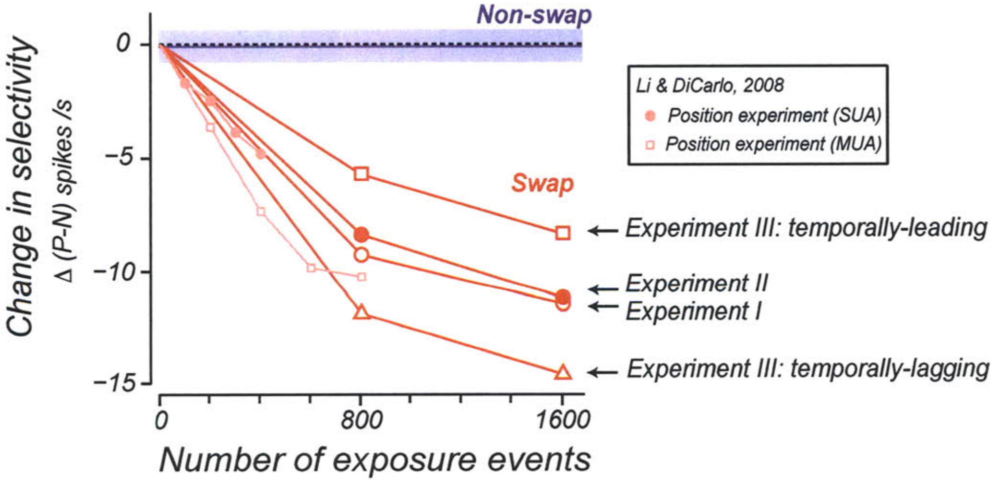


Figure 4-5. Effect Size Comparisons across Different Experience Manipulations.

Mean object selectivity change as a function of the number of swap exposure events for different experiments. For comparison, the data from a position tolerance learning experiment (Li and DiCarlo, 2008) are also shown. Plot format is the same as Figure 4-2A without the error bars. Mean ± SEM Δ(P-N) at the non-swap size/position is shown in blue (all experiments pooled). SUA, single-unit activity; MUA, multi-unit activity.

We next aimed to quantify the absolute magnitude of this size tolerance learning effect across the different experience manipulations deployed here, and to compare that magnitude with our previous results on position-tolerance learning (Li and DiCarlo, 2008). To do this, we plotted the mean selectivity change at the swap size from each experiment as a function of number of swap exposures (Figure 4-5). We found that Experiments I and II produced a very similar magnitude of learning: ~5 spikes/s per 400 swap exposures (also see Discussion for comparison to previous work). This effect grew larger at this approximately constant rate for as long as we could run each experiment, and the magnitude of the size tolerance learning was remarkably similar to that seen in our previous study of position tolerance (Li and DiCarlo, 2008).

4. 3. 4 Size and Position Tolerance Learning: Reversing Old IT Object Selectivity and Building New IT Object Selectivity

The results presented above on size tolerance and our previous study of position tolerance (Li and DiCarlo, 2008) both used the "breaking" of naturally-occurring temporal contiguity experience to discover that we can "break" normal position tolerance and size tolerance (i.e. cause a decrease in adult IT object selectivity in a size- or position-specific manner). While these results are consistent with the inference that naturally-occurring image statistics instruct the original "building" of that normal tolerance (see Introduction), we next sought to test that inference more directly. Specifically, we asked if the temporal contiguity statistics of visual experience can instruct the creation of new IT tolerance (i.e. cause an *increase* in IT object selectivity in a size- or position-specific manner)? Our experimental data offered two ways to test this idea (below) and both revealed that unsupervised temporal contiguity learning could indeed build new IT tolerance. To do these analyses, we took advantage of the fact that we found very similar effects for both size tolerance and position tolerance (Li and DiCarlo, 2008), and we maximized our power by pooling the data across this experiment (Figure 4-5: size experiment I, II, n=42 MUA sites) and our previous position experiment (n=10 MUA sites). This pooling did not qualitatively change the result -- the effects shown in Fig. 4-5 and 4-6 below were seen in the size tolerance data alone (Supplemental Figure 4-S9).

First, as outlined in Figure 4-1C, a strong form of the unsupervised temporal tolerance learning

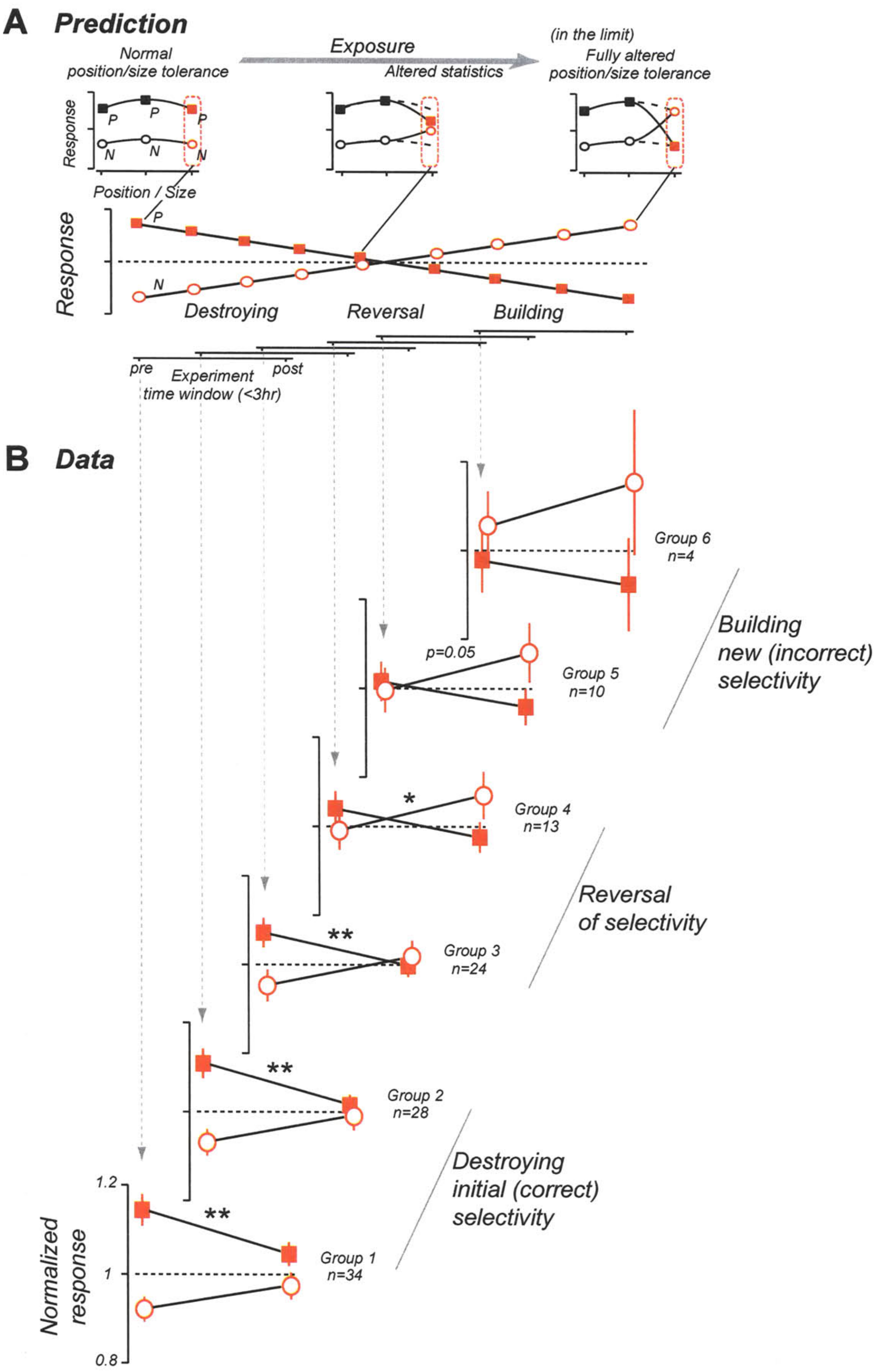

A
Prediction
Normal
position/size tolerance
Exposure
Altered statistics
(in the limit)
Fully altered
position/size tolerance
Response
P
N
Position / Size
Response
P
N
Destroying
Reversal
Building
pre
Experiment
time window (<3hr)
post
B
Data
Group 6
n=4
Building
new (incorrect)
selectivity
p=0.05
Group 5
n=10
*
Group 4
n=13
Reversal
of selectivity
**
Group 3
n=24
**
Group 2
n=28
Destroying
initial (correct)
selectivity
**
Group 1
n=34
1.2
1
0.8
Normalized
response

Figure 4-6. Altered Statistics in Visual Experience Builds Incorrect Selectivity.

(A) Prediction: *top,* most adult IT neurons start with fully position/size tolerant selectivity (left). In the limit of a large amount of altered visual experience, temporal contiguity learning predicts that each neuron will acquire fully altered tolerance (right). *Bottom,* at the swap position/size (red), the selectivity for P over N is predicted to reverse in the limit (prefer N over P). Because we could only record longitudinally from a multi-unit site for less than 3 hours, we do not expect our experience manipulation within a session to produce the full selectivity reversal (pre vs. post) among neuronal sites with strong initial selectivity. However, because different IT sites differ in their degrees of initial selectivity, they start at different distances from selectivity reversal. Thus, our manipulation should produce selectivity reversal among the initially weakly selective sites and build new ("incorrect") selectivity.

(B) Mean ± SEM normalized response to object P and N at the swap position/size among sub-populations of IT multi-unit sites. Sites are grouped by their initial selectivity at the swap position/size using independent data. Data from the size and position tolerance experiments (Li & DiCarlo, 2008) were combined to gain maximal power (size experiment I, II, position experiment, see Supplemental Experimental Procedures). These sites show strong selectivity at the non-swap (control) position/size and no negative change in that selectivity was observed (not shown). ** p<0.01; * p<0.05, one tailed t-test against no change. (Size experiment data only, group 1-6: p<0.01; p<0.01; p<0.01; p=0.02; p=0.07; n.s.).

(UTL) hypothesis predicts that our experience manipulation should not only degrade existing IT selectivity for P over N at the swap size/position, but should eventually reverse that selectivity and then build new "incorrect" selectivity for N over P (Figure 4-1C, note we refer to this as "incorrect" selectivity because the full IT response pattern is inappropriate for the veridical world in which objects maintain their identity across changes in position and size). While the plasticity we discovered is remarkably strong (~5 spikes/s per hour), it did not produce a selectivity reversal for the "mean" IT site within the two-hour recording session (Supplemental Figure 4-S5D). Instead, it only produced a ~50% decrease in selectivity for that "mean" site, which is entirely consistent with the fact that our "mean" IT site had reasonably strong initially selectivity for P over N (mean P-N = ~20 spikes/s). To look more deeply at this issue, we made use of the well-known observation that not all adult IT neurons are identical – some have a large amount of size or position tolerance, while others show a small amount of tolerance (DiCarlo and Maunsell, 2003; Ito et al., 1995; Logothetis and Sheinberg, 1996; Op de Beeck and Vogels, 2000). Specifically, some IT sites strongly prefer object P to N at some sizes/positions, but show only weak (P-N) selectivity at the swap sizes/positions (this neuronal response pattern is illustrated schematically at the top of Figure 4-6). We reasoned that examination of these sites should reveal if our experience manipulation is capable of causing a reversal in selectivity and building of new selectivity. Thus, we used independent data to select neuronal sub-populations from our data pool with varying amounts of initial selectivity at the swap size/position (Supplemental Experimental Procedures). Note that all of these neuronal sites had robust selectivity for P over N at the medium sizes/positions (as schematically illustrated in Figure 4-6A). This analysis revealed that our manipulation caused neuronal sites with weak initial selectivity at the swap size/position to reverse their selectivity, and to build new selectivity (building "incorrect" selectivity for N over P), exactly as predicted by the UTL hypothesis (Figure 4-6).

A second way in which our data might reveal if UTL can build tolerance is to carefully look for any changes in selectivity at the non-swap ("control") size/position. Our experiment was designed to present a large number of normal temporal contiguity exposures at that "control" size/position so that we would perfectly equate its amount of retinal exposure with that provided at the swap size/position. Although some forms of unsupervised temporal contiguity theory might predict that these normal temporal contiguity exposures should increase the (P-N) selectivity at the control size/position, we did not initially make that prediction (Figure 4-1C blue) because we reasoned that most IT sites would already have strong, adult-like selectivity for object P vs. N at that size/position, such that further supporting statistics would have little to teach those IT sites (Figure 4-7A, top right). Consistent with this, we found little mean change in (P-N) selectivity for the "control" condition in either our position tolerance experiment (Li and DiCarlo, 2008) or our size tolerance experiment (Figure 4-2, blue). However, examination of all of our IT sites revealed that some sites happened to have initially weak (P-N) selectivity at the "control" size/position while still having strong selectivity at the medium size/position (Figure 4-7A, top left). This suggested that these sites might be in a more naive state with respect to the particular objects being tested such that our temporal contiguity statistics might expand their tolerance for these objects (i.e. increase their P-N selectivity at the control size/position). Indeed, examination of these sites reveals that our exposure experiment caused a clear, significant building of new "correct" selectivity among these sites (Figure 4-7B), again directly demonstrating that unsupervised temporal contiguity experience can build IT tolerance.

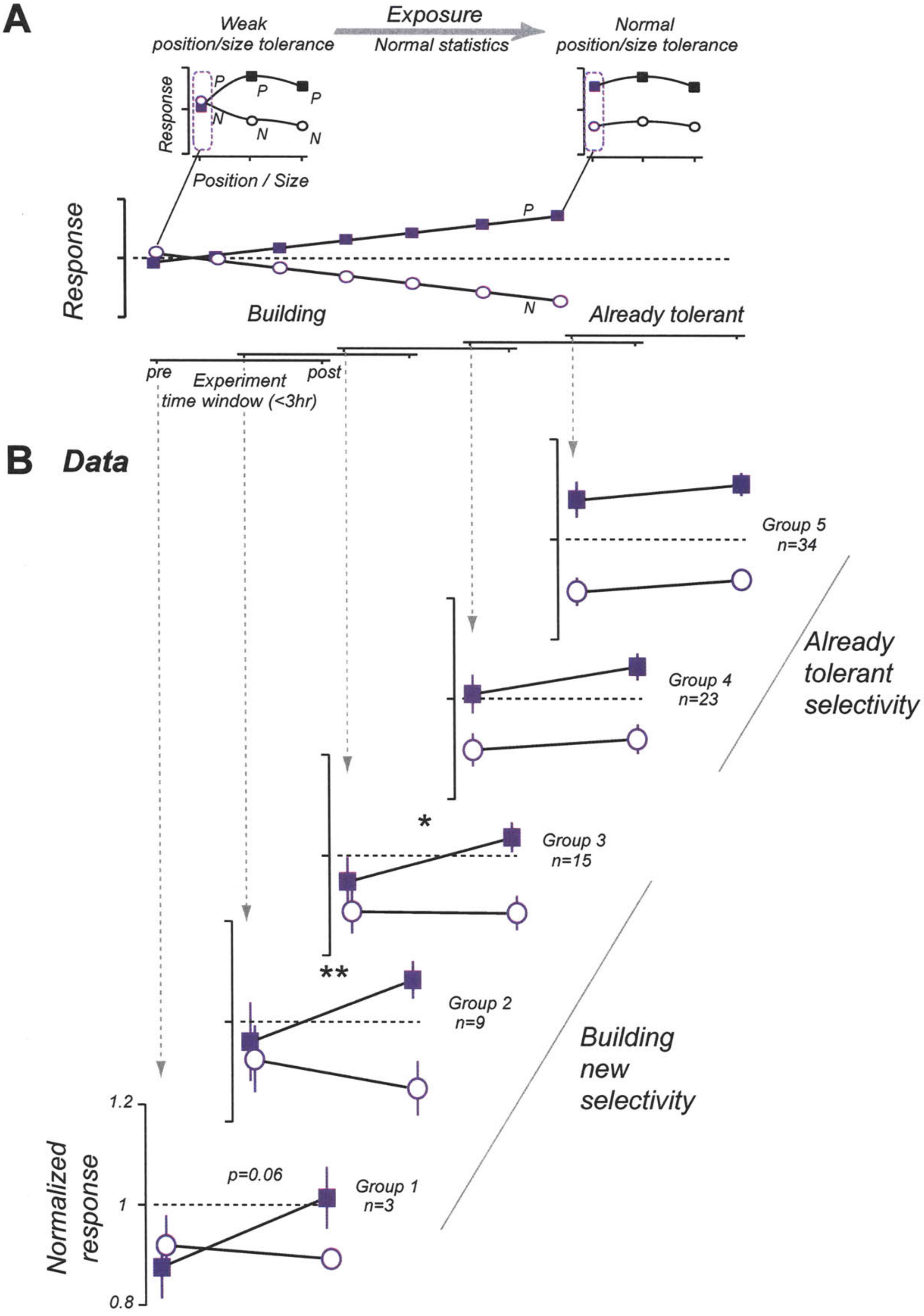

A
Weak
position/size tolerance
Exposure
Normal statistics
Normal
position/size tolerance
Response
P
P
P
N
N
N
Position / Size
Response
P
N
Building
Already tolerant
pre
Experiment
time window (<3hr)
post
B
Data
Group 5
n=34
Already
tolerant
selectivity
Group 4
n=23
*
Group 3
n=15
**
Group 2
n=9
Building
new
selectivity
1.2
1
0.8
Normalized
response
p=0.06
Group 1
n=3

Figure 4-7. Normal ("correct") Statistics in Visual Experience Builds Tolerant Selectivity.

(A) Prediction follows the same logic as in Figure 4-6A, but here for the "control" conditions in which normal temporal contiguity statistics were provided (Figure 4-1). *Top,* temporal contiguity learning predicts that neurons will be taught to build new "correct" selectivity (i.e. normal tolerance), and neurons starting with initially weak position/size tolerant selectivity (left) have the highest potential to reveal that effect. *Bottom,* at the non-swap position/size (blue), our manipulation should build new "correct" selectivity for P over N among IT sites with weak initial selectivity.

(B) Mean ± SEM normalized response to object P and N at the non-swap position/size among sub-populations of IT multi-unit sites. Sites are grouped by their initial selectivity at the non-swap position/size using independent data. Other details same as Figure 4-6B. (Size experiment data only, group 1-5: p=0.06, p<0.01; p=0.05; n.s.; n.s.).

4. 3. 5 Experiment III: Does the Learning Depend on the Temporal Direction of the Experience?

Our results show that targeted alteration of unsupervised natural visual experience rapidly reshapes IT size tolerance -- as predicted by the hypothesis that the ventral stream uses a temporal contiguity learning strategy to build that tolerance in the first place. Several instantiated computational models show how this conceptual strategy can build tolerance (Foldiak, 1991; Masquelier et al., 2007; Masquelier and Thorpe, 2007; Wallis and Rolls, 1997; Wiskott and Sejnowski, 2002; Wyss et al., 2006), and such models can be implemented using variants of Hebbian-like learning rules that are dependent on the timing of spikes (Gerstner et al., 1996; Sprekeler et al., 2007; Wallis and Rolls, 1997; Morrison et al., 2008; Sprekeler and Gerstner, 2009). The time course and task independence of the observed learning are consistent with synaptic plasticity (Markram et al., 1997; Meliza and Dan, 2006), but our data do not constrain the underlying mechanism. One can imagine ventral stream neurons using almost temporally coincident activity to learn which sets of its afferents correspond to features of the same object across size changes. If tolerance learning is spike timing dependent, any experience-induced change in IT selectivity might reflect any temporal asymmetries at the level of the underlying synaptic learning mechanism. For example, one hypothesis is that lingering post-synaptic activity caused by temporally-leading images drives synaptic plasticity in afferents activated by temporally-lagging images. Alternatively, afferents activated by temporally-leading images might be modified by the later arrival of post-synaptic activity caused by temporally-lagging images. Or a combination of both hypotheses. To look for reflections of any such underlying temporal asymmetry, we carried out a third experiment (Experiment III) centered on the question: do temporally-leading images teach temporally-lagging ones, or vice-versa?

We deployed the same experience manipulation as before (linking of different object images across size changes, same as Experiment I), but this time only in one direction (compare single-headed arrows in Figure 4-8A with double headed arrows in Figure 4-1B). For example, during the recording of a particular IT site, the animal only received experience seeing objects temporally transition from small size (arrow "tail" in Figure 4-8A) to large size (arrow "head" in Figure 4-8A), while swapping identity. We strictly alternated the temporal direction of the experience across different IT sites. That is, for the next IT site we recorded, the animal experienced objects transitioning from large size to small size while swapping identity. Thus, object size was counter-balanced across our recorded population, so that we could isolate changes in selectivity among the temporally-leading stimuli (i.e. arrow "tail" stimuli) from changes in selectivity among the temporally-lagging stimuli (i.e. arrow "head" stimuli). As in Experiments I and II, we measured the expression of any experience-induced learning by looking for any change in (P-N) selectivity at each object size measured in a neutral task with all images randomly interleaved (*Test Phase*). We replicated the results in Experiments I and II in that a decrease in (P-N) selectivity was found following swapped experience (red bars are negative in Figure 4-8B). When we sorted our data based on the temporal direction of the animals' experience, we found greater selectivity change (i.e. learning) for the temporally-lagging images (Figure 4-8B). This difference was statistically significant (p=0.038, n=31, two tailed t-test) and cannot be explained by any differences in the IT sites' initial selectivity (Supplemental Figure 4-S4C, also see Supplemental Figure 4-S4B for results with all sites included). This result is consistent with an underlying learning mechanism that favors experience-induced plasticity of the afferents corresponding to temporally-lagging images.

To test if the tolerance learning spread beyond the specifically experienced images, here, we also

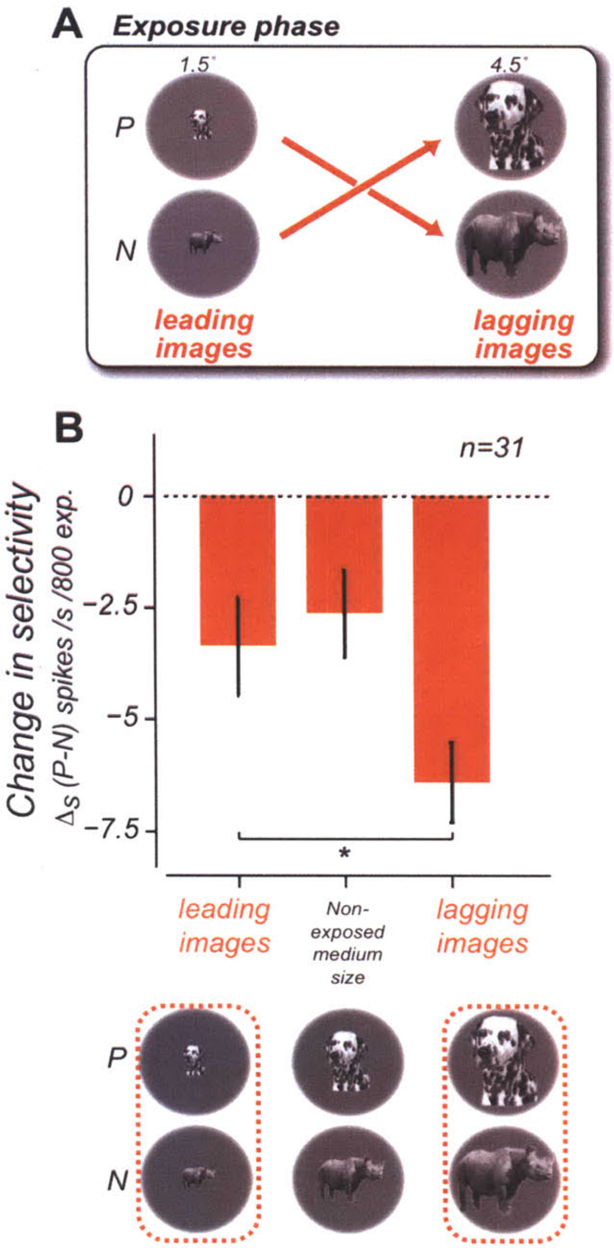


Figure 4-8. Experiment III Exposure Design and Key Results.

(A) *Exposure Phase* design (*top*, same format as Figure 4-1B) and example object images used (*bottom*).

(B) Mean ± SEM selectivity change, Δs(P-N), among the temporally-leading images, the non-exposed images at the medium object size (3°), and the temporally-lagging images. Δs(P-N) was normalized to show selectivity change per 800 exposure events. *p=0.038, two-tailed t-test.

tested object images at an intermediate size (3°) between the two exposed sizes (Figure 4-8). Unlike Experiment I and II, this medium size was not exposed to the animals during the *Exposure Phase* (it was also at a different physical size from the medium size in Experiment I and II). We observed significant selectivity change for the medium size image pairs (8B middle bar; p=0.01, two tailed t-test against zero), which suggests that the tolerance learning has some degree of spread (but not to very different objects, Supplemental Figure 4-S4). Finally, the effect size observed in Experiment III was consistent with, and can explain the effect sizes observed in Experiment I and II. That is, based on the Experiment III effect sizes for the temporally-lagging and -leading images, a first-order prediction of the net effect in Experiments I and II is the average of these two effects (because Experiments I and II employed a 50-50 mix of the experience manipulations considered separately in Experiment III). That prediction is very close to what we found (Figure 4-5).

4. 4 Discussion

The overarching goal of this work is to ask if the primate ventral visual stream uses a general, temporal contiguity driven, learning mechanism to construct its tolerance to object-identity-preserving image transformations. Our strategy was to use experience manipulations of temporally contiguous image statistics to look for changes in IT neuronal tolerance that are predicted by this hypothetical learning mechanism. Here we tested three key predictions that were not answered by previous work (Li and DiCarlo, 2008). First, we asked if these experience manipulations predictably reshaped the size tolerance of IT neurons. Our results strongly confirmed this prediction: we found that the change in size tolerance was large (~5 spikes/s, ~25% IT selectivity change, per hour of exposure) and grew gradually stronger with increasing visual experience. Second, we asked if this tolerance reshaping was induced under visual experience that mimics the common size-tolerance-building statistics in the natural world: temporally contiguous image changes without intervening eye movements, and temporally-smooth dynamics. Our results confirmed this prediction: we found that size tolerance was robustly reshaped in both of these conditions (Figure 4-2), and the magnitude of reshaping was similar to that seen with eye-movement contingent reshaping of IT position tolerance (Li and DiCarlo, 2008, Figure 4-5). Third, we asked if experience with temporal contiguous image statistics could not only "break" existing IT tolerance, but could also "build" new tolerance. Again, our results confirmed this prediction: we found that experience with incorrect statistics can build "incorrect" tolerance (Figure 4-6) and that experience with correct statistics can build correct tolerance (Figure 4-7). Finally, we found that this tolerance learning is temporally asymmetric and spreads beyond the specifically experienced images (Figure 4-8, medium size), results that have implications for underlying mechanisms (see below).

Given these results, it is now highly likely that our previously reported results on eye-movement contingent tolerance learning (Li and DiCarlo, 2008) were only one instance of a general tolerance learning mechanism. Taken together, our two studies show that unsupervised, temporally contiguous experience can reshape and build at least two types of IT tolerance, and that they can do so under a wide range of spatiotemporal regimes encountered during natural visual exploration. In sum, we speculate that these studies are both pointing to the same general learning mechanism that builds adult IT tolerance, and we have previously termed this mechanism "unsupervised temporal slowness learning" (UTL; Li and DiCarlo, 2008).

Our suggestion that UTL is a general tolerance learning mechanism is supported by a number of empirical commonalities between the size tolerance learning here and our previously reported position tolerance learning (Li and DiCarlo, 2008): 1) *object specificity*, the experience-induced changes in IT size tolerance and position tolerance have at least some specificity for the exposed object; 2) *learning induction (driving force)*, in both studies, the magnitude of learning depended on the initial selectivity of the temporally-adjacent images (medium object size here, foveal position in the position tolerance study), which is consistent with the idea that the initial selectivity may provide at least part of the driving force for the learning; 3) *time course of learning expression*, learning increased with increasing amount of experience and changed the initial part of IT response (100 ms after stimulus onset); 4) *response change of learning expression*: in both studies, the IT selectivity change arose from a response decrease to the preferred object (P) and a response increase to the less preferred object (N); 5) *effect size*, our different experience manipulations here as well as our previous position manipulation revealed a similar effect magnitude (~5 spikes/s per 400 swap exposures). More specifically, when measured as learning magnitude per exposure event, size tolerance learning was slightly smaller than that found for position tolerance learning (Figure 4-5), and when considered as learning magnitude per unit time, the results of all three experiments were nearly identical (Supplemental Figure 4-S8). However, we note that our data cannot cleanly de-confound exposure amount from exposure time.

4. 4. 1 Relation to Previous Literature

Previous psychophysical studies have shown that human object perception depends on the statistics of visual experience (e.g. Brady and Oliva, 2008; Fiser and Aslin, 2001; Turk-Browne et al., 2005). Several studies have also shown that manipulating the spatiotemporal contiguity statistics of visual experience can alter the tolerance of human object perception (Cox et al., 2005; Wallis et al., 2009; Wallis and Bulthoff, 2001). In particular, an earlier study (Cox et al., 2005) showed that the same type of experience manipulation deployed here (experience of different object images across position change) produces increased confusion of object identities across position -- a result that qualitatively mirrors the neuronal results reported here and in our previous neuronal study (Li and DiCarlo, 2008). Thus, the available psychophysical data suggest that UTL has perceptual consequences. However, this remains an open empirical question (see *Limitations and Future Direction*).

Previous neurophysiological investigations in the monkey ventral visual stream showed that IT and perirhinal neurons could learn to give similar responses to temporally nearby stimuli when instructed by reward (i.e. so-called "paired associate" learning; Messinger et al., 2001; Miyashita, 1988; Sakai and Miyashita, 1991), or sometimes, even in the absence of reward (Erickson and Desimone, 1999). Though these studies were motivated in the context of visual memory (Miyashita, 1993) and used visual presentation rates of seconds or more, it was recognized that the same associational learning across time might also be used to learn invariant visual features for object recognition (e.g. Foldiak, 1991; Stryker, 1991; Wallis, 1998; Wiskott and Sejnowski, 2002). Our studies provide a direct test of these ideas by showing that temporally contiguous experience with object images can specifically reshape the size and position tolerance of IT neurons' selectivity among visual objects. This is consistent with the hypothesis that the ventral visual stream relies on a temporal contiguity strategy to learn its tolerant object representations in the first place. Our results also demonstrate that UTL is somewhat specific to the experienced objects images (i.e. object, size, position specificity) and operates over natural, very fast time scales (hundreds of ms, faster than those previously reported) in a largely unsupervised manner. This suggests that, during natural visual exploration, the visual system can leverage an enormous amount of visual experience to construct its object invariance.

Computational models of the ventral visual stream have put forms of the temporal contiguity hypothesis to test, and have shown that learning to extract slowly-varying features across time can produce tolerant feature representations with units that mimic the basic response properties of ventral stream neurons (Masquelier et al., 2007; Masquelier and Thorpe, 2007; Sprekeler et al., 2007; Wallis and Rolls, 1997; Wiskott and Sejnowski, 2002; Wyss et al., 2006). These models can be implemented using variants of Hebbian-like learning rules (Masquelier and Thorpe, 2007; Sprekeler and Gerstner, 2009; Sprekeler et al., 2007; Wallis and Rolls, 1997). The time course and task independence of UTL reported here is consistent with synaptic plasticity (Markram et al., 1997; Rolls et al., 1989), and the temporal asymmetry in learning magnitude (Figure 4-8) constrains the possible underlying mechanisms. While the experimental approach used here may seem to imply that experience with all possible images of each object is necessary for UTL to build an "invariant" IT object representation, this is not believed to be true in a full computational model of the ventral stream. For example, V1 complex cells that encode edges may learn position tolerance that ultimately supports the invariant encoding of many objects. Our observation of partial spread of tolerance learning to non-experienced images (Figure 4-8) is consistent with this idea. In particular, at each level of the ventral stream, afferent input likely reflects tolerance already constructed for simpler features at the previous level (e.g. in the context of this study, some IT afferents may respond to an object's image at both the medium size and the swap size). Thus any modification of the swap-size-image-afferents would result in a partial generalization of the learning beyond the specifically experienced images.

4. 4. 2 Limitations and Future Direction

Because the change in object selectivity was *expressed* in the earliest part of the IT response after learning (Supplemental Figure 4-S5A), even while the animal was performing tasks unrelated to the object identity, this rules out any simple attentional account of the effect. However, our data do not rule out the possibility that attention or other top down signals may be required to mediate the learning during the *Exposure Phase*. These potential top-down signals could include non-specific reward, attentional, and arousal signals. Indeed, psychophysical evidence (Seitz et al., 2009; Shibata et al., 2009) and physiological evidence (Baker et al., 2002; Freedman and Assad, 2006; Froemke et al., 2007; Goard and Dan, 2009; Law and Gold, 2008) both suggest that reward is an important factor that can modulate or gate learning. We also cannot rule out the possibility that the attentional or the arousal system may be required for the learning to occur. In our work, we sought to engage the subjects in natural exploration during the *Exposure Phases* under the assumption that visual arousal may be important for ongoing learning, even though we deployed the manipulation during the brief periods of fixation during that exploration. Future experiments in which we systematically control these variables will shed light on these questions, and will help expose the circuits that underlie UTL.

Although the UTL phenomenology induced by our experiments was a very specific change in IT neuronal selectivity, the magnitude of this learning effect was quite large when expressed in units of spikes per second (Figure 4-5: ~5 spikes/s, ~25% change in IT selectivity per hour of exposure). This is comparable to or larger than other important neuronal phenomenology (e.g. attention, Maunsell and Cook, 2002). However, because this effect size was evaluated from the multi-unit signal, without knowledge of how many neurons we are recording from, this effect size should be interpreted with caution. Furthermore, connecting this neuronal phenomenology (i.e. change in IT image selectivity) to the larger problem of size or position tolerance at the level of the IT population or the animal's behavior is not straightforward. Quantitatively linking a neuronal effect size to behavioral effect size requires a more complete understanding of how that neuronal representation is read out to support behavior, and large effects in confusion of object identities in individual IT neurons may or may not correspond to large confusions of object identities in perception. Such questions are the target of our ongoing and future monkey studies in which one has simultaneous measures of the neuronal learning and the animal's behaviors (modeled after those such as Britten et al., 1992; Cook and Maunsell, 2002).

The rapid and unsupervised nature of UTL gives us new experimental access to understand how cortical object representations are actively maintained by the sensory environment. However, it also calls for further characterization of the time course of this learning to inform our understanding of the stability of ventral stream object representations in the face of constantly-available, natural visual experience. This sets the stage for future studies on how the ventral visual stream assembles its neuronal representations at multiple cortical processing levels, particularly during early post-natal visual development, so as to achieve remarkably powerful adult object representation.

4. 5 Experimental Procedures

4. 5. 1 Animals and Surgery

Aseptic surgery was performed on two male Rhesus monkeys (*Macaca mulatta*, 8 and 6 kg) to implant a head post and a scleral search coil. After brief behavioral training (1-3 months), a second surgery was performed to place a recording chamber to reach the anterior half of the temporal lobe. All animal procedures were performed in accordance with National Institute of Health guidelines and the Massachusetts Institute of Technology Committee on Animal Care.

4. 5. 2 General Design

On each experimental day, we recorded from a single IT multi-unit site for 2-3 hours. During that time, the animal was provided with altered visual experience in *Exposure Phases* and we made repeated measurements of the IT site's selectivity during *Test Phases* (Figure 4-1). The study consisted of three separate experiments (Experiments I, II and III), which differed from each other only in the *Exposure Phase* design (described below). We focused on one pair of objects ("swap objects") that the IT site was selective for (preferred object P; and non-preferred object N; chosen using a pre-screening procedure, see Supplemental Experimental Procedures).

Experiment I: Objects (P and N at 1.5°, 4.5°, or 9°) appeared at random positions on a gray computer screen and animals naturally looked to the objects. The image of the just-foveated object was replaced by an image of the other object at a different size (swap exposure event, Figure 4-1A) or an image of the same object at a different size (non-swap exposure event, Figure 4-1A). The image change was initiated 100 ms after foveation and was instantaneous (Figure 4-2, top). We used a fully symmetric design illustrated graphically in Figure 4-1B. This experience manipulation temporally linked pairs of object images (Figure 4-1A shows one such link) and each link could go in both directions (Figure 4-1B shows full design example). For each IT site, we always deployed the swap manipulation at one particular size (referred to as the "swap size": 1.5° or 9°, pre-chosen, strictly alternated between sites), keeping the other size as the exposure-equalized control (referred to as the "non-swap size").

Experiment II: All design parameters were identical to *Experiment I* except that the image changes were smooth across time (Figure 4-2, bottom). The image change sequence started immediately after the animal had foveated the image and the entire sequence lasted for 200 ms (Supplemental Figure 4-S2). Identity-changing morph lines were only achievable on the silhouette shapes. Only Monkey 2 was tested in Experiment II, (given the stimulus class assignment).

Experiment III: We used an asymmetric design that is illustrated graphically in Figure 4-8A: for each IT site, we only gave the animals experience of image changes in one direction (1.5°→4.5° or vice versa, pre-chosen, strictly alternated between sites). The timing of the image change was identical to *Experiment I.*

Another pair of control objects (P' and N', not shown in the *Exposure Phase*) was also used to probe the IT site's responses in the *Test Phase.* The selectivity among the control objects served as a measure of recording stability (below). In each *Test Phase,* the swap and control objects were tested at three sizes (Experiment I and II: 1.5°, 4.5°, 9°; Experiment III: 1.5°, 3°, 4.5°) by presenting them briefly (100 ms) on the animals' center of gaze (50-60 repetitions, randomized) during orthogonal behavioral tasks in which object identity and size were irrelevant. See Supplemental Experimental Procedures for details of the task design and behavioral monitoring.

4. 5. 1 Neuronal Assays

We recorded multi-unit activity (MUA) from the anterior region of IT using standard single microelectrode methods. Our previous study on IT position tolerance learning showed that we could uncover the same learning in both single-unit activity and MUA with comparable effect size (Li and DiCarlo, 2008), thus here, we only recorded MUA to maximize recording time. Over a series of recording days, we sampled across IT and sites selected for all our primary analyses were required to be selective among object P and N (ANOVA, object x sizes, $p<0.05$ for "object" main effect or interaction) and pass a stability criterion (n=27 for Experiment I; 15 for Experiment II; 31 for Experiment III). We verified that the key result is robust to the choice of the stability criteria (Supplemental Figure 4-S4). See Supplemental Experimental Procedures for details of the recording procedures and site selections.

4. 5. 1 Data Analyses

All the analyses and statistical tests were done in MATLAB (Mathworks, Natick, MA) with either custom written scripts or standard statistical packages. The IT response to each image was computed from the spike count in a 150 ms time window (100-250 ms post stimulus onset, data from *Test Phases* only). Neuronal selectivity was computed as the response difference in units of spikes/s between images of object P and N at different object sizes. To avoid any bias in this estimate of selectivity, for each IT site we define the labels "P" (preferred) and "N" by using a portion of the pre-exposure data to determine these labels, and the remaining data to compute the selectivity values reported in the text (Supplemental Experimental Procedures). In cases where neuronal response data was normalized and combined (Figures 4-6, 4-7), each site's response from each *Test Phase* was normalized to its mean response to all objects images in that *Test Phase*. The key results were evaluated statistically using a combination of t-tests and interaction tests (Supplemental Experimental Procedures). For analyses presented in Figure 4-4, we extracted clear single-units from the waveform data of each *Test Phase* using a PCA-based spike sorting algorithm (Supplemental Experimental Procedures).

4. 6 Acknowledgments

We thank Profs. T. Poggio, N. Kanwisher, E. Miller, and the members of our laboratory for valuable discussion and comment on this work. We also thank J. Deutsch, B. Andken, and Dr. R. Marini for technical support. This work was supported by the NIH (grant R01-EY014970 and its ARRA supplement to J.J.D., NRSA 1F31EY020057 to N.L.) and The McKnight Endowment Fund for Neuroscience.

4. 7 References

Afraz, S., Kiani, R., and Esteky, H. (2006). Microstimulation of inferotemporal cortex influences face categorization. Nature.

Baker, C.I., Behrmann, M., and Olson, C.R. (2002). Impact of learning on representation of parts and wholes in monkey inferotemporal cortex. Nat Neurosci *5*, 1210-1216.

Brady, T.F., and Oliva, A. (2008). Statistical learning using real-world scenes: extracting categorical regularities without conscious intent. Psychol Sci *19*, 678-685.

Brincat, S.L., and Connor, C.E. (2004). Underlying principles of visual shape selectivity in posterior inferotemporal cortex. Nat Neurosci *7*, 880-886.

Britten, K.H., Shadlen, M.N., Newsome, W.T., and Movshon, J.A. (1992). The analysis of visual motion: a comparison of neuronal and psychophysical performance. J Neurosci *12*, 4745-4765.

Cook, E.P., and Maunsell, J.H.R. (2002). Attentional modulation of behavioral performance and neuronal responses in middle temporal and ventral intraparietal areas of macaque monkey. J Neurosci *22*, 1994-2004.

Cox, D.D., Meier, P., Oertelt, N., and DiCarlo, J.J. (2005). 'Breaking' position-invariant object recognition. Nat Neurosci *8*, 1145-1147.

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

Erickson, C.A., and Desimone, R. (1999). Responses of macaque perirhinal neurons during and after visual stimulus association learning. J Neurosci *19*, 10404-10416.

Fiser, J., and Aslin, R.N. (2001). Unsupervised statistical learning of higher-order spatial structures from visual scenes. Psychol Sci *12*, 499-504.

Foldiak, P. (1991). Learning invariance from transformation sequences. Neural Computation *3*, 194-200.

Freedman, D.J., and Assad, J.A. (2006). Experience-dependent representation of visual categories in parietal cortex. Nature *443*, 85-88.

Froemke, R.C., Merzenich, M.M., and Schreiner, C.E. (2007). A synaptic memory trace for cortical receptive field plasticity. Nature *450*, 425-429.

Gerstner, W., Kempter, R., van Hemmen, J.L., and Wagner, H. (1996). A neuronal learning rule for sub-millisecond temporal coding. Nature *383*, 76-81.

Goard, M., and Dan, Y. (2009). Basal forebrain activation enhances cortical coding of natural scenes. Nat Neurosci *12*, 1444-1449.

Hung, C.P., Kreiman, G., Poggio, T., and DiCarlo, J.J. (2005). Fast readout of object identity from macaque inferior temporal cortex. Science *310*, 863-866.

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

Kobatake, E., Wang, G., and Tanaka, K. (1998). Effects of shape-discrimination training on the selectivity of inferotemporal cells in adult monkeys. Journal of Neurophysiology *80*, 324-330.

Kreiman, G., Hung, C.P., Kraskov, A., Quiroga, R.Q., Poggio, T., and DiCarlo, J.J. (2006). Object selectivity of local field potentials and spikes in the macaque inferior temporal cortex. Neuron *49*, 433-445.

Law, C.T., and Gold, J.I. (2008). Neural correlates of perceptual learning in a sensory-motor, but not a sensory, cortical area. Nat Neurosci *11*, 505-513.

Li, N., Cox, D.D., Zoccolan, D., and DiCarlo, J.J. (2009). What response properties do individual neurons need to underlie position and clutter "invariant" object recognition? J Neurophysiol *102*, 360-376.

Li, N., and DiCarlo, J.J. (2008). Unsupervised natural experience rapidly alters invariant object representation in visual cortex. Science *321*, 1502-1507.

Logothetis, N.K., and Sheinberg, D.L. (1996). Visual object recognition. Ann. Rev. Neurosci. *19*, 577-621.

Markram, H., Lubke, J., Frotscher, M., and Sakmann, B. (1997). Regulation of synaptic efficacy by coincidence of postsynaptic APs and EPSPs. Science *275*, 213-215.

Masquelier, T., Serre, T., Thorpe, S.J., and Poggio, T. (2007). Learning complex cell invariance from natural video: a plausibility proof. In CBCL Paper (Massachusetts Institute of Technology, Cambridge, MA).

Masquelier, T., and Thorpe, S.J. (2007). Unsupervised learning of visual features through spike timing dependent plasticity. PLoS Comput Biol *3*, e31.

Maunsell, J.H.R., and Cook, E.P. (2002). The role of attention in visual processing. Philos Trans R Soc Lond B Biol Sci *357*, 1063-1072.

Meliza, C.D., and Dan, Y. (2006). Receptive-field modification in rat visual cortex induced by paired visual stimulation and single-cell spiking. Neuron *49*, 183-189.

Messinger, A., Squire, L.R., Zola, S.M., and Albright, T.D. (2001). Neuronal representations of stimulus associations develop in the temporal lobe during learning. Proc Natl Acad Sci U S A *98*, 12239-12244.

Miyashita, Y. (1988). Neuronal correlate of visual associative long-term memory in the primate visual cortex. Nature *335*, 817-820.

Miyashita, Y. (1993). Inferior temporal cortex: where visual perception meets memory. Annual Review of Neuroscience *16*, 245-263.

Morrison, A., Diesmann, M., and Gerstner, W. (2008). Phenomenological models of synaptic plasticity based on spike timing. Biol Cybern *98*, 459-478.

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

Rolls, E.T., Baylis, G.C., Hasselmo, M.E., and Nalwa, V. (1989). The effect of learning on the face selective responses of neurons in the cortex in the superior temporal sulcus of the monkey. Exp Brain Res *76*, 153-164.

Sakai, K., and Miyashita, Y. (1991). Neural organization for the long-term memory of paired associates. Nature *354*, 152-155.

Seitz, A.R., Kim, D., and Watanabe, T. (2009). Rewards evoke learning of unconsciously processed visual stimuli in adult humans. Neuron *61*, 700-707.

Shibata, K., Yamagishi, N., Ishii, S., and Kawato, M. (2009). Boosting perceptual learning by fake feedback. Vision Res *49*, 2574-2585.

Sigala, N., Gabbiani, F., and Logothetis, N.K. (2002). Visual categorization and object representation in monkeys and humans. J Cogn Neurosci *14*, 187-198.

Sprekeler, H., and Gerstner, W. (2009). Robust learning of position invariant visual representations with OFF responses. In COSYNE (Salt Lake City).

Sprekeler, H., Michaelis, C., and Wiskott, L. (2007). Slowness: an objective for spike-timing-dependent plasticity? PLoS Comput Biol *3*, e112.

Stryker, M.P. (1991). Neurobiology. Temporal associations. Nature *354*, 108-109.

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

Turk-Browne, N.B., Junge, J., and Scholl, B.J. (2005). The automaticity of visual statistical learning. J Exp Psychol Gen *134*, 552-564.

Vogels, R., and Orban, G.A. (1996). Coding of stimulus invariances by inferior temporal neurons. Prog Brain Res *112*, 195-211.

Wallis, G. (1998). Spatio-temporal influences at the neural level of object recognition. Network *9*, 265-278.

Wallis, G., Backus, B.T., Langer, M., Huebner, G., and Bulthoff, H. (2009). Learning illumination- and orientation-invariant representations of objects through temporal association. J Vis *9*, 6.

Wallis, G., and Bulthoff, H.H. (2001). Effects of temporal association on recognition memory. Proc Natl Acad Sci U S A *98*, 4800-4804.

Wallis, G., and Rolls, E.T. (1997). Invariant face and object recognition in the visual system. Progress in Neurobiology *51*, 167-194.

Wiskott, L., and Sejnowski, T.J. (2002). Slow feature analysis: unsupervised learning of invariances. Neural Comput *14*, 715-770.

Wyss, R., Konig, P., and Verschure, P.F. (2006). A model of the ventral visual system based on temporal stability and local memory. PLoS Biol *4*, e120.

4. 8 Supplemental Figures

4. 8. 1 Supplemental Figure 4-S1

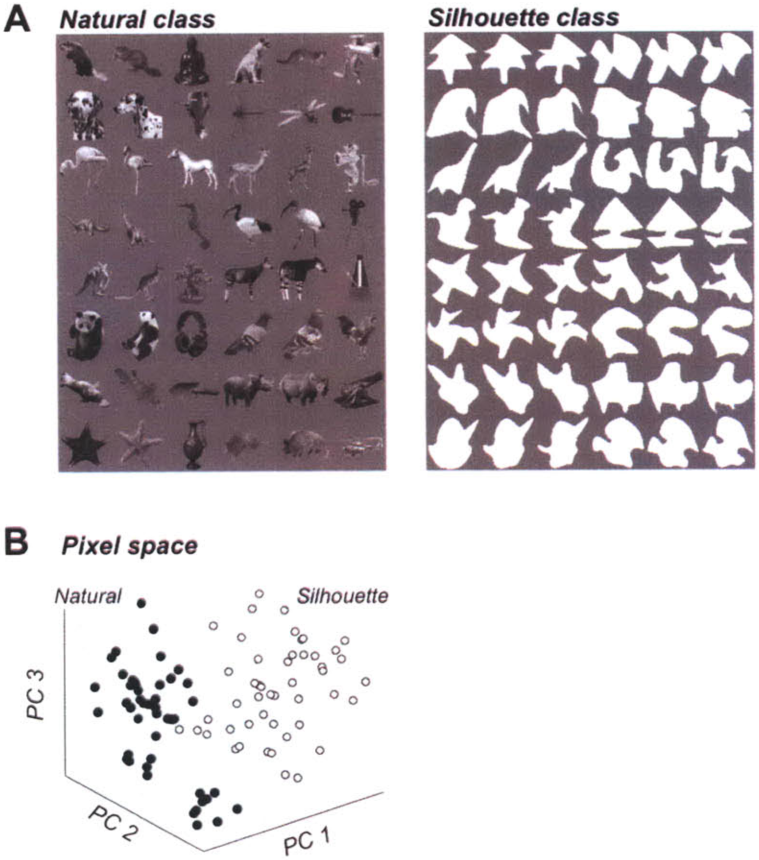


Figure 4-S1. Stimuli and Image Analyses

(A) We selected object pairs from two different stimulus classes (48 cutout natural shapes; 48 silhouette shapes). The swap object pair (P and N used for the key experience manipulation) was always picked from one class for each animal (Monkey 1: natural; Monkey 2: silhouette). The control object pair (P' and N') was always picked from the other stimulus class.

(B) Stimuli from the two classes are quite different from each other in their pixel-wise similarity. This is illustrated when the stimulus images are plotted by scores of the first three principle components (PC) in the pixel space. Principle components were computed from all 96 images. Images were pre-processed to have equal mean and unit variance before image analyses. Solid symbols: natural; open symbols: silhouette.

4. 8. 2 Supplemental Figure 4-S2

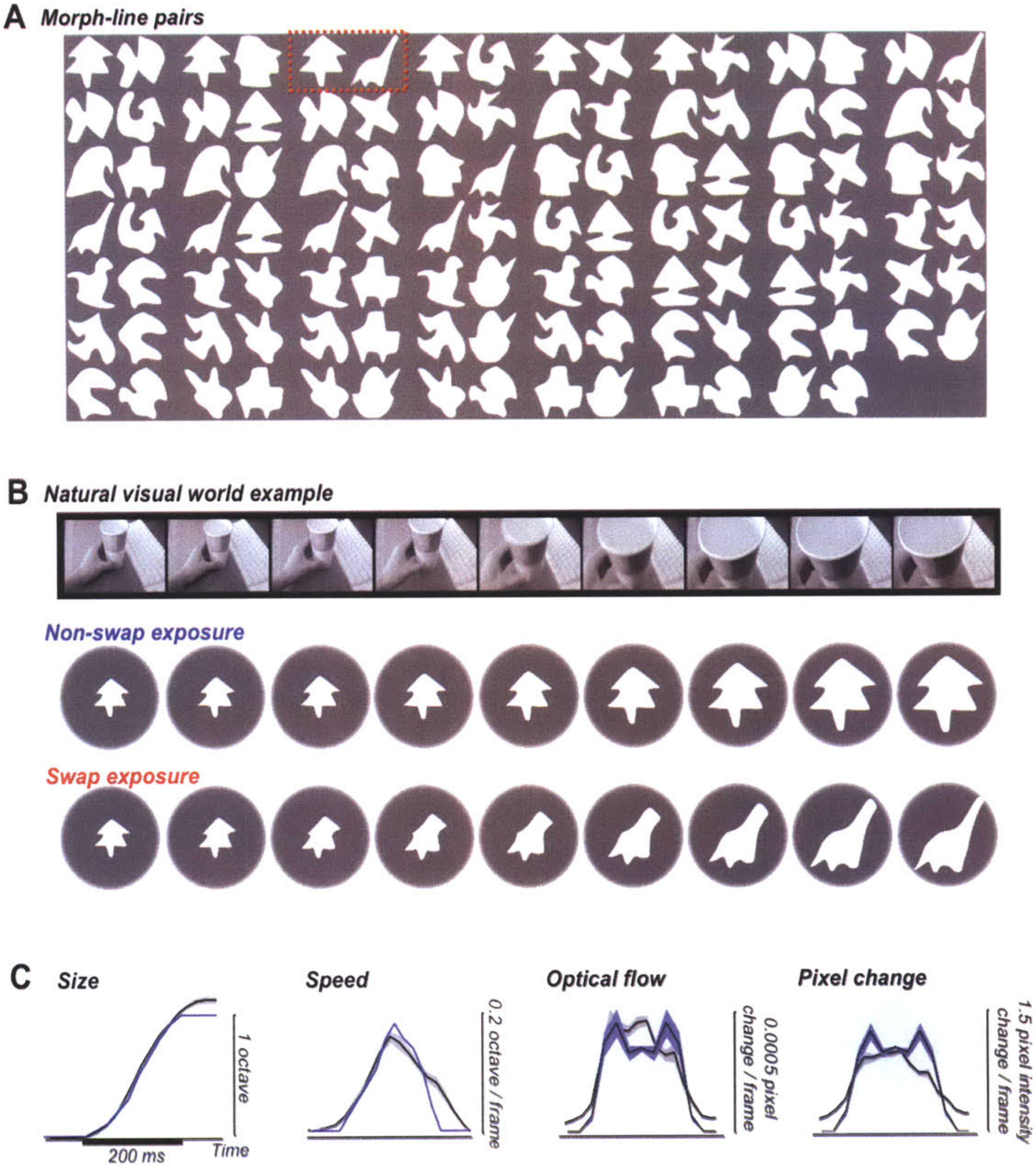


Figure 4-S2. Stimuli from Experiment II and Comparisons to Natural Visual World Example.

(A) Cutout silhouette shapes were rendered using non-uniform rational B-spline. Each shape was rendered from a set of 24 control points. Matching and interpolating between the control points allowed us to parametrically morph between different shapes. Morph-lines were only achievable on a subset of all possible shape pairs in the silhouette class (Figure 4-S1). The figure shows all the morph-line pairs used in Experiment II. Only Monkey 2 was tested in Experiment II given the stimulus class assignment. The example pair in (B) is highlighted.

(B) Top, a real world example of natural visual experience when lifting a cup to drink. Bottom, example exposure events we used in Experiment II (top, non-swap exposure event; bottom, swap exposure event). During each exposure event, the object size change was played out smoothly over a time period of 200 ms (frame rate: 40 frames/sec). We used the same dynamic (i.e. same size change profile but scaled in amplitude) for the two different types of size increase exposure events (1.5°→4.5°, 4.5°→9°, Figure 4-1B). For the object size decrease exposure events (4.5°→1.5°, 9°→4.5°, Figure 4-1B), the reverse sequence was played, which also mimicked the natural visual experience of putting down a cup (not shown).

(C) We quantified the statistics of the visual world example and our movie stimuli by a number of different image measures. Black lines show the visual world example (mean computed from videos of multiple repeats of the same action); blue lines show our movie stimuli (mean computed from all exposure events); shaded areas show SEMs. Object size was measured by the radius of the smallest bounding square around the shape (reported in units of octave, normalized to the initial size). Object size change speed was computed by taking the derivative of the object size measurements. Optical flow was computed using standard computer vision algorithm (Horn, 1986). Brightness patterns in the image move as the objects that give rise to them move. Optical flow quantifies the apparent motion of the brightness pattern. Here, mean optical flow magnitude over the entire image was computed. Pixel change was computed by taking the pixel intensity differences between adjacent video frames and the Euclidean norm of the pixel difference over the entire image was computed. All video frames were pre-processed to have unit variance before image analyses.

4. 8. 3 Supplemental Figure 4-S3

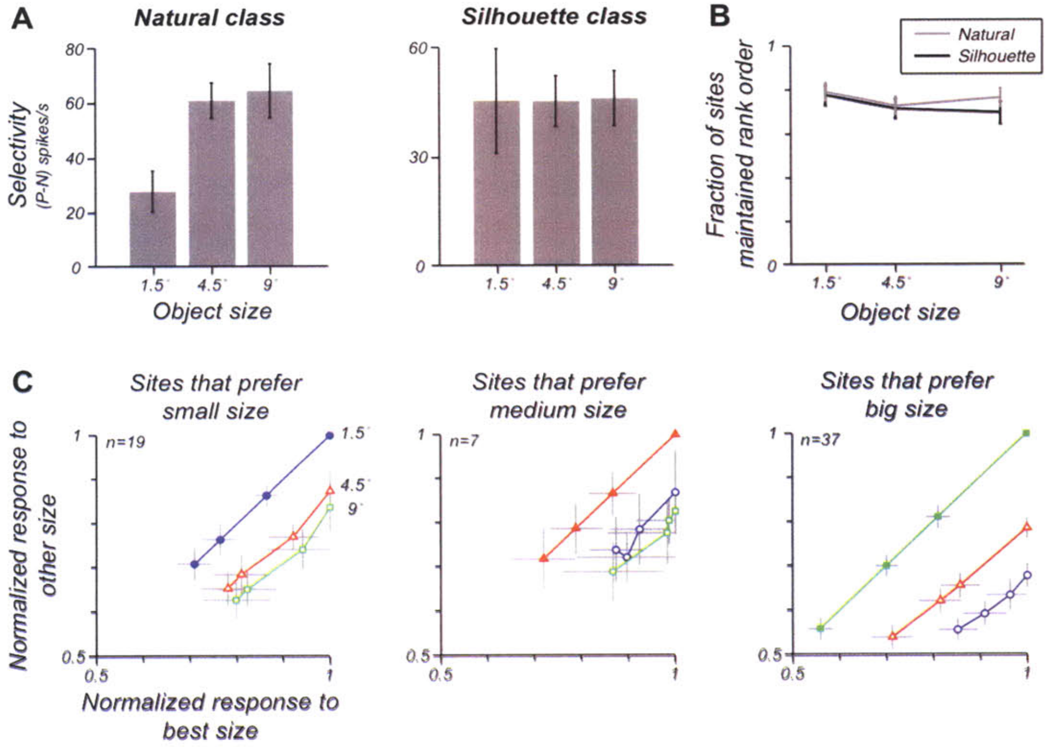


Figure 4-S3. IT Multi-unit Activity Exhibits Size Tolerant Object Selectivity.

(A) IT neurons have object rank order selectivity that is largely unaffected by object size changes (Brincat and Connor, 2004; Ito et al., 1995; Logothetis and Sheinberg, 1996; Vogels and Orban, 1996), and that size tolerance is reflected in the IT multi-unit activity (Hung et al., 2005; Kreiman et al., 2006). Consistent with previous reports, most of the IT sites we recorded maintained their object rank order preference across the range of object size tested here (1.5°~9°). To quantify the degree of IT size tolerance for the swap and control object pairs, for each IT site we determined its preferred (P) and less preferred (N) object within an object pair using a portion of the response data at the medium object size (4.5°). We then used those "P" "N" labels to compute the object selectivity (P-N) from the remaining response data and for other object size. The plots show the mean ± SEM selectivity of all object selective sites from Experiment I and II (n=63). Positive selectivity indicates that IT sites, on average, maintained their object preference across size changes.

(B) Most of the individual IT sites (~80%, n=63) maintained their object rank order preference. The plot shows the fraction of the IT sites in (A) that maintained their object rank order preference at each object size. Errorbars show SEMs.

(C) To summarize the average effect of object size changes on IT object selectivity across all four objects (swap and control object pairs combined), we split the 63 object selective IT sites into three groups based on their size preference. Preferred size for an IT site was defined as the size at which any object evoked the maximum response from the site. We then ranked the object preference based on the response at the preferred size (from best to worst). The abscissa represents the normalized response to the best object at each particular size. The ordinate represents the normalized response to the best object at the preferred size. Each data point shows the mean ± SEM. On average, IT sites maintained their object rank order preference. We found more sites preferring the extremity object sizes (1.5° and 9°) than the medium object size (4.5°), with more sites preferring the big object size (9°).

4. 8. 4 Supplemental Figure 4-S4

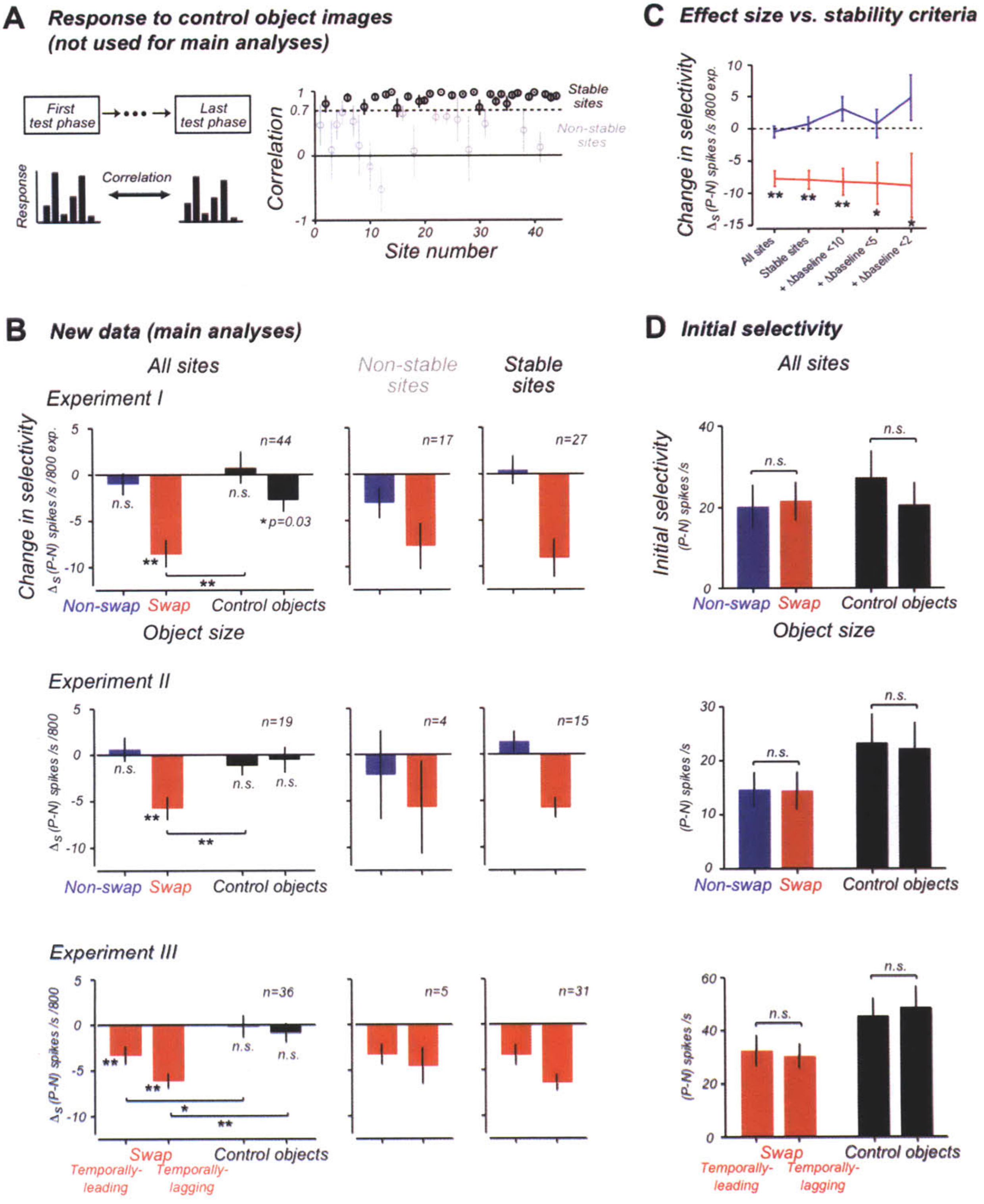


Figure 4-S4. IT Results with All Object Selective Sites before and after Stability Screen.

(A) We deployed the key experience manipulation with a pair of swap objects (P and N) in the *Exposure Phase*. We also measured the IT response to a second pair of control objects (P′ and N′)

along with the swap objects in the *Test Phase* (see Supplemental Experimental Procedures). We were interested in specific selectivity change in IT induced by our experience manipulation. However, there were potential non-specific changes in selectivity (e.g. from electrode drifts in tissue or tissue death) that could contaminate our effect of interest. Unlike traditional single-unit recording where one could judge the stability of long-term recording based on spike waveform, we did not have such measure in multi-unit recording. Thus we sought another independent measure of long-term recording stability (2-3 hours). To do this, we relied on IT selectivity among the images of the control objects (P' and N'). We picked these control objects to be sufficiently different from the swap objects in their pixel-wise similarity (Figure 4-S1). Our analyses (panel B left column) and our previous investigation (Li and DiCarlo, 2008) have revealed that any experience-induced change in selectivity was specific to the swap objects. Leveraging this, we made the assumption that the control objects were far apart from the swap objects in IT shape space, thus they should be little affected by our experience manipulation. For each IT site, we computed Pearson's correlation between its response vectors to these control object images (6 dimensional vector, 2 objects x 3 sizes) measured from the first and last *Test Phase* (right panel: mean ± SEM; data from Experiment I only). A fraction of the sites showed low correlations, meaning their responses to the control object images had deviated from those measured in the first *Test Phase.* Note that a site could also have low correlation from having no tuning among the control object images to begin with, in those cases, we had no power to judge recording stability. In practice, we deemed a site stable if it had a correlation value higher than 0.7.

(B) All the main text results concentrated on the stable IT sites. Here, we present the main IT results from all object selective sites. Left column panels show mean ± SEM selectivity change, Δs(P-N), of the swap objects (red, swap size; blue, non-swap size) and control objects (black, same size as the swap objects). We found the change in IT selectivity was specific to the swap objects at the swap size. Statistically, object specificity of the selectivity change at the swap size was confirmed by a significant "object x exposure" interaction ($p=0.009$, repeated measures ANOVA). Next, we applied the stability screen outlined in (A) using the IT responses to the control object images (not used for the main analyses), we then looked to the change in selectivity, Δs(P-N), among the swap objects at the swap and non-swap size. The stability screen revealed non-specific changes in selectivity of the non-stable IT sites (middle column panels). Among the sites we deemed stable (right column panels), our experience manipulation induced very specific change in selectivity only at the swap size. Δs(P-N) was normalized to show IT selectivity change per 800 exposure events. * $p<0.05$ by t-test; ** $p<0.01$; n.s. $p>0.05$

(C) We also tested more strict forms of stability criteria that included baseline response (change <10, <5, and <2 spikes/s, before vs. after exposure) in addition to the standard stability screen. The plot shows Δs(P-N) at the swap (red) and non-swap size (blue). Data from Experiment I and II are combined (left to right: n=63; n=42; n=18; n=11; n=5). * $p<0.05$; ** $p<0.01$, t-test, swap vs. non-swap size.

(D) Mean ± SEM initial selectivity, (P-N), measured from the first *Test Phase.*

4.8.5 Supplemental Figure 4-S5

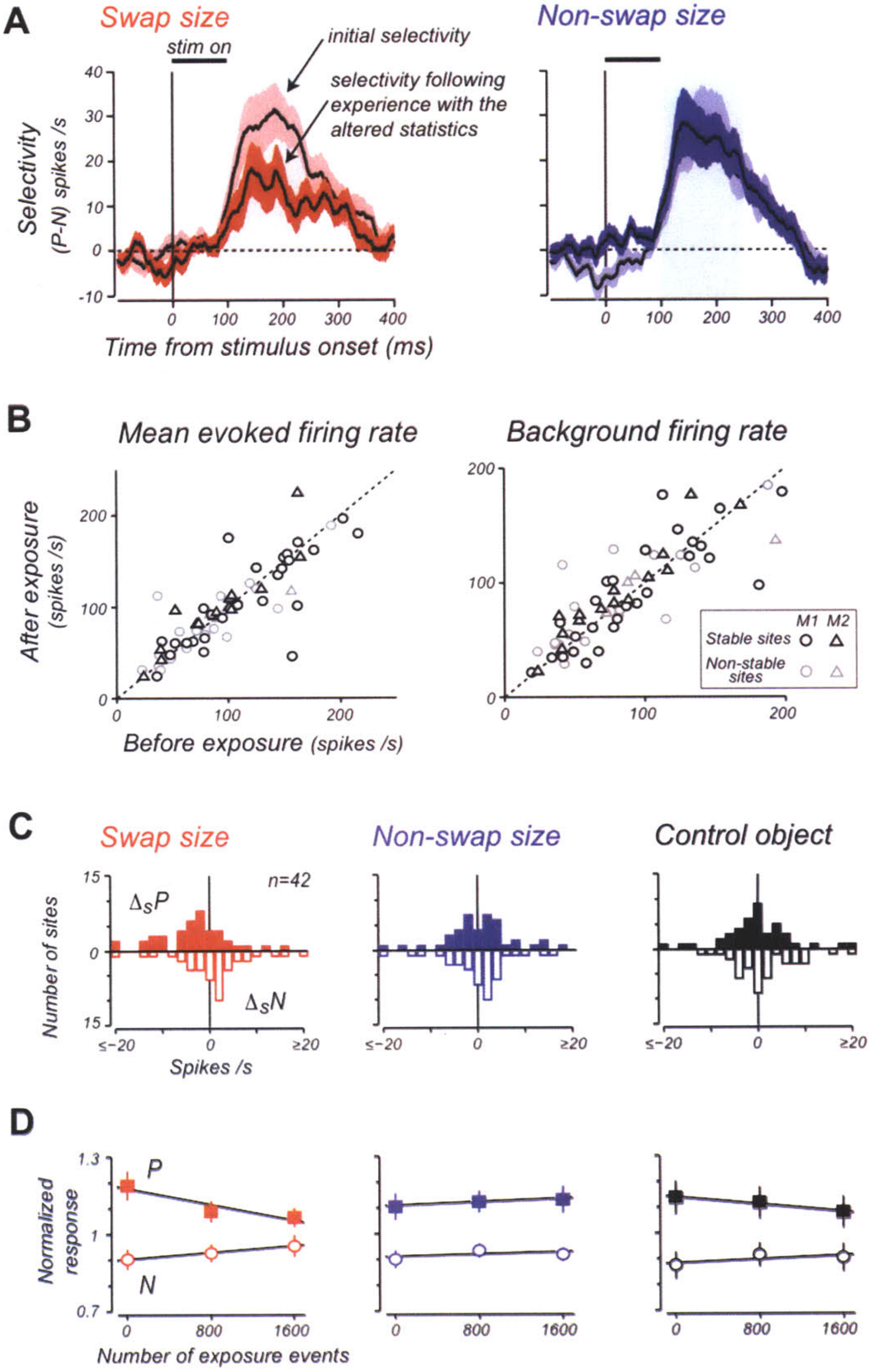


Figure 4-S5. IT Response Changes Induced by Visual Experience.

(A) Mean ± SEM IT selectivity time course at the swap (left) and non-swap size (right) measured in the first (light colored) and last *Test Phase* (dark colored). Data from Experiment I and II are combined (n=42 IT sites). Gray region shows the standard spike count time window we used for all other analyses in the main text.

(B) IT firing rate was not altered by visual experience. For each IT site, we computed its mean evoked firing rate to all object images from the first and last *Test Phase*. All object selective sites were combined from Experiment I and II (n=63). We observed no net change in the mean evoked firing rate before and after our experience manipulation (left panel; p=0.24, two tailed t-test, before versus after). We also observed no net change in IT background firing rate (right panel; p=0.17, two tailed t-test). Background firing was measured from randomly interleaved blank stimulus presentations during the *Test Phases*. A few sites showed large change in their background firing rate even though they were classified as "stable sites" by their selectivity for the control object images (Figure 4-S4). We thus tested more strict forms of stability criteria that included background firing rate with key results unchanged (Figure 4-S4).

(C) We fit standard linear regression to each IT site's responses to object P and N at each object size as a function of the number of exposure events. The slope of the line fits (Δs) provided a measure of the response changes to P and N for each IT site. The histograms show the slope values of all the stable sites from Experiment I and II (n=42). ΔsP and ΔsN were normalized to show response changes per 800 exposure events.

(D) Mean ± SEM normalized responses to object P and N as a function of the number of exposure events. For each IT site, response of each *Test Phase* was normalized to the mean response to all object images in that *Test Phase*.

4. 8. 6 Supplemental Figure 4-S6

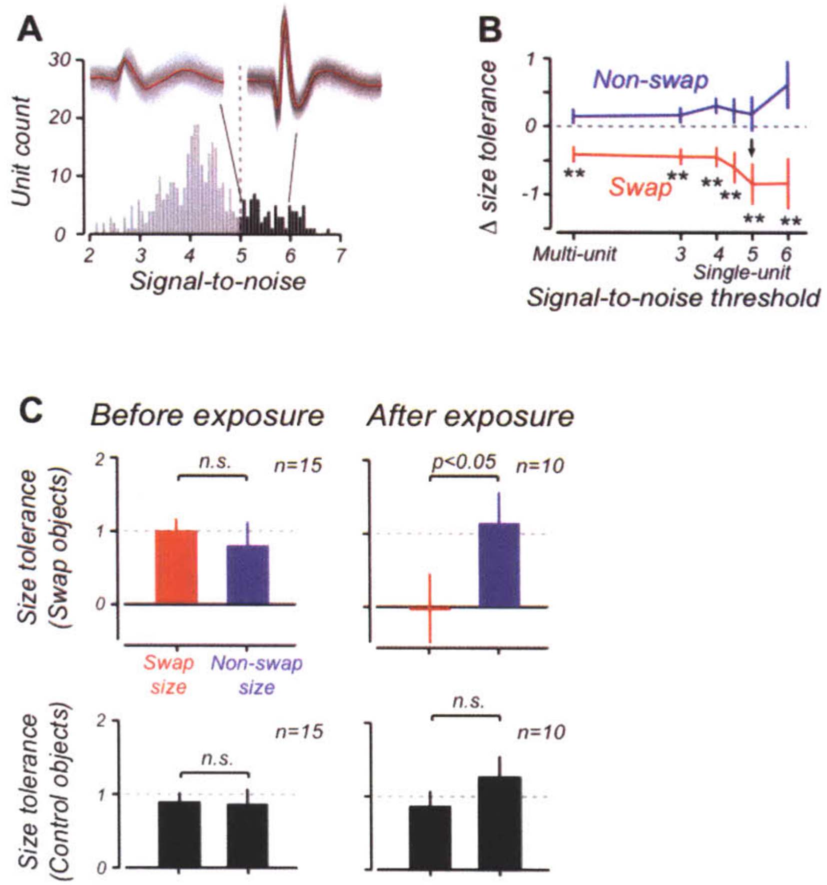


Figure 4-S6. IT Single-Unit result is Robust to Unit Selection Criteria.

(A) We performed PCA-based spike sorting on the waveforms collected during each *Test Phase*, treating each unit as an independent sample from the IT population either before or after the altered visual experience. Each unit obtained from the spike sorting was further evaluated by its signal-to-noise ratio (SNR: ratio of peak-to-peak mean waveform amplitude to standard deviation of the noise). The histogram shows the distribution of SNR for all the units obtained. For all the single-unit analyses in the main text (Figure 4-4), we set a SNR threshold (dash-line: SNR=5.0) above which we will term a unit "single-unit".

(B) To ask if the result was robust to our choice of the single-unit SNR threshold, we systematically varied the threshold and re-performed the same analyses. The plot shows the experience-induced change in size tolerance (Δ size tolerance, same as in Figure 4-4D) at the swap (red) and non-swap (blue) size. We found that the result was highly robust to the single-unit selection criteria, and the experience induced effect at the swap size only grew stronger when we increased the strictness of the single-units criteria. ** $p<0.001$, bootstrap; arrow head shows the single-unit threshold used in Figure 4-4.

(C) Mean ± SEM size tolerance for the swap and control objects measured in the same population of neurons. Same as Figure 4-4B.

4. 8. 7 Supplemental Figure 4-S7

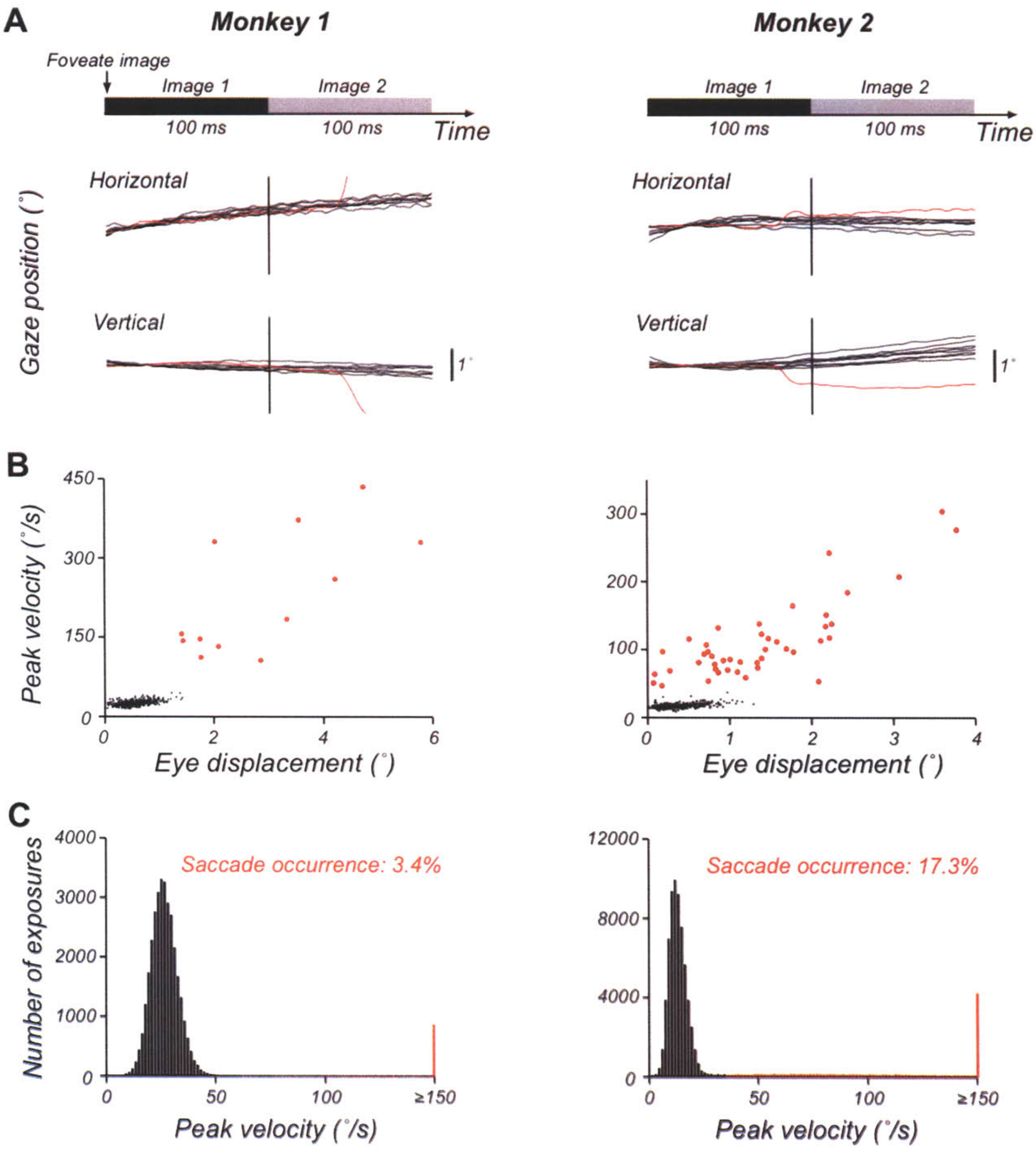


Figure 4-S7. Eye Movement Pattern during Exposure Events.

(A) Our previous study on IT position tolerance learning (Li and DiCarlo, 2008) showed that unsupervised experience of temporally contiguous images coupled by an intervening saccade can reshape IT position tolerance. Here, we showed that unsupervised experience of temporally contiguous images presented on animals' center of gaze is sufficient to induce IT size tolerance learning. The animals freely viewed a gray computer screen on which objects intermittently appeared at random position. We deployed the experience manipulation (i.e. image pairing across time) during brief periods of the animals' fixation. The exposure events were meant to mimic regimes of natural vision where object change size on the retinal due to object motion (de-coupled from intervening eye movements). However, it was possible that the discontinuous image changes we employed always induced small saccades from the animals during the exposure events, hence the observed IT size tolerance learning is simply the same piece of phenomenology as the IT position tolerance learning reported before. Here we examine this possibility by analyzing the *Exposure Phase* eye movement data around the time of image change (±100ms). The plots show the stimulus presentation time sequence (top) and aligned eye position data (bottom) during a few exposure events from one example *Exposure Phase.* The animals were able to maintain their gaze position throughout the periods of image change in most cases, though there were minor drifts (typically <1°). Occasionally, the animals made small saccades (red eye traces), however, these only constituted a small fraction of all exposure events, see (C).

(B) All the eye movement data from the example *Exposure Phase* was plotted in their relationship between the total eye displacement and peak velocity around the time of image change (±100ms). Each data point represents data from one exposure event (i.e. one trace in (A)). For saccades (red dots), there was a systematic relationship between the peak velocity and eye displacement (i.e. main sequence), which distinguished itself from the pattern of fixation eye movement (black dots). There was always good separation between the two types of eye movement pattern, thus we used a peak velocity threshold to define saccades (Monkey 1: ~60°/s; Monkey 2: ~40°/s).

(C) Histograms of eye movement peak velocity during all exposure events (Experiment I population data: all *Exposure Phases* across all recording sessions were combined). Exposure events that contained saccades are shown in red bins and exposure events without saccades are in black bins. The animals made saccades only on a small fraction of all exposure events (Monkey 2 was slightly worse). Given the small occurrence of saccades in comparison to our previous study on position tolerance where saccades accompanied every exposure event (Li and DiCarlo, 2008), we concluded that the possibility of intervening saccades cannot account for the observed IT selectivity change.

4. 8. 8 Supplemental Figure 4-S8

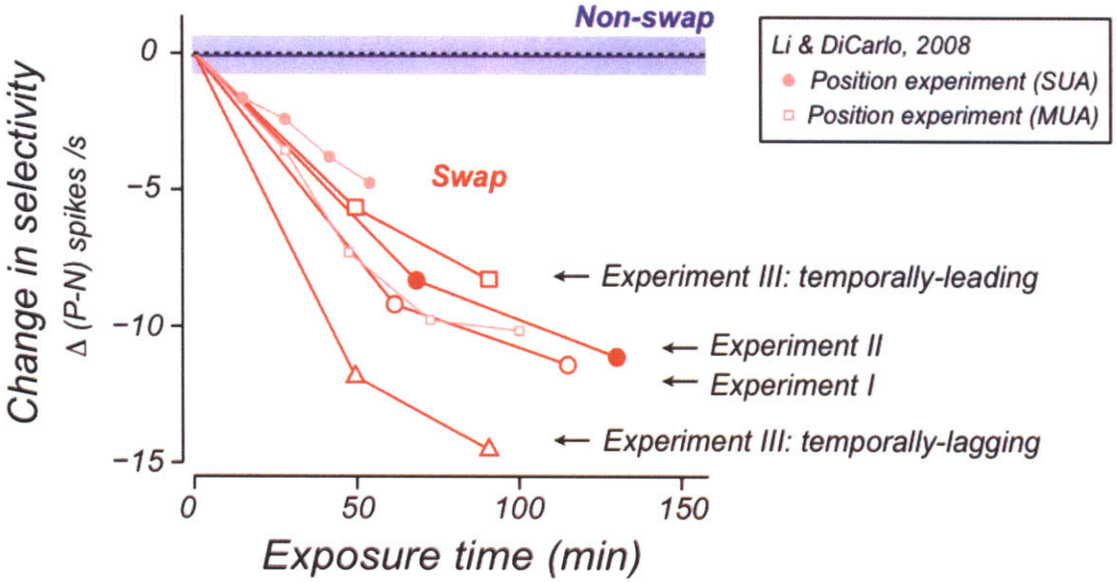


Figure 4-S8. Effect Size Comparisons across Different Experience Manipulations as a Function of Exposure Time.

Mean change in IT object selectivity, Δ(P-N), as a function of swap exposure time for different experience manipulations (i.e. Experiments I, II, III; position experiments: Li and DiCarlo, 2008). Exposure time was determined based on the time *Test Phase* data files were saved. For each data points, we computed the average exposure time across all the neurons/sites (grouped by their *Test Phase* numbers). Plot format is the same as main text Figure 4-5. Mean ± SEM selectivity change at the non-swap size (or position) is shown in blue (pooled across all experiments). SUA: single-unit activity; MUA: multi-unit activity.

4. 8. 9 Supplemental Figure 4-S9

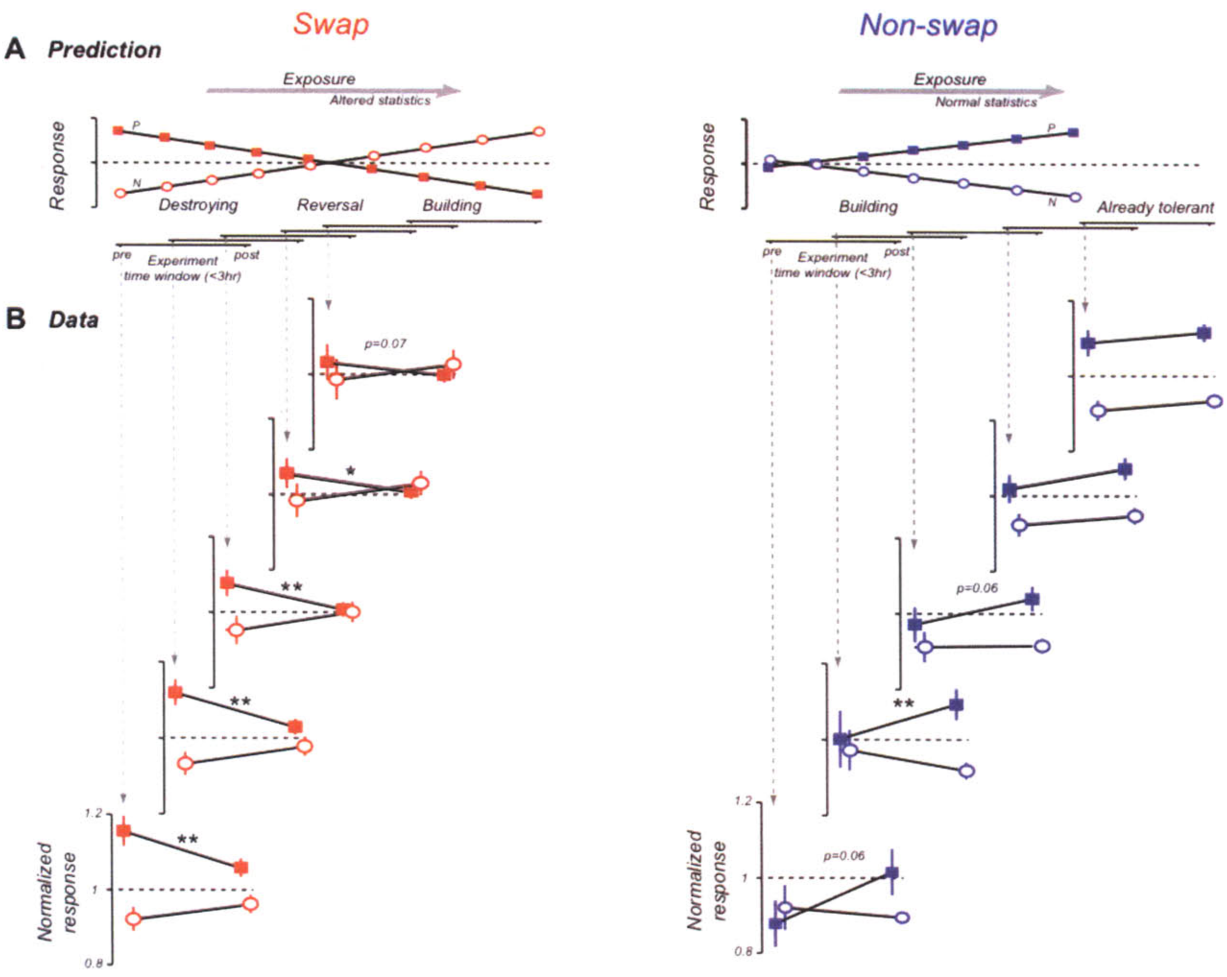


Figure 4-S9. Breaking and Building Tolerant selectivity, Size Experiment Data.

Mean ± SEM normalized response to object P and N at the swap size (A) and non-swap size (B) among sub-populations of IT multi-unit sites. Other details same as Figure 4-6 and 4-7. Size experiment data only (Experiment I and II).

4. 9 Supplemental Experimental Procedures

4. 9. 1 Visual Stimuli

Stimuli were presented on a 21" CRT monitor (85 Hz refresh rate, ~48 cm away, background gray luminance: 22 Cd/m2, max white: 46 Cd/m^2). We used 96 achromatic images from two classes of visual stimuli: 48 cutout natural objects and 48 silhouette shapes, both presented on gray background (Figure 4-S1). We chose these two classes of stimuli to be sufficiently different from each other in their pixel-wise similarity (Figure 4-S1), so that neuronal plasticity induced among one object class would be unlikely "spill-over" to the other class (our results and previous work confirmed this assumption, see Figure 4-S4 and Li and DiCarlo, 2008). All stimuli were presented on the animal's center of gaze during IT selectivity testing. In all experiments, we always used three object sizes (1.5°, 4.5°, 9°, in Experiment I and II; 1.5°, 3°, 4.5° in Experiment III). Object size was defined as the width of the smallest bounding square to contain the object. The medium object sizes were used to pick preferred (P) and non-preferred (N) objects for an IT site in an initial screening (see *Neuronal Assays* below), but we designed our manipulations and analyses to focus on the two extremity sizes (Figures 4-1B, 4-1C, 4-8A).

In Experiment II, to create the smoothly-varying identity-changing movie stimuli, we created morph lines between a subset of the silhouette shapes. Seven intermediate morphs were created in-between each object pairs. The movie stimuli were created to match the dynamics of object size changes that could be encountered in the natural world (see Figure 4-S2).

4. 9. 2 Behavioral Assay

Custom software controlled the stimulus presentation and behavioral monitoring. Eye position was monitored in nearly real-time (lag of ~3 ms) using standard sclera coil technique (Robinson, 1963) and in-house software, and saccades >0.2° were reliably detected (DiCarlo and Maunsell, 2000).

Test Phase: During each *Test Phase* (~10 minutes), IT neuronal selectivity was probed in two different tasks. Monkey 1 freely searched an array of eight small dots (size 0.2°) vertically arranged 3° apart. The dots never changed in appearance, but on each "trial", one dot would be randomly baited in that a juice reward was given when the animal foveated that dot, and the next "trial" continued uninterrupted. Typically, the monkey saccaded from one dot to another (not always the closest dot) looking for the hidden reward. During this task, object images were presented (100 ms duration) on the animal's center of gaze, (onset time was the detected end of a saccade; approximately one such presentation every other saccade, never back-to-back saccades). Thus, the monkey's task was unrelated to these test stimuli. To limit unwanted experience with the visual stimuli, each such presented object was immediately removed upon detection of any saccade and these aborted presentations were not included in the offline analyses. Monkey 2 performed a more standard fixation task in which it foveated a single, central dot (size 0.2°, ±1.5° fixation window) while object images were presented at a natural, rapid rate (5 images/s; 100 ms duration, 100 ms blank intervals). Reward was given at the end of the trial (5-8 images presented per trial). Upon any break in fixation, any currently present object image was immediately removed (and not included in the analyses), and the trial aborted. The animal could typically maintain fixation successfully in >75% of the trials. Aside from the task differences (free-viewing search vs. fixation), retinal stimulation in the two tasks was essentially identical. ~60 (±2) repetitions of each image were collected in the first *Test Phase* and ~50 (±2) repetitions in all the later *Test Phases*.

Exposure Phase: During each *Exposure Phase* (~1.5 hr), the animal freely viewed the monitor while object images (pseudo-randomly chosen) intermittently appeared at random positions on the screen. Because foveating a suddenly appearing object is a natural, automatic behavior, essentially no training was required, and the monkey almost always looked directly to the object (>90% of the time). 100 ms after the animal had foveated the object (defined by a saccade offset criteria of eye velocity<10°/s and a ±1.5° window centered on the object), the object underwent a size change on the animal's center of gaze. Importantly, some of the object size changes were accompanied by identity changes (i.e. our key manipulation, see details of specific experiments in the main text Experimental Procedures). The free viewing was meant to keep the monkey engaged in natural visual exploration, but the manipulation of object size statistics was always deployed during the brief intervals of fixation during natural exploration (see eye movement analyses in Figure 4-S7). The animal was only rewarded for looking to the object to encourage exploration, thus no explicit supervision was involved. There were a total of 8 different exposure event types in the full design (illustrated by the eight arrows in Figure 4-1B). One *Exposure Phase* consisted of 1600 exposure events: 200 exposure events per arrow exactly.

4. 9. 3 Neuronal Assay

Muti-unit activity (MUA) was gathered from 154 IT sites (n=44 for Experiment I; 19 for Experiment II; 91 for Experiment III) by randomly sampling over a ~4x4 mm area of the ventral STS and ventral surface lateral to the AMTS (Horsey-Clark coordinates: AP 13-17 mm; ML 18-22 mm at recording depth) from the left hemispheres of two monkeys. MUA was defined as all the signal waveforms in the spiking band (300 Hz – 7 kHz) that crossed a threshold set to ~2 s.d. of the background activity. That threshold was held constant for the entire session. A snippet of waveform data sampled at 0.07 ms intervals was recorded for 8 ms around each threshold-triggering event and saved for offline spike sorting (see *Data Analyses* below).

Each day, a glass shielded platinum-iridium microelectrode wire was introduced into the brain *via* a guide-tube and advanced to the ventral surface of the temporal lobe by a hydraulic microdrive (guided by anatomical MRI). We then advanced the microelectrode while the 96 object images (Figure 4-S1) were pseudo-randomly presented on the animals' center of gaze (animal tasks identical to those in the *Test Phases*). Once a visually driven recording site was found (based on online inspection), we stopped advancing and left the electrode in the brain to allow for tissue settling (up to 2 hours) before the recording session started. Each recording session began with an initial screening in which the IT sites were probed with the same object set (96 objects, ~10 repetitions per object, all presented on the center of gaze) for object pair selection:

Main Test Objects (Swap Objects): Among the objects that drove the site significantly above its background response (t-test against randomly interleaved blank presentation, p<0.05, not corrected for multiple tests), the most preferred (P) and least preferred (N) objects were chosen as a pair. Thus, both objects tended to drive the neuronal recording site, and most sites had selectivity for one (P) over the other (N). These two objects were chosen subject to the condition that both objects were from the natural object class (Monkey 1) or both were from the silhouette object class (Monkey 2; see Figure 4-S1).

Control Objects: For each recorded IT site, we also used the same initial screening (above) to choose a second pair of control objects (P' and N'). Our goal was to choose two objects the IT site was selective for but were very distant from the swap objects in IT shape space. Because we do not know the dimensions of IT shape space, we cannot strictly enforce this. In practice, we simply ensured that the control objects were always chosen from the object class that was not used for the swap objects (i.e. the silhouette object class for Monkey 1, and the natural object class for Monkey 2, see Figure 4-S1). Within this constraint, the control objects were chosen using the exact same responsivity and selectivity criteria as the swap objects (described above).

Once the initial screening and object selection was completed, we then carried out the *Test* and *Exposure Phases* in alternation while making continuous recording from the IT site for the entire recording session (~3 hours). The swap objects and control objects were each tested at all three sizes in each *Test Phase* but only the swap objects were shown and manipulated during the *Exposure Phases.*

4. 9. 4 Data Analyses

Neuronal data recorded from the 154 IT sites was first tested for their object selectivity. Offline analyses revealed that a fraction of the sites were not significantly selective among the swap object pairs (two-way ANOVA, 2 object x 3 sizes, p>0.05 for both "object" main effect and "object x size" interaction), probably because only a limited number of response repetitions were collected during the initial screening and we selected the objects to both produce a statistically significant response (as described above). We excluded those sites and only concentrated on the remaining object-selective sites (n=43 for Experiment I; 19 for Experiment II; 36 for Experiment III, many sites from Experiment III showed significant selectivity only for the swap object pair or only for the control object pair, but we concentrated on the sites that showed significant selectivity for both the swap and control object pairs). These sites were subject to one more screening for recording stability (see below) and all the results presented in the main text were from the object-selective and stable sites (n=27 for Experiment I; 15 for Experiment II; 31 for Experiment III).

Recording Stability Screen: We were interested in specific selectivity changes induced by our experience manipulation. However, we were concerned that non-specific selectivity changes (e.g. resulting from electrode drifts in tissue or neuronal injury) could potentially contaminate our effect of interest. Our controls were designed to make sure that we would not interpret any such effects as evidence of learning, but we still wanted to do our best to insure that any non-specific effects would not mask the size of our effect of interest. Unlike single-unit recording where one can judge the stability of recording based on spike waveform isolation, we do not have such measures in multi-unit recording. Thus we sought another independent measure of recording stability. To do this, we relied on the selectivity among the control object images (see above). We proceeded under the assumption that these control object images were far apart from the swap object pairs in the IT shape space, there should be little change in the selectivity among these control object images induced by our experience manipulation (our results and previous work confirmed this assumption; see Supplemental Figure 4-S4 and Li and DiCarlo, 2008). That is, the response to these objects provides a gauge of any non-specific changes in IT selectivity. To quantify that gauge, we computed Pearson's correlation between the control image response vectors (6 dimensional vector, 2 objects x 3 sizes) measured from the first and last *Test Phases*. We deemed an IT site "stable" if it had a correlation value higher than 0.7 (Figure 4-S4). In the main text, we only present results from these stable sites because they provide the cleanest look at our data and the best quantitative measure of learning magnitude. Critically, this site selection procedure relies only on data that is fully independent of our key exposure condition and key control condition (e.g. Figure 4-1B), so there is no selection bias. Nevertheless, we also repeated the same analyses on all of the recorded IT sites and found that the main results were qualitatively unchanged (see Figure 4-S4). We also tested more strict forms of stability criteria that included background activity change (<10, <5, and <2 spikes/s). With these stability criteria, all the key results also remained the same (Figure 4-S4C).

Computing (P-N) neuronal selectivity: To avoid any bias in this estimate of selectivity, for each IT site, we set aside an independent set of response data from the first *Test Phase* (10 response repetitions to each object in each size) and used those data only to define the labels "P" and "N" ("P" was taken as the object that elicited a bigger overall response pooled across object size). We recorded 10 extra response repetitions in the first *Test Phase* in anticipation of this need for independent data (60 repetitions in the first *Test Phase*, 50 repetitions in the later *Test Phases*). The label "P" and "N" for the site was then held fixed across object size and later *Test Phases*, and all remaining data was used to compute the selectivity (P-N) using these labels. This procedure ensured that any observed response difference between object P and N reflected true selectivity, not selection bias. Because different splitting of screen and remaining data may not result in consistent "P" "N" label, for each IT site this procedure was performed 100 times (different splitting of screen and remaining data in the first *Test Phase*) to obtain an averaged selectivity estimate (P-N). Variability arising from this procedure is reflected in the error bars of Figure 4-2C and 3B for each IT site.

Statistical Tests for the "Size x Exposure" Interaction: The key part of our experimental prediction is that any change in object selectivity should be found predominantly at the swap size (Figure 4-1C). To directly test for such an interaction between object size and our independent variable (exposure), we performed two different statistical tests on the neuronal selectivity measurements (P-N, in units of spikes/s). This main prediction and statistical results are from pooling across neurons (i.e. pooled "subjects" design with counterbalance).

First, we applied a two-factor repeated measures ANOVA. To design the test, we treated each IT site as one repeated measurement (i.e. one subject) with two within-group factors ("exposure" and "size"). Repeated measures ANOVA expects that all subjects are measured across the same number of conditions, however, our data was such that each IT site was tested for differential amount of time: some IT sites had three *Test Phases* while others only had two (due to different rates of experimental progress on each day and normal variation in the animal's daily work ethic). To get around this problem, for each IT site, we simply used the data only from the first and last *Test Phase*, omitting the data from the intermediate *Test Phases* for some IT sites. Thus in our ANOVA design, the "exposure" factor had two levels, and the "size" factor also had two levels: swap and non-swap. Our main focus was on the significant interactions between "exposure" and "size" (see main text). Our data also revealed significant main effects of "exposure" (Experiment I: p=0.0004; Experiment II: p=0.014) and no significant main effect of "size" (p = 0.72; p = 0.32). Given our experience manipulation and counterbalanced experience design across object size, this pattern of main effects is expected under the temporal contiguity hypothesis (see Figure 4-1C).

We also carried out a second, more non-parametric statistical test for the interaction of "exposure" and "size" by applying a general linear model. The formulation is similar to ANOVA. However, it is not subject to assumptions about the form of the trial-by-trial response variability. We have previously used the same method in our study on IT position tolerance learning (Li and DiCarlo, 2008) and simulations with Poisson spiking neurons have confirmed the correctness of our analysis code (~5% significant occurrence at p<0.05 with null effects). The model had the following form:

$$(P-N)_{neuron=n,size=s,\exp osure=e} = a_n + b_1 \cdot s + b_2 \cdot e + b_3 \cdot (s \cdot e)$$

The three independent variables of the model were: "size" (*s*), "exposure" (*e*), and their interaction (i.e. their product, $s{\cdot}e$). The "size" factor had two levels (i.e. $s = 1$ for swap size, -1 for non-swap size) the "exposure" factor had up to three levels depending how long a site was tested, (i.e. $e = 0$ for pre-exposure, and could be up to 1600 exposures in increments of 800's). Each a_n was the selectivity offset specific to each IT site; b_1, b_2, and b_3 were slope parameters that were shared among all the sites (i.e. within subject factors). Thus, the complete model for our population of *n* sites (*n*=27, Experiment I; n=15, Experiment II) contained a total of *n*+3 parameters that were fit simultaneously to our entire data set. The a_n's absorbed the site-by-site selectivity differences that were not of interest here, and the remaining three parameters described the main effects in the population, with b_3 of primary interest (interaction).

We fit the linear model to the data (standard least squares), and then asked if the observed value of the interaction parameter (b_3) was statistically different from 0. To do this, we obtained the variation of the b_3 estimate *via* bootstrap over both IT sites and repetitions of each site's response data. The exact procedure was done as follows: for each round of bootstrap over IT sites, we randomly selected (with replacement) *n* sites from our recorded *n* sites, so a site could potentially enter one round of bootstrap multiple times. Once the sites were selected, we then randomly selected (with replacement) the response repetitions included for each site (our unit of data here was a scalar spike rate in response to a single repetition of one object image in one size). Importantly, the selection of the response repetitions was done after we have excluded 10 response repetitions reserved for determining object labels ("P" and "N"). This absolute independence of the data allowed us to obtain unbiased selectivity estimates. Each site's (P-N) was computed from its selected response repetitions. The linear model was then fit to the data at the end of these two random samples to obtain a new b_3 estimate. This procedure was repeated 1000 times yielding a distribution of b_3 estimates, and the final p-value was computed as the fraction of that distribution that was less than 0. This p-value was interpreted as: if we were to repeat this experiment, with both the variability observed in the neuronal responses as well as the variability in which IT sites were sampled, what is the chance that we would *not* see the interaction observed here? In effect, this bootstrap procedure allowed us to derive a confidence interval on the model parameter estimate (b_3), and the duality of confidence intervals and hypotheses testing allowed us to report that confidence interval as a p-value (Efron and Tibshirani, 2003).

Statistical Tests for the Response Change in Single Sites: We evaluated each IT multi-unit site's selectivity (P-N) change by fitting linear regression as a function of the number of exposure events to obtain a slope, Δs(P-N). The statistical significance of the response change for each IT site was evaluated by permutation test. Specifically, for each site, we randomly permuted the *Test Phase* label of the response data (i.e. which *Test Phase* each sample of P and N response data belonged to, our unit of data here was a scalar spike rate in response to a single repetition of one object image in one size). We then re-computed the (P-N) selectivity on the permuted data and fit the linear regression. The permutation procedure was performed 1000 times to yield a distribution of slopes (empirical "null distribution" of Δs(P-N)). The p-value was determined by counting the fraction of the null distribution that exceeded the linear regression slope obtained from the data. All sites with $p < 0.05$ were deemed significant (see main text).

Combining the Position and Size Tolerance Learning Data: In main text Figures 4-6 and 4-7, we pooled the data from size experiment I, II, (n=42 MUA sites), and our previous position tolerance experiment (n=10 MUA sites collected using the same method described above, see Li and DiCarlo, 2008) because the two experiments used similar experience manipulations and the effect magnitude was comparable (Figure 4-5). To enter this analysis, we required that the sites had (P-N) selectivity at the medium object size/position (>5 spikes/s and <50 spikes/s, n=34). This was done under the logic that such selectivity is needed to provide a driving force for learning. We then used independent data to divide the sites into different groups based on the selectivity at the swap position/size in Figure 4-6 (Group 1: all sites; Group 2: <40; Group 3: <20; Group 4: <10; Group 5: <5; Group 6: <0) or at the non-swap position/size in Figure 4-7 (Group 1: <0; Group 2: <5; Group 3: <10; Group 4: <20; Group 5: all sites). We used independent data to select these sub-populations so that any stochastic fluctuations in site-by-site selectivity would produce no average selectivity change.

Single-unit Sorting and Analyses: We performed principle component analyses (PCA) based spike sorting on the waveform data collected during each *Test Phase*. K-mean clustering was performed in the PCA feature space to yield multiple units. The number of clusters was determined automatically by maximizing the distances between points of different clusters. Each unit obtained from the clustering was further evaluated by its signal-to-noise ratio (SNR: ratio of peak-to-peak mean waveform amplitude to standard deviation of the noise). For the analyses presented in Figure 4-4, we set a SNR threshold of 5.0, above which we will term a unit "single-unit". We verified that the key result is robust to the choice of this threshold (Figure 4-S6).

Because there was a great amount of cell-to-cell variability in IT neurons' selectivity, we computed a normalized selectivity measure for each neuron (Figure 4-4). Each neuron's size tolerance was computed as (P-N)/(P-N)$_{\mathrm{medium}}$, where (P-N) is the selectivity among the two objects at the tested size and (P-N)$_{\mathrm{medium}}$ is the selectivity at the medium object size. A size tolerance of 1.0 means that a neuron perfectly maintained its selectivity across the size variations spanned here. Because not all the single-units had object selectivity, only units that showed selectivity at the medium size were included ((P-N)$_{\mathrm{medium}}$>1 spikes/s).

4. 10 Supplemental References

Brincat, S.L., and Connor, C.E. (2004). Underlying principles of visual shape selectivity in posterior inferotemporal cortex. Nat Neurosci 7, 880-886.

DiCarlo, J.J., and Maunsell, J.H.R. (2000). Form representation in monkey inferotemporal cortex is virtually unaltered by free viewing. Nat Neurosci 3, 814-821.

Efron, B., and Tibshirani, R.J. (2003). An Introduction to the Bootstrap. (Chapman & Hall).

Horn, B. (1986). Robot Vision (MIT Press).

Hung, C.P., Kreiman, G., Poggio, T., and DiCarlo, J.J. (2005). Fast readout of object identity from macaque inferior temporal cortex. Science *310*, 863-866.

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

Kreiman, G., Hung, C.P., Kraskov, A., Quiroga, R.Q., Poggio, T., and DiCarlo, J.J. (2006). Object selectivity of local field potentials and spikes in the macaque inferior temporal cortex. Neuron *49*, 433-445.

Li, N., and DiCarlo, J.J. (2008). Unsupervised natural experience rapidly alters invariant object representation in visual cortex. Science *321*, 1502-1507.

Logothetis, N.K., and Sheinberg, D.L. (1996). Visual object recognition. Ann. Rev. Neurosci. *19*, 577-621.

Robinson, D.A. (1963). A method of measuring eye movements using a scleral search coil in a magnetic field. IEEE Transactions on Biomedical Engineering *101*, 131-145.

Vogels, R., and Orban, G.A. (1996). Coding of stimulus invariances by inferior temporal neurons. Prog Brain Res *112*, 195-211.

Chapter 5

General discussion

5. 1 Acknowledgements

During the past five years at MIT, I have benefited tremendously from conversations with my advisor, Dr. James DiCarlo, and many other brilliant scientists who have come through our lab and outside of our lab. These conversations, aided by my reading of the literature, taught me how to think about the problem of visual object recognition and they shaped the way I think about and approach any scientific problem. The majority, if not all, of my understanding of the problem of visual object recognition stem from these conversations, and I will summarize them in discussion format in this final chapter of my thesis.

5. 2 Summary of the key findings and looking forward

In this thesis, I have described results from two lines of work examining the neuronal representation underlying visual object recognition in the ventral visual stream of non-human primates. The work is particularly focused on the last anatomical stage of the ventral stream, the inferior temporal cortex (IT), where we believe such an object representation lives (DiCarlo and Cox 2007; Gross 2002; Logothetis and Sheinberg 1996; Tanaka 1996; Vogels and Orban 1996). We are particularly interested in the representation's ability to tolerate image variations arising from object identity preserving transformations ("invariance" or "tolerance").

In the first part of this thesis, I used computer simulation to show that preservation of rank-order object selectivity (but not response magnitude) in the face of identity preserving transformation is a single neuron response property that correlated well with the population's ability to support invariant recognition tasks (Li et al. 2009). I will refer to this single neuron response property as "single-unit separability". This result makes a number of assumptions. First, it assumes a definition of the recognition task. By task, what I mean is a set of images (arising from a set of objects undergoing a set of potential identity-preserving transformations) that operationally defines the recognition problem. These are likely "natural images" that are likely well represented by the ventral stream (as supposed any random image sets like white noise patterns). Second, the result assumes that units of the population cover the stimulus space we defined, meaning that the population has high discriminatory power for each individual exemplar image. Finally, there is assumed to be a fixed amount of resources, meaning there are a finite number of units in the population and each unit produces a finite number of spikes.

In addition to consolidating existing experimental data, single neuron metrics like single-unit separability likely inform us about the computational goals of single neurons along the ventral stream. Quantitative measurements of such metrics will likely connect electrophysiology data to computational models and constrain the possible operations the models are allowed to have. For example, improvements in single-unit separability result in gains in population performance that cannot be achieved with linear operations on the input alone, such as pooling of units with small receptive fields from the previous layer to make larger receptive fields without formatting the content of that data (Rust and Dicarlo 2010). Future electrophysiology experiments should make systematic measurements of single-unit separability with a large set of images across many neurons along the successive cortical stages of the ventral visual stream (this type of work is already being carried out in some cortical areas, e.g. Rust and Dicarlo 2010). Quantifying gains in single-unit separability across the measured neuronal distributions and comparing such gains to computational models will inform us which class of operations the models should implement to give rise to the pattern of single-unit separability gains measured experimentally.

In a real biological system such as the ventral stream, there are likely additional constraints such as wiring limitations (i.e. number of afferent connection per neuron allowed) and homeostatic mechanisms. These real system constraints are not considered in our simulation, however, they likely limit the class of possible operations further. Future simulations could be done to explore the effect of these constraints on the final measure of recognition performance in a full computational model.

A single neuron metrics such as single-unit separability is advantageous because it is a local metric, meaning it can be computed on a per-neuron-basis without further knowledge about the rest of the population, but it connects to the final measure of population performance on the recognition tasks. Individual neurons also have to work with information that is local to its input as well, and collectively as a population they can support the recognition tasks. Thus local metrics likely connect to quantities that are mechanistically computable by real neurons. Ultimately, we hope to discover the simple principles that guide the ventral stream neurons to become the way they process and pass on information. The second (and majority) part of this thesis is concerned with finding neuronal evidences for one such simple principle that could guide the ventral stream neurons to set up their tolerant response property through experience.

In the second part of this thesis, I described some neuronal evidence of single neurons in the ventral stream building preservation of object selectivity across transformations (i.e. single-unit separability) by relying on the temporal contiguity statistics of natural visual experience. In two experiments, I manipulated the temporal contiguity of the animals' visual experience by temporally coupling images of different objects together across object position and size change. Experience in this altered visual world predictably reshaped IT neurons' position and size tolerant selectivity and created a confusion of the temporally linked objects across the position and size change that was manipulated. Furthermore, experience with normal temporal contiguity statistics could build normal position and size tolerance. These neuronal changes could be induced under spatiotemporal regimes that mimicked the naturally occurring visual experience and they rapidly reshaped IT position and size tolerance with an hour of experience. The size of the effect increased with experience and it was comparable for both the position tolerance reshaping and size tolerance reshaping. From these results, we infer that the ventral visual stream is relying on a general temporal contiguity based learning mechanism to build and maintain all types of tolerance.

Discovering this novel form of learning opens up the possibility that the tolerant response phenomenology observed in adult IT might be acquired through visual development. Along with that possibility comes many questions. For example, we do not know the relationship between this IT neuronal learning and its perceptual consequence in the tested animals. We do not yet know how far we can push this learning by providing the animals with even more experience, (the effect seemed to be growing larger the more experience we provided the animals with for the duration of the experimental sessions), and whether the learning sustains across days. Most importantly, we do not know whether this learning is in fact used by IT to set up its neuronal tolerance during development. The rapid and unsupervised nature of the tolerance learning gives us an accessible experimental doorway to answer some of these questions. With the tools we have at hands (awake-primate acute physiology), we can start linking the neuronal changes to the animals' perceptual changes by training the animals to make perceptual judgments on the same set of stimuli we manipulate. Questions about the time course of the learning will have to be addressed with new techniques that enable stable and long-term recording of the same neurons. The most crucial question of whether the ventral stream uses the temporal contiguity learning to set up its neuronal tolerance during visual development will have to be addressed in new experimental preparations such as experience manipulations in infant animals.

5. 3 The larger problem of invariant object recognition

The work in this thesis is motivated by the "invariance problem" (DiCarlo and Cox 2007; Pinto et al. 2008; Riesenhuber and Poggio 2000, see Chapter 1). The problem is posed at the level of the whole system. No single neurons is strictly solving the "invariance problem" and it is difficult to imagine how to map such a computational goal onto single neurons in a complex information processing system like the ventral visual stream. Rather, each single neuron is likely performing operations on its inputs in a way that obeys very local rules. For systems neuroscientists, the experimental questions are posed at the level of single neurons, because neurons are likely the basic units of information processing and it is what our tools grant us access to. Thus there remains a gap between the physiological findings in single neurons and the system that as a whole solves the "invariance problem".

This thesis provides two key results at the level of single neurons. First, there is a relationship between the single-unit separability (preservation of rank order object selectivity) and the goodness of the population in supporting invariant recognition tasks (Li et al. 2009; Rust and Dicarlo 2010). Based on this relationship, we think single-unit separability is one potential proxy measure for invariance at the level of single neurons, and temporal contiguity can build such response property in real ventral stream neurons (Li and DiCarlo 2008; 2010). These results still remain at a qualitative level because they are not yet in a form that can be handed to a theorist and inform better building of recognition systems in ways that go beyond what has already been attempted.

Ultimately, the gap between single-neuron-operations and computations to solve the "invariance problem" has to be bridged by sophisticated computational models (Perry et al. 2010; Pinto et al. 2009; Riesenhuber and Poggio 1999; Serre et al. 2007). In a full-scale model, different local measures can be formalized as model parameters and be thoroughly explored to make explicit the effect of varying these parameters on the computation. At the same time, we enjoy the benefit of being able to compare the values of these parameters to what has been empirically well measured in the real brain. These parameters will reflect underlying operations single neurons are performing on their inputs or operations on a small pool of neurons. For example, different forms of Hebbian learning rules (rate based vs. timing based) that may lead to temporal contiguity learning can be implemented and tested in a computational model to examine their consequent behaviors, and those behaviors can generate new hypotheses for future physiology experiments.

With a more quantitatively framed hypothesis space by theories and models, there is likely still a need for collecting more physiology data to better constrain theories in the domain of object recognition. For experimenters, the current need is to collect data with more images per neuron (ideally with carefully-chosen, standardized image sets) rather than collecting a moderate amount of data from many neurons. Beyond the view of single neurons, there is likely still room for new discoveries of phenomenology at a larger anatomical scale with new tools. New tools are coming online now to allow scientists to ask questions that could not be asked before: microscopy imaging methods such as two-photon now allow large scale, *in vivo*, chronic monitoring of neuronal activity (Dombeck et al. 2007; Komiyama et al. 2010; Ohki et al. 2005; Svoboda and Yasuda 2006); optogenetic tools are starting to be applied *in vivo* in the primate visual system (Han et al. 2009), and can be used to ask causal questions and to dissect neuronal circuits (Luo et al. 2008; O'Connor et al. 2009; Petreanu et al. 2007). These new methods will likely reveal new neuronal phenomenology which were previously not visible or inaccessible to experimenters. The volume of the data acquisition is drastically increasing with the technology, however, soon, large volume of data can only be examined with specific hypothesis in mind and more sophisticated theories will be needed to guide the formulation of scientific questions.

One important contribution of systems neuroscience is perhaps to bridge levels, and map a set of neurally implemented operations onto interesting computational problems such as the "invariance problem".

5. 4 Closing

Temporal contiguity learning in the context of learning invariance is getting at the heart of two basic functions all sensory systems have to perform: grouping of some stimuli to be the "same" (invariance), and differentiating of some stimuli to be "different" (selectivity). There are two remaining questions that puzzled me and inspired my interest to continue on to pursue work in sensory learning after my Ph.D. study.

First, for invariance learning, temporal contiguity learning outlined here alone may not be enough to set up the IT neuronal tolerance. Temporal contiguity instructs the neurons to respond more similarly to things that are coupled in time (i.e. to associate stimuli together; also previously reported in the context of "paired associates" learning; Erickson and Desimone 1999; Messinger et al. 2001; Miyashita 1988). Taking this simple form of learning to the extreme, neurons should eventually lose their selectivity because everything is coupled in time across some timescale. How do neurons maintain their selectivity given a lifetime of experience (i.e. ability to still differentiate stimuli apart)? Theoretical forms of temporal contiguity learning such as slow feature analyses often invoke additional competitive mechanism between units to ensure coverage of the stimulus space in the process of minimizing slowness (Wiskott and Sejnowski 2002). This is to prevent all the units from arriving at the same, non-interesting solutions such as having no tuning at all. Thus, there might exist neuronal mechanisms that allow interaction or competition between different neurons in a local pool, so that neurons compete with each other locally for global coverage of the stimulus space and to maintain some form of homeostatic balance. As a very indirect evidence of this, in our studies, I found that not all neurons learn the same way under identical learning-induction protocols, suggesting that other factors beyond simple temporal contiguity learning were at play. For example, if IT neurons already exhibit good tolerance (i.e. strong selectivity across transformations), we could not build any more tolerance into these neurons by providing the animal with even more normal statistics (see Chapter 4). It seems that these neurons have reached a balanced state and cannot be perturbed further.

The second question that interests me is about the rapid nature of the learning we were able to induce in our experiments. Though it is difficult to map our experience manipulation onto experience time in real life, the fact that one could alter and even reverse IT neurons' tuning in just hours raises questions about how the cortical representation is actively maintained and stabilized in the face of constantly available experience. The answer to this question may have to do with the speed and specificity of the learning versus the size of the stimulus space that are represented by the neuronal population. Any local perturbation induced by learning may or may not impact the representation of other stimuli.

Answers to these questions, I hope, are going to inform us about how a set of local learning rules (e.g. association, competition) may manifest themselves at the level of neuronal population to set up or shape a set of local operations onto a computation. To study these questions, I feel that one needs to venture beyond quantifying learning phenomenology in single neurons and start to focus on the mechanisms at circuitry level (on the anatomical scale of a few hundred microns or within a cortical column) that underlie the single neuron learning phenomenology. Studying these questions requires an experimental preparation where one has an operational definition of sensory learning, and tools that allow one to simultaneously examine a large population of neurons during the defined learning process. I am excited by the experimental possibilities the new imaging and genetic tools bring. Two-photon microscopy now allows *in vivo,* chronic monitoring of neuronal activity across a large, genetically identified population of neurons (Dombeck et al. 2010; O'Connor et al. 2009). These new methods allow experimenters to ask questions about learning in single neurons and in interactions between neurons. Leveraging these recent advances in tools, a very first step is to systematically quantify any learning effect across different neuronal types. Having such a knowledge base will well position one to ask the next set of mechanistic questions outlined above. I am eager to participate in the scientific pursue that works toward identifying simple, local principles that govern local sensory learning that ultimately lead to high performance in computationally challenging global tasks, such as object recognition.

5. 5 References

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

Dombeck DA, Harvey CD, Tian L, Looger LL, and Tank DW. Functional imaging of hippocampal place cells at cellular resolution during virtual navigation. *Nat Neurosci* 2010.

Dombeck DA, Khabbaz AN, Collman F, Adelman TL, and Tank DW. Imaging large-scale neural activity with cellular resolution in awake, mobile mice. *Neuron* 56: 43-57, 2007.

Erickson CA, and Desimone R. Responses of macaque perirhinal neurons during and after visual stimulus association learning. *J Neurosci* 19: 10404-10416, 1999.

Gross CG. Genealogy of the "grandmother cell". *Neuroscientist* 8: 512-518, 2002.

Han X, Qian X, Bernstein JG, Zhou HH, Franzesi GT, Stern P, Bronson RT, Graybiel AM, Desimone R, and Boyden ES. Millisecond-timescale optical control of neural dynamics in the nonhuman primate brain. *Neuron* 62: 191-198, 2009.

Komiyama T, Sato TR, O'Connor DH, Zhang YX, Huber D, Hooks BM, Gabitto M, and Svoboda K. Learning-related fine-scale specificity imaged in motor cortex circuits of behaving mice. *Nature* 464: 1182-1186, 2010.

Li N, Cox DD, Zoccolan D, and DiCarlo JJ. What response properties do individual neurons need to underlie position and clutter "invariant" object recognition? *J Neurophysiol* 102: 360-376, 2009.

Li N, and DiCarlo JJ. Unsupervised natural experience rapidly alters invariant object representation in visual cortex. *Science* 321: 1502-1507, 2008.

Li N, and DiCarlo JJ. Unsupervised natural visual experience rapidly reshapes size-invariant object representation in inferior temporal cortex. *Neuron* 67: 1062-1075, 2010.

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

Luo L, Callaway EM, and Svoboda K. Genetic dissection of neural circuits. *Neuron* 57: 634-660, 2008.

Messinger A, Squire LR, Zola SM, and Albright TD. Neuronal representations of stimulus associations develop in the temporal lobe during learning. *Proc Natl Acad Sci U S A* 98: 12239-12244, 2001.

Miyashita Y. Neuronal correlate of visual associative long-term memory in the primate visual cortex. *Nature* 335: 817-820, 1988.

O'Connor DH, Huber D, and Svoboda K. Reverse engineering the mouse brain. *Nature* 461: 923-929, 2009.

Ohki K, Chung S, Ch'ng YH, Kara P, and Reid RC. Functional imaging with cellular resolution reveals precise micro-architecture in visual cortex. *Nature* 433: 597-603, 2005.

Perry G, Rolls ET, and Stringer SM. Continuous transformation learning of translation invariant representations. *Exp Brain Res* 204: 255-270, 2010.

Petreanu L, Huber D, Sobczyk A, and Svoboda K. Channelrhodopsin-2-assisted circuit mapping of long-range callosal projections. *Nat Neurosci* 10: 663-668, 2007.

Pinto N, Cox DD, and DiCarlo JJ. Why is real-world visual object recognition hard? *PLoS Comput Biol* 4: e27, 2008.

Pinto N, Doukhan D, DiCarlo JJ, and Cox DD. A high-throughput screening approach to discovering good forms of biologically inspired visual representation. *PLoS Comput Biol* 5: e1000579, 2009.

Riesenhuber M, and Poggio T. Hierarchical models of object recognition in cortex. *Nat Neurosci* 2: 1019-1025, 1999.

Riesenhuber M, and Poggio T. Models of object recognition. *Nat Neurosci* 3 Suppl: 1199-1204., 2000.

Rust NC, and Dicarlo JJ. Selectivity and tolerance ("invariance") both increase as visual information propagates from cortical area V4 to IT. *J Neurosci* 30: 12978-12995, 2010.

Serre T, Oliva A, and Poggio T. A feedforward architecture accounts for rapid categorization. *Proc Natl Acad Sci U S A* 104: 6424-6429, 2007.

Svoboda K, and Yasuda R. Principles of two-photon excitation microscopy and its applications to neuroscience. *Neuron* 50: 823-839, 2006.

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

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

Wiskott L, and Sejnowski TJ. Slow feature analysis: unsupervised learning of invariances. *Neural Comput* 14: 715-770, 2002.